РОЗДІЛ 8
СТРУКТУРНА, ПРИКЛАДНА ТА МАТЕМАТИЧНА ЛІНГВІСТИКА

UDC 811.161'32]:004.032.26:004.055:004.423.46
DOI https://doi.org/10.32782/tps2663-4880/2023.31.29

POLITICAL INTERNET DISCOURSE AS A SOURCE OF DATA FOR LINGUISTIC ANALYSIS

ПОЛІТИЧНИЙ ІНТЕРНЕТ-ДИСКУРС ЯК ДЖЕРЕЛО ДАНИХ ДЛЯ ЛІНГВІСТИЧНОГО АНАЛІЗУ

Dovhan O.V.,
orcid.org/0000-0002-6728-818X
PhD in Philology,
Doctoral student at the Department of Slavic, Romance and Oriental Languages
Mykhailo Dragomanov Ukrainian State University

The article highlights the resource-based approach to the study of political Internet discourse, which is productive due to its representativeness of the peculiarities of the communication process within it. The author emphasizes that, first of all, it is about the desire to influence the interlocutor (reader, listener, viewer), which is determined through explicit and implicit suggestibility, which prevails over evaluation. The author presents political Internet discourse as a special type of discourse within which emotionality is actualized (although rational arguments are present). This, according to the author, leads to the understanding of the latter not so much as a sphere of persuasion as of influence and manipulation.

The article emphasizes that within the framework of the above-mentioned discourse, it is productive to study the mechanisms of linguistic influence on the addressee (mostly mass, sometimes group or individual), who is in conditions of conflictogenic communication. The author emphasizes that it is also advisable to take into account the work of influence mechanisms actualized by the chosen strategy when studying political Internet discourse as a source of data for linguistic analysis. Thus, the communicative and pragmatic aspect of the use of the language poly system in the political sphere is in line with current trends in the study of its structure and other features.

The researcher notes that modern political communication is characterized by a continuous struggle of ideologies, represented in the clash of opposing interests and the increased distortion of data actualized within it. Thus, the work with the latter aims to create a positive/negative attitude towards the subjects of the russian government's activities – Ukrainians, their worldview and intentions, results, and cultural achievements. The author emphasizes that political Internet discourse's communicative-pragmatic and resource aspect has several gaps. This state of affairs is explained by the fact that each variant of political language is distinctive, not to mention the limited nature of its research by certain approaches. The above-mentioned makes the present study relevant, as it presents an innovative (resource-based) approach to the language poly system (in particular, political Internet discourse) as a data center that can be verified, validated, etc. for further linguistic analysis.

Key words: discourse, political Internet discourse, analysis of Internet discourse, text analysis, machine learning, artificial neural networks.

Стаття висвітлює ресурсний підхід до вивчення політичного інтернет-дискурсу, який є продуктивним через свою репрезентативність щодо особливостей комунікативного процесу у його межах. Автор підкреслює, що, першою чергою, мовиться про прагнення впливати на співбесідника (читача, слухача, глядача), що визначається через експліфіковану й імпліфіковану сугестивність, що превалює над оцінністю. Представлено політичний інтернет-дискурс як особливий вид дискурсу, у межах якого відбувається актуалізація емотивності (хоча раціональні аргументи мають місце). Це, на думку автора, спричиняє розуміння останнього не стільки сферою переконання, скільки – впливу та маніпуляцій.

У статті підкреслено, що у межах вищезазначеного дискурсу продуктивним є вивчення механізмів мовного впливу на адресата (переважно масового, часом – групового або індивідуального), який знаходиться в умовах конфліктогенного спілкування. Автор підкреслює, що доцільно також враховувати при дослідженні політичного інтернет-дискурсу як джерела даних для лінгвістичного аналізу роботу механізмів впливу, актуалізовану обраною стратегією. Отже, комунікативно-прагматичний аспект використання мовної полісистеми у політичній сфері суголосний актуальним тенденціям дослідження її структури та інших особливостей.

Дослідник зазначає, що для сучасної політичної комунікації притаманною є неперервна боротьба ідеологій, репрезентована у зіткненні протилежних інтересів і посиленому викривленні актуалізованих у її межах даних. Так, робота з останніми має за мету створення позитивного/негативного ставлення до суб'єктів діяльності російського уряду – українців, їх світогляду і намірів, результатів та культурних надбань. Автор акцентує увагу на тому, що комунікативно-прагматичний та ресурсний аспект політичного інтернет-дискурсу має низку лакун. Такий стан справ пояснюється тим, що кожен варіант політичної мови є самобутнім, не кажучи вже про обмеженість його досліджень певними підходами. Вищезазначене продукує актуальність представленого дослідження, в межах якого репрезентовано інноваційний (ресурсний) підхід до мовної полісистеми (зокрема, політичного інтернет-дискурсу) саме як осередку даних, які можна верифікувати, валідувати тощо з метою подальшого лінгвістичного аналізу.

Ключові слова: дискурс, політичний інтернет-дискурс, аналіз інтернет-дискурсу, текстовий аналіз, машинне навчання, штучні нейронні мережі.

Statement of the problem in general terms and its connection with important scientific or practical tasks. An integral element of the political system is factually motivated political Internet discourse (structurally, it is a construct created from political and Internet discourse itself), the peculiarities of which can be illustrated by the example of the full-scale russian-Ukrainian war. Within the framework of the latter, the authorities of the above-mentioned country have actualized several narratives aimed at motivating the phenomenon of russian ontology, in particular, in the context of the interaction between "I – Others" or "Ours – Others".

Naturally, the narratives actualized by the russian authorities are purely political, and thus play a worldview role, substantiating the idea of the "russian world" and constructing the "correct" perception of events. At the same time, the above-mentioned narratives, i.e., how processes in Ukraine and Ukrainians are represented in the political environment and the media, shape the attitude of russian society toward them. First of all, such narratives affect the development and implementation of russia's policy, as well as the results of the upcoming elections in this country, and the attitude of pro-russian countries to our state and citizens. In turn, this affects the opportunities available to Ukrainians abroad, the ability of the latter to exercise their rights, and affects their integration processes in the most general sense.

It is noteworthy that the actualization of russian narratives is naturally fueled by an information campaign that includes a whole galaxy of false information (propaganda, misinformation, and disinformation). At the same time, globalization and integration processes are developing, which in some way resonate with the above: for example, political communication is developing rapidly. The genesis of the latter produces a change in the stylistic structure of political texts, the emergence of new and more effective ways of influencing the consciousness of its objects (listeners), the actualization of allusions and various globalization manifestations (intertextuality, intermediality and interaudience).

The pivotal role in the above-mentioned communication is played by texts in this area, which is the object of study of political linguistics, which usually conducts a comparative analysis of political (Internet) discourse. The latter allows us to determine the dominant trends of both global, national, and local-state character. Naturally, political texts are often representatives of certain narratives (in the context of the russian-Ukrainian confrontation, we are talking about toxic narratives). In turn, such narratives coexist with hate speech (recall the sentiment analysis of the text), intolerance and discrimination (study of vocabulary and semantics), and the actualization of several false data (linguistic analysis of propaganda, misinformation, and disinformation).

Usually, the wording used in discussions of certain issues that naturally affect the understanding of social processes, existing problems, and public policy issues is indicative of the linguistic side. For example, russian Internet discourse is characterized by the use of vocabulary that dehumanizes Ukrainians and depersonalizes them, turning them into a resource, mass, etc., helping to justify the government's aggression. In particular, statements that represent Ukrainians as "Others" and "Not like us" and so on produce discrimination against other nationalities (and not only minorities) in russian society and lead to the expansion of intolerant thinking and behavior in general. One of the manifestations is the genesis of hate speech in russian Internet discourse alongside the established discrimination of the "Other/Others" and fellow citizens.

In our opinion, the effectiveness of the analysis of both political and Internet discourse (in particular, political Internet discourse) is directly related to the approaches to it as an object of study. The above-mentioned parameterization of this process within the framework of linguistics (in particular, political linguistics) is possible, but we consider the *resource approach* to be productive. The latter allows us to expand the specific linguistic tools with the methodology of data science, statistics, etc. In this approach, a political text is perceived deeper than just a set of certain elements of the language poly system, each of which can be studied.

Instead, the perception of political Internet discourse as a resource or source of data modifies the entire causal series and the logic of its processing and research within linguistics. Under such conditions, a text is studied not only in the context of its parameterization (language levels, units, structure, etc.), discursively (as a logical and natural result of a broader context), but also resourcefully (features of the data represented, validation, updating, representation of the latter and methods of their processing in the context of data science, machine learning, etc.).

Analysis of the latest research and publications that have initiated the solution of this problem and on which the author relies. The problem of political Internet discourse as a source of data for linguistic analysis is quite complex, which produces its multilayered nature and leads to its actualization in several interdisciplinary and integrated researches. First of all, the nature of the above-mentioned problem is related to the analysis of the data of this discourse, which, in turn, produces appeals to data science, analytical and synthetic information processing, machine learning, and the like.

In particular, the socio-cultural aspect of the problem was studied in the research of P. Lorenz-Spreen et al. [5], in which the authors analyzed the causal relationship between the global spread of digital media and the decline of democracy. The researchers tracked the occurrence of causal and correlation data (496 items), highlighting the connection between the mainstreaming of digital media and several political variables.

A criticism of the analysis of social networks in the context of considering their users as certain nodes or data centers is made by N. Botzer and T. Weninger [6], in which the authors investigated the specifics of online communication using natural language processing (hereinafter – NLP). The scientists emphasize that the established positioning of users as the above-mentioned cells leads to the perception of concepts or narratives actualized in social networks as certain streams. Accordingly, the latter pass through or past such social network centers, being actualized immediately (in the first case) or delayed (in the second case).

The study of sentiment analysis (emotional vocabulary) is devoted to the research of T. Widmann and M. Wich [20], in which the authors note that such studies are often based on sentiment dictionaries that focus on positive or negative tone. According to the scientists, this approach has several disadvantages, as the above-mentioned dictionaries are adapted to non-political areas and actualize the “Bag of words” approach in their research.

The analysis of the functioning of political blogs as components of political Internet discourse is contained in the research of M. Cornfield et al. [7]. Scientists position the functioning of political blogs and mass media as elements of information noise in the context of increased correlation of their data representation. The issue of actualization of false data (propaganda, misinformation, and disinformation) in political Internet discourse is further explored in the research of M. Gupta et al. [8], in which the authors emphasize that fake news creates a polarization of society.

Scientists argue that people’s political beliefs and cultural values usually correlate with the extent to which they believe in false content shared on social media. The researchers are studying the impact of people’s political beliefs and cultural values on the likelihood of fake news using a repeated measures method (which exposes people to different fake news scenarios). The results of an online survey based on questionnaires collected from participants in the United States of America and India confirm that conservative people tend to be more likely to validate fake news. The authors note that this study adds to the knowledge of the characteristics that make people more likely to trust fake news.

The peculiarities of using neural network models (in particular, large language models; hereinafter – LLM) are highlighted in the research of S. Feng et al. [9] in which the authors identify data sources for their training (news, discussion forums, and online encyclopedias). They note that most of the above data sources and the data itself are politically and socially biased. The researchers emphasize that the analyzed paper develops new methods for measuring political bias in LLMs trained on the above data (socioeconomic) and subsequent LLMs trained based on the first models with specific political bias.

The above issue in the context of the phenomenon of digital literacy is studied in the research of A. Guess and K. Munger [10], where the authors highlight the correlation between the spread of disinformation and other artifacts and media literacy. The researchers emphasize that there is a significant difference in the levels of digital literacy among the population, which has an age correlation. The genesis of opinion formation in Internet discourse (based on the example of comments on various YouTube videos related to COVID-19) is presented by S. Gupta, G. Jain, A. Tiwari [11]. In the analyzed research, the authors note that polarization caused by social media through selective access to information on the Internet during the COVID-19 pandemic has become a major cause of concern for countries around the world. The researchers analyze the temporal dynamics of polarization in the COVID-19-related Internet discourse and emphasize that the degree of polarization in the Internet discourse has grown exponentially with the pandemic.

The tools for studying political Internet discourse are discussed in the research of M. Osnabrügge, E. Ash, M. Morelli [14], in which the authors present and evaluate the use of supervised learning for cross-domain topic classification. The authors note that in this approach, the algorithm learns to classify topics in a labeled source corpus and then extrapolates them to an unlabeled target corpus from another domain. The researchers emphasize that the ability to use existing training data makes this method much more efficient than intra-domain supervised learning.

The potential of digitalization processes (in particular, their tools) for studying political Internet discourse is illustrated by the research of S. Srivastava [18], in which the author demonstrates the features of classification algorithms. According to scientist, the latter, which are actively used by companies such as Facebook, Google, and Amazon, actualize unsupervised and semi-supervised machine learning on huge databases to detect objects (faces, text processing, etc.) to model predictions for commercial and political purposes. The researcher focuses on the phenomenon of algorithmic control, which has been criticized by a wide interdisciplinary scientific community. This resonance was caused by several shortcomings identified by scientists, including mass surveillance, information pollution, behavioral herding, bias, and discrimination.

A logical continuation of the previous researches is the study of political discourse by E. Hyvönen et al. [16], in which the authors present a new open infrastructure called ParliamentSampo. The latter, according to the scientists, is a tool for studying parliamentary culture, language, and activities of politicians in Finland. The researchers emphasize that for the first time, the entire time series of about one million plenary speeches of the Finnish Parliament has been converted into data and unified formats, including CSV, Parla-CLARIN, ParlaMint, and RDF Linked Open Data (LOD).

The study of the phenomenon of prejudice in the context of social ontology and cultural genesis is continued by S. Schweitzer, K. Dobson, A. Waytz [17], which analyzes 4 national representative research. The latter found evidence of bias in people's perception of opposing points of view expressed in Internet discourse. In particular, the researchers summarize the results of the analyzed studies, focusing on the peculiarities of attributing the authorship of certain tweets to bots: for example, American Democrats and Republicans tend to attribute tweets to bots when they express counter-ideological views. Thus, the researchers identify a persistent bias that has implications for online political discussion and political polarization in general.

The use of hate speech in the context of the actualization of far-right discourse in Internet discourse is studied in the research of S. Hagen and D. De Zeeuw [12], in which the authors note that most current research focuses on the dangers of normalizing such language. The scholars emphasize that most of the above research are based on the assumption that far-right terms retain problematic meanings over time and on different platforms. The researchers consider contextual changes in meaning to be a pivotal factor in assessing the normalization of problematic but vague terms with the dynamics of their spread on the Internet.

Thus, the authors highlight the changing meaning of the term “based”, a word that was borrowed from Black Twitter and became a staple of the far-right's online vernacular in the mid-2010s. Using a qualitative and quantitative cross-platform approach, the researchers analyzed the evolution of the term between 2010–2021 on Twitter, Reddit, and 4chan. The researchers found that although the far-right meaning of the aforementioned term “based” has partially been preserved, it has become diffuse in the process of being actualized by other communities. This happened due to the processing of the core meaning: “not worrying about other people's opinions”, and “staying true to yourself” with a layer of political connotations. In turn, the above, according to the authors, calls into question the understanding of far-right memes and hate speech as having a single and stable problematic meaning. The various meanings and subcultural functions of the analyzed word in specific online communities prove the veracity of the scientists' hypothesis.

The study of digital sovereignty as a component of European information policy is localized in the research of D. Lambach and K. Oppermann [13], in which the authors highlight the functioning of the phenomenon in the example of Germany. The researchers localize several meanings attributed to digital sovereignty in German political Internet discourse. First of all, they talk about the digital structure for reconstructing the narratives involved in the formation of the above meanings.

Thus, the analysis of the historiography on the study of political Internet discourse as a source of data for linguistic analysis allowed us to track the core trends in contemporary scientific discourse and localize gaps in existing research. Despite the leading role of political Internet discourse in several social sciences (sociology, political science, linguistics, etc.), natural sciences (physics, biology, chemistry, etc.), and sciences of thought (philosophy, logic, psychology, etc.), the analyzed historiography shows a lack of research on this type of data in the context of its analysis, processing, and actualization (for example, in the context of data science, machine learning, etc.). Our research aims to fill this gap and study the peculiarities of the above-mentioned type of discourse in the context of data research.

Highlighting the previously unresolved parts of the general problem to which this article is devoted. The aforementioned resource-based approach to the study of political Internet discourse is productive because of its representativeness in terms of the peculiarities of the communication process within it. First of all, we are talking about the desire to influence the interlocutor (reader, listener, viewer), which is determined through explicit and implicit suggestibility [19], which prevails over evaluation. The above produces the positioning of political Internet discourse as a special kind of discourse, mainly actualizing emotionality (although rational arguments do take place). This, in turn, leads to the understanding of the latter not so much as a sphere of persuasion as a sphere of influence and manipulation.

Thus, within its framework, it is productive to study the mechanisms of linguistic influence on the addressee (mostly mass, sometimes group or individual), who is in conditions of conflictogenic communication. When studying political Internet discourse as a source of data for linguistic analysis, it is also advisable to take into account that the study of influence mechanisms is actualized by the chosen strategy. Thus, the communicative and pragmatic aspect of the use of the language poly system in the political sphere is in line with the current trends in the study of its structure and other features.

Contemporary political communication is characterized by the continuous struggle of ideologies (agonality), which is represented in the clash of opposing interests and the increased distortion of data. Working with the latter aims to create a positive/negative attitude towards the subjects of the russian government's activities – Ukrainians, their worldview and intentions, results, and cultural achievements. Several integrated, relevant researches are not included in our review of historiography in this research, as it has certain space requirements.

However, the communicative-pragmatic and resource aspect of political Internet discourse has several gaps. This state of affairs is explained by the fact that each variant of political language (texts) is distinctive, not to mention the limited nature of its research by certain approaches. The above-mentioned determines the relevance of our research, in which we present an innovative approach to the language poly system as a source of data that can be verified, validated, etc.

Formulation of the article's objectives (statement of the task). *The purpose* of the article is to consider the peculiarities of analyzing the texts of political Internet discourse as an object of research. *The subject* is the specifics of the above phenomenon in the context of data processing and the functioning of the artificial neural network as an innovative tool of linguistic science.

Presentation of the main research material with full justification of the scientific results obtained. The resource-based approach to political Internet discourse as a source of data (in particular, for linguistic analysis) naturally produces preprocessing of the latter. The point is that to validate certain data that we plan to use for linguistic analysis (classical or mediated by neural network modeling), it is necessary to understand their specifics and features.

If we are talking about Internet discourse (in particular, political discourse), then it is necessary to first understand what kind of media, according to the Law of Ukraine "On Media" are considered online media [4]. First and foremost, these are media that disseminate information while performing a media function. Thus, according to the Recommendation of the Committee of Ministers of the Council of Europe (hereinafter – the Recommendation) [2] on the definition of:

a) *the intention to act as a media outlet* (self-identification as a media outlet, established working methods, adherence to journalistic standards, etc);

b) *the purpose and goals of the activity corresponding to the outlined specifics* (breadth of information presented: typological and species diversity of content);

c) *editorial control is in place* (there is a certain editorial policy documented: it refers to the regulatory and legal support of the activity);

d) *actualization of methods and tools for disseminating information* (based on the thematic focus, target audience, etc.);

e) *meeting public expectations* (first of all, this is a product activity that involves customer orientation: accessibility, pluralism, reliability, transparency, etc.).

The above makes it possible to localize the parameterization of online media in Ukraine, which, in turn, will make the preprocessing stage more efficient. Thus, in our country, we will distinguish the following types of online media:

1. *News agencies* (for example, Interfax, UNIAN, etc.).

Such agencies are distinguished by the orientation of the content they produce, which is primarily oriented inwardly to the journalistic community (i.e., to journalists) rather than outwardly to external users (readers, listeners, viewers). Thus, such sources are characterized by the availability of certain data, as news agencies accumulate, analyze, and present them primarily in the format of news (there is no specific diversity: news, analytics, interviews, etc.). In the context of true/false data, such agencies are not a verified source of information, as they publish both "free" news (familiar to us) and "paid" news on a completely legal basis.

2. *Socio-political media* (for example, "Ukrayinska Pravda", etc.).

They present different types of information (news, analytics, interviews, etc.), the information data is differentiated (topics, types, etc.), and has a wide thematic focus (from political news to weather forecasts). As for the validation of the data presented, it is much higher in such sources (due to species diversity, reputational risks, etc.), so they can be positioned as reliable sources.

3. *Secondary (relative to the mainstream) media* (for example, 112.ua, nv.ua, Radio Liberty, etc.).

This group includes various branches, the so-called "subsidiaries", i.e., business entities controlled by other such entities. Naturally, in the matter of validation of such sources, it is advisable to focus on the peculiarities of the "main" source: in particular, the specifics of compliance with journalistic standards, editorial policy, owners, etc. It should be noted that this type of media has its peculiarities: for example, TV channel branches are mostly representative of the content of the notional primary source and do not often produce separate content. As for print media branches, they are gradually disappearing, changing places: a magazine becomes an appendix to a website or something else. Instead, foreign media branches are simply a representation of their content with the integration of the Ukrainian context (mostly formal).

4. *News aggregators* (for example, ukr.net, etc.).

In fact, like the next position, they are not media, as they do not fall under the above points of the Recommendation, as they simply accumulate information on a certain platform without adherence to certain standards and validation. The above makes it impossible to use such sources as true data, which is why their actualization (as well as the next item) is not directly productive for linguistic research.

5. *Scavengers sites* (for example, ua-vestnik.com, ukrainianwall.com, newsua.biz, politeka.net, etc. [3]).

Most of these sites contain emotionally colored information and a specific domain name (*.cc, *.com, *.pp.ua, *.biz.ua), as well as an original presentation of data because instead of journalists, the authors are philologists, copywriters, and others. Sometimes the source of the information is anonymous or has been repeatedly identified as a center of false information, and the editorial contacts are sent to Yandex, Mail.ru, Rambler, etc. servers [15]. To implement the strategy of discrediting the Ukrainian authorities, such sources have updated a well-established set of speech tactics: accusations, negative injection, stringing together negative events or consequences, indirect insults, labeling, humiliating comparisons, etc. The use of metaphors with a negative assessment of events in Ukraine, irony, and sarcasm also contribute to the achievement of the goal of the scavenger sites.

The process of validating political Internet discourse data for linguistic analysis is directly affected by the peculiarities of online media funding as a source of objective data. Let's divide the above-mentioned features into three zones (white, gray, and black – based on journalistic integrity and for ethical reasons, we will not provide examples for the last two zones):

White zone

1. *Differentiation of data* (for example, "Liga. Zakon", etc.).

The essence of this approach to media financing is that some of the information on the resource is free, while others are paid for. For example, the full text or the text without advertising of a legal document on "Liga.Zakon" is paid for, but you can read an incomplete version for free. This is a productive feature for the companies that own the resources, but extremely destructive for the research process, since the incomplete text cannot be representative, and therefore its use is not advisable.

2. *Financial support from patrons, audience, etc.: grants or donations* (for example, Texty.org.ua, etc.).

It does not affect the validity of the resource's data, so it is a productive source for conducting linguistic analysis using innovative technologies.

3. *Placement of advertising* (almost any resource).

It is noteworthy that there is a difference between the editorial policy of the resource and the advertising policy of the service customers, and therefore the data on this source may have the opposite meaning. Naturally, the actualization of innovative tools of data science, machine learning, etc. has several drawbacks. For example, it is quite difficult to build a neural network model training process in such a way that it can distinguish between "specific" resource material and "advertising" material.

4. *Placement of native or natural advertising* (almost any resource).

Advertising material from a professional journalist that does not contain manipulation, propaganda, misinformation, or disinformation. It contains verified facts but is written by order of the business owner or his employees, and the resource directly informs about its advertising origin. The main feature is that it fits into the context of the resource, which makes the material not striking and looks organic in the context of the topic, time, and style. It is quite harmful if considered as valid data, as the financial component makes it invalid about the reputed data: it is written to order, and therefore not representative of the research.

Gray zone

5. *Work on the principle of clickbait*.

The essence of this approach is the use of headlines that provoke the reader to click on the link and get to a certain resource. It is noteworthy that the content of the headline and the material on the site where we get to are different. Thus, the reader is disoriented: not valid for any research, especially in the context of data preprocessing for linguistic analysis.

Black zone

6. *Use of "jeansa"*.

Placing advertising, i.e. paid material without indicating this, leads to misleading users. Not valid.

7. *Actualization of black PR*.

The peculiarity of this approach is the paid placement of false or true, but discreditable facts about a certain person to receive a monetary reward from him or her for removing the material. Not valid.

8. *Use of stop-list technology*.

Receiving a monetary reward for not publishing certain information. Not valid.

Thus, the analysis of the peculiarities of political Internet discourse resources for linguistic analysis has shown several trends in the existence of modern scientific discourse: in particular, its role in shaping worldview (public opinion and political beliefs). It is noteworthy that linguistic research is becoming an important tool for studying this environment in terms of understanding its content in the context of working with false data, manipulations, etc.

At the same time, an important component of the effectiveness of such studies is the validity of the collected data, the means of its verification, as well as several discursive knowledge that allows for a qualitative analysis process. Since the Internet discourse is a medium, a space of fluctuations in the meaning assigned (it is equally dominated by official statements of politicians and public discussions), the process of the methodology of its data comes to the fore. The latter is connected with the need for their representativeness and reliability for further analysis since the lack of validation will lead to distortion of the results and produce incorrect and unreliable interpretations of any study.

We consider the main methods of the above process to be: *content analysis* (will allow us to identify the core themes and patterns of political texts and ensure the objectivity and reliability of linguistic analysis); *cross-checking* (will produce a study of the results of linguistic analysis in the context of their correlation with other sources of political information, confirming their reliability); *expert evaluation* (to assess the compliance of the results with the linguistic patterns and cultural genesis of political texts); *scalability* (to ensure a wide sample (taking into account the volume and representativeness), which will determine the validity of the results).

Conclusions from this research and prospects for further research in this area. Thus, the study of political Internet discourse as a source of data for linguistic research is a core element of ensuring the reliability and objectivity of such works. The actualization of several methods (content analysis, cross-validation, expert evaluation, scalability) produces a reliable basis for conducting representative, reliable, and knowledge-intensive linguistic research.

Perspectives for further research on the analyzed issues are as follows: *automated processing of language data* (improvement of machine learning methods, work with neural network models of various types and with different layers and NLP); *study of a number of socio-cultural correlations* (tracking the influence of cultural and social contexts and their correlation on the linguistic tactics and strategies implemented in the above-mentioned discourse); *analysis of transformational changes in political Internet discourse* (highlighting the genesis of linguistic patterns, influence strategies and changes in approaches to building the communication process); *development of validation standards* (to promote consistency and interdisciplinarity of research implemented within the subject of study); *integrated content research* (actualization of video speeches, audio recordings of speeches, etc. using neural network modeling); *localization of the value of social media filters and algorithms* (studying the impact of the latter on the representation and dissemination of political content, researching additional aspects of the problem of generating false data and manipulations in Internet discourse).

REFERENCES:

1. Про медіа : Закон України від 13 грудня 2022 р. № 2849-IX. *Верховна Рада України* : вебпортал. URL: https://goo.su/lSP5y (дата звернення: 25.12.23).

2. Щодо просування сприятливого середовища для якісної журналістики у цифрову еру : [неофіційний переклад] : Рекомендація Комітету Міністрів від 17 березня 2022 року № CM/Rec(2022)4. *Національна*

рада України з питань телебачення і радіомовлення : вебсайт. URL: https://goo.su/4mnjca (дата звернення: 25.12.23).

3. Астрологи, Ванга та Безугла. Моніторинг дези в Україні за 1–29 листопада 2023 року. *Texty.org.ua* : вебсайт. URL: https://goo.su/OLG6qN (дата звернення: 25.12.23).

4. Володовська В. Хто такі онлайн-медіа в законопроєкті про медіа. *Національна рада України з питань телебачення і радіомовлення* : вебсайт. 2020. URL: https://goo.su/MEhs2S6 (дата звернення: 25.12.23).

5. A systematic review of worldwide causal and correlational evidence on digital media and democracy / P. Lorenz-Spreen et al. *Nature human behavior*. 2023. Volume 7, Issue 1. P. 74–101. *Nature Human Behavior* : website. URL: https://goo.su/iIBKuW (date of application: 25.12.23).

6. Botzer N., Weninger T. Entity graphs for exploring online discourse. *Knowledge and Information Systems*. 2023. P. 1–19. *Springer Link* : website. URL: https://goo.su/IsaSaf9 (date of application: 25.12.23).

7. Buzz, blogs, and beyond : The Internet and the national discourse in the fall of 2004 / M. Cornfield et al. In *The Political Communication Reader*. Routledge, 2023. P. 296–305. *Taylor & Francis Online* : website. URL: https://goo.su/PRxm (date of application: 25.12.23).

8. Fake news believability : The effects of political beliefs and espoused cultural values / M. Gupta et al. *Information & Management*. 2023. Volume 60, Issue 2. https://doi.org/10.1016/j.im.2022.103745 *Science Direct* : website. URL: https://goo.su/cZ0gol6 (date of application: 25.12.23).

9. From Pretraining Data to Language Models to Downstream Tasks : Tracking the Trails of Political Biases Leading to Unfair NLP Models / S. Feng et al. *Cornell University* : website. 2023. URL: https://goo.su/1z9sJ (date of application: 25.12.23).

10. Guess A. M., Munger K. Digital literacy and online political behavior. *Political science research and methods*. 2023. Volume 11, Issue 1. P. 110–128. *Cambridge University Press* : website. URL: https://goo.su/hQsA64 (date of application: 25.12.23).

11.Gupta S., Jain G., Tiwari A. A. Polarised social media discourse during COVID-19 pandemic : evidence from YouTube. *Behaviour & Information Technology.* 2023. Volume 42, Issue 2 : Social Response to the Covid-19 Pandemic. P. 227–248. https://doi.org/10.1080/0144929X.2022.2059397 *Taylor & Francis Online* : website. URL: https://goo.su/QE0laDt (date of application: 25.12.23).

12. Hagen S., de Zeeuw D. Based and confused : Tracing the political connotations of a memetic phrase across the web. *Big Data & Society*. 2023. Volume 10, Issue 1. https://doi.org/10.1177/20539517231163175 *Sage Journals* : website. URL: https://goo.su/HbYIt (date of application: 25.12.23).

13. Lambach D., Oppermann K. Narratives of digital sovereignty in German political discourse. *Governance*. 2023. Volume 36, Issue 3. P. 693–709. https://doi.org/10.1111/gove.12690 *Wiley* : online library. URL: https://goo.su/WLzy5b (date of application: 25.12.23).

14. Osnabrügge M., Ash E., Morelli M. Cross-domain topic classification for political texts. *Political Analysis*. 2023. Volume 31, Issue 1. P. 59–80. *Cambridge University Press* : website. URL: https://goo.su/mXKL (date of application: 25.12.23).

15. Overconfidence in news judgments is associated with false news susceptibility / B. A. Lyons et al. *Proceedings of the National Academy of Sciences.* 2021. Volume 118, Issue 23. https://doi.org/10.1073/pnas.2019527118 *PNAS* : website. URL: https://goo.su/zVlmf (date of application: 25.12.23).

16. ParliamentSampo infrastructure for publishing the plenary speeches and networks of politicians of the Parliament of Finland as open data services. / E. Hyvönen et al. In Proceeding: *International Workshop on Knowledge Graph Generation from Text (TEXT2KG), co-located with ESWC 2023 conference*. CEUR Workshop Proceedings. 2023. *Semantic Computing Research Group (SeCo)* : website. URL: https://goo.su/QRYIV (date of application: 25.12.23).

17. Schweitzer S., Dobson K. S., Waytz A. Political Bot Bias in the Perception of Online Discourse. *Social Psychological and Personality Science*. 2023. https://doi.org/10.1177/19485506231156020 *Sage Journals* : website. URL: https://goo.su/xRoiyY (date of application: 25.12.23).

18. Srivastava S. Algorithmic governance and the international politics of Big Tech. *Perspectives on politics*. 2023. Volume 21, Issue 3. P. 989–1000. *Cambridge University Press* : website. URL: https://goo.su/LIMji (date of application: 25.12.23).

19. Suggestion. *Cambridge dictionary* : website. URL: https://goo.su/oe59 (date of application: 25.12.23).

20. Widmann T., Wich M. Creating and comparing dictionary, word embedding, and transformer-based models to measure discrete emotions in German political text. *Political Analysis*. 2023. Volume 31, Issue 4. P. 626–641. *Cambridge University Press* : website. URL: https://goo.su/CfXyX (date of application: 25.12.23).