UDC 811.161’32]:004.032.26:004.055:004.423.46
DOI https://doi.org/10.52726/as.humanities/2023.4.8

O. V. DOVHAN
PhD in Philology,
Doctoral Student at the Department of Slavic, Romance and Oriental Languages,
Drahomanov Ukrainian State University, Kyiv, Ukraine
E-mail: dovgan396@gmail.com
https://orcid.org/0000-0002-6728-818X

FEATURES OF NEURAL NETWORK MODELING OF THE PROCESSES OF RECOGNIZING LINGUISTIC MARKERS OF THE CATEGORIES OF SENSE AND ABSURDITY IN THE CONTEXT OF WORKING WITH FALSE DATA

The full-scale invasion of Ukraine by the russian federation had a number of geopolitical and domestic consequences. One of the manifestations of such consequences on both of the above levels was the change in the modern information environment under the influence of the actualization of numerous russian narratives. Naturally, the consequence of such false information should be the weakening and vulnerability of Ukrainians to manipulation, disinformation, etc.

The latter is represented in the peculiarities of updating the above-mentioned data with false, exaggerated, distorted, etc. information. The above, in turn, determines the relevance and urgency of the problem of their existence in the context of russian disinformation, misinformation propaganda, and fakes (as well as diplomatic fakes), as well as the development of tools for their analysis to prevent their destructive impact.

We are talking about the nature of the language poly system and Natural Language Processing (hereinafter – NLP) as tools for processing language data for further use in the process of Machine Learning (hereinafter – ML) and analysis using an artificial neural network. These include vectorization, tokenization, lemmatization, speech interface (Google Assistant and others), automatic translation (Google Translate, Reverso, DeepL, etc.), and other areas. Naturally, the aforementioned study is of particular relevance in the context of the hybrid nature of the Russian-Ukrainian war, in which the media space (online discourse) is home to a whole bunch of fake, distorted, actually false, and other data.

That is why in our research we will analyze the main issues related to the problem of methodological features of neural network modeling of the processes of recognizing linguistic markers of the categories of sense and absurdity, paying attention to their nature, mutual influence, and interdependence, priority and socio-cultural significance for Ukrainian society. In addition, we will consider the specifics of the above process in the context of working with false data (disinformation), as well as the impact of the latter on its course.

Key words: linguistics, natural language processing, text data, artificial neural networks, machine learning, neural network modeling methodology.

Introduction. The full-scale invasion of Ukraine by the russian federation had a number of geopolitical and domestic consequences. One of the manifestations of such consequences on both of the above levels was the change in the modern information environment under the influence of the actualization of numerous russian narratives. We will use this term in the sense of narratives, explanations, etc. that are used without reference to specific data or facts. It is noteworthy that such narratives are essentially interpretations, an act of creativity that nevertheless has a specific purpose.

In our case, russian narratives are aimed at atomizing Ukrainian society and stratifying it (let us recall numerous fake news about certain current events: mobilization (news about a unit with Ukrainian children), military operations (a series of videos on Tik Tok telling about "terrible" losses and defeats of our defenders), military administration (news about the death of the Commander-in-Chief of the Armed Forces of Ukraine Valeriy Zaluzhnyi or his serious condition, etc.) Naturally, such false information should result in the weakening and vulnerability of Ukrainians to manipulation, disinformation, etc. The latter is represented in the peculiarities of updating the above-mentioned data with false, exaggerated, distorted, etc. information. The above, in turn, determines the relevance and urgency of the problem of their existence in the context of russian disinformation, misinformation, propaganda, and fakes (as well as diplomatic fakes), as well as the development of tools for their analysis to prevent their destructive impact.

At the same time, it is advisable to take into account the functional nature of the language poly system, which, being the core means of human communication, simultaneously poses the greatest threat in terms of dealing with false data. First and foremost, we are talking about the nature and dynamics of such processes within the designated construct as the generation, transformation, modification, etc. of the underlying meanings. The latter is relevant for human ontology (existence in the context of the noosphere), epistemology (processes of actualization of background knowledge: gaps, realities, etc.), axiology (value ranking of certain life events), etc. In turn, the destruction, substitution, distortion, misrepresentation, etc. of the above elements causes a number of national security problems for modern states, in light of which fact-checking initiatives in Ukraine (Ministry of Digital Transformation, Center for Countering Disinformation, Ukraine Crisis Media Center, etc.) and the world are gaining particular importance.

An important feature of Russian disinformation campaigns, compared to their counterparts implemented by other states, is that they are not limited to certain chronological limits. Thus, in practice, this means that they are permanent, functioning not sporadically (for example, the election campaign of a candidate in the United States, France, etc.), but purposefully and without regard to specific socio-political changes. Therefore, the study of the methodological features of neural network modeling of the processes of recognizing linguistic markers of the categories of sense and absurdity in the context of working with false data is particularly relevant.

It is about the nature of the language poly system and Natural Language Processing (hereinafter – NLP) as tools for processing language data for further use in the process of Machine Learning (hereinafter – ML) and analysis with the help of an artificial neural network. These include vectorization, tokenization, lemmatization, speech interface (Google Assistant and others), automatic translation (Google Translate, Reverso, DeepL, etc.), and other areas [Fazil, M. et al., 2023]. Naturally, the above-mentioned study is of particular relevance in the context of the hybrid nature of the russian-Ukrainian war, within which whole clusters of fake, distorted, actually false, and other data function in the media space (online discourse).

That is why in our research we will analyze the main issues related to the problem of methodological features of neural network modeling of the processes of recognizing linguistic markers of the categories of sense and absurdity, paying attention to their nature, mutual influence, and interdependence, priority and socio-cultural significance for Ukrainian society. In addition, we will consider the specifics of the above process in the context of working with false data (disinformation), as well as the impact of the latter on its course.

The problem analyzed in the research is integrated, interdisciplinary, etc., which produces its actualization in a number of scientific researches. For example, the research of NLP in the context of continental philosophy and philosophy of language (M. Heidegger, S. Freud, J. Lacan, and others) is presented in the research of M. Heimann, A. Hübener (2023), which discusses the problem of negation and negativity, which is central to continental discourse but is absent in studies of the Large Language Model (hereinafter – LLM). The latter is mostly used for a number of text-processing tasks (generation, translation, paraphrasing, classification, etc.).

An analysis of the peculiarities of using NLP is presented in the research by S. Chung et al. (2023), in which the authors identified four main areas of research on the above problem, as well as reviewed representative areas of its application. The results of the researchers' work showed a narrowing of the gap between NLP and various areas of its actualization while localizing a number of gaps in certain areas of the above process and its methodologies.

F. Guo (2023) continues to study the problem of NLP, whose research emphasizes the pivotal role of understanding the language poly system for various psychological measurements, as well as the place of semantic cues in this process. The author notes that with the development of computer science in general and Data Science (hereinafter – DS) in particular, NLP has become an effective method of analyzing textual data and representative indicators of psychological measurements. The scientist emphasizes that the work demonstrated the effectiveness of finely tuned NLP models for classifying pairwise relations into trivial/low or moderate/high empirical relations, providing preliminary validity evidence without manual data collection.

The problem of Natural language generation (NLG) is addressed by O. Nikula (2023), which analyzes the creation of neural fake news. The author used the Grover neural network model to generate a set of articles based on both real and fake news written by humans. S. Rastogi, D. Bansal (2023) continue to study the problem of fake news, noting that a number of researches offer a solution to the above problem of fake news based on Machine Learning (hereinafter – ML). In their research, the authors present a typology of false data, the time of its detection, a taxonomy for classifying researches of this issue, and the prospect of studying the above issue.

The problems associated with updating neural network models to create and detect false data are also analyzed in the research by Ș. Repede (2023). The analyzed research considers a range of issues related to the above-mentioned issues: from the complexities of the terminology system to the specification of databases in the context of fake news research and the peculiarities of their labeling.

The features of online social networks (OSNs) and the specifics of false data, hate speech, etc. in their environment are presented in the research by M. Fazil et al. (2023). In the analyzed research, hate speech in the above networks is studied using the Bidirectional long short-term memory networks layer (hereinafter – BiLSTM) – a layer of the Recurrent neural network (hereinafter – RNNs), which studies bidirectional long-term dependencies between time steps of time series or data sequences. The aforementioned neural network model updates the existing methods of representing words in a multichannel environment with multiple filters with different kernel sizes to capture semantic relations in different windows.

P. Kar, S. Debbarma (2023) continue to study the peculiarities of hate speech in the online environment and the specifics of its detection using artificial neural networks. In the analyzed research, the authors propose Optimal feature extraction and Hybrid diagonal gated RNNs (FE-DGRNN), which includes a three-stage methodology. Thus, at the first stage, after preprocessing, they update improved seagull optimization, which is used to extract multiple features from a text data set mixed with codes. At the second stage, the quantum search optimization algorithm is used to optimize the extracted features, which proactively solves the problem of data dimensionality at the stage of further detection. At the third stage, Hybrid diagonal gated RNNs (Hyb-DGRNNs) are used to detect hate speech and analyze sentiment based on it.

The research by E. Park, V. Storey (2023) organically continues the above work. In the analyzed research, the authors analyze sentiment, which is considered in the context of the Emotion Ontology Framework (EOF). The latter can be updated to develop an ontology of emotions in the context of improving understanding of the role of affect, context, and behavioral information in human moods, and is also one of the representative features of reliable information.

D. Arnfield (2023) analyzes the relative performance of multilayer perceptron, random forest, and multinomial naive Bayesian classifiers trained on the basis of "bag of words" and frequency-inverse dense frequency transforms of documents in Fake News Corpus and Fake and Real News Dataset. The author emphasizes that the aforementioned training included updating contextually categorized fake news to improve the effectiveness of the tools, with Fake News Corpus showing much better results than Fake and Real News Dataset.

Thus, our bibliometric analysis has shown that neural network modeling of the processes of recognizing linguistic markers of the categories of sense and absurdity in the context of working with false data is based on NLP [V. Derbentsev et al., 2023]. An important task of the latter is sentiment analysis, which is mostly studied using Deep learning (hereinafter – DL) models. In particular, based on Deep neural network (hereinafter – DNNs): Convolutional neural networks (hereinafter – CNNs), RNNs, their varieties – Long short-term memory (hereinafter – LSTM) with layers (hereinafter – CNNs-LSTM) and BiLSTM with CNNs layers (BiLSTM-CNNs).

Aim and objectives. *The purpose* of the article is to study the methodological features of neural network modeling of speech marker recognition processes. *The subject* is the specificity of the aforementioned phenomenon in the context of analyzing the categories of sense and absurdity as an innovative tool of linguistic science. In turn, the purpose and subject of the study allowed us to formulate its *objectives*:

1. To systematize theoretical achievements in the field of neural network modeling of language

processes and apply them to the recognition of linguistic markers of the categories of sense and absurdity.

2. Analyze the existing algorithms and methods of computer linguistics, ML, DS, etc., and develop a methodological framework for neural network modeling of the processes of recognizing language markers of the categories of sense and absurdity.

3. To present the work with linguistic markers of the categories of sense and absurdity of the neural network as a means to detect and prevent the spread of false data.

Methods. *The research design* has a clear structure, which is related to the stages of its implementation:

1. *Preprocessing (collection and preparation) of textual data (2022).* This stage includes updating the following ML methods and algorithms, in particular:

a) *vectorization* ("Bag of words", its use is related to the fact that the data is represented by the occurrence of words, but does not take into account their position in the above data, in turn, this allows you to record the number of occurrences of each word in the analyzed array), i.e. transformation of text data (text input) into vectors (vector representation), which can be potentially processed by an artificial neural network;

b) *tokenization* (instance of a sequence of characters in the part of the analyzed data set that is accumulated for use in semantic processing);

c) *normalization* (processing of text data to match keywords by fields labeled as "filtered", "aspectualized" or "sorted");

d) *cleaning the language data from non-alphabetic characters* (replacing all non-lexical characters with spaces, etc.);

e) *lemmatization* (the process of converting word forms into lemmas – their primary dictionary form, similar to the selection of the base of each lexical unit in a sentence: for example, for nouns and adjectives – nominative case, and for verbs, participles, adverbs – verbs in the infinitive);

e) *removal of stop words* (removal of articles, interjections, conjunctions, etc. that are not representative of the meaning), etc.

1.1. Formal verification of URLs with a neural network model (2023). Updating the fact-checker toolkit (data from Hostmaster (https://goo.su/NomD) and Who.Is (https://goo.su/SrJZRFM), which contains registration data of websites registered in Ukraine (Hostmaster) and the world (Who.Is)), which is a reliable basis for the proper operation of the artificial neural network.

1.2. Work on the principle of plagiarism detection programs (2023). Updating the descriptive and receptive methods (content definition), as well as the entire methodological potential of DS: *clustering* (allocation of homogeneous segments), *use of neural networks* (the process of using software to solve applied problems), *boosting* (a means of enhancing the detection of certain features), *trees* (an algorithmic way of representing data in the form of hierarchical relationships – branching), *NLP* (classification, content representation and sentiment analysis), *scraping* (conversion of non-relational data into relational data), *frameworks* (software tools to simplify and accelerate the achievement of the necessary planned program goals).

1.3. Checking the analyzed data for the presence of certain language phrases and constructions (2023). The method of analyzing scientific research was updated, which led to the search and analysis of scientific publications related to neural network modeling (in particular, language units).

1.4. Determination of the percentage of expression (partially using the data from 1.3., 2023). Methods of computer, mathematical, and corpus linguistics, combinatorial methods, structural methods, and the above ML and DS methods are updated.

1.5. Analysis of the authenticity of available multimedia objects (2023). Fact-checking methods for checking *images* (Google Images (https://goo.su/X3xmFA), TinEye (https://goo.su/oCk8) – tools for checking images; Image Edited? (https://goo.su/QsjMFO) – a tool for checking images for possible editing; ImgOps (https://goo. su/9toqLU) – a tool for comprehensive verification of images and other operations with them, etc.) and *videos* (YouTube DataViewer (https://goo.su/xrFbRVg) – a tool for verifying videos from YouTube; InVID (https://goo.su/07uY7FE) – a browser extension that allows you to frame an existing video and search for its original sources or the date of first publication using screenshots, etc.) were updated.

1.6. Study of the specifics of factuality in the analyzed textual data (2023). The descriptive and receptive methods (content definition), as well as all the above-mentioned methodological potential of DS, have been updated.

1.7. Checking the time of publication of a text (2023). The above-mentioned ML potential has been updated.

2. Process of preparation (presentation) of language data (2023). The ML potential has been updated, in particular, *the feature description of an object* (the features that are fed as input should be relevant to the label we get as output – the result of a neural network or a complex of neural networks: for example, morphological features will not be or are not sufficiently representative for lexicographic research, while they are relevant for semantics research) and *the selection of key features* (the decision of relevant features about the research goal is individualized since tracking correlations between certain features is an assumption of the developer's linguistic research, within which some features are suitable for prediction and others are not).

3. Choosing the architecture (type) of a neural network (2023). The methodological potential of ML is updated, in particular, artificial neural networks are used (for context research, it is advisable to use RNNs, for text understanding – CNNs or Transformer with attenuation, and to study the methodological features of neural network modeling of the processes of recognizing language markers of the categories of sense and absurdity in the context of working with false data, a combination of RNNs and CNNs will be productive).

4. Algorithm design process (2023). In this case, a productive type of ML is *learning with the teacher*, when an artificial neural network model receives a set of examples, each of which has the correct answer specified by the developer. Thus, the artificial neural network will be trained based on a data set with true and false information sources. It is noteworthy that ML is based on training data, which must also be updated using methods such as backpropagation and optimization algorithms (e.g., gradient descent, etc.) to improve the results.

5. The process of training the algorithm on available (prepared) data with its subsequent validation on them (2023). After the artificial neural network has been built, it needs to be trained on two samples: *training* (designed to calibrate the system's weights, when the required value is reached, it is advisable to determine *the loss function* (representing the difference between the expected and correct results of the neural network model), as it is representative of the error in the classification of markers) and *testing* (during which we validate the results of the neural network model, which can be determined using special metrics: accuracy, F1-mean, and confusion matrix, which represent the system's performance).

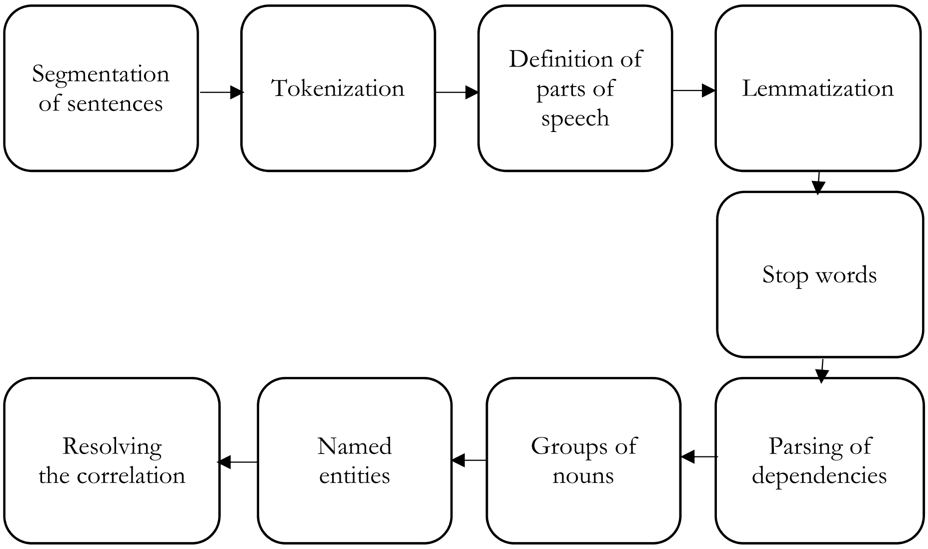
6. Interpretation of results (2023). The descriptive and receptive methods (content definition), as well as the entire methodological potential of DS, are updated: the identification of patterns, narratives, factors, etc. that represent the unreliability of textual data analyzed by an artificial neural network.

Results. The above makes it possible to position neural network modeling as a relevant tool for processing linguistic data, carried out to further update the latter in several areas of linguistic research (analysis, classification, structuring, and detection of hidden patterns in large volumes of heterogeneous complexly structured data) [M. Heimann, A. Hübener, 2023]. At the same time, the realization of any complex task, including the construction of methodological features of neural network modeling of linguistic markers of the categories of sense and absurdity, involves the construction of a pipeline (conveyor), which is shown in Scheme 1:

Thus, the pivotal stages of methodological approaches to neural network modeling of the processes of recognizing linguistic markers of the categories of sense and absurdity in the context of working with false data are as follows:

1. Preprocessing (collection and preparation) of text data (2022). During this stage, the linguist-developer processes the data set in a certain way, highlighting the aspects relevant to his/her research. This stage precedes the direct classification of speech data, consisting of the above-mentioned processes of vectorization, tokenization, normalization, cleaning of non-alphabetic characters, lemmatization, removal of stop words, etc. (see METHODS and Scheme 1).

In addition, at this stage, it is advisable to accumulate carefully filtered data, according to the researcher's algorithm. This stage is perhaps the most important since the ability to distinguish reliable



Scheme 1. Conveyor NLP

information from manipulated and false information correlates with the accuracy of the neural network model's results. Thus, we propose to implement in such a model the ability to search for data on the Internet, based on the principle of operation of most plagiarism-checking programs. The latter is based on analyzing a large amount of textual data and finding similarities with the text under analysis, while in our case, the neural network model will rely on the data we have prepared, comparing texts from the Web with them, forming a conclusion about their veracity based on the degree of coincidence with them, among other things:

1.1. Formal verification of URLs with a neural network model (2023). This means analyzing a particular source based on the peculiarities of its email address. Thus, we check the website address and the degree of coincidence with the list prepared by us (for example, in our opinion, relevant information for comparison is data from official sources: websites of ministries, departments, etc.)

1.2. Work on the principle of plagiarism detection programs (2023) will allow an artificial neural network *to compare the text under analysis with a similar text available on an official source*. Thus, the degree of similarity, which is a disadvantage in the case of analyzing a scientific text, will allow such data to be formally classified as true at this particular stage.

1.3. Checking the analyzed data for the presence of certain language phrases and constructions (2023). This means that a linguist developer prepares a list of linguistic phrases and constructions that, taken together, will allow the analyzed text to be positioned as false. First of all, we are talking about the degree of emotionality of the text (a number of emotional digressions, comments, the presence of interjections, exclamatory forms, etc.); the presence of interrogative structures that are not typical of the official style; the discursiveness of certain data in relation to the general interpretation; melodrama and artistry of the analyzed texts, etc. We also consider the language of the analyzed document to be the pivotal one (especially in the context of the hybrid Russian-Ukrainian war): it is said that the occupiers' manipulative practices are commonly based on the actualization of the russian language as a form for presenting transformed narratives by default, etc.

1.4. Determination of the percentage of expression (partially using the data from 1.3, 2023). Formal grounds for positioning the analyzed data as expressive are the presence of emotionally marked vocabulary, exclamatory and interrogative sentences, etc.

1.5. Analysis of the authenticity of available multimedia objects (2023). It is not exactly a linguistic component of verification, but it is effective in combination with textual data verification, as it allows you to position certain data as true/false with greater certainty.

1.6. Study of the specifics of factuality in the analyzed textual data (2023). This refers to the absence of specifics (names, places, dates, and other information) as a pivotal milestone of false information. In this case, it is advisable to include the search for such data in the algorithm of the neural network model.

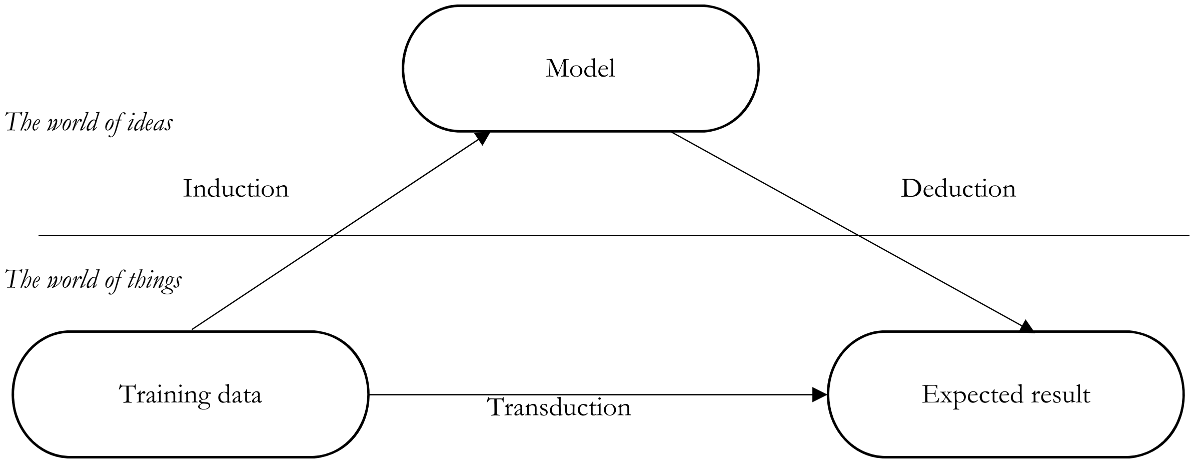
1.7. Checking the time of publication of a text (2023). A common manipulative practice is when official sources do not comment on a nighttime attack on Ukrainian cities, but certain sources (on the same night) publish information about it. That is why the time of publication in relation to the data provided for training the neural network model and the verified sources is a relevant feature, since usually data published from official (truthful) sources have approximately the same time of publication (for example, the next morning), while false data will be more differentiated in time.

2. Process of preparation (presentation) of language data (2023). It is a key stage of the neural network modeling process because the correctness of the algorithm and the ranking of text data on the truth/falsity scale depends on the representativeness of the data set containing sense and absurd language markers. Naturally, an artificial neural network needs to be trained on representative data that is fundamental in terms of covering the problem under study. It is noteworthy that the data representation correlates with a specific application task because for each of them, it is necessary to choose specialized methods that are productive for its solution. The most common methods of data representation in the process of training a neural network or ML (see Scheme 2) are *feature description of an object* and selection of key features.

3. Choosing the architecture (type) of a neural network (2023). This is the second most important stage after the data preparation (presentation) process because the success of neural network modeling depends on the choice of the tool for analysis, as well as its feasibility, functional characteristics, etc. That is why it is necessary to initially consider the type of artificial neural network, which is determined based on the typological and functional specifics. In particular, it is advisable to use RNNs for contextual research, CNNs or Transformer with attentional processing for text understanding, and a combination of RNNs and CNNs will be productive for studying the methodological features of neural network modeling of the processes of recognizing language markers of the categories of sense and absurdity in the context of working with false data.

The aforementioned artificial neural network models are productive for working with textual data, as they are able to track certain patterns, narratives, etc., and analyze sequences and correlations in the analyzed data. In addition, in the above modeling, lexical items are represented as points in hyperspace that correspond to certain senses, and sentence construction is positioned by moving along the above points in a multidimensional space. This is the aforementioned vectorization of language data, i.e. transformation of textual data (textual input) into vectors (vector representation) that can be potentially processed by a neural network.

4. Algorithm design process (2023). The choice of actualized methods correlates with the specific task of the research carried out by the developer-linguist: for example, to study the recognition



Scheme 2. Three types of ML

of linguistic markers of the categories of sense and absurdity, classification, clustering, regression, ranking, and genetic algorithms will be productive (despite the significant drawback of the latter in the form of the lack of induction (data analysis and model building on their basis) learning) [P. Kar, S. Debbarma, 2023] (see Diagram 1).

5. The process of training the algorithm on available (prepared) data with its subsequent validation on them (2023). After the artificial neural network has been built, it needs to be trained on two samples: *training* and *testing*. The above is intended to assess whether this neural network model is able to distinguish true from false data during text analysis.

6. Interpretation of results (2023). This stage is productive in identifying patterns, narratives, factors, etc. that represent the unreliability of the textual data analyzed by the artificial neural network. In turn, such information is useful for raising public awareness (in particular, in Ukraine and Europe) of potential russian manipulative practices, identifying false or distorted information, and assisting the above-mentioned communities with an automated tool for detecting them – a neural network.

Discussion. The problem of methodological features of neural network modeling of the processes of recognizing linguistic markers of the categories of sense and absurdity in the context of working with false data analyzed in this study is an integrated one. In turn, this leads to the controversial nature of its theoretical interpretation and practical implementation: undoubtedly, one of the main tasks of NLP is the recognition and classification of language categories (classification, content representation, and sentiment analysis) E. Park, V. Storey (2023), and the ML and DS tools – artificial neural network models – are best suited for the above task in general and for working with false data in particular.

Thus, in the study by S. Chung et al. (2023) revealed the existence of a gap between NLP itself and the practical implementation of its implementation, while M. Heimann, A. Hübener (2023) highlighted a number of problems with LLM, which is a core component of ML on large corpora of texts and is relevant to a number of text data processing tasks (generation, translation, paraphrasing, classification, etc.). The thesis is that, according to O. Nikula (2023), the most productive method for studying news is the use of NLG. At the same time, F. Guo (2023) considers fine-tuned models for classifying pairwise connections into trivial/low or moderate/high empirical connections to be a universal NLP tool.

M. Fazil et al. (2023) position BiLSTM-RNNs as such a tool, which updates existing methods of representing words in a multichannel environment with multiple filters with different kernel sizes to capture semantic relations in different windows. It is noteworthy that D. Arnfield (2023) proves in this case the effectiveness of multilayer perceptron, random forest, and multinomial naive Bayesian classifiers trained on a bag of words and frequency-inversion dense frequency transform terms of documents in Fake News Corpus and Fake and Real News Dataset.

At the same time, P. Kar, S. Debbarma (2023) argue for the relevance of a three-stage methodology

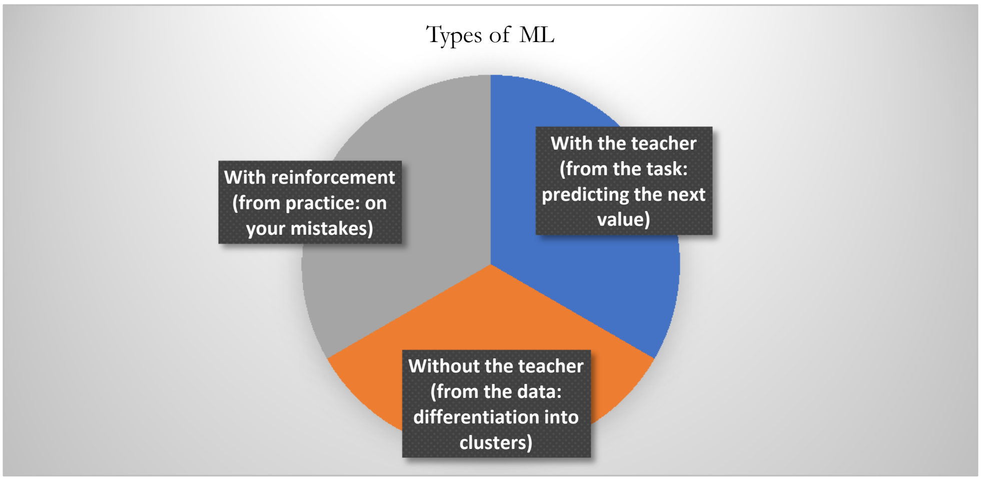


Diagram 1. Types of ML

based on FE-DGRNN: a) extracting multiple features from a text data set mixed with codes; b) optimizing the extracted features, which produces a preventive solution to the problem of data dimensionality at the stage of further detection; c) detecting hate speech and analyzing sentiment based on it. In addition, the typology of false data by S. Rastogi, D. Bansal (2023) is also quite controversial, as it seems incomplete (the latter can be explained by the rapid development of such data), and the study by Ș. Repede (2023) is too generalized due to certain methodological gaps.

Thus, neural network modeling of the processes of recognizing linguistic markers of the categories of sense and absurdity in the context of working with false data is based on NLP [V. Derbentsev et al., 2023]. An important task of the latter is sentiment analysis, which is mostly studied using Deep Learning (hereinafter – DL) models. In particular, on the basis of Deep neural network (hereinafter – DNN): CNNs, RNNs, and their variants – Long short-term memory (hereinafter – LSTM) with layers (hereinafter – CNN-LSTM) and BiLSTM with CNN layers (BiLSTM-CNN).

From the perspective of linguistic pragmatics, neural network models are able to recognize, analyze and process streaming speech, literary studies – to study the degree of expression and features of semantic representation (morphological, lexical, syntactic design of semantic cells), science – to determine the primary source (author) of text data, preventing academic dishonesty, etc. In addition, the above-mentioned artificial neural networks will significantly speed up scientific linguistic research (in particular, work on research topics of departments, divisions, etc.). This is especially true in the context of rapid technological development and the growth of research materials and the specifics of their processing.

That is why the actualization of artificial neural network models in text data mining is a modern and promising area of linguistic (and other) research. For instance, a thorough and comprehensive analysis of large corpora of texts has an impact on social, economic, educational, and other spheres. In addition, the use of neural network models in the process of processing data of various types and kinds (including text) will help improve the quality and accuracy of automatic translation and verification of the accuracy of the information presented. It is also about taking into account the peculiarities of recognizing language categories and the specifics of the process of distinguishing lexical items, semantics, etc. (in particular, in the case of large amounts of analyzed data).

The theoretical significance of this research will be a comprehensive study of the processes of recognizing linguistic markers of the categories of sense and absurdity in the context of working with false data using neural network modeling (in particular, the linguistic stylistic and lexical-semantic features of the construction of linguistic markers of the categories of sense and absurdity). A systematic approach to this problem makes it possible to generalize the practical principles of linguistic pragmatics and outline the prospects for studying the lexical and phraseological level and style of such texts. The results of the study are a contribution to computer and mathematical linguistics, data science. At the same time, they can serve as a basis for further development of the methodology of linguistic and contrastive studies, as well as the development of interdisciplinary projects that will allow for more frequent use of tools, methods, and theories of related subjects, deepening their effectiveness.

Practical value. The observations and results of the research can be used to study the methodology for recognizing linguistic markers of sense and absurdity in the context of working with false data; in comparative studies of European and Ukrainian strategies for countering disinformation; to analyze the effectiveness of ML in analyzing, identifying, and countering russian narratives; in studying the use of various strategies for countering disinformation in the context of ML and DS tools.

The main limitations of the research. Artificial neural network models can make mistakes, giving inaccurate results for certain tasks, since in order to work properly with such models, it is necessary to train them on a number of typical examples. This is despite the fact that the same artificial neural network models are also used to generate the aforementioned data, which leads to the latter's adaptability, accumulation speed, etc. Thus, the process of training an artificial neural network is quite lengthy and complex in preparing datasets (it is important to understand the variety of false data generated by the russian federation, which is also constantly changing).

As for the above-mentioned linguistic categories, in particular, the methodological features of neural network modeling of the processes of recognizing linguistic markers of the categories of sense and absurdity in the context of working with false data, artificial neural network models achieve greater accuracy. However, we are only talking about effective recognition and analysis of language categories, sentiment analysis, hate speech extraction, etc., but the truthfulness/falsity of data is more difficult to implement. For example, RNNs and a number of their modifications (LSTMs) and networks with reproducible attention (Attention) should be used for recognizing and classifying lexical items, syntactic units, and parts of speech.

Instead, neural network algorithms (e.g., Word2Vec, GloVe) and transformer models (e.g., BERT) should be used to work with semantics (in particular, sense). The former is capable of performing the aforementioned vectorization of textual data (transforming lexical units into vectors representing the input information). In turn, this allows us to study the peculiarities of semantic representation and relationships between lexical items, which is important in the context of studying the coherence and factuality of textual data. In addition, updating the above algorithms is productive for contextual processing of such data, which allows the researcher to visualize the network of correlations between lexical items, phrases, phrases, etc. Thus, this will allow us to implement most of the above steps directly and, indirectly, as one of the remaining ones.

Conclusions

Relevance. The full-scale invasion of Ukraine by the russian federation has led to a change in the political context in general and the information field in particular. First and foremost, this concerns the circulation of a large amount of false information (disinformation, misinformation, propaganda, fakes, and, in the future, deepfakes).

Narratives that do not contain facts, evidence, etc. have become a productive tool for actualizing the above-mentioned type of data in the information field. In turn, this makes them an important component of the hybrid war against Ukraine and an element of reflexive governance (influence on decision-making through the above manipulations). The above demonstrates the relevance of the study, namely, the need for modern tools for analyzing, tracking, counteracting, etc. false data in the form of neural network modeling of language marker recognition processes.

The obtained results demonstrate the potential of neural network modeling in recognizing linguistic markers of the categories of sense and absurdity in the process of working with false data in the context of NLP. This is because the above-mentioned artificial neural network models allow us to more thoroughly track the specifics of fluctuating changes in the intended sense: we are talking about a system of correlations between words, phrases, sentences, etc. In turn, the latter will facilitate the identification and study of linguistic categories in the context of linguistic pragmatics, in particular, the peculiarities of their existence in the environment of true/false data.

Main conclusions. The methodology of neural network modeling of the processes of recognizing the linguistic categories of sense and absurdity in the process of working with false data is a complex problem that has no clear solution. This is due to the existing poly-instrumentalism when the solution of a particular problem can be achieved in different ways (types of artificial neural networks, types of actualized layers, originality of the combination, specifics of the learning process, etc.) The presented study represents one of the ways to build this process, which the author considers promising for further work in the context of combating data manipulation and falsification.

The research is useful for use in the educational process: theoretical and special courses on ML, DS, NLP, text linguistics, lexicology, lexicography, comparative stylistics, language culture, and philosophy of language.

Prospects for further research lie in the fundamental analysis of the possibilities of using neural network modeling of the linguistic categories of sense and absurdity in the process of working with false data, in particular: a) to identify artifacts (anomalies) in text and multimedia data: thus, artificial neural networks can be trained to track such artifacts and correlations in data, with subsequent detection of deviations from the built templates of certain data, which is essential for this process; b) to generate the above artifacts (anomalies) in text and multimedia data: the process is the opposite of the first point, but productive for it, since the ability to produce artifacts produces the possibility of their more accurate localization.

REFERENCES

1. Arnfield, D. (2023). Enhanced Content-Based Fake News Detection Methods with Context-Labeled News Sources.

2. Chung, S., Moon, S., Kim, J., Kim, J., Lim, S., & Chi, S. (2023). Comparing natural language processing (NLP) applications in construction and computer science using preferred reporting items for systematic reviews (PRISMA). *Automation in Construction*, *154*, 105020.

3. Derbentsev, V. D., Bezkorovainyi, V. S., Matviychuk, A. V., Pomazun, O. M., Hrabariev, A. V., & Hostryk, A. M. (2023). A comparative study of deep learning models for sentiment analysis of social media texts. In *CEUR Workshop Proceedings* (pp. 168-188).

4. Fazil, M., Khan, S., Albahlal, B. M., Alotaibi, R. M., Siddiqui,T., & Shah, M. A. (2023). Attentional multi-channel convolution with bidirectional LSTM cell toward hate speech prediction. *IEEE Access*, *11*, 16801-16811.

5. Guo, F. (2023). *Revisiting Item Semantics in Measurement: A New Perspective Using Modern Natural Language Processing Embedding Techniques* (Doctoral dissertation, Bowling Green State University).

6. Heimann, M., & Hübener, A. F. (2023). Circling the Void: Using Heidegger and Lacan to think about Large Language Models.

7. Kar, P., & Debbarma, S. (2023). Multilingual hate speech detection sentimental analysis on social media platforms using optimal feature extraction and hybrid diagonal gated recurrent neural network. *The Journal of Supercomputing*, 1-32.

8. Nikula, O. (2023). Linguistic Feature Analysis of Real and Fake News: Human-written vs. Grover-written.

9. Park, E. H., & Storey, V. C. (2023). Emotion Ontology Studies: A Framework for Expressing Feelings Digitally and its Application to Sentiment Analysis. *ACM Computing Surveys*, *55*(9), 1-38.

10. Rastogi, S., & Bansal, D. (2023). A review on fake news detection 3T's: Typology, time of detection, taxonomies. *International Journal of Information Security*, *22*(1), 177-212.

11. Repede, Ș. E. (2023). Researching disinformation using artificial intelligence techniques: challenges. *Bulletin of "Carol I" National Defence University*, *12*(2), 69-85.

О. В. ДОВГАНЬ
кандидат філологічних наук,
докторант кафедри слов'янських, романських і східних мов,
Український державний університет імені Михайла Драгоманова, м. Київ, Україна
Електронна пошта: dovgan396@gmail.com
https://orcid.org/0000-0002-6728-818X

ОСОБЛИВОСТІ НЕЙРОМЕРЕЖЕВОГО МОДЕЛЮВАННЯ ПРОЦЕСІВ РОЗПІЗНАВАННЯ ЛІНГВІСТИЧНИХ МАРКЕРІВ КАТЕГОРІЙ СМИСЛУ ТА АБСУРДУ В КОНТЕКСТІ РОБОТИ З НЕПРАВДИВИМИ ДАНИМИ

Повномасштабне вторгнення російської федерації на територію України мало цілу низку наслідків геополітичної і внутрішньодержавної природи. Одним з проявів таких наслідків обох вищезазначених рівнів стала зміна сучасного інформаційного середовища під впливом актуалізації численних російських наративів. Закономірно, що наслідком функціонування таких неправдивих даних має стати ослаблення, вразливість українців до маніпуляцій, дезінформації тощо.

Останні репрезентовано у особливостях актуалізації вищезазначених даних з неправдивою, гіперболізованою, перекрученою тощо інформацією. Це, своєю чергою, зумовлює актуальність та нагальність проблеми їх побутування в контексті російської дезінформації, мізінформації пропаганди та фейків (а також дипфейків), а також вироблення інструментарію для їх аналізу задля попередження їх деструктивного впливу.

Йдеться про природу мовної полісистеми та Natural Language Processing (далі – NLP) як інструменти для обробки мовних даних з метою подальшого використання у процесі Machine Learning (далі – ML) та аналізу за допомогою штучної нейронної мережі. Це векторизація, токенізація, лематизація, мовленнєвий інтерфейс (Google Assistant та інші), автоматичний переклад (Google Translate, Reverso, DeepL тощо) та інші напрями. Природно, що вищезазначене дослідження набуває особливої актуальності ще й у контексті означеної гібридної природи російсько-української війни, в межах якої у медіапросторі (інтернет-дискурсі) функціонують цілі грона фейкових, викривлених, власне неправдивих та інших даних.

Саме тому у своєму дослідження ми проаналізуємо основні питання дотичні до проблеми методологічних особливостей нейромережевого моделювання процесів розпізнавання мовних маркері категорій смислу й абсурду,

звернувши увагу на їх природу, взаємовплив і взаємозумовленість, першорядність та соціокультурну значущість для українського суспільства. Окрім того, нами буде розглянуто специфіку вищезазначеного процесу в контексті роботи з неправдивими даними (дезінформацією), а також вплив останніх на його перебіг.

Ключові слова: лінгвістика, обробка природної мови, текстові дані, штучні нейронні мережі, машинне навчання, методологія нейромережевого моделювання.