DOI: https://doi.org/10.26565/2304-6201-2022-56-03

UDC 004.93

Malyha Ihor Yevheniyovych *postgraduate*
Kharkiv National University of V.N. Karazin,4 Svobody Square, Kharkiv, Ukraine, 61077
e-mail: igormalyga@gmail.com;
https://orcid.org/0000-0002-5708-7739

Shmatkov Serhiy Ihorovych *doctor of science, professor; Head of the Department of Theoretical Theoretical and Applied System Engineering*
Kharkiv National University of V.N. Karazin,4 Svobody Square, Kharkiv, Ukraine, 61077
e-mail: s.shmatkov@karazin.ua
Scopus Author ID: 57203141869

Machine learning methods for solving semantics and context problems in processing textual data

Topicality. As machine learning capabilities expand and impact many aspects of modern life, such as natural language processing, understanding semantics and context in textual data is becoming increasingly important. Semantics and context play a significant role in the ability of machines to understand human language. They are central elements in various applications such as machine translation, sentiment analysis, spam detection, voice recognition, and others. However, these aspects are often neglected or underestimated when processing textual data. Despite significant progress in this area, the problem of semantics and context remains unresolved, which reduces the efficiency and accuracy of many machine learning systems.

Goal: The main goal of this article is to investigate the problem of understanding semantics and context in machine learning in the textual data processing. The article aims to identify the main challenges associated with understanding semantics and context, and how they affect various aspects of text processing. Additionally, current techniques and approaches used in the field of machine learning for solving those problems have been analyzed and their limitations identified.

Research methods. Analysis, explanation, classification.

The results. It has been found that despite significant advances in machine learning technologies, problems of semantics and context in processing textual data are still existing. They affect the quality and accuracy of decisions made by machine learning based systems, which can lead to incorrect analysis and distortion of data. It has been found that even modern transformer-based models can face challenges in understanding semantics and context, especially in complex and multi-valued scenarios.

Conclusions. On the basis of the conducted research, it has been concluded that the problem of semantics and context in the processing of textual data is significant and requires further study. The existing methods and technologies show high results in some cases, but may be insufficient in others, especially complex ones. It is proposed to continue research in this area, to develop new methods and approaches that will be able to effectively solve these problems. It is also important to study how different contextual factors affect the semantics of textual data and how these effects can be taken into account when designing and using machine learning systems.

Keywords: *Machine learning, natural language processing, semantics, context, textual data, neural networks, transformers, BERT, GPT-3, data analysis, sentiment analysis, semantic analysis.*

How to quote: I. Malyha, and I. Shmatkov, "Machine learning methods for solving semantics and context problems in processing textual data." *Bulletin of V.N. Karazin Kharkiv National University, series "Mathematical modelling. Information technology. Automated control systems*, vol. 56, pp. 35-42, 2022. https://doi.org/10.26565/2304-6201-2022-56-03

1. Introduction

Machine learning is one of the fastest growing sectors of the technology industry, where textual data processing plays an important role. Applying machine learning to textual data, including sentiment analysis, emotion recognition, automatic translation, information retrieval, and more, opens up endless possibilities for advancing science, technology, and enterprise.

However, understanding and processing a text in machine learning faces several challenges, particularly in understanding semantics and context. Semantics, that is, understanding the meaning of a word or phrase, is one of the biggest problems in processing textual data. Intelligent systems often have difficulty in understanding that words can have different meanings in different contexts.

In addition, the problem of the context is also a big challenge for machine learning. Understanding how the meaning of a word can change depending on its context in the text is a complex task that machines are not yet able to perform effectively.

Given this, the relevance of those problems is topical. Solving them will improve the ability of machine learning to process textual data, paving the way for further progress.

2. Statement of the problem in a general form and its connection with important scientific or technical tasks. Review of topical publications.

Machine learning and artificial intelligence create many opportunities to transform large amounts of unstructured textual data into valuable information. However, applying these technologies to real world situations often leads to complex problems, especially in the context of natural language processing (NLP).

One of the main problems lies in semantic understanding. Artificial intelligence cannot yet fully comprehend the meaning of words, especially when used in different contexts. This is a natural for humans, but machines cannot yet sufficiently reproduce this ability.

The issue of context is also critical. Context can greatly affect the meaning of words or phrases, and it is difficult for machines to take this into account. For example, they may not understand cultural or historical context, which can lead to misinterpretation of textual data.

These issues limit the potential effectiveness and accuracy of machine learning algorithms that process textual data. Solving those problems will open new opportunities for the application of machine learning in various fields, including text analysis, automatic translation, text generation, and many others.

In the field of machine learning and natural language processing, the importance of semantics and context has been emphasized in a number of important studies.

First of all, a significant contribution to the study of semantics in machine learning is the work of Tomas Mikulov and his colleagues. They developed the word2vec model, which became one of the main methods of representing words in a vector space. The word2vec model uses large volumes of text for training aimed at learning high-dimensional vector representations of words that can represent semantic relations between them. However, this method still has limitations because it does not take into account the different meaning of a word in different contexts. This model uses neural network architectures to learn high-dimensional vector word representations from a large dataset. An important point is that Word2Vec can learn the semantic relations between words by displaying them in a vector space [1].

In addition, Bahdanau, Cho, and Bengio have made significant contributions to our understanding of how machines can better learn to recognize context. They proposed an attention model that allows focusing attention on specific parts of the input while generating the output. This provides more flexibility in context understanding and can help in various tasks including machine translation and automatic generalization [2].

In recent years, the BERT (Bidirectional Encoder Representations from Transformers) model, developed by Devlin and colleagues, has attracted special attention. BERT uses bidirectional transformers, which allows the model to better understand the context of a word because it analyzes the text in both directions. This made it possible to significantly improve the results in some NLP tasks, but at the same time new challenges arose, such as the large quantity of resources required for training the model and the weak ability to interpret [3].

We should also note the work of Vaswani et al., who developed the Transformer architecture for machine translation. This architecture, based on self-attention mechanisms, favored several previous approaches to semantics and context, including recurrent neural networks (RNNs) and convolutional neural networks (CNNs). However, Transformer also has its limitations, especially regarding the processing of long texts and taking into account the global context [4].

3. Highlighting previously unresolved parts of the general problem, to which the article is devoted, with justification of the relevance of the solution. Research of other authors on the topic.

The main theoretical approaches to understanding semantics and context in textual data processing are based on word embedding models and neural networks.

However, despite the success of the Word2Vec model, it struggled to cope with polysemy - cases where one word has several meanings depending on the context. This has led to the development of context-aware models, such as the GloVe (Global Vectors for Word Representation) model, which additionally uses statistical information from the word correlation matrix [5].

A more recent advance has been the development of models that can better understand the contextual information of a word. One such model is BERT (Bidirectional Encoder Representations from Transformers), proposed by Devlin. BERT uses bidirectional transformers, which allows the model to better understand the context of a word by analyzing the text in both directions.

These and other studies have made valuable contributions to understanding the challenges of semantics and context in machine learning. However, those problems are still open and require further research.

3. Statement of the task.

The purpose of the article is to analyze the problem of semantics and context when processing textual data in machine learning. The article highlights how relevant this problem is, how it affects the quality of machine learning systems, and what possible consequences may arise from underestimating this issue.

It is necessary to identify the main problem situations where semantics and context are crucial, and to analyze how effectively modern machine learning techniques and algorithms can cope with these cases.

The latest theories and practices in this area need to be reviewed to determine whether they can offer meaningful progress to the solution of this problem.

The final purpose of this paper is to formulate recommendations for further research in this area and ways that can improve the use of machine learning for semantic and context-aware processing of textual data.

To achieve this goal, it is necessary to solve the following tasks.

- To reveal the relevance and importance of the problem of semantics and context. To show how deep this problem is, it is necessary to identify the key problematic points that researchers and practitioners face when processing textual data in machine learning.
- Identify and analyze the main problem situations. It is intended to consider typical situations where the consideration of semantics and context is particularly important, and to analyze how effectively modern machine learning techniques and algorithms can cope with these cases.
- Overview of modern methods and techniques of textual data processing. This requires analyzing how existing techniques and technologies handle semantics and context to identify their potential weaknesses and opportunities for improvement.
- Identify possible directions for further research. Based on the analysis, it is planned to identify the key areas, in which the further research needs to be conducted, and formulate the practical recommendations.

4. Presentation of the main material with a full justification of the obtained scientific results.

Increasingly, machine learning-based systems depend on the efficient understanding and processing of textual data. However, there are significant challenges related to semantics and context that still require further research and development.

Polysemy, the problem of distinguishing different meanings of the same word in different contexts, is one of the most obvious challenges in this area. Standard natural language processing (NLP) models such as Word2Vec or GloVe, which are based on vector representations of words, do not take this problem into account. They create one vector for each word, regardless of the context in which it is used. Therefore, words with multiple meanings can be problematic [6].

Another problem that arises when processing textual data is the identification and processing of sarcasm and irony. This is difficult because sarcasm often requires a deep understanding of the context and situation in which it is used. A study conducted on the Twitter sources revealed that most machine learning models cannot effectively cope with this task [7].

Please note that the meaning of words may change depending on the context in which they are used. In particular, context can be related to culture, social conditions, era or even individual characteristics of the person using the language. This creates problems for machine learning models trying to learn the general rules of language, because these models may not take such variations in context into account.

That becomes particularly important when dealing with such phenomena as slang or dialects, where the use and meaning of words may differ significantly from those adopted in the standard language. For example, the study of Curtis and colleagues has shown that machine learning displays difficulties when trying to understand and generate texts written in African American Variants of English (AAVE) [8].

In addition, there are problems with processing implicit content such as metaphors, allegories, and other figurative expressions. These expressions usually require a deep understanding of the context and

cultural background to be correctly interpreted, which is beyond the capabilities of most modern machine learning systems.

The issue of processing texts that contain descriptive, emotional or subjective information is especially difficult. These types of information often require a deep understanding of context and personal experience, which is difficult to model in machine learning systems.

To solve these problems, a number of problems and technologies have been developed recently, namely.

- Using contextual vector representations of words. To solve the problem of polysemy, so-called contextual vector word representations are used in machine learning and NLP research. One such model is BERT (Bidirectional Encoder Representations from Transformers), which was introduced in 2018 by Google. It uses an attention mechanism that allows the model to look at the context from both sides of a word to determine its meaning.
- Using fine-tuning of models. Researchers have also begun to use a technique known as "fine-tuning" to adapt general machine learning models to specific tasks or domains. This may involve pre-training a model on a large dataset and then fine-tuning it on a smaller, more specialized dataset. This approach shows promising results in solving the problem of language variability in different contexts [9-10].
- Use of ensembles and multimodal models. Researchers also use ensemble techniques and multimodal models to combine different types of information and solve complex language processing problems. For example, a combination of text, audio and visual information can be used to better understand the context in a dialogue [11-13].
- Using deep learning to understand implicit content. Recent research shows that deep learning can help in understanding implicit content in texts, such as identifying sarcasm or irony. This can be achieved using complex neural network architectures, which include recurrent neural networks (RNNs), long-term memory networks (LSTMs), and the attention mechanism [14-16].
- Using transformative models for speech generation: Transformers such as OpenAI's GPT (Generative Pretrained Transformer) use the attention mechanism and other state-of-the-art techniques for high-quality text generation. These models can generate much more naturalistic text sequences compared to older models, but they also have their own challenges in terms of context processing and semantics [17-18].
- Semantic multi-task learning: Some researchers are working on using multi-task learning to solve multiple language processing problems simultaneously. This means that the model tries to perform several tasks at the same time, using a common semantic space. For example, the model can simultaneously try to classify emotions in the text, determine sentiment and recognize named entities [19].
- Generalized machine learning models: Some researchers are also working on generalized models that can adapt to a wide range of language processing tasks without the need for fine-tuning. These models use universal architectures and algorithms that can effectively handle a variety of language tasks and domains [20].

However, it should be noted that all these methods have their own limitations and challenges. For example, contextual models of dictionary representations often require enormous computing resources and data to train. Nevertheless, these approaches open up new possibilities for improving the processing of textual data in machine learning.

Research findings show that to improve understanding semantics and context, practical research can be conducted in the following direction.

- Selection of dataset. Selecting a corpus of text that includes a wide range of contexts and semantic interactions that machine learning models often encounter in the real world. This data set should include a variety of language styles, from academic texts to everyday spoken language, allowing us to evaluate how the models perform on a variety of tasks.
- Data pre-processing. Before training the models, several important data preprocessing steps must be performed. This includes removing noise (such as non-standard characters or URLs), using tokenization to break text into individual words, and using stemming or lemmatization to reduce words to their base form.

- Model training. To train the models, it is proposed to use several types of machine learning models to understand the semantics and context in our textual data. This includes lexical representation models such as Word2Vec and GloVe, contextual lexical representation models such as BERT and GPT-3, and transformative models for language generation.
- Assessment of models. Evaluation of the models will include the use of a set of metrics commonly used to evaluate the quality of machine learning in processing textual data. These can be, for example, accuracy, completeness (recall), precision, F1-score, or metrics considered in the context of tasks, such as BLEU for translation tasks or ROUGE for text generation tasks. Additionally, semantic-specific metrics such as the distance between word vectors in the representation space can be considered.
- Statistical analysis. After evaluating the models, it is necessary to conduct a statistical analysis of the results, using tests to compare the results of the models and determine the statistical significance of the difference between them. Tests such as Student's t-test, ANOVA, or non-parametric tests can be used for this, depending on the nature of our data.
- Processing of results. Finally, one can start processing and analyzing the results, interpreting them in the context of the study. It is necessary to compare the results with existing research and theories, and to determine what implications these results have for the field of machine learning and textual data processing.

5. Conclusions

In this article, the problems of semantics and context in textual data when using machine learning have been considered and analyzed. The article identified the main challenges associated with understanding semantics and context, and how they affect various aspects of text processing. In addition, the current techniques and approaches used in the field of machine learning to solve these problems and their limitations have been considered.

The analytical nature of this article allows us to better understand the current problems that arise when processing textual data in machine learning. The issues of semantics and context in processing textual data in machine learning are becoming increasingly relevant at the intersection of disciplines such as computer science and linguistics. This article analyzed those problems, examining them from different angles and trying to find possible ways to solve them.

From a practical point of view, the research is important for developers of machine learning systems. Challenges they may face have been identified, and directions for the further research that may help address these challenges are suggested. One such direction is the development of new machine learning methods and techniques that relies more on the semantics and context. Another direction may focus on finding effective ways to incorporate those aspects into the existing models.

REFERENCES

1. Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space.
   Available at: https://arxiv.org/abs/1301.3781.
2. Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural Machine Translation by Jointly Learning to Align and Translate.
   Available at: https://arxiv.org/abs/1409.0473.
3. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Available at: https://www.aclweb.org/anthology/N19-1423/.
4. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is All you Need. Available at: https://papers.nips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.
5. Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global Vectors for Word Representation. Available at: https://www.aclweb.org/anthology/D14-1162/.
6. Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed Representations of Words and Phrases and their Compositionality.

Available at: https://papers.nips.cc/paper/2013/hash/9aa42b31882ec039965f3c4923ce901b-Abstract.html.
7. Davidov, D., Tsur, O., & Rappoport, A. (2010). Semi-supervised recognition of sarcastic sentences in Twitter and Amazon.
Available at: https://aclanthology.org/W10-2914/.
8. Blodgett, S. L., Green, L., & O'Connor, B. (2018). Demographic Dialectal Variation in Social Media: A Case Study of African-American English.
Available at: https://aclanthology.org/D16-1120/.
9. Howard, J., & Ruder, S. (2018). Universal Language Model Fine-tuning for Text Classification. Available at: https://www.aclweb.org/anthology/P18-1031/.
10. Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. Available at: https://www.aclweb.org/anthology/N18-1202/.
11. Huang, P. S., He, X., Gao, J., Deng, L., Acero, A., & Heck, L. (2013). Learning Deep Structured Semantic Models for Web Search using Clickthrough Data.
Available at: https://posenhuang.github.io/papers/cikm2013_DSSM_fullversion.pdf.
12. Xu C., McAuley J., (2018). The Importance of Generation Order in Language Modeling. Available at: https://www.aclweb.org/anthology/D18-1324/.
13. Suzuki M., Matsuo Y., (2020). A survey of multimodal deep generative models. Available at: https://arxiv.org/abs/2207.02127.
14. Felbo, B., Mislove, A., Søgaard, A., Rahwan, I., & Lehmann, S. (2017). Using Millions of Emoji Occurrences to Learn Any-domain Representations for Detecting Sentiment, Emotion and Sarcasm. Available at: https://www.aclweb.org/anthology/D17-1169/.
15. Reyes A., Rosso P., (2016). Mining Subjective Knowledge from Customer Reviews: A Specific Case of Irony Detection. Available at: https://aclanthology.org/W11-1715.pdf.
16. Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., & Hovy, E. (2016). Hierarchical Attention Networks for Document Classification.
Available at: https://www.aclweb.org/anthology/N16-1174/.
17. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I. (2019). Better language models and their implications.
Available at: https://openai.com/blog/better-language-models/.
18. Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., ... Amodei, D. (2020). Language Models are Few-Shot Learners Available at: https://proceedings.neurips.cc/paper/2020/file/1457c0d6bf5478631ec67e564d04505b-Paper.pdf.
19. Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. R. (2019). GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. Available at: https://openreview.net/pdf?id=rJ4km2R5t7.
20. Lu, X., Xiong, C., Parikh, A. P., & Socher, R. (2019). ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-
Language Tasks. Available at: https://arxiv.org/abs/1908.02265.

ЛІТЕРАТУРА

1. Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space.
Available at: https://arxiv.org/abs/1301.3781.
2. Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural Machine Translation by Jointly Learning to Align and Translate.
Available at: https://arxiv.org/abs/1409.0473.
3. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Available at: https://www.aclweb.org/anthology/N19-1423/.

4. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is All you Need. Available at: https://papers.nips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.
5. Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global Vectors for Word Representation. Available at: https://www.aclweb.org/anthology/D14-1162/.
6. Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed Representations of Words and Phrases and their Compositionality. Available at: https://papers.nips.cc/paper/2013/hash/9aa42b31882ec039965f3c4923ce901b-Abstract.html.
7. Davidov, D., Tsur, O., & Rappoport, A. (2010). Semi-supervised recognition of sarcastic sentences in Twitter and Amazon.
Available at: https://aclanthology.org/W10-2914/.
8. Blodgett, S. L., Green, L., & O'Connor, B. (2018). Demographic Dialectal Variation in Social Media: A Case Study of African-American English.
Available at: https://aclanthology.org/D16-1120/.
9. Howard, J., & Ruder, S. (2018). Universal Language Model Fine-tuning for Text Classification. Available at: https://www.aclweb.org/anthology/P18-1031/.
10. Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. Available at: https://www.aclweb.org/anthology/N18-1202/.
11. Huang, P. S., He, X., Gao, J., Deng, L., Acero, A., & Heck, L. (2013). Learning Deep Structured Semantic Models for Web Search using Clickthrough Data. Available at: https://posenhuang.github.io/papers/cikm2013_DSSM_fullversion.pdf.
12. Xu C., McAuley J., (2018). The Importance of Generation Order in Language Modeling. Available at: https://www.aclweb.org/anthology/D18-1324/.
13. Suzuki M., Matsuo Y., (2020). A survey of multimodal deep generative models.
Available at: https://arxiv.org/abs/2207.02127.
14. Felbo, B., Mislove, A., Søgaard, A., Rahwan, I., & Lehmann, S. (2017). Using Millions of Emoji Occurrences to Learn Any-domain Representations for Detecting Sentiment, Emotion and Sarcasm. Available at: https://www.aclweb.org/anthology/D17-1169/.
15. Reyes A., Rosso P., (2016). Mining Subjective Knowledge from Customer Reviews: A Specific Case of Irony Detection.
Available at: https://aclanthology.org/W11-1715.pdf.
16. Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., & Hovy, E. (2016). Hierarchical Attention Networks for Document Classification.
Available at: https://www.aclweb.org/anthology/N16-1174/.
17. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I. (2019). Better language models and their implications.
Available at: https://openai.com/blog/better-language-models/.
18. Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., ... Amodei, D. (2020). Language Models are Few-Shot Learners Available at: https://proceedings.neurips.cc/paper/2020/file/1457c0d6bf5478631ec67e564d04505b-Paper.pdf.
19. Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. R. (2019). GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. Available at: https://openreview.net/pdf?id=rJ4km2R5t7.
20. Lu, X., Xiong, C., Parikh, A. P., & Socher, R. (2019). ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-
Language Tasks. Available at: https://arxiv.org/abs/1908.02265.

Малига Ігор Євгенійович *аспірант*
Харківський Національний Університет ім. В.Н. Каразіна, площа Свободи 4, Харків, Україна, 61077
e-mail: igormalyga@gmail.com;
https://orcid.org/0000-0002-5708-7739

Шматков Сергій Ігорович *д.т.н., професор; завідуючий кафедри теоретичної теоретичної та прикладної системотехніки*
Харківський Національний Університет ім. В.Н. Каразіна, площа Свободи 4, Харків, Україна, 61077
e-mail: s.shmatkov@karazin.ua
Scopus Author ID: 57203141869

Методи машинного навчання для вирішенні проблем семантики та контексту при обробці текстових даних

Актуальність. З розширенням можливостей машинного навчання та його впливом на багато аспектів сучасного життя, включаючи обробку природної мови, розуміння семантики та контексту в текстових даних стає все більш актуальним. Семантика та контекст відіграють значну роль у здатності машин розуміти людську мову. Вони є центральними елементами в різних програмах, таких як машинний переклад, аналіз настроїв, виявлення спаму, розпізнавання голосу тощо. Однак цими аспектами часто нехтують або недооцінюють під час обробки текстових даних. Незважаючи на значний прогрес у цій галузі, проблема семантики та контексту залишається невирішеною, що знижує ефективність і точність багатьох систем машинного навчання.

Мета: Основна мета цієї статті — дослідити проблему семантики та контексту в машинному навчанні, а саме в обробці текстових даних. Стаття має на меті визначити основні проблеми, пов'язані з розумінням семантики та контексту, а також те, як вони впливають на різні аспекти обробки тексту. Крім того, буде проаналізовано поточні методи та підходи, які використовуються в галузі машинного навчання для вирішення цих проблем, і визначено їх обмеження.

Методи дослідження. Аналіз, пояснення, класифікація.

Результати. Було встановлено, що незважаючи на значні досягнення в технологіях машинного навчання, проблеми семантики та контексту в обробці текстових даних все ще існують. Вони впливають на якість і точність рішень, що приймаються системами на основі машинного навчання, що може призвести до некоректного аналізу та спотворення даних. Було виявлено, що навіть сучасні моделі на основі трансформаторів можуть зіткнутися з проблемами розуміння семантики та контексту, особливо в складних і багатозначних сценаріях.

Висновки. На основі проведеного дослідження зроблено висновок, що проблема семантики та контексту при обробці текстових даних є суттєвою та потребує подальшого вивчення. Існуючі методи і технології, хоч і показують високі результати в одних завданнях, можуть виявитися недостатніми в інших, особливо складних, ситуаціях. Пропонується продовжити дослідження в даному напрямку, розробити нові методи та підходи, які б змогли ефективно вирішити ці проблеми. Також важливо вивчити, як різні контекстуальні фактори впливають на семантику текстових даних і як ці ефекти можна врахувати при проектуванні та використанні систем машинного навчання.

Ключові слова: *Машинне навчання, обробка природної мови, семантика, контекст, текстові дані, нейронні мережі, трансформатори, BERT, GPT-3, аналіз даних, аналіз настроїв, семантичний аналіз.*

Як цитувати: І. Є. Малига, С. І. Шматков. «Методи машинного навчання для вирішенні проблем семантики та контексту при обробці текстових даних». *Вісник В.Н. Каразіна Харків Національний університет, серія "Математичне моделювання. Інформаційні технології. Автоматизоване управління системи*, вип. 56. с.35-42, 2022.
https://doi.org/10.26565/2304-6201-2022-53-03

Надійшла у першій редакції 12.08.2022, в останній - 11.11.2022.