UDC 004.4 DOI https://doi.org/10.35546/kntu2078-4481.2025.2.2.39

R. S. SAVITSKYI
Postgraduate Student,
Senior Lecturer at the Department of Software Engineering
Zhytomyr Polytechnic State University
ORCID: 0000-0001-9804-3604

A. YU. MIANOVSKA
Master's Student at the Department of Software Engineering
Zhytomyr Polytechnic State University
ORCID: 0009-0002-9904-3002

# INTEGRATING NLP LIBRARIES AND PRE-TRAINED MODELS FOR AUTOMATED TEXT COMPLEXITY ASSESSMENT

*The rapid development of intelligent multi-agent systems has significantly influenced the field of Natural Language Processing (NLP), enabling more efficient and scalable text analysis methods. This study explores a modular approach to NLP, emphasising the integration of specialised libraries and models to enhance text processing capabilities. Modern NLP solutions streamline tasks like text complexity assessment by leveraging machine learning (ML) and deep learning techniques, particularly through pre-trained models and structured linguistic pipelines. Such integration allows for addressing not only syntactic and lexical aspects but also deeper semantic relationships within texts. The modular system proposed in this research allows for the seamless combination of various NLP components, ensuring adaptability to different analytical tasks. The study focuses on widely used NLP tools, including SpaCy for linguistic analysis and BERT (Bidirectional Encoder Representations from Transformers) for deep contextual understanding. The approach integrates traditional linguistic analysis with advanced neural network-based models to facilitate the assessment of text difficulty. The research presents a systematic approach to categorising texts of varying complexity, ranging from essential children's stories to advanced legal documents.*

*The research underscores the effectiveness of modular NLP systems in addressing the growing demand for automated text analysis. Combining structured linguistic features with deep-learning-based contextual embeddings enables accurate classification of text complexity, facilitating language learning and computational text processing applications. A key advantage of the modular approach is its flexibility and scalability, allowing researchers and developers to integrate customised NLP solutions for diverse applications. By adopting a modular approach, NLP continues to evolve, providing scalable and adaptable solutions for the ever-increasing challenges of text analysis.*

***Key words:*** *modular approach, natural language processing, text analysis, text complexity, text classification, pre-trained model.*

Р. С. САВІЦЬКИЙ
аспірант,
старший викладач кафедри інженерії програмного забезпечення
Державний університет «Житомирська політехніка»
ORCID: 0000-0001-9804-3604

А. Ю. М'ЯНОВСЬКА
студентка кафедри інженерії програмного забезпечення
Державний університет «Житомирська політехніка»
ORCID: 0009-0002-9904-3002

# ІНТЕГРАЦІЯ БІБЛІОТЕК NLP ТА ПОПЕРЕДНЬО НАВЧЕНИХ МОДЕЛЕЙ ДЛЯ АВТОМАТИЧНОЇ ОЦІНКИ СКЛАДНОСТІ ТЕКСТУ

*Стрімкий розвиток інтелектуальних мультиагентних систем суттєво вплинув на сферу обробки природної мови (NLP), дозволяючи більш ефективні та масштабовані методи аналізу тексту. Це дослідження досліджує модульний підхід до NLP, наголошуючи на інтеграції спеціалізованих бібліотек і моделей для покращення можливостей обробки тексту. Сучасні рішення NLP спрощують такі завдання, як оцінка складності тексту, використовуючи методи машинного навчання (ML) і глибокого навчання, зокрема, за допомогою попередньо навчених моделей та структурованих лінгвістичних конвеєрів. Така інтеграція дозволяє враховувати не лише синтаксичні та лексичні аспекти, але й глибші семантичні зв'язки в текстах. Модульна система, запропонована в цьому дослідженні, дозволяє бездоганно поєднувати різні компоненти NLP, забезпечуючи адаптивність до різних аналітичних завдань. Дослідження зосереджено на широко використовуваних інструментах NLP включаючи SpaCy*

*для лінгвістичного аналізу та BERT (Bidirectional Encoder Representations from Transformers) для глибокого розуміння контексту. Підхід поєднує традиційний лінгвістичний аналіз із передовими моделями на основі нейронних мереж, щоб полегшити оцінку складності тексту. Дослідження представляє системний підхід до класифікації текстів різної складності, починаючи від простих дитячих історій і закінчуючи складними юридичними документами.*

*Дослідження підкреслює ефективність модульних систем NLP у вирішенні зростаючого попиту на автоматизований аналіз текстів. Поєднання структурованих лінгвістичних характеристик із контекстними вбудовуваннями, заснованими на глибокому навчанні, забезпечує точну класифікацію складності текстів, що сприяє їх використанню у вивченні мов та комп'ютерній обробці текстів. Ключовою перевагою модульного підходу є його гнучкість і масштабованість, що дозволяє дослідникам і розробникам інтегрувати індивідуальні рішення NLP для різних застосувань. Завдяки впровадженню модульного підходу NLP продовжує розвиватися, пропонуючи масштабовані та адаптивні рішення для дедалі складніших завдань аналізу тексту.*

***Ключові слова:*** *модульний підхід, обробка природної мови, аналіз тексту, складність тексту, класифікація тексту, попередньо навчена модель.*

**Problem statement**

In the contemporary world, the volume of textual information is rapidly growing, creating challenges for efficient and meaningful analysis. In domains such as education, business, and social sciences, the ability to process large amounts of unstructured textual data has become a critical task. The increasing fragmentation and inconsistency of data further complicate comprehension and interpretation.

Manual processing of such vast data sets often exceeds human capabilities, making automation essential. Technologies such as Artificial Intelligence (AI), Machine Learning (ML), and Natural Language Processing (NLP) offer powerful solutions for text processing, classification, key information extraction, and the generation of analytical conclusions. However, the growing diversity and complexity of digital content introduce new obstacles, such as the lack of modularity and adaptability in current NLP tools, which hinder scalability and reuse across diverse tasks.

There is a clear need for modular systems that simplify the development of NLP workflows, ensure interoperability, and adapt to various use cases such as sentiment analysis, text summarisation, and complexity assessment. The problem of automating the evaluation of text complexity is particularly relevant, as subjective human judgment is difficult to scale and standardise.

**Analysis of the latest research and publications**

With the exponential rise of AI and digital content, where the internet is rich with millions of texts and diverse content, Flayeh A. K. et al. (2022) investigated applying NLP techniques for analysing large volumes of text data. They sought to solve the problem of locating hidden information in enormous text databases using artificial intelligence methods [1].

Xipeng Q. et al. (2020) conducted a comprehensive survey on pre-trained models (PTMs) and their impact on advancing NLP, aiming to provide a practical guide for understanding, utilising, and improving PTMs in deep learning and neural networks [2].

Shikun K. et al. (2024) researched the robustness of the Latent Dirichlet Allocation (LDA) model, a widely used Bayesian algorithm for text analysis, by addressing its sensitivity to the choice of prior distributions. The authors demonstrated that the model's parameters are not fully identified. They developed two algorithms to quantify how changes in priors affect posterior means. The approach was applied to study the effects of increased transparency on U.S. monetary policy discussions, showcasing the importance of robust modelling in text analysis [3].

The research increasingly seeks to bridge qualitative and quantitative approaches; Parks L. and Peters W. (2022) explored integrating NLP techniques with qualitative methods. They created a mixed-methods text analysis workflow to address the challenge of balancing depth and breadth in analysing diverse text datasets. The method focuses on the iterative interaction between manual analysis and computational tools to gain deeper insights through interdisciplinary dialogue [4].

In the context of growing challenges in analysing unstructured text, Roldan-Baluis et al. (2022) conducted a systematic review to explore the impact of NLP techniques and ML algorithms on addressing societal issues. By analysing articles published in SCOPUS in 2021, they identified Term Frequency-Inverse Document Frequency (TF-IDF) as the most utilised NLP technique, Support Vector Machines (SVM) as the leading supervised learning algorithm, and Long Short-Term Memory (LSTM) as the predominant deep learning approach [5].

Benicio et al. (2022) developed a tool leveraging Text Mining and NLP techniques to transform unstructured free-text data into a structured format. They conducted a comparative analysisThey conducted a comparative analysis between data manually retrieved by healthcare professionals and data processed by their tool, with statistical evaluation using the Kruskal-Wallis test. Results demonstrate no significant difference between the manual and automated methods, showcasing the tool's effectiveness in facilitating accurate data retrieval from unstructured medical records [6].

Fate V. D. (2021) investigated how text analytics can help derive valuable insights from various types of textual data, such as support tickets, medical records, and product descriptions. The study outlines the difficulties of working with

unstructured data, as it lacks consistent formats or structures. Text analytics will provide a new dimension to knowledge management and many applications across the spectrum, for instance, from automatic systems to chatbots. The research emphasises the growing significance of text analytics regarding the challenges posed by modern data architectures [7].

**Purpose of the article**

The purpose of this article is to develop and demonstrate a modular system for automated text complexity analysis using advanced Natural Language Processing (NLP) technologies. Given the exponential growth of digital textual content and the limitations of manual processing, the proposed approach focuses on building scalable, flexible, and adaptable tools that support the structuring, classification, and interpretation of unstructured text data. The emphasis is placed on combining pre-trained models and modular components to enhance the automation and accuracy of text analysis workflows.

This study aims to:

1. Explore modularity as a design principle for developing flexible and reusable NLP-based systems.
2. Analyse the capabilities of existing NLP libraries in tasks such as text preprocessing, linguistic analysis, and feature extraction.
3. Assess the effectiveness of pre-trained models in preserving semantic context and supporting complex tasks like classification and semantic analysis.
4. Demonstrate how modular NLP components can be integrated into a unified framework to streamline development, ensure interoperability, and facilitate system scalability.

**Presentation of the main material**

With the rapid development of AI technologies, Natural Language Processing has become one of the most rapidly evolving branches of modern information technology. NLP enables computers to understand and interpret human language, either fully automatically or with some human intervention [8]. The field is multidisciplinary and closely related to computer science and linguistics. NLP employs various methods, ranging from sophisticated machine learning algorithms, notably deep learning, to more conventional rule-based approaches, enabling systems to perform tasks such as syntactic parsing, named entity recognition, and text summarization.

Computers achieving human-like proficiency in language processing would mark the advent of intelligent machines. This belief stems from the idea that language use is deeply tied to human cognitive abilities [9]. As researchers develop increasingly sophisticated models, the gap between human and machine language understanding narrows. Opportunities are unlocking innovation in education and other fields by empowering engines to respond to natural language effectively.

Conventionally, the text analysis process in NLP is divided into several stages. One of the key components is syntax, which defines how words are arranged to form grammatically correct sentences [8]. This stage ensures the structural integrity of language by considering grammatical rules and the relationships between words. Another crucial aspect is semantics, which focuses on how words and phrases convey meaning, often represented in NLP through vector semantics [9]. Finally, pragmatics examines meaning in context, determining how language is interpreted based on situational factors [8].

A straightforward text analysis method starts by looking at the syntactic structure of phrases to create a structured framework that may be used to interpret literal or semantic meaning. A pragmatic analysis step, which determines the text's or utterance's context-specific meaning, usually comes next. While syntax and semantics focus more on sentence-level specifics, pragmatic analysis frequently concentrates on discourse-level comprehension. The difference between them can occasionally make it challenging to comprehend how natural language processing works [10].

The tripartite division into syntax, semantics, and pragmatics is a preliminary framework for analysing the processing of natural language texts. In addition to the previously established stages, we can introduce two earlier phases to refine the process further. These are tokenisation, which involves breaking the surface text into smaller, manageable units, and lexical analysis, where individual words or tokens are identified and categorised based on linguistic properties [10]. These initial steps provide a foundation for more advanced analysis in NLP.

Defining fundamental units like characters, words, and sentences is essential in the textual analysis of digital natural language texts. However, the complexity of this task is compounded by the diversity of languages, writing systems, and document sources. Resolving these ambiguities is a core focus of NLP [10].

Each piece of collected text data should have a distinct identifier for easy reference. These pieces are called documents in text analytics, usually consisting of several characters. Multiple documents together form a document collection or corpus. The growth in corpus size and variety has driven the development of automated techniques for harvesting and preparing text for NLP tasks. Characters are grouped to create words or terms in a particular language, and these words are the central focus of our analysis [12].

The first step in any NLP pipeline is text preprocessing, which involves tokenisation, stemming and lemmatisation. Libraries streamline these tasks, allowing developers to focus on higher-level objectives rather than reinventing these basic operations. Text preprocessing ensures that the input data is clean, standardised, and ready for advanced NLP tasks. One main factor for text preprocessing is eliminating noise and redundant text information, for example, special characters,

stop words, and punctuation. The practice helps reduce the data's dimensionality and, at the same time, improves the efficiency of subsequent analysis.

The following example elucidates the primary purpose. The raw data is "Artificial intelligence and machine learning are transforming industries.» The processed sentence is ['artificial', 'intelligence', 'and', 'machine', 'learning', 'be', 'transform', 'industry'].

Text preprocessing takes raw text as input and produces cleaned tokens as output. Tokens are individual words or phrases that are counted based on word frequency and act as the key features for analysis. The flow of characters in a natural language text should be partitioned into distinct, meaningful tokens before any language processing can be performed. Perfect punctuation would make tokenisation straightforward, but real languages are imperfect.

The utilisation of data in machine learning (ML) models is a process proven to yield efficient outcomes and facilitate comprehension. The process encompasses a series of methodologies employed to transform, clean and develop raw text data into a format appropriate for tasks involving NLP or ML. The objective of text preprocessing is to enhance the quality and usability of text data for subsequent analysis or modelling.

Since NLP relies upon sophisticated computational skills, developers require optimal tools to use NLP approaches and algorithms to produce services that can manipulate natural languages. Developers can use ready-made tools that facilitate text preprocessing, allowing them to concentrate more on building robust ML models.

NLP libraries are indispensable for automating tasks since they offer practical algorithms and pre-made solutions. They help transform raw text into meaningful insights, handling tasks like text preparation and building models to analyse language interactions.

Due to its extensive library and frameworks, which allow developers to design sophisticated language-driven applications quickly, Python is one of the most popular languages for NLP [11]. Its ease of usage will enable developers to quickly test and adjust various strategies, making it perfect for ML tasks where experimentation and rapid prototyping are key. Advancements in NLP tools and libraries have greatly improved text analysis and language understanding, making complex linguistic tasks more accessible.

The growing demand for efficient and scalable NLP tools has led to the development of libraries that balance performance and ease of use. One such solution is SpaCy, a high-performance NLP library designed for real-world applications. SpaCy is designed to bridge the gap between industrial applications and academic research in NLP. Its speed, pre-trained features, and user-friendly design tackle common challenges in modern text analysis.

One of SpaCy's key strengths is its ability to provide high-performance NLP processing while maintaining ease of use. It offers an optimised tokenisation system that efficiently segments text into words or sentences, enabling streamlined analysis [9]. Unlike some earlier libraries, SpaCy includes robust support for stemming and lemmatisation, reducing words to their root forms while preserving contextual accuracy [10].

Another advantage of SpaCy is its built-in part-of-speech (POS) tagging, which helps identify grammatical roles within a sentence, supporting syntactic analysis and sentence structure comprehension [13]. Additionally, SpaCy enables parsing techniques that analyse sentence grammar in depth, facilitating detailed linguistic evaluation [8]. The library also excels in Named Entity Recognition (NER), leveraging pre-trained models to identify and categorise entities such as individuals, organisations, and locations [10]. This feature allows for more practical information extraction and data-driven decision-making. Furthermore, SpaCy employs advanced dependency parsing techniques to determine relationships between words in a sentence, enhancing grammatical understanding and semantic analysis [7].

SpaCy's modular pipeline architecture promotes customisation. Users can add their processing components to specialised tasks such as sentiment analysis or domain-specific preprocessing. This feature ensures that SpaCy is a standalone tool and part of a broader ML or NLP workflow.

NLP models have become central to many AI applications, helping businesses and researchers understand and generate human language. They are widely used for tasks such as automatic translation, chatbot interactions, and even content creation. Over the years, advancements in ML and neural networks have propelled the evolution of NLP models, leading to more accurate and efficient language understanding systems. A key breakthrough in this evolution has been the development of pre-trained models, eliminating the need to build and train models from scratch.

Deep learning models trained on large-scale datasets to perform specific NLP tasks are known as pre-trained models for NLP [11]. Language models are trained on immense amounts of data, enabling them to achieve accuracy that is sensitive to the context. One of the reasons deep learning has become so effective is its layered architecture, which mimics the hierarchical nature of language processing. The initial layers process raw text by transforming words into numerical representations, often using embeddings, while deeper layers analyse syntactic dependencies and semantic meaning. By leveraging multiple processing layers, models continuously refine their understanding of human language, improving contextual awareness [14].

Once trained, a pre-trained language model can be applied to various tasks, like answering questions or summarising content, without needing to start from scratch every time. This saves effort since they do not always require large amounts of labelled data or custom training for each new project.

Pre-trained language models provide a flexible and scalable way to handle NLP tasks. They can work across various formats like text, images, and audio, and their ability to adapt through fine-tuning or tailored prompts has significantly improved the field. Pre-trained models have become essential to modern NLP workflows, driving innovation and improving efficiency in applications such as translation tools and summarization systems. Their ability to generate high-quality translations and concise summaries with minimal human intervention has made them indispensable in automating language processing tasks [15].

NLP models are essential for text analysis because they offer efficient and accurate solutions for processing human language. Pre-trained models, in particular, are highly effective as they come pre-trained on extensive datasets, saving time and computational resources compared to building models from scratch. One key advantage of using models is their adaptability; they can be fine-tuned for specific analytic tasks.

BERT has emerged as a groundbreaking model in the field of NLP. Developed by Google in 2018, BERT introduced a paradigm shift by enabling models to deeply understand the context of words in a sentence by processing them bi-directionally. Unlike traditional NLP models that read text in one direction, either left-to-right or right-to-left, BERT considers the entire context, making it robust for understanding complex linguistic nuances [9].

The model is trained on a large corpus using unsupervised learning tasks, such as Masked Language Modelling (MLM) and Next Sentence Prediction (NSP). MLM helps BERT predict masked words in a sentence, while NSP trains it to understand relationships between sentence pairs. After pre-training, BERT is fine-tuned on explicit tasks like text classification or question answering using smaller, labelled datasets [9]. BERT's remarkable success stems from training on an extensive dataset comprising 3.3 billion words. The dataset includes 2.5 billion words from Wikipedia and 800 million from the Google BooksCorpus. Vast sources of information have significantly enhanced BERT's deep comprehension of the English language.

In analysing language difficulty, BERT has been shown to model sentence complexity by identifying lexical richness, grammatical structures, and inter-sentence dependencies. BERT is an excellent tool for tasks such as evaluating educational materials, grading essays and adapting content for varying reading levels [15]. In evaluating the complexity of English texts, the BERT model has demonstrated its capacity to analyse patterns in sentence construction and vocabulary usage to generate insights about readability. It can also adapt to niche domains when combined with other models or datasets, enhancing its reliability and interpretability.

BERT's role in NLP extends beyond mere complexity evaluation. As part of foundational NLP methodologies, it serves as a blueprint for modern models that integrate nuanced textual understanding with large-scale applications, offering immense potential for research and industry.

The modular approach to text complexity analysis leverages the synergy between advanced NLP models and libraries. It decomposes the task into independent modules – each addressing a specific aspect of text complexity – allowing for flexibility, scalability, and precise customisation. The proposed modular system combines SpaCy for traditional NLP preprocessing and feature extraction with BERT for contextual understanding.

SpaCy is a fast and flexible tool for preparing text for NLP tasks. Some of its key features include POS identifying distributions, building dependency trees, and analysing sentence structures. These capabilities create a structured linguistic representation of the text, ensuring that subsequent modules can effectively perform deeper contextual analysis and more advanced NLP tasks. By segmenting the text into manageable components, SpaCy ensures that other modules can work with well-organised data. One standout feature of SpaCy is its ability to integrate with transformer models like BERT through the spacy-transformers module, combining traditional linguistic analysis with modern contextual embeddings.

In a modular system, BERT takes on two primary roles. Firstly, it generates detailed representations of words and sentences by analysing context, offering a deep understanding of linguistic features. Secondly, it can be trained on datasets that align with CEFR levels, from A1 to C2. This additional training allows BERT to categorise text complexity with high effectiveness, aligning its assessments with widely recognised language proficiency standards. BERT's bidirectional approach to understanding context makes it particularly good at analysing complex sentence patterns. Its multilingual support is also suitable for comparing texts in different languages or working with translations.

Determining text complexity is critical in language learning and education. Combining SpaCy with BERT provides a robust framework for classifying text based on CEFR levels. By leveraging pre-trained embeddings and fine-tuning them on annotated datasets, texts can be classified based on syntactic, lexical, and semantic features.

The modularity of the NLP-based text classification system lies in the seamless integration of SpaCy's preprocessing pipeline with BERT's contextual embeddings. Preprocessed linguistic features from SpaCy are combined with BERT embeddings to form a comprehensive feature set. These features are then used in a downstream classifier to assign CEFR levels to input texts. SpaCy and BERT collaborate to classify the complexity of three distinct text types – children's stories, exam passages, and legal documents – into appropriate CEFR levels.

Children's stories often feature short, simple sentences, basic vocabulary, and a repetitive structure that reinforces principal concepts for beginner readers. SpaCy's syntactic parsing reveals that declarative phrases with few dependent clauses are more common. POS tagging identifies nouns, verbs, and adjectives characteristic of A1-level texts.

Furthermore, SpaCy computes lexical density and sentence length, confirming that they match beginner-level proficiency. BERT captures the semantic simplicity of children's stories by embedding their contextual patterns. For example, repetitive phrases are analysed to detect the straightforward semantic relationships and literal expressions indicative of early language acquisition stages. BERT excels at distinguishing A1 texts by recognising these characteristic patterns.

Exam passages combine a variety of syntactic patterns with a moderate amount of lexical variation. SpaCy's dependency parsing uncovers compound sentences and intermediate-level grammatical constructions, while its POS tagging reveals balanced distributions of academic nouns, verbs, and connectors. Additionally, SpaCy calculates lexical sophistication metrics, identifying less frequent but general-use terms. BERT evaluates the semantic and contextual properties of exam passages. By analysing sentence embeddings, it identifies the academic tone and multi-word expressions. Contextual understanding allows BERT to categorise texts at B1-level, reflecting intermediate proficiency requirements.

Legal documents exhibit high levels of syntactic complexity, formal register, and specialised terminology. SpaCy's syntactic parsing identifies nested clauses, conditional statements, and other markers of complex syntax. Its vocabulary analysis flags rare and domain-specific terms, while text statistics demonstrate above-average sentence length and lexical density. These characteristics strongly correlate with C2-level texts. BERT has been developed to address the semantic complexities of legal texts by embedding the relationships between abstract phrases and technical terms. For example, BERT can capture some terms' formal tone and intent since they are contextualised inside the legal framework. By leveraging embeddings from optimised legal datasets, BERT accurately classifies texts to C2-level.

Combining SpaCy's ability to analyse linguistic structure with BERT's contextual understanding creates a system well-suited for categorising texts by CEFR levels. It addresses both the computational and linguistic challenges of text complexity. Its adaptable design ensures reliable performance with different text types and languages, conferring significant utility in language learning and research domains.

Table 1

**Comparative table of responsibilities and CEFR level determining**

| Type text | SpaCy identifies | BERT identifies | CEFR level |
|---|---|---|---|
| Children's story | Simple sentences, basic vocabulary, present tense | Semantic simplicity, repetitive patterns | A1 |
| Exam passage | Academic vocabulary, moderate syntactic complexity | Contextual usage of academic terms, intermediate tone | B1 |
| Legal Document | Long sentences, formal register, domain-specific terms | Semantic density, legal jargon, abstract structures | C2 |

## Conclusion

Implementing the modular approach to NLP provides a powerful way to handle the increasing complexity of textual data. The modularity makes implementing NLP systems easier, encourages broader adoption, and drives innovation in academic applications. Integrating SpaCy for text preprocessing and BERT for contextual understanding enables a more versatile approach to various text analysis problems.

The study emphasises the merits of employing pre-trained models, which attain elevated levels of accuracy in complex tasks such as text categorisation and semantic analysis. Through the systematic combination of SpaCy's traditional linguistic features and BERT's contextual embeddings, the proposed modular framework demonstrates robustness in classifying texts based on their complexity and aligning them with established benchmarks like CEFR levels.

This study, furthermore, fosters interdisciplinary cooperation by bridging the gap between theoretical developments in NLP and real-world applications. Incorporating modularity improves the scalability of NLP systems while also speeding up development processes. The thorough approach described here highlights how modularity facilitates the use of state-of-the-art technology and makes text analysis workflows accurate and effective.

Future research could build upon this modular architecture by adding domain-specific features, improving multilingual capabilities, and investigating real-time applications in dynamic contexts. By further honing and expanding the modular approach, NLP can develop further to satisfy the ever-increasing needs of processing and comprehending textual material in a world that is becoming increasingly information-driven. This work provides a foundation for such advancements, fostering innovation and accessibility in modern language processing.

## Bibliography

1. Flayeh A. K., Hamodi Y. I., Zaki N. D. Text analysis based on natural language processing (NLP) // Proceedings of the 2nd International Conference on Advanced Engineering and Smart Technology (AEST). 2022. URL: https://doi.org/10.1109/AEST55805.2022.10413039 (date of access: 15.06.2025).

2. Qiu X., Sun T., Xu Y., Shao Y., Dai N., Huang X. Pre-trained models for natural language processing: A survey // Science China Technological Sciences. 2020. Vol. 63, No. 10. P. 1872–1897. DOI: https://doi.org/10.1007/s11431-020-1647-3 (date of access: 15.06.2025).

3. Shikun K., Montiel Olea J. L., Nesbit J. Robust machine learning algorithms for text analysis // Quantitative Economics. 2024. Vol. 15, No. 4. P. 939–970. DOI: https://doi.org/10.3982/QE1825 (date of access: 15.06.2025).

4. Parks L., Peters W. Natural language processing in mixed-methods text analysis: A workflow approach // International Journal of Social Research Methodology. 2022. Vol. 26, No. 1. P. 1–13. DOI: https://doi.org/10.1080/13645579.2021.2018905 (date of access: 15.06.2025).

5. Roldan-Baluis W. L., Zapata N. A., Mañaccasa Vásquez M. S. The effect of natural language processing on the analysis of unstructured text: A systematic review // International Journal of Advanced Computer Science and Applications. 2022. Vol. 13, No. 5. P. 43–51. DOI: https://doi.org/10.14569/IJACSA.2022.0130507 (date of access: 15.06.2025).

6. Benício D. H. P., Júnior J. C. X., Paiva K. R. S., de Camargo J. D. A. S. Applying text mining and natural language processing to electronic medical records for extracting and transforming texts into structured data // Research, Society and Development. 2022. Vol. 11, No. 6. P. e37711629184. DOI: https://doi.org/10.33448/rsd-v11i6.29184 (date of access: 15.06.2025).

7. Fate V. D., Deshmukh A. B., Watane H. N. Text analytics: An approach to artificial intelligence // Journal of Emerging Technologies and Innovative Research. 2021. Vol. 8, No. 7. P. f104–f111. URL: https://www.jetir.org/papers/JETIR2107644.pdf (date of access: 15.06.2025).

8. Copestake A. Natural Language Processing: Lectures. Cambridge, U.K.: University of Cambridge, Computer Laboratory, 2004. URL: https://www.cl.cam.ac.uk/teaching/2002/NatLangProc/revised.pdf (date of access: 15.06.2025).

9. Jurafsky D., Martin J. H. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. Upper Saddle River, NJ: Prentice Hall, 2008. URL: https://web.stanford.edu/~jurafsky/slp3/ed3book.pdf (date of access: 15.06.2025).

10. Indurkhya N., Damerau F. J. Handbook of Natural Language Processing. 2nd ed. Boca Raton, FL: Chapman & Hall/CRC, 2010. DOI: https://doi.org/10.1201/9781420085938 (date of access: 15.06.2025).

11. Montesinos López O. A., Montesinos López A., Crossa J. Fundamentals of artificial neural networks and deep learning // In: Multivariate Statistical Machine Learning Methods for Genomic Prediction. Springer, 2022. DOI: https://doi.org/10.1007/978-3-030-89010-0_10 (date of access: 15.06.2025).

12. Pawłowski A., Walkowiak T. NLP for digital humanities: Processing chronological text corpora // Proceedings of the 4th International Conference on Natural Language Processing for Digital Humanities. 2024. P. 105–112. DOI: https://doi.org/10.18653/v1/2024.nlp4dh-1.10 (date of access: 15.06.2025).

13. Partalidou E., Spyromitros-Xioufis E., Doropoulos S., Vologiannidis S., Diamantaras K. Design and implementation of an open source Greek POS tagger and entity recogniser using spaCy // arXiv Preprint. 2019. DOI: https://doi.org/10.1145/3350546.3352543 (date of access: 15.06.2025).

14. Camacho-Collados J., Täckström O. Embeddings in natural language processing // Proceedings of the 28th International Conference on Computational Linguistics: Tutorial Abstracts. 2020. P. 10–15. DOI: https://doi.org/10.18653/v1/2020.coling-tutorials.2 (date of access: 15.06.2025).

15. Paaß G., Giesselbach S. Foundation Models for Natural Language Processing: Pre-trained Language Models Integrating Media. Cham, Switzerland: Springer, 2023. DOI: https://doi.org/10.1007/978-3-031-23190-2 (date of access: 15.06.2025).

**References**

1. Flayeh, A. K., Hamodi, Y. I., & Zaki, N. D. (2022). Text analysis based on natural language processing (NLP). Proceedings of the 2nd International Conference on Advanced Engineering and Smart Technology (AEST). https://doi.org/10.1109/AEST55805.2022.10413039

2. Qiu, X., Sun, T., Xu, Y., Shao, Y., Dai, N., & Huang, X. (2020). Pre-trained models for natural language processing: A survey. Science China Technological Sciences, 63(10), 1872–1897. https://doi.org/10.1007/s11431-020-1647-3

3. Shikun, K., Montiel Olea, J. L., & Nesbit, J. (2024). Robust machine learning algorithms for text analysis. *Quantitative Economics*, 15(4), 939–970. https://doi.org/10.3982/QE1825

4. Parks, L., & Peters, W. (2022). Natural language processing in mixed-methods text analysis: A workflow approach. *International Journal of Social Research Methodology*, 26(1), 1–13. https://doi.org/10.1080/13645579.2021.2018905

5. Roldan-Baluis, W. L., Zapata, N. A., & Mañaccasa Vásquez, M. S. (2022). The effect of natural language processing on the analysis of unstructured text: A systematic review. *International Journal of Advanced Computer Science and Applications*, 13(5), 43–51. https://doi.org/10.14569/IJACSA.2022.0130507

6. Benício, D. H. P., Júnior, J. C. X., Paiva, K. R. S., & de Camargo, J.1 D. A. S. (2022). Applying text mining and natural language processing to electronic medical records for extracting and transforming texts into structured data. *Research, Society and Development*, 11(6), e37711629184. https://doi.org/10.33448/rsd-v11i6.29184

7. Fate, V. D., Deshmukh, A. B., & Watane, H. N. (2021). Text analytics: An approach to artificial intelligence. *Journal of Emerging Technologies and Innovative Research*, 8(7), f104–f111. https://www.jetir.org/papers/JETIR2107644.pdf

8. Copestake, A. (2004). *Natural Language Processing: Lectures*. Cambridge, U.K.: University of Cambridge, Computer Laboratory. https://www.cl.cam.ac.uk/teaching/2002/NatLangProc/revised.pdf

9. Jurafsky, D., & Martin, J. H. (2008). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Upper Saddle River, NJ: Prentice Hall. https://web.stanford.edu/~jurafsky/slp3/ed3book.pdf

10. Indurkhya, N., & Damerau, F. J. (2010). *Handbook of Natural Language Processing* (2nd ed.). Boca Raton, FL: Chapman & Hall/CRC. https://doi.org/10.1201/9781420085938

11. Montesinos López, O. A., Montesinos López, A., & Crossa, J. (2022). Fundamentals of artificial neural networks and deep learning. In *Multivariate Statistical Machine Learning Methods for Genomic Prediction*. Springer. https://doi.org/10.1007/978-3-030-89010-0_10

12. Pawłowski, A., & Walkowiak, T. (2024). NLP for digital humanities: Processing chronological text corpora. *Proceedings of the 4th International Conference on Natural Language Processing for Digital Humanities*, 105–112. https://doi.org/10.18653/v1/2024.nlp4dh-1.10

13. Partalidou, E., Spyromitros-Xioufis, E., Doropoulos, S., Vologiannidis, S., & Diamantaras, K. (2019). Design and implementation of an open source Greek POS tagger and entity recogniser using spaCy. *arXiv Preprint*. https://doi.org/10.1145/3350546.3352543

14. Camacho-Collados, J., & Täckström, O. (2020). Embeddings in natural language processing. *Proceedings of the 28th International Conference on Computational Linguistics: Tutorial Abstracts*, 10–15. https://doi.org/10.18653/v1/2020.coling-tutorials.2

15. Paaß, G., & Giesselbach, S. (2023). *Foundation Models for Natural Language Processing: Pre-trained Language Models Integrating Media*. Cham, Switzerland: Springer. https://doi.org/10.1007/978-3-031-23190-2