Технічні науки

УДК 004.42

Багнюк Володимир Сергійович

аспірант

ПВНЗ «Європейський університет»

Bahniuk Volodymyr

Postgraduate of the

PHEI "European University"

ORCID: 0009-0008-6196-3878

Баланюк Наталія Юріївна

кандидат юридичних наук

Глухівський національний педагогічний університет ім. О. Довженко

Balaniuk Nataliia

Candidate of Law

Oleksandr Dovzhenko Hlukhiv National Pedagogical University

ORCID: 0009-0005-5376-3612

ТЕОРЕТИЧНІ ОСНОВИ АНАЛІЗУ МОВНОЇ ІНФОРМАЦІЇ У ВЕЛИКИХ ОБСЯГАХ ДАНИХ
THEORETICAL FOUNDATIONS OF LINGUISTIC INFORMATION ANALYSIS IN BIG DATA

Анотація. *У статті проведено дослідження інформаційних технологій, зокрема методів Data Mining, Natural Language Processing (NLP), методів опорних векторів (SVM), а також глибоких нейронних мереж (наприклад, CNN та LSTM) для обробки мовної інформації у великих обсягах даних, що дозволяють здійснювати ефективний семантичний аналіз великих обсягів мовної інформації із використанням методів машинного навчання.*

Зосереджено увагу на понятійному апараті та специфічних особливостях мовної інформації як об'єкта обробки.

Проаналізовано перелік існуючих сучасних підходів до інтелектуального аналізу текстів, зокрема класифікацію, кластеризацію, тематичне моделювання, ідентифікація ключових концептів та інтерпретація значень у контексті. Висвітлено перелік основоположних проблем, які тісно супроводжують безперервний процес обробки природномовних даних, таких як: полісемія, контекстна обумовленість значень та вплив соціокультурних чинників на семантичну інтерпретацію. Зроблено заключні висновки щодо подальшої перспективи застосування Data Mining, Natural Language Processing (NLP), векторів (SVM), а також глибоких нейронних мереж (наприклад, CNN та LSTM), зокрема, у прикладних рішеннях, що передбачають підтримку актуальних процесів з прийняття рішень, пошукову обробку інноваційних інформаційних запитів та автоматизовану семантичну обробку певного обсягу текстових даних.

Вступ. У сучасну епоху діджиталізації, інформаційне оновлення масиву мовної інформації, що виробляється щоденно, збільшується в арифметичній прогресії. Суспільні мережі, онлайн-медіа, форуми, цифрова документація – усе це вже сформована для подальшої діджиталізації сукупність джерел неструктурованих текстових даних, що згодом потребуватимуть ретельного вивчення. Опрацювання подібних масивів даних досить звичними методами не забезпечують належної ефективності, що, у свою чергу, ставить завдання впровадити інноваційні технології оброблення інформації, здатних у повному обсязі оптимізувати процес масштабованої, швидкої та точної обробки мовних даних.

Особливо набуває вагомості впровадження інноваційних підходів, побудованих на технологіях машинного навчання, нейронних моделей та

відповідно хмарних сервісів обробки даних, що дають змогу враховувати контекст при аналізі мовних даних, змістовних аспектів і структурних особливостей.

Мета роботи – дослідити перелік інноваційних інформаційних технологій, зокрема методів Data Mining, Natural Language Processing (NLP), векторів (SVM), а також глибоких нейронних мереж (наприклад, CNN та LSTM) для обробки мовної інформації у великих обсягах даних, що дозволяють здійснювати ефективний семантичний аналіз великих обсягів мовної інформації із використанням методів машинного навчання. Зосередити увагу на понятійному апараті та специфічних особливостях мовної інформації як об'єкта обробки.

Матеріали і методи. У межах даного дослідження було використано комплексний підхід, який поєднує класичні методи обробки природної мови (NLP) та сучасні алгоритми машинного навчання: векторів (SVM), глибоких нейронних мереж (наприклад, CNN та LSTM), а також міждисциплінарний підхід, що поєднує методи інтелектуального аналізу даних (Data Mining), технології обробки природної мови (Natural Language Processing (NLP), а також елементи статистичного й машинного навчання.

Результати. Перспективний напрямок дослідження у відповідній сфері передбачає формування гібридних моделей, що синтезують традиційні алгоритми Data Mining із сучасними нейромережевими архітектурами. Досліджені інформаційні технологія на сучасному етапі демонструють високі показники, паралельно забезпечуючи високий рівень точності автоматизованого аналізу мовної інформації та надає можливість наочно візуалізувати результати, що підвищує ефективність їх використання аналітиками та дослідниками. Інтерфейс системи підтримує наочне відображення результатів, що спрощує їх інтерпретацію та використання в аналітичних та наукових цілях.

Застосування сучасних моделей глибокого навчання суттєво покращує якість семантичного аналізу порівняно з традиційними статистичними методами, забезпечуючи більш глибоке та контекстно-орієнтоване розуміння мовних даних.

Необхідно також удосконалювати методики анотації даних та розробляти автоматизовані інструменти для ефективного формування мовних корпусів, що забезпечує підвищення якості навчання моделей і точності їх семантичного аналізу.

Перспективи. У статті досліджено інформаційну технологію аналізу великих обсягів мовної інформації, що поєднує машинне навчання, сучасні методи обробки природної мови та хмарні інфраструктури.

Розроблене рішення демонструє високу точність і швидкодію, що дозволяє використовувати його в практичних сферах, таких як:

- *моніторинг соціальних мереж,*
- *автоматичне сортування документації,*
- *аналітика користувацьких відгуків,*
- *інфомоніторинг медіа.*

Перелік подальших досліджень, на нашу думку, має бути спрямований на адаптацію технологічних процесів до багатомовних корпусів та інтеграцію з системами штучного інтелекту для генерації тексту.

Ключові слова: *Data Mining, обробка природної мови, NLP, великі дані, тематичне моделювання, семантичний аналіз, класифікація текстів, мовна інформація, IT-рішення.*

Summary. *The article presents a study of information technologies, in particular Data Mining methods, Natural Language Processing (NLP), Support Vector Machines (SVM), as well as deep neural networks (e.g., CNN and LSTM) for processing linguistic information in large-scale data environments.*

These approaches enable efficient semantic analysis of extensive language data using machine learning techniques.

The focus is placed on the conceptual framework and specific characteristics of linguistic information as an object of processing. The study analyzes a range of existing modern approaches to intelligent text analysis, including classification, clustering, topic modeling, key concept identification, and contextual meaning interpretation.

A set of fundamental challenges that continuously accompany the processing of natural language data is outlined, such as polysemy, context-dependent meaning, and the influence of sociocultural factors on semantic interpretation.

Final conclusions are drawn regarding the future prospects of applying Data Mining, NLP, SVM, and deep neural networks (e.g., CNN and LSTM) in applied solutions aimed at supporting decision-making processes, handling innovative information queries, and enabling automated semantic processing of textual data.

Introduction. In the modern era of digitalization, the informational expansion of language data produced daily is increasing at an arithmetic progression. Social networks, online media, forums, and digital documentation have already formed a complex of unstructured textual data sources that are subject to further digital processing and detailed study. Traditional methods of handling such data volumes no longer provide sufficient efficiency, which, in turn, necessitates the implementation of innovative information processing technologies capable of fully optimizing the scalable, rapid, and accurate handling of linguistic data.

Particularly significant is the adoption of innovative approaches based on machine learning technologies, neural models, and corresponding cloud-based data processing services, which enable contextual analysis of linguistic information while considering semantic content and structural features.

Purpose - to investigate a range of innovative information technologies, particularly Data Mining methods, Natural Language Processing (NLP), Support Vector Machines (SVM), as well as deep neural networks (e.g., CNN and LSTM) for processing linguistic information in large-scale data. These technologies enable effective semantic analysis of extensive language data using machine learning techniques. Additionally, to focus on the conceptual framework and specific characteristics of linguistic information as an object of processing.

Materials and methods. This study employed a comprehensive approach combining classical natural language processing (NLP) methods with modern machine learning algorithms: Support Vector Machines (SVM), deep neural networks (e.g., CNN and LSTM), as well as an interdisciplinary approach integrating data mining techniques, natural language processing technologies, and elements of statistical and machine learning.

Results. A promising research direction in this field involves the development of hybrid models that synthesize traditional Data Mining algorithms with modern neural network architectures. The investigated information technologies currently demonstrate high performance, simultaneously ensuring a high level of accuracy in automated linguistic information analysis and providing the capability for clear visualization of results, which enhances their usability for analysts and researchers. The system interface supports intuitive presentation of outcomes, simplifying their interpretation and application in analytical and scientific contexts.

The application of modern deep learning models significantly improves the quality of semantic analysis compared to traditional statistical methods, enabling a deeper and more context-aware understanding of linguistic data.

It is also necessary to improve data annotation techniques and develop automated tools for the efficient creation of linguistic corpora, which in turn enhances model training quality and the accuracy of semantic analysis.

Discussion. The article investigates an information technology for analyzing large volumes of linguistic data, combining machine learning, modern natural language processing methods, and cloud infrastructures.

The developed solution demonstrates high accuracy and performance, enabling its application in practical areas such as:

– social media monitoring,

– automatic document classification,

– user feedback analytics,

– media infomonitoring.

We believe that future research should focus on adapting technological processes to multilingual corpora and integrating with artificial intelligence systems for text generation.

Key words: *Data Mining, Natural Language Processing, NLP, Big Data, Topic Modeling, Semantic Analysis, Text Classification, Linguistic Information, IT Solutions*

Постановка проблеми. У сучасному інноваційному світі, що має шалену активність власного розвитку, пов'язаному із науково-технологічним прогресом, обробка великих обсягів інформаційних даних є основоположним вектором розробок у сфері комп'ютерних наук. Ключове місце відводиться мовній інформації – фундаментальному комунікаційному засобу, що вміщує не тільки певний обсяг фактичних даних, а й відповідний перелік смислових, емоційно-вольових та соціальних значень. Ефективність в детальній обробці усіх параметрів мовної інформації сприяє подальшому зростанню обсягів використання штучного інтелекту «споживачами», і, як наслідок, розвитку інтелектуальних інформаційних механізмів, комп'ютерного перекладу, рішень для підтримки управлінських дій.

Беручі до уваги семантичну й синтаксичну складові мови, її

невизначеність і контекстуальну залежність, з'являється нагальна потреба в розробленні вузькоспеціалізованих IT-рішень.

Невирішеність актуальних проблем в галузі смислової та контекстуальної інтерпретації тексту обумовлює необхідність подальших досліджень у галузі природної обробки мови (NLP), що є одним з ключових напрямів спеціальності 122 «Комп'ютерні науки».

У міру зростання обсягів неструктурованих текстових даних у віртуальному просторі набирає актуальності використання автоматизованих методів аналізу. Лінгвістичний матеріал, виконаний у текстовому представленні, аудіозаписів і різноманітних мультимедійних форматів, вимагає використання індивідуалізованих підходів, що характеризуються варіативним специфічним підходом порівняно з обробкою структурованих даних. Сучасний перелік виникаючих труднощів пов'язаний із семантичною неоднозначністю лексичних одиниць, складністю та різноплановістю граматичних конструкцій, а також соціокультурними особливостями мовного середовища.

Аналіз останніх досліджень і публікацій. Реальна проблематика ефективного опрацювання мовних даних у великих обсягах стає об'єктом зростаючого інтересу дослідників у галузі штучного інтелекту та обчислювальної лінгвістики. У центрі дослідницької діяльності – розробка інноваційних стратегій моделювання лінгвістичного напрямку, з помірно низьким рівнем лінгвістичної ресурсної доступності, а також поєднання етично стих аспектів, принципів підзвітності та універсальності вищенаведених систем.

Так, у наукових розробках Magiff J. та Nikolov N. [15], Manning C., Raghavan P. & Schütze H. [12], Jurafsky D. & Martin J. [8], Feldman, R. & Sanger J. [7], здійснено процедуру поглибленого огляду переліку методики усунення браку даних у контексті генеративного моделювання, а саме: back-translation, диференціація текстових корпусів і багатоаспектне

навчання. У дослідженнях авторами зроблено спробу доведення наступного факту: за браком лінгвістичних даних можливо на належному рівні забезпечити покрокову ефективність результатів за допомогою крос-лінгвальних інноваційних методик.

У дослідженні Dossou B. та Aidaso E. [5], Tan P.-N., Steinbach M. & Kumar V. [18], Devlin J., Chang M.-W., Lee K., & Toutanova K. [4] передбачено коригування методології розвитку NLP для мов із обмеженим ресурсним напрямом – поетапний перехід від пасивного формування корпусів до безперебійної взаємодії з носіями мов. Відповідний метод сприятиме у довгостроковій перспективі не лише створенню сучасних адаптованих моделей, а й запровадженню на практиці динамічного навчання із культурним контекстом.

Питання етики у мовному моделюванні розглянуто в роботі Kaneko M. [9], Mikolov T., Chen K., Corrado G. & Dean, J. та співавт. [14], де науковцями сформовано корпус цільових даних, застосувавши оновлений механізм із безпосередньою взаємодією користувачів з відповідними значними обсягами лінгвістичних моделей. Даний модернізаційний підхід сприятиме виявленню й аналізу можливих упереджень у нетиповій поведінці розроблених моделей, токсичність або морально сумнівні відповіді, що є особливо важливим у контексті застосування великих мовних моделей у суспільній комунікативній галузі. У праці Pakray P. [17], Bird S., Klein E., & Loper E. [2], Aggarwal C. та співавт. [1] присвячені покроковому аналізу інноваційних прикладів запровадження NLP у мовах, для яких бракує цифрових/лінгвістичних ресурсів. Дослідники акцентують увагу на безпечній розробці модернізаційних моделей корпусів переважно у адміністративній, медичній та освітній галузях, з подальшою адаптацією функціонуючих лінгвістичних моделей з огляду на специфічні особливості таких мов.

Практичне застосування крос-лінгвальної аугментації в якості ефективного інструментарію для розпізнавання іменованих сутностей (NER) підтверджено емпіричними даними дослідження Ehsan T. [6] та Blei D., Ng A., & Jordan M. [3], Liu B. [5]. В тому числі вченими досягнуто підвищення точності NER у пакистанській лінгвістиці, паралельно застосовуючи додаткові дані із суміжної комплексної лінгвістики тієї ж мовної групи.

Ще один підхід, що заслуговує на увагу, був запропонований дослідниками: Pandya G. та Bhatt B. [16], Marivate V., Sefara T. [13], які з'ясували безпосередній факт впливу навчання на комплекс споріднених мов в області автоматизованих питань і відповідей. З'ясувалося, що лінгвістична спорідненість може на достатньо ефективному рівні виступати посередником для поліпшення результатів, особливо за наявності синтаксичної подібності.

Немаловажним є також формування дослідницьких інноваційних комплексних спільнот з покроковим обміном накопиченими обсягами знань. Під час вступного воркшопу LoResLM (2025) [11] був продемонстрований перелік численних досліджень, спрямованих на поетапне створення модернізаційних корпусів, методів трансферного навчання та мультилінгвістичних моделей, які націлені на мови, що ще не включені до великих комерційних LLM.

Отже, перелік інноваційних сучасних досліджень підтвердив суттєву актуальність проблематики подальшої обробки лінгвістичних інформаційних даних в умовах у середовищі Big Data. У центрі сучасної уваги – комплексне поєднання мовознавчих, технічних і етичних засад з подальшою ефективною метою об'єднання адаптивних та інклюзивних лінгвістичних моделей.

Технології Data Mining зазвичай застосовуються у завданнях, пов'язаних із класифікацією текстів та тематичним моделюванням

(наприклад, Latent Dirichlet Allocation), кластеризація документів, виявлення інформаційних патернів, побудова онтологій. Роботи Feldman & Sanger [7], Tan, Steinbach & Kumar [18] лежать в основі підходів до аналізу текстових корпусів засобами інтелектуального аналізу даних.

Самостійним напрямом є Text Mining – підгалузь Data Mining, що сконцентрована саме на аналізі лінгвістичних інформаційних даних. Її завдання включають вилучення сутностей, емоційний аналіз (sentiment analysis), визначення авторства, анотування текстів. Сучасні дослідження активно залучають глибинне навчання (deep learning), трансформери (зокрема моделі на кшталт BERT, GPT), що дозволяє досягати високих результатів при роботі з великими текстовими корпусами.

Водночас наявний сучасний комплекс емпіричних даних демонструє низку проблемних аспектів, що впливають на ефективність запровадження Data Mining у лінгвістичному середовищі, серед яких: відсутність достовірної розмітки, проблеми з урахуванням контексту та наявність полісемії, морфосинтаксичні варіаційні компоненти і вплив соціокультурних чинників на семантику. Ці умови обумовлюють необхідність розробки інтегрованих (або міждисциплінарних) підходів, які базуються на комплексному міждисциплінарному підході, що охоплює лінгвістичний напрямок, інформатику, когнітивістику та взаємодію із штучним інтелектом.

Мета роботи полягає у дослідженні переліку інноваційних інформаційних технологій, зокрема методів Data Mining, Natural Language Processing (NLP), векторів (SVM), а також глибоких нейронних мереж (наприклад, CNN та LSTM) для обробки мовної інформації у великих обсягах даних, що дозволяють здійснювати ефективний семантичний аналіз великих обсягів мовної інформації із використанням методів машинного навчання.

Наукова новизна дослідження полягає у тому, що в ній дістало

подальшого опрацювання узагальнення методик анотації даних з подальшим удосконаленням автоматизованого інструментарію для ефективного формування мовних корпусів, що в майбутньому суттєво вплине на забезпечення підвищення якості навчання моделей і точності їх семантичного аналізу.

Матеріалами дослідження є:

– **корпуси текстових даних** (відкриті набори даних, такі як Reuters-21578, 20 Newsgroups, Wikipedia dumps, що містять великі обсяги текстової інформації різної тематики. Україномовні корпуси, наприклад, тексти зі ЗМІ, офіційні документи та наукові публікації. Корпуси із соціальних мереж для аналізу сучасної живої мови та її варіативності);

– **аудіо- та мультимедійні дані** (аудіозаписи, які підлягали автоматичній транскрипції для подальшого текстового аналізу. Мультимедійні джерела, що комбінують текстову і звукову інформацію);

– **апаратне забезпечення** (сервери з графічними процесорами (GPU) для тренування моделей на основі трансформерів. Комп'ютерні кластери для паралельної обробки великих масивів даних).

Методи дослідження:

– **попередня обробка даних** (токенізація, лемматизація та нормалізація текстів);

– **класифікація текстів** (використання алгоритмів наївного байєсівського класифікатора, методів опорних векторів (SVM), а також глибоких нейронних мереж (наприклад, CNN та LSTM). Навчання моделей на розмічених корпусах із подальшою оцінкою якості за допомогою метрик точності, повноти та F1-міри);

– **кластеризація** (застосування алгоритмів k-середніх, ієрархічної кластеризації та DBSCAN для групування документів за змістом. Оцінка результатів кластеризації за допомогою внутрішніх метрик, таких як силуетний коефіцієнт);

– **тематичне моделювання** (використання моделі латентного розподілу Діріхле (LDA) для виявлення основних тем у корпусах. Візуалізація тематичних розподілів для аналізу змісту);

– **вилучення ключових понять** (застосування методів TF-IDF і графових алгоритмів (TextRank) для визначення найбільш релевантних термінів і фраз);

– **семантичний аналіз** (використання трансформерних моделей, таких як BERT, для контекстного аналізу значень слів і фраз. Інтерпретація результатів із урахуванням лексичної неоднозначності та соціокультурного контексту);

– **оцінка результатів** (використання стандартних метрик для оцінки якості класифікації, кластеризації і тематичного моделювання. Порівняння отриманих результатів із базовими моделями та альтернативними підходами).

Виклад основного матеріалу. У сучасному інноваційному середовищі спостерігається значне розширення обсягів процесу діджиталізації, через що окрема частина відповідного явища фіксується у вигляді мовної інформації: документів, звукових файлів і медіа-контенту. Як правило, зазвичай йдеться про інформацію з відсутньою або обмеженою логічною й послідовною структурою, що знижує відсоток ефективності стандартної методології з її відповідного оброблення та подальшої інтерпретації. У сучасних умовах швидкої трансформації підходів до обробки масивів даних (Big Data), специфічної актуальності набувають інструменти, здатні забезпечувати автоматизоване виявлення закономірностей, формальних моделей і ключових відношень у великих корпусах мовних даних. Застосування методів Data Mining є одним із найперспективніших підходів до аналізу великих даних.

Data Mining – це інтелектуальний аналіз даних, тобто процес виявлення у великих масивах даних раніше невідомих, неочевидних, але

потенційно корисних закономірностей, залежностей, шаблонів чи знань з використанням математичних, статистичних, алгоритмічних та машинних методів [2]. Методи інтелектуального аналізу даних (Data Mining) застосовуються в процесі обробки як текстових, так і мультимедійних джерел мовної інформації, паралельно виокремлюючи провідні ключові концепти, виявляючи основоположну тематику, перелік емоційних маркерів, авторські забарвлення та будь-які визначальні релевантні характеристики.

Обробка лінгвістичної інформації відмічає достатньо високу складність з огляду на низку важливих чинників: неоднозначністю словникового запасу, варіативністю синтаксичних конструкцій, а також впливом контексту й соціокультурних чинників, що в комплексі впливають безпосередньо на значення та контекстуальне застосування мовних елементів. Враховуючи відповідну проблематику, відчувається необхідність у створенні і впровадженні спеціальних моделей і алгоритмів, що напряму залежать від лінгвістичних особливостей інформації в якості об'єкта аналізу.

Матеріал роботи спрямований на дослідженні перспективних напрямків запровадження методів Data Mining для аналізу мовних елементів у великих інформаційних масивах в розрізі актуальних викликів, включаючи комплекс інноваційних підходів у напрямі їх нейтралізації, разом з тим демонструються приклади доцільного застосування обраних технологій у прикладних ситуаціях.

У галузі сучасних ІТ-рішень значне зростання масштабів неструктурованих інформаційних потоків, головним чином текстових, вимагає новітніх підходів до подальшої ефективної обробки та аналізу. Лінгвістична інформація, що вміщує комплекс текстових, аудіо- та мультимедійних форматів, є основоположним носієм освітніх ресурсів і обміну комунікативними даними у цифровому середовищі. Водночас

індивідуальні характеристики природної мови, такі полісемія, умовність, а також культурно-суспільна визначеність, утворюють певні труднощі для подальшого автоматизованого аналізу.

Методика Data Mining, що охоплює численні інноваційні підходи та техніки для аналізу важливої інформації з величезних масивів даних, складає ключовий інструментарій для опрацювання сформованих цілей. Реалізація Data Mining в області лінгвістичної обробки даних надає можливість поетапного автоматичного впорядкування, групування отриманих даних і побудови тематичних моделей з поступовим виокремленням важливих концептуальних елементів тексту.

Автоматичне розпізнання та об'єднання текстових даних

Класифікація виокремлює в якості основоположного фундаментального методу – Data Mining, що має за функціональну мету автоматичний розподіл певного обсягу документації за заздалегідь обумовленими алгоритмами. Для досягнення вище встановлених цілей застосовують певний спектр відпрацьованих алгоритмів, серед яких: такі як наївний байєсівський класифікатор, метод опорних векторів (SVM), дерева рішень та нейронні мережі. Запровадження на практиці чітко визначеного класифікатору значно підвищує результативність пошукових заходів інформаційних обсягів з подальшим впорядкуванням значних текстових форматів.

Метод кластеризації порівняно із класифікацією, є методом без відповідного нагляду, що в майбутньому надає дозвіл у покроковому специфічному виявленні сукупності природних групових складників текстових файлів, беручі за основу параметри схожості їхнього змістовного наповнення відповідно. Додатковий параметр являється корисним при фактичному здійсненні аналізу об'ємних корпусів, де можуть бути відсутній перелік розмічених даних. Популярними алгоритмами є: k-середніх, ієрархічна кластеризація, DBSCAN.

Тематичне моделювання

З подальшою метою здійснення глибинного аналізу змісту текстових обсягів матеріалу використовують метод тематичного моделювання, що у свою чергу сприяє виявленню основоположних тем, що здійснюють специфічний шлях у проходженні крізь об'ємний масив документів. Метод латентного розподілу Діріхле (LDA) є одним із найбільш популярних у галузі. Так, LDA володіє специфічною функцією у моделюванні текстових файлів у вигляді суміші тематичних напрямків, в яких кожна індивідуальна тема представлена корегуючим набором слів із 90% ймовірністю. Дана програма спрямована на допомогу в автоматичному класифікаторі інформаційних обсягів, створювати оглядові тематичні майданчики, а також здійснювати роботу у знаходженні оновлених патернів у лінгвістичних складниках.

Вилучення ключових понять і семантичний аналіз

Здійснення поетапного виокремлення у основоположних судженнях: виокремлення найпоширеніших дефінітивних понять та змістовних фраз, що сприяє у покроковому обмеженні інформаційних обсягів з неструктурованих текстових файлів. З цією метою віднайшов широке застосування саме алгоритмічний порядок на основі статистичних метрик, таких як TF-IDF, а також графові методи типу TextRank.

Найскладнішим аспектом з усіх в обробці природної лінгвістики є семантичний аналіз, який паралельно спрямований на урахування контексту. Інноваційний спектр трансформерних моделей, наприклад BERT, спрямований на ефективне розпізнання дефінітивних навантажень слів у відповідності від сформованого оточення, що бере активну участь у процесі вирішення проблематичного питання з полісемії та неоднозначності. Це у значному обсязі покращує відповідну якість інтерпретації текстових файлів з подальшим поетапним застосуванням сукупності автоматизованих систем для вирішення інноваційних завдань,

таких як відповіді на звернення, синтез текстів, реферативний виклад, Al - переклад.

Висновки та перспективи подальших досліджень

Попри досягнуті результати, масштабна обробка лінгвістичної інформації у великих масштабах даних характеризується наявністю певних труднощів. Передусім, ретельна підготовка й анотування навчальних даних є достатньо складним і вимогливим процесом. Другим важливим аспектом є робота з аудіо- і візуальною інформацією, яка потребує розширених засобів розпізнавання і транскрипції, що ускладнює взаємодію з текстовими даними відповідно. По-третє, сучасні моделі відзначаються значною ресурсоємністю під час навчання, що зумовлює певні бар'єри для їх практичного застосування у межах типових організаційних структур. Важливим аспектом є урахування соціокультурних детермінант мови, які визначають характер смислових інтерпретацій та комунікативних нюансів. Для реалізації поставлених завдань необхідною є міждисциплінарна інтеграція лінгвістичних, культурологічних та комп'ютерних наук, що забезпечує комплексний підхід до дослідження мовних процесів..

Ефективне розроблення сучасних систем штучного інтелекту (ШІ) у сфері оброблення мовних одиниць потребує міждисциплінарної взаємодії знань з лінгвістики, культурології та комп'ютерних наук, що, у свою чергу, надає змогу досягти локальної точності та адаптивності інноваційних алгоритмів.

Подальший перелік структурованих досліджень у галузі 122 «комп'ютерні науки» має бути спрямований на розробку модернізаційних гібридних моделей, що об'єднають класичні алгоритми Data Mining, Natural Language Processing (NLP), методів опорних векторів (SVM), а також глибоких нейронних мереж (наприклад, CNN та LSTM із сучасними нейромережевими архітектурами. Немаловажно паралельно працювати

над процесом з подальшого удосконалення методів анотації даних з метою створення ефективного автоматизованого інструментарію для підготовки корпусів.

Оптимальне безпомилкове керування обчислювальними ресурсами й інтеграція паралельних обчислювальних інноваційних технологій підвищать здатність системи працювати з великими лексичними обсягами. Адаптація моделей до лінгвістичного й культурного розмаїття є ключовою умовою для досягнення їхньої модернізаційної універсальності та практичної цінності.

Запровадження методів Data Mining у обробці мовної інформації на практиці у сучасному діджиталізованому світі має надважливе значення для розвитку систем підтримки прийняття обґрунтованих рішень, оскільки воно забезпечує якісніший аналіз лінгвістичних файлів, підвищує закономірний відсоток точності прогнозування та поступово сприяє ухваленню обґрунтованих управлінських або аналітичних рішень. Відповідний напрям є особливо актуальним у галузі бізнес-аналітики, суспільно-політичних досліджень, інформаційної безпеки та цифрової гуманітаристики.

Література

1. Aggarwal, C. C. (2015). *Data Mining: The Textbook*. Springer. https://doi.org/10.1007/978-3-319-14142-8

2. Bird, S., Klein, E., & Loper, E. (2009). *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. O'Reilly Media. *arXiv preprint*, arXiv:1406.6790. aclanthology.org. (дата звернення: 29.09.2025).

3. Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, *3*, 993–1022.

https://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf (дата звернення: 15.10.2025).

4. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 4171–4186). https://doi.org/10.48550/arXiv.1810.04805

5. Dossou, B. F. P., Aïdasso, H. (2025). Towards Open-Ended Discovery for Low-Resource NLP. URL: https://arxiv.org/abs/2510.01220 (дата звернення: 16.10.2025).

6. Ehsan T., Solorio T. (2025). Enhancing NER Performance in Low-Resource Pakistani Languages using Cross-Lingual Data Augmentation. URL: https://arxiv.org/abs/2504.08792 (дата звернення: 16.10.2025).

7. Feldman, R., & Sanger, J. (2007). *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*. Cambridge University https://www.wongpartnership.com/upload/medias/KnowledgeInsight/document/17634/TheInsolvencyReview10thedition-Singapore Chapter.pdf (дата звернення: 29.09.2025).

8. Jurafsky, D., & Martin, J. H. (2021). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition* (3rd ed., draft). Retrieved from https://web.stanford.edu/~jurafsky/slp3 (дата звернення: 29.09.2025).

9. Kaneko, M., Bollegala, D., Baldwin, T. (2025). An Ethical Dataset from Real-World Interactions Between Users and Large Language Models. IJCAI URL: https://www.ijcai.org/proceedings/2025/1082 (дата звернення: 16.10.2025).

10. Liu, B. (2012). *Sentiment Analysis and Opinion Mining*. Morgan & Claypool Publishers. https://doi.org/10.2200/S00416ED1V01Y201204HLT016 (дата звернення: 15.10.2025).

11. LoResLM Workshop Organizers. (2025). Proceedings of the First Workshop on Language Models for Low-Resource Languages (LoResLM 2025). COLING / ACL. URL: https://eprints.lancs.ac.uk/id/eprint/227367 (дата звернення: 16.10.2025).

12. Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press. Accounting and Business Research - ACCOUNT BUS RES. 37. 285-300. http://doi.org/10.1080/00014788.2007.9663313 (дата звернення: 29.09.2025).

13. Marivate, V., Sefara, T. (2020). Improving Short Text Classification Through Global Augmentation Methods. У збірці Lecture Notes in Computer Science, vol 12279 (CD-MAKE 2020). DOI: 10.1007/978-3-030-57321-8_21 repository.up.ac.za

14. Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint*, arXiv:1301.3781. https://doi.org/10.48550/arXiv.1301.3781 (дата звернення: 16.10.2025).

15. Magiff, J., Nikolov, N. S. (2025). Overcoming Data Scarcity in Generative Language Modelling for Low-Resource Languages: A Systematic Review. URL: https://arxiv.org/abs/2505.04531 (дата звернення: 16.10.2025).

16. Pandya, H. A., Bhatt, B. S. (2024). Does Learning from Language Family Help? A Case Study on a Low-Resource Question-Answering Task. Natural Language Processing. URL: https://core.ac.uk/download/pdf/234156317.pdf (дата звернення: 16.10.2025).

17. Pakray, P., Gelbukh, A., Bandyopadhyay, S. Natural Language Processing Applications for Low-Resource Languages. Cambridge University Press / Assessment, онлайн 28 лютого 2025. DOI: 10.1017/nlp.2024.33 Cambridge University Press & Assessment.

18. Tan, P.-N., Steinbach, M., & Kumar, V. (2019). *Introduction to Data Mining* (2nd ed.). Pearson. MDPI. repository.up.ac.za (дата звернення:

29.09.2025).