DOI 10.31891/2307-5732-2024-333-2-38
УДК 004.8

СТАХІВ МИКОЛА
Національний університет «Львівська політехніка»
https://orcid.org/0009-0008-9550-5736
e-mail: mykola.a.stakhiv@lpnu.ua
СКОПІВСЬКИЙ СТАПАН
Національний університет «Львівська політехніка»
https://orcid.org/0009-0008-4668-0140
e-mail: stepan.y.skopivskyi@lpnu.ua

ВІД АНАЛІЗУ ТЕКСТУ ДО МОДЕЛЮВАННЯ ПРИРОДНОЇ МОВИ: КОМПЛЕКСНЕ ДОСЛІДЖЕННЯ

У статті представлено комплексний аналіз сучасних прийомів і технологій, що використовуються для аналізу тональності текстових даних. Завдяки серії експериментів було розглянуто базові алгоритми машинного та глибокого навчання, такі як аналіз настроїв, щоб автоматично виявляти та класифікувати емоційні тони в текстах. Також досліджуються практичні застосування цих методів, починаючи від моніторингу соціальних мереж і закінчуючи експертним аналізом та аналізом відгуків користувачів. Далі стаття заглиблюється в класифікацію текстових даних за допомогою сучасних методів машинного та глибокого навчання. У цьому розділі статті розглядаються кілька моделей класифікації, включаючи наївний байєсів класифікатор, методи опорних векторів і нейронні мережі, підкреслюючи їхні переваги та обмеження. Крім того, наголошується на важливості впровадження класифікації тексту в різних областях, включаючи аналіз соціальних мереж, категоризацію статей новин і автоматизовану обробку документів. Крім того, розглядається кластеризація схожих текстових даних для подальшого аналізу. Порівнюються різні алгоритми кластеризації, такі як k-середні, ієрархічна кластеризація та спектральна кластеризація, з особливим акцентом на їх застосуванні до великих текстових корпусів. Також проведено демонстрацію практичного застосування кластеризації тексту, зокрема організації даних, пошуку тем та ідентифікації стилів у письмових роботах. Рухаючись далі, у статті розглядається процес виявлення тематичної структури в текстових даних та її подальший аналіз. Досліджено поглиблений аналіз методів тематичного моделювання, таких як модель прихованого розподілу Діріхле (LDA), а також її можливості та обмеження. Практичне застосування тематичного моделювання демонструється в різних областях, включаючи аналіз колекції текстів, виявлення новинних тенденцій і автоматичну категоризацію документів. Нарешті, у статті обговорюються потенційні виклики та майбутні перспективи розвитку тематичного моделювання.

Ключові слова: обробка природної мови, аналіз тональності, класифікація текстів, кластеризація текстів та тематичне моделювання.

STAKHIV MYKOLA, SKOPIVSKYI STEPAN
Lviv Polytechnic National University

FROM TEXT ANALYSIS TO NATURAL LANGUAGE MODELING: A COMPREHENSIVE EXPLORATION

The article presents a comprehensive analysis of modern techniques and technologies used to analyze the tonality of textual data. Through a series of experiments, basic machine and deep learning algorithms such as sentiment analysis were developed to automatically detect and classify emotional tones in texts. Practical applications of these methods are also explored, ranging from social network monitoring to expert analysis and analysis of user reviews. Next, the article delves into the classification of textual data using modern machine and deep learning methods. This section of the paper reviews several classification models, including the Bayesian most-narrow classifier, support vector methods, and neural networks, highlighting their advantages and limitations. In addition, it emphasizes the importance of implementing text classification in various fields, including social network analysis, news article categorization, and automated document processing. In addition, clustering of similar textual data is considered for further analysis. Various clustering algorithms, such as k-means, hierarchical clustering, and spectral clustering, are compared, with special emphasis on their application to large text corpora. A demonstration of the practical application of text clustering, including data organization, topic search, and style identification in paper works, was also demonstrated. Moving on, the article considers the processing of individual thematic structures in textual data and its further analysis. An in-depth analysis of topic modeling techniques, such as the realized Dirichlet distribution (LDA) model, as well as its capabilities and limitations, is explored. Practical applications of topic modeling are demonstrated in various ways, including text collection analysis, news trend detection, and automatic document categorization. Finally, the article discusses some challenges and future prospects for the development of thematic modeling.

Keywords: natural language processing, tonality analysis, text classification, text clustering, and thematic modeling.

Постановка проблеми

Сфера обробки природної мови (NLP) має вирішальне значення для розвитку інтелектуальних технологій у наш час. Незважаючи на значний прогрес у цій галузі, все ще існують критичні проблеми, які потребують уваги та досліджень. У статті висвітлено ці проблеми та їхній вплив на такі аспекти NLP, як тональний аналіз, класифікація тексту, кластеризація тексту та моделювання теми. Буде розглянуто перешкоди, з якими стикається NLP, і пропонуємо життєздатні рішення для вдосконалення та розвитку технологій NLP.

Аналіз останніх досліджень і публікацій

В роботі[1] розглядається чутливість глибоких нейронних мереж (DNN), що використовуються в обробці природної мови (NLP), до суперечливих прикладів, які є модифікованими вхідними даними, призначеними для введення моделі в оману. Незважаючи на успіх DNN в NLP, їхня вразливість до

суперечливих прикладів перешкоджає розгортанню в критично важливих для безпеки середовищах. Робота містить огляд сучасного використання змагальних прикладів для обману DNN і пропонує вичерпний огляд їх застосування для підвищення надійності DNN в NLP. Автори підсумовують останні методи створення змагальних текстів, пропонують таксономію для їх класифікації та переглядають стратегії захисту проти змагальних прикладів. У документі також розглядаються проблеми та пропонуються майбутні напрямки досліджень у цій галузі.

В роботі [2] обговорюються останні досягнення в обробці NLP, підмножині штучного інтелекту, зосередженій на комп'ютерному розумінні людських мов. За останні п'ять років відбувся швидкий прогрес у таких завданнях NLP, як машинний переклад, відповіді на запитання та машинне розуміння читання, завдяки глибокому навчанню та великим наборам даних. У роботі розглядаються інструменти NLP на основі нейронної мережі з трьох точок зору: моделювання, навчання та міркування. Він охоплює різні парадигми моделювання та методи навчання, наголошуючи на новій області міркувань для нейронного NLP. Підкреслюється важливість міркувань для побудови інтерпретованих і орієнтованих на знання моделей для вирішення складних завдань.

В дослідженні [3] визначаються основні фактори, що сприяють перешкодам у багатомовному машинному перекладі. Систематичні експерименти показують, що на інтерференцію або синергію в першу чергу впливають розмір моделі, обсяг даних і частка кожної мовної пари в загальному наборі даних. Примітно, що значні перешкоди виникають, коли модель дуже мала відносно доступних навчальних даних. Стандартні конфігурації трансформатора з менш ніж одним мільярдом параметрів ефективно зменшують перешкоди та підвищують синергію. Дослідження також підкреслює важливість налаштування температури вибірки для контролю пропорції мовних пар у даних, демонструючи її ключову роль у балансуванні перешкод між мовними парами з низьким і високим ресурсом, що призводить до чудової загальної продуктивності.

Мета і завдання статті

Проаналізувати та систематизувати інформацію про ключові аспекти NLP, такі як аналіз тональності, класифікація тексту, кластеризація тексту та моделювання теми.

Виклад основного матеріалу

Завдяки використанню обробки природної мови (англ. Natural Language Processing, NLP) комп'ютерні системи можуть сприймати й аналізувати людську мову. Це потужний інструмент, який розробники використовують для організації та структурування знань, що дозволяє комп'ютерам виконувати складні завдання, такі як автоматичне підсумовування, мовний переклад, ідентифікація іменованих об'єктів, аналіз тональності, розпізнавання мовлення та тематична сегментація.

NLTK, що означає Natural Language Toolkit, є першокласною платформою, спеціально розробленою для створення програм NLP за допомогою Python. Він може похвалитися зручними для користувача інтерфейсами, які підтримують різноманітні мовні корпуси та бібліотеки обробки тексту, які виконують такі завдання, як класифікація, токенізація, формування коренів, тегування частини мови (POS-тегування), синтаксичний аналіз і семантичний аналіз.

Аналіз тональності. Було використано два інструменти: VADER (англ. Valence Aware Dictionary and sEntiment Reasoner), який видає оцінки настроїв для вхідних даних, та word_tokenize, що дозволяє розбивати великий текст на послідовність більш дрібних одиниць, таких як речення або слова. Перший, на відміну від інших методів, що вимагають навчання зв'язаного тексту перед використанням, готовий до аналізу без будь-якого попереднього налаштування та є унікальним завдяки тому, що чітко розрізняє позитивність і негативність різного ступеня.

```
0 Більшість "слуг" за повернення закону про мобілізацію на доопрацювання – джерела
1 У Греції легалізують одностатеві шлюби й усиновлення
2 OpenAI офіційно запустила маркетплейс чат-ботів GPT Store
3 У Раді Європи визначились із кандидатами на генсека
4 Армія РФ вдарила по вугільному підприємству у Мирнограді на Донеччині, загинула людина
5 Дохід "Укрпошти" торік зріс до майже 12 мільярдів після падіння у 2022-му
6 Прикордонник "Смурф" про оборону Маріуполя, "Азовсталь", полон, "втому від війни"
7 У Києві знайшли мертвими ексзаступницю генпрокурора та її доньку
8 Окупанти обстріляли Дніпропетровщину, є поранені
9 На Раді "Україна-НАТО" Києву пообіцяли в цьому році зброї на мільярди євро
10 У Кривому Розі через серйозну аварію частина районів без світла – Вілкул
11 Стефанчук пояснив, чому Рада не стала звільняти Безуглу
12 Вчені виявили, що відмова від м'яса знижує ризик інфікування коронавірусом
13 Росіяни вкрили вогнем село на Херсонщині, постраждало подружжя
14 Складається, як Lego: українець створив сталеве укриття, яке нібито витримує понад 43 тонни
15 ОАЕ заборонили пропускати танкери під прапором Камеруну: перевозять російську нафту
16 Зведення Генштабу: РФ здійснила 10 спроб вибити українські війська з лівого берега Дніпра
17 Повітряні сили ліквідували російську ракету, яка летіла на Дніпропетровщину
18 Верховний суд оскаржуватиме поновлення на посаді судді Богдана Львова
19 Вогонь HIMARS зійшов на ворожі "Торнадо-Г" і "Мста-С"
```

Рис. 1. Список заголовків новин

Алгоритм VADER виводить оцінки настроїв для 4 класів настроїв:

- neg: негативний;
- neu: нейтральний;
- pos: позитивний;
- compound: складний (сукупний бал).

Нижче подано список заголовків новин, які будуть піддаватися аналізу тональності з використанням вищезазначених інструментів.

Для проведення даного типу аналізу необхідно завантажити тональний словник відповідної мови, здійснити попередню обробку тексту (тобто провести очищення від стоп-слів та приведення слів в нормальну форму) для поліпшення роботи алгоритмів.

```
0 Більшість "слуг" за повернення закону про мобілізацію на доопрацювання – джерела RAW:  0.0 NORM:  0.25
1 У Греції легалізують одностатеві шлюби й усиновлення RAW:  0.0 NORM:  0.0
2 OpenAI офіційно запустила маркетплейс чат-ботів GPT Store RAW:  0.0 NORM:  0.0
3 У Раді Європи визначились із кандидатами на генсека RAW:  0.0 NORM:  0.0
4 Армія РФ вдарила по вугільному підприємству у Мирнограді на Донеччині, загинула людина RAW:  0.0 NORM:  -0.4588
5 Дохід "Укрпошти" торік зріс до майже 12 мільярдів після падіння у 2022-му RAW:  -0.25 NORM:  -0.25
6 Прикордонник "Смурф" про оборону Маріуполя, "Азовсталь", полон, "втому від війни" RAW:  -0.4588 NORM:  -0.6124
7 У Києві знайшли мертвими ексзаступницю генпрокурора та її доньку RAW:  0.0 NORM:  -0.4588
8 Окупанти обстріляли Дніпропетровщину, є поранені RAW:  0.0 NORM:  -0.25
9 На Раді "Україна–НАТО" Києву пообіцяли в цьому році зброї на мільярди євро RAW:  0.0 NORM:  -0.25
10 У Кривому Розі через серйозну аварію частина районів без світла – Вілкул RAW:  0.0 NORM:  0.0
11 Стефанчук пояснив, чому Рада не стала звільняти Безуглу RAW:  0.25 NORM:  0.25
12 Вчені виявили, що відмова від м'яса знижує ризик інфікування коронавірусом RAW:  0.0 NORM:  0.0
13 Росіяни вкрили вогнем село на Херсонщині, постраждало подружжя RAW:  0.0 NORM:  -0.25
14 Складається, як Lego: українець створив сталеве укриття, яке нібито витримує понад 43 тонни RAW:  0.25 NORM:  0.25
15 ОАЕ заборонили пропускати танкери під прапором Камеруну: перевозять російську нафту RAW:  0.0 NORM:  0.0
16 Зведення Генштабу: РФ здійснила 10 спроб вибити українські війська з лівого берега Дніпра RAW:  0.0 NORM:  0.0
17 Повітряні сили ліквідували російську ракету, яка летіла на Дніпропетровщину RAW:  0.0 NORM:  0.0
18 Верховний суд оскаржуватиме поновлення на посаді судді Богдана Львова RAW:  0.25 NORM:  0.0
19 Вогонь HIMARS зійшов на ворожі "Торнадо-Г" і "Мста-С" RAW:  0.0 NORM:  -0.25
```

Рис. 2. Результат тональності заголовків статей

Класифікація текстів. Важливо розрізняти класифікацію та кластеризацію, оскільки остання передбачає групування текстів на основі критеріїв без попередньо визначених категорій. Класифікацію тексту можна виконувати вручну або автоматично, використовуючи набір правил або методи машинного навчання, на різних документах, таких як веб-сторінки, книги та медіа-статті тощо.

Класифікація тексту має багато застосувань, включаючи аналіз настроїв, класифікацію новин, визначення мови, виявлення спаму, маршрутизацію клієнтів, класифікацію резюме тощо. Однак, щоб використовувати текстові значення в моделі машинного навчання, їх потрібно спочатку перетворити на числові значення або функції. Цей процес, відомий як витяг і розробка ознак, включає вилучення інформації з даних у числовому форматі. Зазвичай використовується простий та інтуїтивно зрозумілий метод перетворення тексту в корисні функції TF-IDF (англ. TF — term frequency, IDF — inverse document frequency), щоб уникнути ускладнення завдання.

```
Водонапірну вежу у Хмельницькому перетворять на туристичний об'єкт,Культура,uk
Шону Коннери – 90,Культура,uk
Стрічка Слабошпицького про Чорнобиль перемогла на фестивалі в Белграді,Культура,uk
Табачник просить вчителів звозити школярів до Шевченка,Культура,uk
Російська компанія контролюватиме 50% книжкового ринку України,Культура,uk
До кінця року знімуть 10 нових повнометражних українських фільмів,Культура,uk
Завтра у Києві нагородять лауреатів книжкового рейтингу НСПУ,Культура,uk
Завтра в Луцьку відкриють відреставроване приміщення художнього музею,Культура,uk
У Сумах відкрилася унікальна виставка акварелей,Культура,uk
800 студентів зібрали кошти на меморіальну дошку Назарію Яремчуку,Культура,uk
Фільми Тарантіно та Стівенсона стали національним надбанням США,Культура,uk
ООН схвалила революцію про нагляд в Інтернеті,Культура,uk
"У Вінниці відбудеться фестиваль, присвячений Михайлу Коцюбинському",Культура,uk
Кабмін створив оргкомітет з підготовки і проведення Євробачення-2017,Культура,uk
Протягом трьох місяців в Одесі триватиме літературний фестиваль,Культура,uk
Україна готується до ювілею видатного актора Богдана Ступки,Культура,uk
"Вінницький театр готує прем'єру мюзиклу ""Красуня і чудовисько"" в авторському прочитанні",Культура,uk
Український національний стенд на European Film Market розпочинає свою роботу,Культура,uk
У Центрі сучасного мистецтва M17 відбулась церемонія нагородження  M17 Sculpture Prize,Культура,uk
"Дві українські стрічки потрапили в довгий перелік номінантів на ""Оскар""",Культура,uk
"На Вінниччині видали мартиролог із повним списком подолян, які були вислані на Соловки",Культура,uk
До Вінниці повертається відреставрований стіл Симона Петлюри,Культура,uk
"Бібліотека ""Гомера Гуцульщини"" повернулася до Львівського історичного музею",Культура,uk
```

Рис. 3. Текстовий набір даних для подальшої класифікації

Після перетворення тексту у функції наступним кроком є вибір моделі для класифікації. Важливо спробувати різні моделі, перш ніж вибрати найкращу, і scikit-learn полегшує це, дозволяючи користувачам порівнювати результати різних моделей за допомогою функцій fit() і predict().

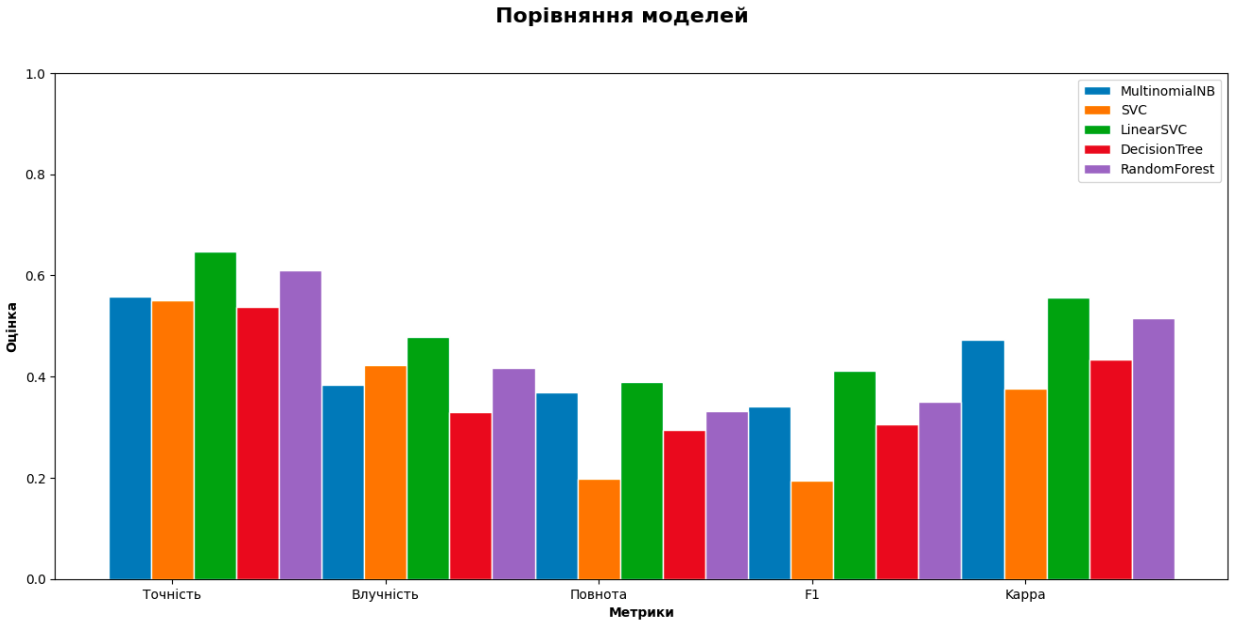

Рис. 4. Результат порівняння запропонованих моделей

Для класифікації тексту було реалізовано кілька моделей:

- Multinomial Naive Bayes
- Support Vector Machines (SVM, LinearSVM)
- Neural Network with Softmax Layer
- Decision Trees
- Random Forests

Оцінка моделей класифікаторів передбачає обчислення таких показників:

- Точність
- Влучність
- Повнота
- F1 міра
- Cohens Kappa міра
- Матриця невідповідностей

Оцінюється здатність кожного класифікатора правильно класифікувати тексти, а для вивчення результатів і розрахунку показників використовується матриця невідповідностей. Зрештою, щоб визначити, яка модель є найефективнішою, ми повинні розрахувати та порівняти точність кожного класифікатора. Визначення функції для оцінки класифікаторів спрощує ітераційний процес.

Коли найкращу модель знайдено та налаштовано, важливо зберегти її для майбутніх прогнозів без необхідності перенавчати класифікатор. Векторизатор, який використовується для зіставлення даних, також має залишатися однаковим для тестових і прогнозованих даних.

```
Точність : 64.68%
Категорія:  Культура
Заголовок:  Стрічка Слабошпицького про Чорнобиль перемогла на фестивалі в Белграді
```

Рис. 5. Результат класифікація одного із елементів з набору даних

Кластеризація текстів. Передбачає групування колекції непозначених текстів разом на основі їх подібності. Мета полягає в тому, щоб тексти в одному кластері були більш схожі один на одного, ніж на тексти в інших кластерах. Кілька алгоритмів кластеризації доступні в Scikit-Learn та інших бібліотеках, причому кластеризація k-середніх є однією з найпростіших для розуміння.

Алгоритм шукає заздалегідь визначену кількість кластерів у немаркованому багатовимірному масиві даних, використовуючи припущення, що кожен кластер має «центр кластера», який є середнім арифметичним усіх точок, що належать кластеру. Крім того, кожна точка в кластері знаходиться ближче до свого центру, ніж до інших центрів кластера.

Для кластеризації тексту ми будемо використовувати список новин, що складається з необроблених заголовків статей (табл. 1). Однак попередня обробка тексту необхідна для підвищення продуктивності алгоритму. Це передбачає очищення стоп-слів і переведення слів у їхню звичайну форму, щоб зменшити шум і переконатися, що алгоритм не зважує минулий час над іменниками теперішнього чи множини над іменниками однини.

Таблиця 1

Фрагмент набору необроблених заголовків статей

Заголовок	Категорія	Мова
У Луцьку протестують медики клінічного пологового будинку	Коронавірус	uk
Німеччина вводить обов'язковий карантин для громадян з "червоних" країн	Коронавірус	uk
Wizz Air скасував більше 20 рейсів з України на листопад	Коронавірус	uk
Іран готовий забезпечити однаковий розмір компенсації родичам всіх жертв катастрофи МАУ, - Єнін	Політика	uk
Facebook заблокував 2,2 млн рекламних постів, що стосувались виборів у США	Політика	uk
У ПАРЄ вважають поганим сигналом вихід Росії з переговорів про МН17	Політика	uk
Зеленський ушанував пам'ять воїнів, загиблих за незалежність України	Події	uk
На Вінниччині відкрили меморіал "Борцям за незалежність України"	Події	uk
Українська Рада бізнеса – керівнику офісу Президента: не варто прислухатися до тіньовиків, почуйте доброчесний бізнес	Суспільство	uk
На Вінниччині вже розпочався продаж новорічних ялинок	Суспільство	uk
На Вінниччині з'явився пам'ятник томату	Суспільство	uk
Тираж книги Кіпіані "Правда Василя Стуса" розкупили за лічені хвилини після вироку суду про заборону її розповсюдження, - видавництво	Культура	uk
Видавництво подасть апеляцію на рішення Дарницького суду щодо заборони книги Кіпіані про Стуса	Культура	uk
У Луцьку відкрили виставку друкованих матеріалів УПА	Культура	uk
Медзакупівлі України заощадили понад 1 млрд грн на закупівлях лікзасобів у ProZorro	Економіка	uk
Нафтогаз підвищив ціни на газ для виробників тепла ще на чверть	Економіка	uk
Україна експортувала вже 15 мільйонів тонн зерна	Економіка	uk
Канада, Норвегія і Литва рекомендують не інвестувати в Україну через кризу в "зеленій" енергетиці	Екологія	uk
На півдні Туреччини вирує масштабна лісова пожежа	Екологія	uk

Після попередньої обробки ми створимо два списки словників: лише бази та лише токенізація. Потім ми запровадимо алгоритм TF-IDF Vectorizer, щоб отримати результуючу матрицю (табл. 2). Може бути складно визначити кількість кластерів для призначення даним. За допомогою KMeans дані надто сильно кластеризовані для ієрархічної кластеризації або будь-якого алгоритму, який знаходить кластери самостійно. Тому ми використаємо метод для оцінки кількості кластерів, які з'являться в наборі даних.

Після визначення кластерів кожну статтю можна пов'язати з відповідним кластером.

Таблиця 2

Фрагмент набору нормалізованих та кластеризованих заголовків статей

Індекс	Заголовок	Категорія	Мова	Нормалізація	Кластер
0	У Луцьку протестують медики клінічного пологового будинку	Коронавірус	uk	луцький протестувати медик клінічний пологовий будинок	3
1	Німеччина вводить обов'язковий карантин для громадян з "червоних" країн	Коронавірус	uk	німеччина вводити обов ' язковити карантин громадянин `` червоний " країна	5
2	Wizz Air скасував більше 20 рейсів з України на листопад	Коронавірус	uk	wizz air скасувати 20 рейс україна листопад	10
3	Іран готовий забезпечити однаковий розмір компенсації родичам всіх жертв катастрофи МАУ, - Єнін	Політика	uk	іран готовий забезпечити однаковий розмір компенсація родич жертва катастрофа мау єнін	9
4	Facebook заблокував 2,2 млн рекламних постів, що стосувались виборів у США	Політика	uk	facebook заблокувати 2,2 рекламний пост стосуватися вибір сша	8
5	У ПАРЄ вважають поганим сигналом вихід Росії з переговорів про МН17	Політика	uk	парє вважати поганий сигнал вихід росія переговори мн17	6
6	Зеленський ушанував пам'ять воїнів, загиблих за незалежність України	Події	uk	зеленський ушанувати пам ' ять воїн загиблий незалежність україна	24
8	На Вінниччині відкрили меморіал "Борцям за незалежність України"	Події	uk	вінниччина відкрити меморіал `` борець незалежність україна "	20
9	Українська Рада бізнеса – керівнику офісу Президента: не варто прислухатися до тіньовиків, почуйте доброчесний бізнес	Суспільство	uk	український рад бізнес – керівник офіс президент варта прислухатися тіньовик почути доброчесний бізнес	13
10	На Вінниччині вже розпочався продаж новорічних ялинок	Суспільство	uk	вінниччина розпочатися продаж новорічний ялинка	22
11	На Вінниччині з'явився пам'ятник томату	Суспільство	uk	вінниччина ' явитися пам ' ятникнути томат	4

Нарешті, ми можемо надрукувати кластери та пов'язані назви статей (рис. 6). Деякі кластери можуть бути кращими за інші, але, виходячи з назв, K-Means, здається, розділив статті на відносно послідовні кластери.

```
Топ слова у кластері:

Кластер 0 слова: луцький, виставка, матеріал, упа, друкований, медик,

Кластер 0 назви:
 – У Луцьку протестують медики клінічного пологового будинку
 – У Луцьку відкрили виставку друкованих матеріалів УПА


Кластер 1 слова: вінниччина, пам, незалежність, бізнес, україна, томат,

Кластер 1 назви:
 – Німеччина вводить обов'язковий карантин для громадян з "червоних" країн
 – Іран готовий забезпечити однаковий розмір компенсації родичам всіх жертв катастрофи МАУ, – Єнін
 – Facebook заблокував 2,2 млн рекламних постів, що стосувались виборів у США
 – У ПАРЄ вважають поганим сигналом вихід Росії з переговорів про МН17
 – Зеленський ушанував пам'ять воїнів, загиблих за незалежність України
 – Харьков попал в "красную" зону карантина
 – На Вінниччині відкрили меморіал "Борцям за незалежність України"
 – Українська Рада бізнеса – керівнику офісу Президента: не варто прислухатися до тіньовиків, почуйте доброчесний бізнес
 – На Вінниччині вже розпочався продаж новорічних ялинок
 – На Вінниччині з'явився пам'ятник томату
 – Нафтогаз підвищив ціни на газ для виробників тепла ще на чверть
 – На півдні Туреччини вирує масштабна лісова пожежа
 – Рибоохоронний патруль зафіксував масову загибель хамси у Молочному лимані
 – Старікова виграла п'яту медаль чемпіонату Європи з велотреку в Пловдиві


Кластер 2 слова: суд, стус, видавництво, книга, кіпіані, заборонити,

Кластер 2 назви:
 – Тираж книги Кіпіані "Правда Василя Стуса" розкупили за лічені хвилини після вироку суду про заборону її розповсюдження, – видавництво
 – Видавництво подасть апеляцію на рішення Дарницького суду щодо заборони книги Кіпіані про Стуса


Кластер 3 слова: транспорт, громадський, призвести, повний, степанов, локдаун,

Кластер 3 назви:
 – Зупинка громадського транспорту призвела б до повного локдауну, – Степанов
 – У вінницькому громадському транспорті почали встановлювали лічильники, щоб порахувати "зайців"
 – Відсутність державної політики у транспорті веде галузь до занепаду, – Олександр Скічко


Кластер 4 слова: україна, збірний, 15, мільйон, експортувати, тонна,

Кластер 4 назви:
 – Wizz Air скасував більше 20 рейсів з України на листопад
 – Медзакупівлі України заощадили понад 1 млрд грн на закупівлях ліказасобів у ProZorro
 – Україна експортувала вже 15 мільйонів тонн зерна
 – Канада, Норвегія і Литва рекомендують не інвестувати в Україну через кризу в "зеленій" енергетиці
 – Ще троє гравців збірної України з футболу захворіли на коронавірус
 – Баскетболістки збірної України перемогли Фінляндію у відборі Євро-2021
```

Рис. 6. Результат кластеризації набору тексту заголовків статей

Загалом, кластеризація тексту є корисним інструментом для групування непозначених текстів разом і виявлення шаблонів або подібностей у них.

Тематичне моделювання. Передбачає захоплення семантичної інформації, яка виходить за рамки окремих слів, виявлення прихованих тем або тем у документах, відповідне анотування документів і використання цих анотацій для керування, узагальнення, пошуку та рекомендації вмісту.

В сферах машинного навчання та обробки природної мови тематичне моделювання — це статистична модель, яка використовується для ідентифікації абстрактних «тем» у колекції документів. Це інструмент, який часто використовується для аналізу тексту, що дозволяє ефективно аналізувати великі обсяги тексту шляхом групування документів у кластери та виявлення прихованих семантичних структур у них.

Оскільки кількість доступних текстових даних дуже велика, традиційні підходи до навчання під наглядом не застосовуються, оскільки до набору текстових даних немає зручної позначки. Замість цього використовується тематичне моделювання, щоб розкрити ці ярлики та виявити різні категорії в даних, наприклад, ті, що містяться в газетній статті.

Поширений метод тематичного моделювання, прихований розподіл Діріхле (англ. Latent Dirichlet allocation, далі LDA), використовує матричну факторізацію та зазвичай використовується в цьому дослідженні. Представляючи корпус у вигляді матриці термінів-документів, важливо зазначити, що оцінювання неконтрольованого навчання може бути складним, оскільки немає визначеної правильної теми чи відповіді. Однак ми можемо ідентифікувати схожі тематичні ідеї, спільні для згрупованих документів, і визначити їх значення самостійно.

З точки зору підготовки даних векторизатор налаштовується з певними параметрами. Параметр max_df=.2 ігнорує терміни, які зустрічаються в більш ніж 20% документів, тоді як max_df = 3 ігнорує терміни, які зустрічаються в більш ніж трьох документах. Крім того, max_features встановлює розмір випадкових підмножин функцій, які слід враховувати під час розділення вузла, по суті видаляючи терміни, які зазвичай використовуються в багатьох документах. Корисно виключити часті та нерелевантні слова, а також можливі опечатки чи орфографічні помилки.

Для ефективного використання LDA для тематичного моделювання ми можемо зробити два основних

припущення. По-перше, документи схожої тематики будуть містити схожі групи слів. Це розумне припущення, оскільки воно передбачає, що документи, які стосуються однієї теми, як-от бізнес чи економіка, міститимуть такі слова, як гроші, ціна та частки ринку. По-друге, приховані теми можна виявити, досліджуючи групи слів, які зазвичай зустрічаються разом у всьому корпусі. Ці два припущення можна змоделювати математично, розглядаючи документи як розподіли ймовірностей прихованих тем, а самі теми — як розподіли ймовірностей слів. Кожному документу можна призначити розподіл ймовірностей за встановленою кількістю прихованих тем, наприклад, вісім, що забезпечує можливість належності до кожної теми. В таблиці 3 подано найпопулярніші слова у темах, а на рисунку 7 – як саме перетинаються теми і категорії.

Таблиця 3

Найпопулярніші слова у темах

	Тема 1	Тема 2	Тема 3	Тема 4	Тема 5	Тема 6	Тема 7	Тема 8
0	іран	covid	та	карантин	пам	український	світ	карантинний
1	виставка	19	сша	день	ять	2021	створити	червоний
2	літак	випадок	підписати	локдаун	рік	львівський	почати	зона
3	мау	доба	фото	вихідний	єр	голова	кількість	обмеження
4	сша	коронавірус	зеленський	ввести	загиблий	міський	планувати	та
5	запустити	новий	померти	ситуація	українка	рок	скасувати	зон
6	катастрофа	виявити	посол	проти	провести	кінець	мільйон	область
7	мзс	зафіксувати	закликати	щодо	президент	продовжити	місяць	штат
8	відкритися	вакцина	закон	зеленський	засідання	та	працювати	кий
9	ракета	кий	український	люди	відставка	лікар	відбутися	байден
10	український	вакцинація	белт	санкція	закликати	чехія	та	сполучений
11	українець	моз	версиин	жорсткий	вінниччина	коронавирус	пандемія	сша
12	супутник	вооз	лучшиий	вакцинація	зеленський	вінницький	візит	людина
13	транспорт	хворий	недела	рішення	вшанувати	рад	рок	11
14	кримський	щеплення	мова	час	оон	шмигаль	млрд	відкрити
15	музей	підтвердити	вибух	вважати	формат	питання	китайський	карантин
16	змі	понад	указ	план	жертва	головний	відвідати	оновити
17	spacex	степанов	національний	уряд	оголосити	перший	законопроєкт	кличко

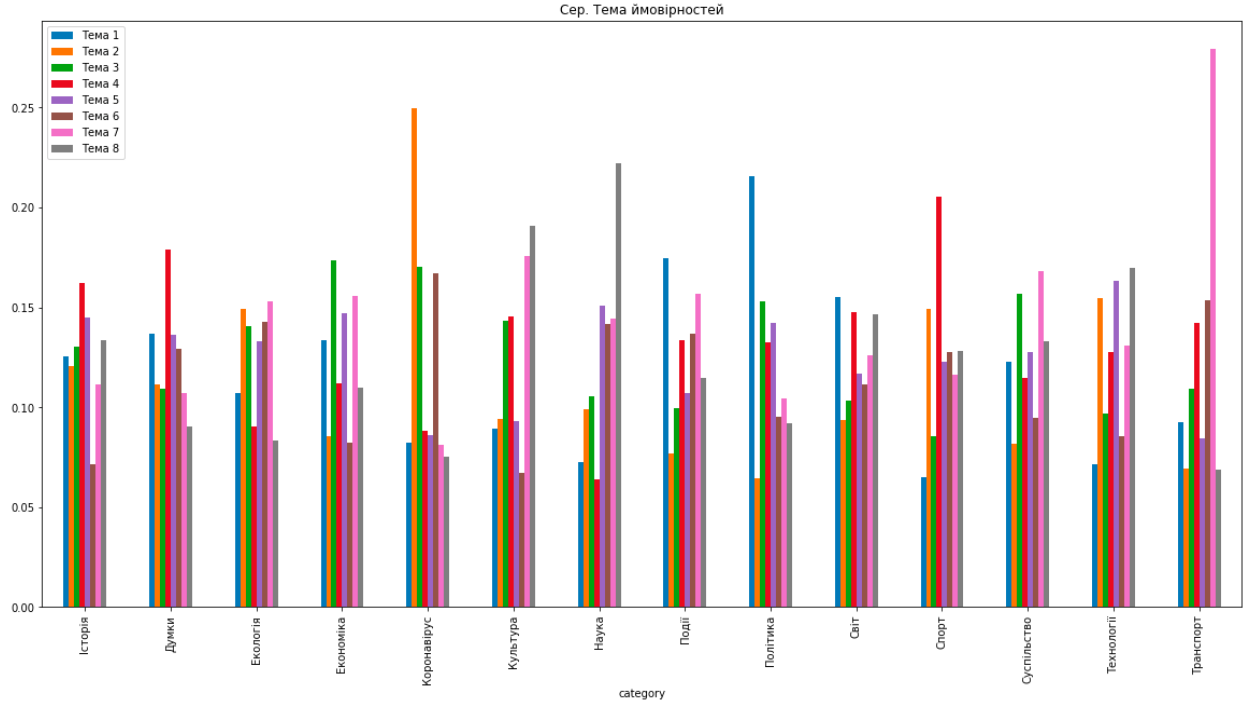

Рис. 7. Перетин теми і категорії

Відмінність категорій «Коронавірус», «Наука», «Політика», «Спорт» і «Транспорт» очевидна, і кожна з них має чіткий зв’язок з іншими темами. Оцінку цих тем можна покращити, перевіряючи їх за набором критеріїв і використовуючи відповідну інформацію з теми та гіпотези. Оцінюючи таким чином якість інформації, можна критикувати тему та додатково аналізувати вихідний текст (рис. 8).

За допомогою інтерактивного інструменту pyLDAvis, який дозволяє досліджувати значення об’єктів і взаємозв’язків, зображено подання (рис. 9) розподілу набору термінів за темами (ліворуч) та співвідношення кількості термінів у темі до їх загальної кількості (праворуч). Це надасть користувачам можливість зрозуміти загальні зв’язки та спосіб, у який предмети стосуються кожного з них.

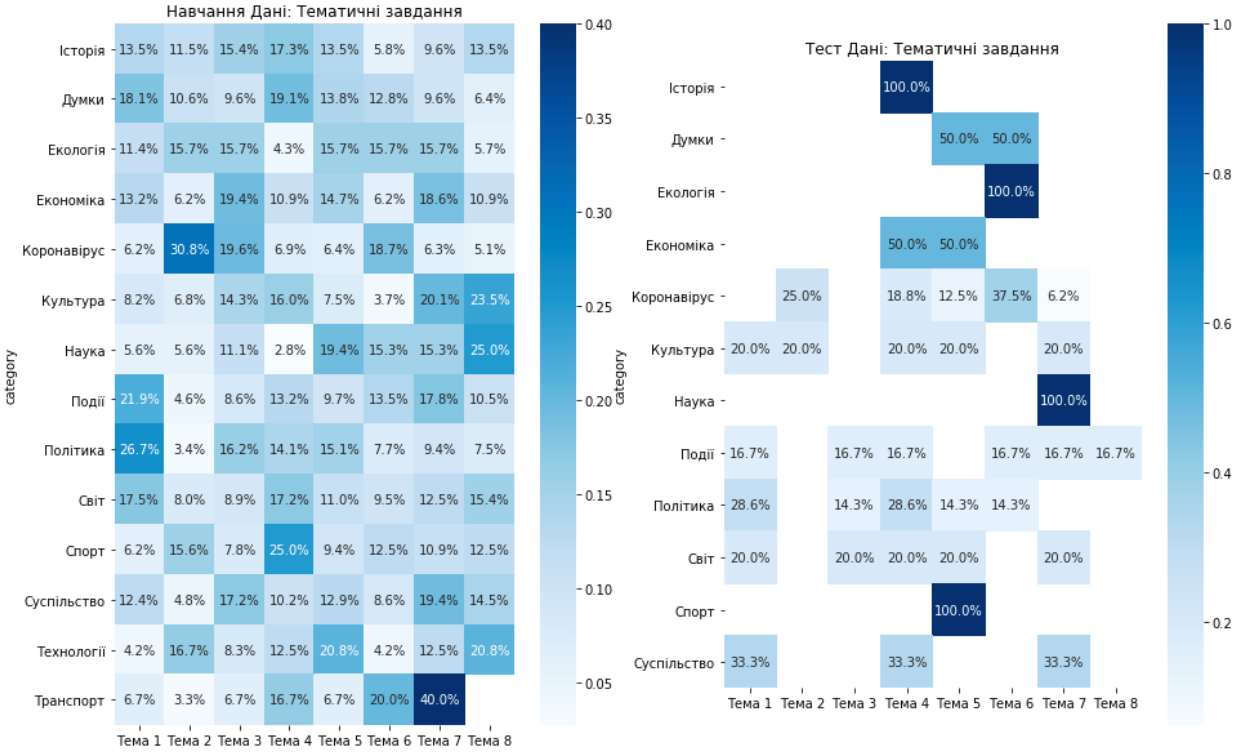

Рис. 8. Оцінка даних для навчання (ліворуч) та тестових даних (праворуч)

Рис. 9. Візуалізація залежності термінів в кожній з тем відносно їх загальної кількості

Висновки

За отриманими результатами можна стверджувати, що обробка природної мови відіграє значну роль у розвитку інтелектуальних технологій і має великий потенціал для подальшого розвитку. Визначення ключових викликів та напрямків дослідження для інновацій дозволяє краще зрозуміти поточний стан та майбутні перспективи NLP.

References

1. Alshemali, B., & Kalita, J. (2020). Improving the Reliability of Deep Neural Networks in NLP: A review. Knowledge-Based Systems, 191, 105210. https://doi.org/10.1016/j.knosys.2019.105210.

2. Zhou, M., Duan, N., Liu, S., & Shum, H. Y. (2020). Progress in neural NLP: modeling, learning, and

reasoning. Engineering, 6(3), 275–290. https://doi.org/10.1016/j.eng.2019.12.014.

3. Shaham, U., Elbayad, M., Goswami, V., Levy, O., & Bhosale, S. (2023). Causes and Cures for Interference in Multilingual Translation. Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). https://doi.org/10.18653/v1/2023.acl-long.883.

4. Turchin, A., & Builes, L. F. F. (2021). Using natural language processing to measure and Improve quality of diabetes care: A Systematic review. Journal of Diabetes Science and Technology, 15(3), 553–560. https://doi.org/10.1177/19322968211000831.

5. Fritz. (2023, September 21). The 7 NLP techniques that will change how you communicate in the future (Part I). Fritz Ai. https://fritz.ai/top-nlp-techniques/.

6. Onyshchenko A., Onyshchenko K. Adaptive method of training neural networks. Technique and technology. Science, Research, Development #29. Gdansk, 2020. p. 9-11.