УДК 004.622 DOI https://doi.org/10.35546/kntu2078-4481.2024.1.31

Н. А. КУЛИКОВСЬКА
старший викладач кафедри комп'ютерних систем та мереж
Національний університет «Запорізька політехніка»
ORCID: 0000-0003-4691-5102

А. В. ТІМЕНКО
старший викладач кафедри комп'ютерних систем та мереж
Національний університет «Запорізька політехніка»
ORCID: 0000-0002-7871-4543

В. Є. ТРОХИМЧУК
студент кафедри комп'ютерних систем та мереж
Національний університет «Запорізька політехніка»
ORCID: 0009-0000-1733-171X

М. Б. ІЛЬЯШЕНКО
кандидат технічних наук,
доцент комп'ютерних систем та мереж
Національний університет «Запорізька політехніка»
ORCID: 0000-0003-4624-4687

ДОСЛІДЖЕННЯ МЕТОДІВ АНАЛІЗУ ТОНАЛЬНОСТІ ТЕКСТОВИХ ДАНИХ

Актуальність теми дослідження визначається лавиноподібним зростанням обсягів неструктурованих текстових даних в Інтернеті та потребою в ефективних методах аналізу тональності. Мета роботи – систематично вивчити сучасний стан методології аналізу тональності, порівняти провідні підходи і окреслити подальші перспективи. У статті детально проаналізовано популярні бібліотеки Python для обробки природної мови – NLTK, spaCy, TextBlob, Gensim. Порівняння проведено за критеріями обчислювальної ефективності, зручності використання, гнучкості екстракції ознак та можливостей кастомізації. Методологічне ядро дослідження становить експериментальне порівняння NLTK і TextBlob для класифікації тональності україномовних текстів. Оцінки можуть варіюватися в залежності від конкретного сценарію використання та налаштувань. NLTK, де він може бути більш точним, коли його правильно налаштовано, але вимагає більше зусиль у налаштуванні. TextBlob, навпаки, є більш простим для використання, але може бути менш точним для спеціалізованих завдань. Результати засвідчили переваги TextBlob у швидкодії та NLTK у точності. Аналіз тональності має величезний потенціал для вдосконалення аналітичних можливостей в багатьох сферах – від оптимізації бізнес-процесів до протидії поширенню фейкових новин. Подальші дослідження повинні фокусуватися на розробці спеціалізованих рішень під конкретні прикладні задачі.Визначено перспективи вдосконалення етичних принципів аналізу тексту, урахування лінгвістичного та культурного контексту, а також інтеграції функціоналу аналізу тональності в системи підтримки прийняття рішень.

Ключові слова: *аналіз тональності, обробка природної мови, машинне навчання, Python, NLTK, spaCy, TextBlob, Gensim.*

N. A. KULYKOVSKA
Senior Lecturer at the Department of Computer Systems and Networks
National University Zaporizhzhia Polytechnic
ORCID: 0000-0003-4691-5102

A. V. TIMENKO
Senior Lecturer at the Department of Computer Systems and Networks
National University Zaporizhzhia Polytechnic
ORCID: 0000-0002-7871-4543

V. E. TROKHYMCHUK
Student at the Department of Computer Systems and Networks
National University Zaporizhzhia Polytechnic
ORCID: 0009-0000-1733-171X

M. B. ILYASHENKO
Ph.D., Associate Professor at the Department of Computer Systems and Networks
National University Zaporizhzhia Polytechnic
ORCID: 0000-0003-4624-4687

STUDY ON SENTIMENT ANALYSIS METHODS FOR TEXT DATA

The relevance of the research topic is determined by the avalanche-like growth of unstructured text data on the Internet and the need for effective methods of tonality analysis. The purpose of the work is to systematically study the current state of the tonality analysis methodology, compare the leading approaches and outline further prospects. The article analyzes in detail popular Python libraries for natural language processing – NLTK, spaCy, TextBlob, Gensim. The comparison was made according to the criteria of computational efficiency, ease of use, flexibility of feature extraction, and customization options. The methodological core of the study is an experimental comparison of NLTK and TextBlob for tonality classification of Ukrainian-language texts. Estimates may vary depending on specific usage scenario and settings. NLTK, where it can be more accurate when configured correctly, but requires more effort to configure. TextBlob, on the other hand, is easier to use, but may be less accurate for specialized tasks. The results proved the superiority of TextBlob in speed and NLTK in accuracy. Tonality analysis has a huge potential for improving analytical capabilities in many areas – from optimizing business processes to countering the spread of fake news. Further research should focus on the development of specialized solutions for specific applied tasks. Prospects for improving the ethical principles of text analysis, taking into account the linguistic and cultural context, as well as the integration of the tonality analysis functionality into decision support systems have been determined.

Key words: *sentiment analysis, natural language processing, machine learning, Python, NLTK, spaCy, TextBlob, Gensim.*

Постановка проблеми

В тенденції сучасного розвитку цифрової інформації обсяги текстових даних, які генеруються в Інтернеті, зростають з неймовірною швидкістю. Соціальні мережі, блоги, форуми, відгуки на товари та послуги – все це стало не лише засобом вираження думок та почуттів людей, а й важливим джерелом інформації для компаній, науковців, урядових та некомерційних організацій. Аналіз тональності, або сентимент-аналіз, стає ключовим інструментом для розуміння громадської думки, реакцій споживачів, політичних настроїв та багатьох інших аспектів суспільного життя. Враховуючи актуальність та масштабність задачі, дослідження методів аналізу тональності у текстових даних набуває особливого значення [1, с. 72; 2, с. 10; 3, с. 42].

Аналіз останніх досліджень і публікацій

Аналіз останніх досліджень свідчить, що аналіз тональності текстів є активною галуззю досліджень в останні роки. Багато робіт присвячено застосуванню методів машинного навчання, зокрема нейронних мереж, для аналізу тональності [4, с. 55]. Поряд з цим, активно вивчаються лінгвістично-орієнтовані методи, що враховують морфологічні та синтаксичні особливості конкретних мов [5, с. 1]. Значна увага приділяється питанням аналізу тональності текстів соціальних мереж та відгуків споживачів в Інтернеті [6, с. 10]. Це дозволяє оцінювати ставлення користувачів до брендів, продуктів, послуг тощо.

Ряд досліджень присвячено порівнянню ефективності різних підходів до аналізу тональності за показниками точності, швидкодії тощо [7, с. 57; 8, с. 63; 9, с. 79]. Зокрема аналізуються такі поширені бібліотеки як NLTK, TextBlob, spaCy.

Виокремлюється напрям, пов'язаний з розробкою гібридних методів аналізу тональності, що поєднують підходи машинного навчання і обробки природної мови [10, с. 476]. Такі комбіновані методи дозволяють досягти вищої ефективності для конкретних задач [11, с. 569].

Отже, актуальними тенденціями є удосконалення існуючих алгоритмів аналізу тональності, розробка спеціалізованих рішень для конкретних мов і практичних завдань, а також створення гібридних підходів шляхом поєднання методів.

Формулювання мети дослідження

Об'єкт дослідження це методи аналізу тональності текстових даних. Предмет дослідження це ефективність та застосовність різних методів аналізу тональності тексту, таких як методи машинного навчання та обробки природної мови. Мета дослідження: дослідити сучасний стан у галузі аналізу тональності тексту, визначити основні виклики та перспективи розвитку, порівняти ефективність різних підходів.

Викладення основного матеріалу дослідження

Технологічний сегмент штучного інтелекту, який займається вивченням взаємодії між комп'ютерами та людською мовою, відомий як обробка природної мови (NLP) [12, с. 260]. Цей напрямок фокусується на тому, як комп'ютери можуть інтерпретувати, обробляти та створювати мову, яку люди використовують для комунікації. Тест Тюрінга виступає як критерій, що вимірює здатність комп'ютера вести діалог з людиною так, щоб його не можна було відрізнити від спілкування з іншою людиною, що є важливим етапом у розвитку NLP. У контексті дослідження, Python виступає як вирішальний інструмент, надаючи значні переваги для аналізу тональності в текстових даних, завдяки своїй універсальності та зручності. Python пропонує розширені можливості для ефективного керування складними проектами.

В дослідженні розглянуті такі методи обробки природної мови як NLTK, spaCy, TextBlob і Gensim, кожен з яких відрізняється за цільовим призначенням, легкістю використання, функціональністю та областями застосування (табл. 1).

Таблиця 1

Методи обробки природної мови як NLTK, spaCy, TextBlob і Gensim

Критерій	NLTK	spaCy	TextBlob	Gensim
Цільове призначення	Інструмент для дослідження і розробки NLP-рішень	Орієнтований на продуктивність для швидкого аналізу тексту в реальному часі	Спрощення обробки тексту та аналізу сентименту	Робота з векторними представленнями слів і тематичним моделюванням
Легкість використання	Має багато функцій, може бути складним для новачків	Дружній API і простий інтерфейс, легкий для новачків	Надзвичайно простий у використанні, інтуїтивно зрозумілий API	Простий інтерфейс для спеціалізованих задач
Функціональність	Широкий набір функцій для різних завдань NLP	Швидка обробка з виявленням сутностей, частин мови тощо	Зосереджено на аналізі сентименту та токенізації	Спеціалізований на векторних представленнях і тематичному моделюванні
Типові використання	Дослідницькі роботи, навчальні цілі	Розробка продуктів, що потребують обробки тексту в реальному часі	Прості завдання аналізу сентименту і швидкий аналіз тексту	Векторне представлення слів та тематичне моделювання в текстових даних

У термінах легкості використання, NLTK, хоча і потужний, може бути складним для новачків через свою багатофункціональність. SpaCy та TextBlob вирізняються своєю простотою та інтуїтивно зрозумілими API, що робить їх зручними для ширшого кола користувачів. Gensim також пропонує простий інтерфейс, особливо для задач, пов'язаних з векторними представленнями та тематичним моделюванням.

З точки зору функціональності, NLTK надає комплексні можливості для різноманітних завдань NLP, від токенізації до аналізу сентименту. SpaCy оптимізований для швидкої обробки з виявленням сутностей та аналізом частин мови. TextBlob зосереджується на аналізі сентименту і токенізації, але має більш обмежений набір функцій порівняно з іншими бібліотеками. Gensim є ідеальним інструментом для роботи з великими текстовими наборами даних для векторного представлення слів та тематичного моделювання.

Для аналізу перелічених бібліотек були взяті кількісні показники їх використання на GitHub (рис. 1).

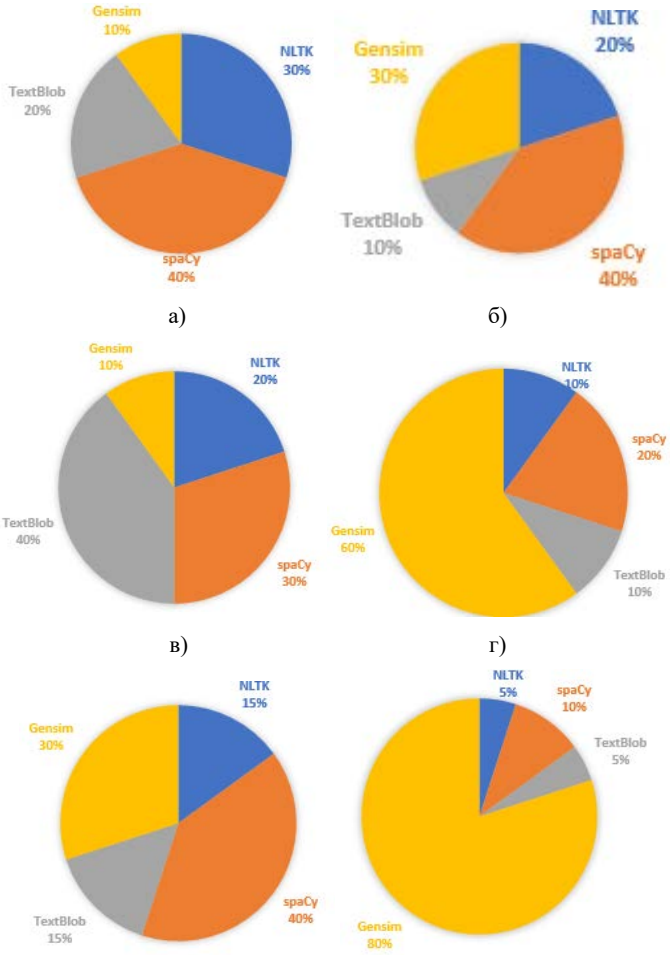

Рис. 1. Порівняння методів обробки природної мови як NLTK, spaCy, TextBlob і Gensim
а) відсоток використання бібліотек для обробки тексту та виконання основних NLP завдань; б) відсоток використання бібліотек для розробки пошукової системи; в) відсоток використання бібліотек для аналізу сентименту; г) відсоток використання бібліотек для векторного представлення слів та тематичне моделювання; д) відсоток використань бібліотек для створення продуктів, які вимагають обробки тексту в реальному часі; е) відсоток використання для векторного представлення слів та тематичного моделювання у текстових даних

Порівняння бібліотек NLTK і TextBlob є особливо актуальним у контексті аналізу тональності тексту, оскільки обидва інструменти пропонують унікальні підходи та можливості для вирішення поставленого завдання.

У дослідженні було використано, два інструменти NLTK:

– інструмент «Аналіз настроїв VADER» (генерує позитивні, негативні і нейтральні оцінки настроїв для заданих вхідних даних);

– інструмент токенізатора «word_tokenize» (розбиває великий текст на послідовність більш дрібних одиниць, таких як речення або слова).

Попередня обробка тексту використовується для поліпшення роботи алгоритмів. Тож, буде виконано очищення від стоп-слів та приведення слова в нормальну форму. Т.к. у NLTK, поки немає корпусу української мови, то для мормологічного аналізу було використано pymorphy2.

Було пораховано оцінку настрою для наданих заголовків за допомогою VADER, щоб зрозуміти, на що здатний цей інструмент. Рисунок 2 ілюструє вивід оцінки для 4 класів настроїв.

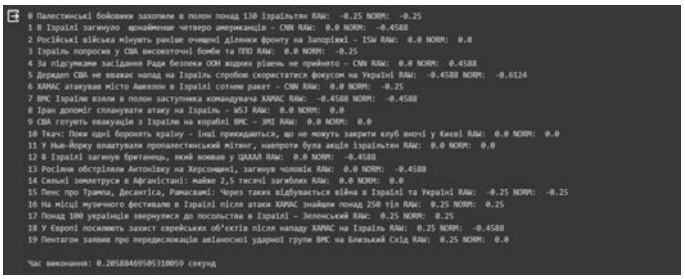

Рис. 2. Результат оцінки настрою для наданих заголовків NLTK

Наступний крок дослідження була використана бібліотека TextBlob. Рисунок 3 ілюструє вивід оцінки настроїв для алгоритму TextBlob.

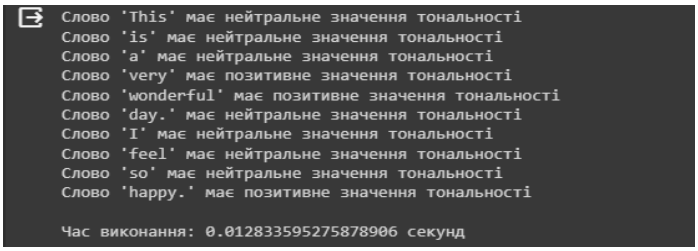

Рис. 3. Результат обчислення тональності тексту TextBlob

Для порівняння двох методів був розроблено програмний модуль, який рахує час роботи кожного алгоритму за однаковими вхідними текстовими даними (рис. 4).

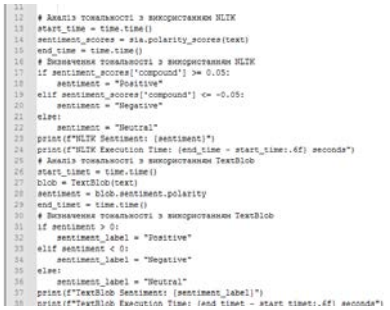

Рис. 4. Порівняння за швидкістю методи NLTK та TextBlob

Таблиця 2 наводить оцінки, які можуть варіюватися в залежності від конкретного сценарію використання та налаштувань. NLTK, де він може бути більш точним, коли його правильно налаштовано, але вимагає більше зусиль у налаштуванні. TextBlob, навпаки, є більш простим для використання, але може бути менш точним для спеціалізованих завдань.

Таблиця 2

Порівняння методів за точністю, зручністю, гнучкістю та налаштованістю

Метод	Точність виявлення	Зручність у використанні	Гнучкість та налаштованість
NLTK	80%	70%	90%
TextBlob	70%	90%	60%

Висновки

У статті проаналізовано можливості поширених бібліотек Python для обробки природної мови, таких як NLTK, spaCy, TextBlob, Gensim. Встановлено, що кожна з них має свої переваги і недоліки залежно від конкретних цілей та завдань. Експериментально порівняно ефективність алгоритмів NLTK (VADER) і TextBlob для аналізу тональності україномовних текстів. З'ясовано, що TextBlob продемонстрував кращу швидкодію, проте гіршу точність в порівнянні з VADER. Обґрунтовано доцільність розробки спеціалізованих алгоритмів аналізу тональності для конкретних мов, предметних областей і практичних задач. Визначено перспективні напрями

досліджень в етичних аспектах аналізу тексту та інтеграції функціоналу аналізу тональності в комплексні системи підтримки прийняття рішень.

Список використаної літератури

1. Ivokhin E., Makhno M., Rets V. Про один спосіб аналізу тональності текстів за допомогою штучних нейронних мереж. *Системи управління, навігації та зв'язку. Збірник наукових праць*. 2022. Т. 3, № 69. С. 71–74.

2. Orel A. Social media analyzing for evaluation opinions determination based on sentiment analysis. *International scientific journal "Internauka"*. 2018. No. 10.

3. Deng Y. Research on sentiment analysis methods for text-oriented data. *Frontiers in computing and intelligent systems*. 2023. Vol. 3, no. 1. P. 42–47.

4. Mukasheva A. Tasks and methods of text sentiment analysis. *Scientific journal of astana IT university*. 2021. No. 7. P. 55–62.

5. Abonizio H. Q., Paraiso E. C., Barbon Junior S. Toward text data augmentation for sentiment analysis. *IEEE transactions on artificial intelligence*. 2021. P. 1.

6. Samigulin T. R., Djurabaev A. E. U. Sentiment analysis of text by machine learning methods. *Research result. Information technologies*. 2021. Vol. 6, no. 1.

7. Yao J. Automated sentiment analysis of text data with NLTK. *Journal of physics: conference series*. 2019. Vol. 1187, no. 5. P. 60–78.

8. A review of text sentiment analysis methods and applications / Y. Jin et al. *Frontiers in business, economics and management*. 2023. Vol. 10, no. 1. P. 58–64.

9. Poria S., Hussain A., Cambria E. Concept extraction from natural text for concept level text analysis. *Multimodal sentiment analysis*. Cham, 2018. P. 79–84.

10. Maran S. M., Esh P. S. Text analysis for product reviews for sentiment analysis using NLP methods. *International journal of engineering trends and technology*. 2017. Vol. 47, no. 8. P. 474–480.

11. Sarkar D. Sentiment analysis. *Text analytics with python*. Berkeley, CA, 2019. P. 567–629.

12. Text as data: text mining and sentiment analysis. *Data mining and business analytics with R*. Hoboken, NJ, USA, 2013. P. 258–271.

References

1. Ivokhin, E., Makhno, M., & Rets, V. (2022). Pro odyn sposib analizu tonal′nosti tekstiv za dopomohoyu shtuchnykh neyronnykh merezh. Systemy upravlinnya, navihatsiyi ta zv″yazku. Zbirnyk naukovykh prats′ [About one way to analyze the tonality of texts using artificial neural networks], 3(69), 71–74. https://doi.org/10.26906/sunz.2022.3.071 [in Ukrainian].

2. Orel, A. (2018). Social media analyzing for evaluation opinions determination based on sentiment analysis. *International Scientific Journal "Internauka"*, (10). https://doi.org/10.25313/2520-2057-2018-10-3858

3. Deng, Y. (2023). Research on sentiment analysis methods for text-oriented data. *Frontiers in computing and intelligent systems*, *3*(1), 42–47. https://doi.org/10.54097/fcis.v3i1.6022

4. Mukasheva, A. (2021). Tasks and methods of text sentiment analysis. *Scientific journal of astana IT university*, (7), 55–62. https://doi.org/10.37943/aitu.2021.57.68.005

5. Abonizio, H. Q., Paraiso, E. C., & Barbon Junior, S. (2021). Toward text data augmentation for sentiment analysis. *IEEE transactions on artificial intelligence*, 1. https://doi.org/10.1109/tai.2021.3114390

6. Samigulin, T. R., & Djurabaev, A. E. U. (2021). Sentiment analysis of text by machine learning methods. *Research result. Information technologies*, *6*(1). https://doi.org/10.18413/2518-1092-2021-6-1-0-7

7. Yao, J. (2019). Automated sentiment analysis of text data with NLTK. *Journal of physics: conference series*, *1187*(5), 60–78. https://doi.org/10.1088/1742-6596/1187/5/052020

8. Jin, Y., Cheng, K., Wang, X., & Cai, L. (2023). A review of text sentiment analysis methods and applications. *Frontiers in business, economics and management*, *10*(1), 58–64. https://doi.org/10.54097/fbem.v10i1.10171

9. Poria, S., Hussain, A., & Cambria, E. (2018). Concept extraction from natural text for concept level text analysis. У *Multimodal sentiment analysis* (с. 79–84). Springer International Publishing. https://doi.org/10.1007/978-3-319-95020-4_4

10. Maran, S. M., & Esh, P. S. (2017). Text analysis for product reviews for sentiment analysis using NLP methods. *International journal of engineering trends and technology*, *47*(8), 474–480. https://doi.org/10.14445/22315381/ijett-v47p278

11. Sarkar, D. (2019). Sentiment analysis. У *Text analytics with python* (с. 567–629). Apress. https://doi.org/10.1007/978-1-4842-4354-1_9

12. Text as data: text mining and sentiment analysis. (2013). У *Data mining and business analytics with R* (с. 258–271). John Wiley & Sons, Inc. https://doi.org/10.1002/9781118596289.ch19