УДК 004.8

Віт Р.В.

Хмельницький національний університет

ПРАКТИЧНА РЕАЛІЗАЦІЯ МЕТОДУ ВИЯВЛЕННЯ ЦИФРОВОГО ВИСНАЖЕННЯ ЗА АНАЛІЗОМ ЦІЛЬОВИХ ОБ'ЄКТІВ МНОЖИНИ ПОВІДОМЛЕНЬ ЛЮДИНИ

У роботі розглянуто поточний стан наукового напряму виявлення цифрового виснаження за аналізом цільових об'єктів множини повідомлень людини у контексті пошуку іменованих сутностей та пошуку ключових слів, та на основі опрацьованого матеріалу запропоновано власний метод виявлення цифрового виснаження за аналізом цільових об'єктів множини повідомлень людини. Цей метод використовує алгоритми машинного навчання для адаптивного розпізнавання об'єктів, враховуючи специфіку предметної області, що дозволяє значно скоротити час обробки даних і знизити ризик втрати важливої інформації. Розроблений метод виявлення цифрового виснаження за аналізом цільових об'єктів множини повідомлень людини призначений для автоматизації процесу ідентифікації ключових елементів у великих масивах текстових даних, й спрямований на підвищення точності та ефективності аналізу текстової інформації.

The paper reviews the current state of the scientific direction of digital exhaustion detection by analyzing target objects of a set of human messages in the context of named entity search and keyword search, and based on the developed material, proposes its own method of digital exhaustion detection by analyzing target objects of a set of human messages. This method uses machine learning algorithms for adaptive object recognition, taking into account the specifics of the subject area, which allows significantly reducing data processing time and reducing the risk of losing important information. The developed method of digital exhaustion detection by analyzing target objects of a set of human messages is designed to automate the process of identifying key elements in large arrays of text data, and is aimed at increasing the accuracy and efficiency of text information analysis.

Методи виявлення цільових об'єктів у предметній області є критично важливими для ефективного аналізу та обробки великих обсягів інформації. В умовах зростаючої складності даних, які охоплюють різноманітні предметні області, необхідність розробки та вдосконалення методів автоматизованого виявлення цільових об'єктів стає все більш актуальною [1]. Це особливо важливо в таких сферах, як штучний інтелект, а саме системи обробки природної мови та інформаційний пошук. Відсутність надійних та ефективних методів виявлення цільових об'єктів може призвести до втрати важливої інформації, зниження точності прийняття рішень та збільшення витрат на аналіз даних. Враховуючи

швидкий розвиток технологій та постійне зростання обсягів інформації, дослідження методів виявлення цільових об'єктів набуває особливої ваги.

Виявлення цільових об'єктів у заданій предметній області передбачає застосування спеціальних алгоритмів та методів, спрямованих на ідентифікацію та класифікацію елементів, які мають ключове значення для аналізу конкретної задачі. У роботі цільові об'єкти будуть шукатись у текстових даних, а під терміном «цільові об'єкти» буде матись на увазі сукупність множини ключових слів та множини NER з групуванням шляхом лематизації.

Виявлення цільових об'єктів у системах NLP, зокрема розпізнавання іменованих сутностей, відіграє важливу роль у багатьох завданнях аналізу тексту та обробки інформації. Основна мета NER полягає в ідентифікації і класифікації значущих елементів тексту, таких як імена людей, назви організацій, географічні назви, дати та інші сутності, які мають специфічне значення для конкретного контексту. Це завдання є ключовим для ряду практичних задач, таких як інформаційний пошук, машинний переклад, обробка юридичних документів та аналіз даних у соціальних медіа.

Одним із перспективних напрямків для задачі виявлення цільових об'єктів є використання методів машинного навчання, які дозволяють автоматично адаптуватися до особливостей даних та поліпшувати точність виявлення об'єктів з часом [2].

З проведеного аналізу, запропоновано автоматизувати виявлення цільових об'єктів предметної області з використанням підходів машинного навчання. Автоматизація виявлення цільових об'єктів предметної області сприятиме значному підвищенню ефективності та точності ідентифікації релевантних об'єктів у великих обсягах даних.

Проблему виявлення цільових об'єктів предметної області варто розглядати у контексті пошуку іменованих сутностей та пошуку ключових слів [2]. Даними задачами широко займаються науковці як по всьому світу, так і в Україні [3, 4].

Стрімке зростання обсягів цифрової комунікації та постійна взаємодія з інформаційними системами формують нові виклики для психоемоційного стану користувачів [5]. Одним із таких викликів є феномен цифрового виснаження – стану, що характеризується перевантаженням інформацією, зниженням мотивації, когнітивною втомою та зменшенням залученості у цифрові взаємодії [6]. На відміну від класичних проявів стресу, цифрове виснаження формується у контексті щоденного використання гаджетів і цифрових сервісів, що ускладнює його своєчасне та об'єктивне виявлення традиційними методами психологічної діагностики.

Текстові дані, які продукує користувач – повідомлення у месенджерах [7], коментарі на платформах, електронні листи, внутрішні освітні та корпоративні комунікації – відображають його когнітивні та емоційні стани [8]. Саме тому аналіз множини повідомлень людини, зокрема цільових об'єктів (тематичних доменів, адресатів, намірів, інформаційних фокусів), відкриває можливість раннього

виявлення цифрового виснаження шляхом виявлення змін у структурі, частотності та семантиці цифрової взаємодії [9].

Метою роботи є практична реалізація та дослідження методу виявлення цифрового виснаження за аналізом цільових об'єктів множини повідомлень людини.

Метод виявлення цифрового виснаження за аналізом цільових об'єктів множини повідомлень людини призначений для автоматизації процесу ідентифікації ключових елементів у великих масивах текстових даних, спрямований на підвищення точності та ефективності аналізу текстової інформації. Схема та кроки методу виявлення цифрового виснаження за аналізом цільових об'єктів множини повідомлень людини наведені на рис. 1.

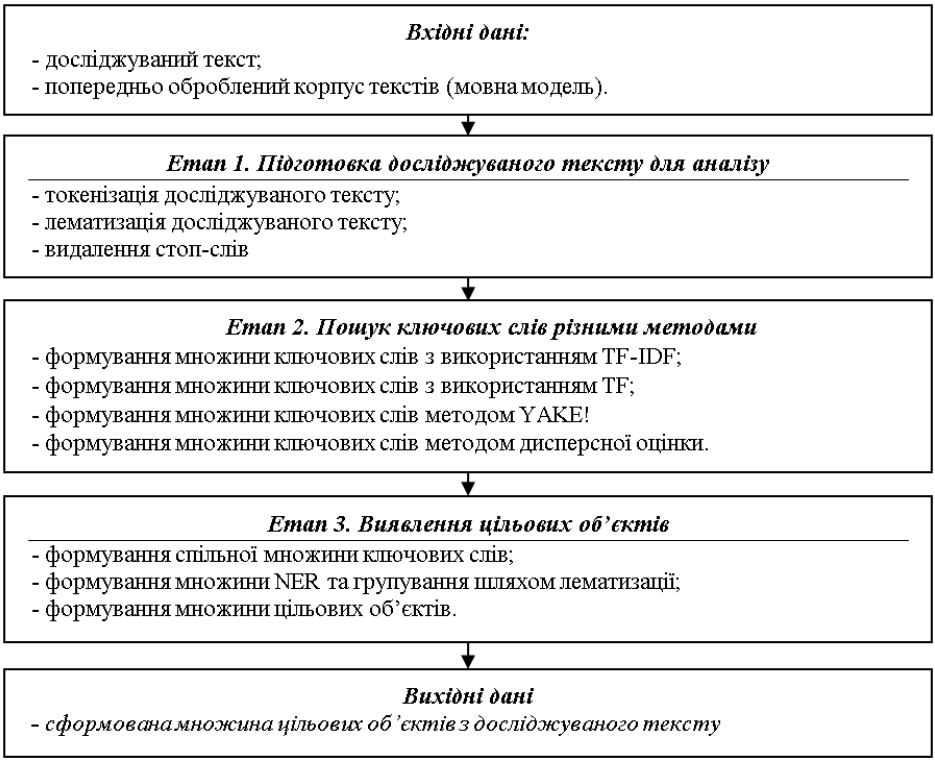

Рисунок 1 – Етапи роботи методу виявлення цифрового виснаження за аналізом цільових об'єктів множини повідомлень людини

Запропонований підхід ґрунтується на використанні алгоритмів машинного навчання, які забезпечують адаптивне розпізнавання релевантних об'єктів із

урахуванням специфіки предметної області та особливостей множини повідомлень користувача. Такий підхід дає змогу істотно прискорити обробку текстових даних і зменшити ймовірність втрати змістовно значущої інформації [10]. Розроблений метод цифрової діагностики, побудований на аналізі цільових об'єктів у масиві повідомлень людини, здійснює перетворення вхідних даних – досліджуваного тексту та збалансованого корпусу відповідної предметної області – у вихідний набір цільових об'єктів. Цей набір формується шляхом об'єднання ключових слів, отриманих різними підходами без дублювання, і NER-сутностей, що пройшли лематизацію.

Метод оперує двома типами вхідних даних:

1) текстом, що підлягає аналізу;

2) попередньо підготовленим корпусом текстів певної предметної області.

Перший етап передбачає лінгвістичну нормалізацію вхідного тексту, що включає токенізацію, лематизацію та усунення стоп-слів. На другому етапі виконується виявлення ключових слів за допомогою кількох незалежних методів, таких як TF-IDF, TF, YAKE! та дисперсійний підхід [11], після чого для кожного методу формується відповідна множина ключових термінів.

Третій етап спрямований на побудову остаточного набору цільових об'єктів. Він включає об'єднання всіх отриманих множин ключових слів без повторів та інтеграцію NER-сутностей, нормалізованих шляхом лематизації. Результатом є структурований набір цільових елементів, який відображає предметно-орієнтовані характеристики досліджуваного тексту.

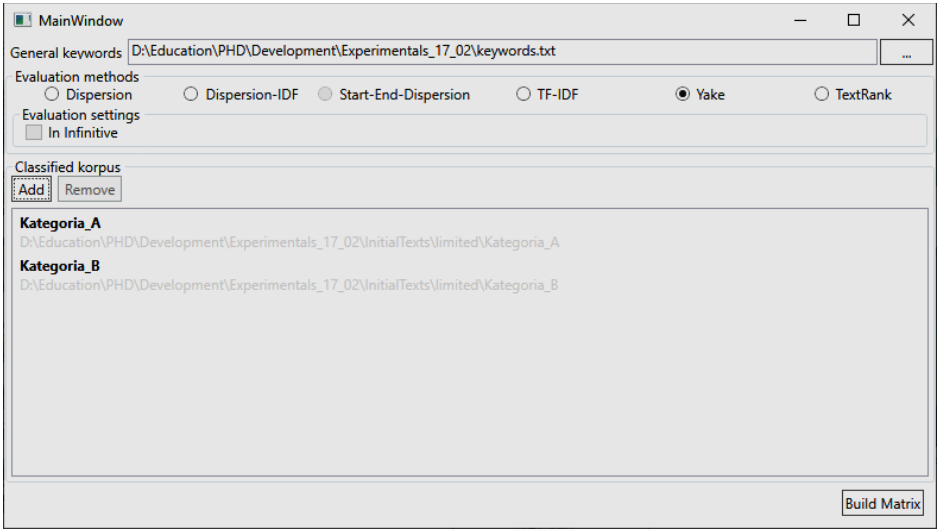

Рисунок 2 – Експериментальний застосунок для пошуку цільових об'єктів множини повідомлень людини досліджуваними методами

Для перевірки працездатності та валідності запропонованого методу було створено програмний застосунок мовою C#, призначений для автоматичного перетворення текстових файлів тестового набору у відповідні множини цільових об'єктів предметної області. Інтерфейс головного вікна розробленого застосунку наведено на рис. 2.

Оскільки українська мова повсякденного спілкування значно відрізняється від літературної через велику кількість діалектів, слів-запозичень та слів-покручів, наявні частотні словники не здатні охопити всю множину української мови [12]. Для створення вектора значущих слів українською мовою було вирішено об'єднати кілька частотних словників, з відсіканням стоп-слів. Після об'єднання та фільтрації довжина вектора значущих слів склала 1500 елементів.

Для дослідження ефективності запропонованого підходу було створено окреме консольне програмне забезпечення мовою Python, яке передбачає використання отриманого списку цільових об'єктів для досліджуваних текстів, та словників для окреслених тем. Відповідно, знайдені цільові об'єкти були переведені у векторне представлення розміром 1500 (як розмір словника) методом One-Hot Encoding. Надалі було перевірено Евклідові відстань між текстами одного спрямування, а також були обраховані Евклідові відстані між векторами протилежних категорій.

Матриця відстаней рис. 3 демонструють чітке розділення текстів на дві основні групи з різним змістом [8]. Перша група текстів (1–5) має тісніші зв'язки між собою, аналогічно як друга група (6–10) також має менші внутрішні відстані, але водночас має великі відстані до текстів з першої групи, що свідчить про те, що ці групи належать до різних тематик. Тексти всередині кожної групи мають невеликі відстані, що свідчить про їхню тематичну схожість.

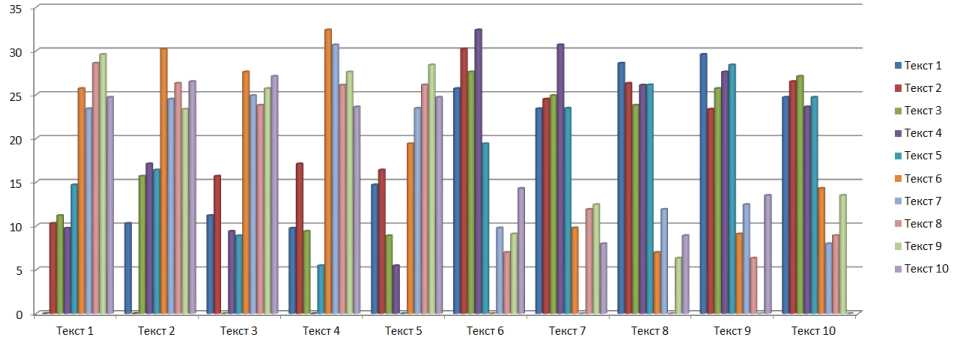

Рисунок 3 – Евклідові відстані між тестовими текстами двох категорій

Таким чином, було розглянуто сучасний стан досліджень, пов'язаних із виявленням цифрового виснаження на основі аналізу множини текстових

повідомлень користувача, зокрема через процедури пошуку ключових слів та ідентифікації іменованих сутностей. На підставі опрацьованих джерел запропоновано власний підхід до визначення ознак цифрового виснаження, що ґрунтується на аналізі цільових об'єктів текстового контенту. Запропонований метод використовує алгоритми машинного навчання для адаптивного виявлення релевантних об'єктів з урахуванням галузевої специфіки, що дає змогу зменшити час обробки даних та знизити ймовірність втрати змістовно важливої інформації. Розроблений підхід орієнтований на автоматизацію процесу виділення значущих елементів із великих текстових масивів, підвищуючи точність та загальну ефективність аналізу.

Метод трансформує вхідні текстові дані та підготовлений збалансований корпус предметної області у множину цільових об'єктів, що включає об'єднання ключових слів, отриманих різними підходами без дублювання, а також згрупованих шляхом лематизації сутностей NER. На відміну від відомих рішень, метод інтегрує як ключові слова, так і предметно орієнтовані іменникові сутності, що дозволило покращити точність виявлення цільових елементів.

Для перевірки ефективності запропонованого підходу сформовано навчальний датасет із 400 текстів українською мовою розмовного стилю. Крім того, створено програмний застосунок для генерування множини цільових об'єктів за текстами тестової вибірки, а також окремий консольний інструмент для роботи з отриманими списками цільових об'єктів і словниками предметних областей, що відповідали структурі датасету.

Проведені експерименти показали, що сформовані методом цільові об'єкти придатні для виконання подальших класифікаційних завдань. На основі оцінювання за метрикою Евклідових відстаней продемонстровано чітке групування текстів певної категорії та збільшення відстані відносно ортогональних класів, що свідчить про практичну цінність та ефективність розробленого підходу. Подальші дослідження доцільно спрямувати на розширення набору категорій та випробування інших метрик оцінювання якості виявлених цільових об'єктів, а також на порівняння результатів із ефективністю сучасних великих мовних моделей.

Перелік посилань

1. Молчанова М.О., Мазурець О.В., Собко О.В., Віт Р.В., Назаров В.В. Алгоритм виявлення аб'юзивного вмісту в україномовному аудіоконтенті для імплементації в об'єктно-орієтовану інформаційну систему. Науковий журнал «Вісник Хмельницького національного університету» серія: Технічні науки. Хмельницький, 2024. №1 (331). С. 101-106.
2. Mazurets O., Sobko O., Vit R., Pasternak V. Practical Approach for Detection by Deep Learning of Target Objects of Subject Area Based on Semantic Connectivity Indicators in Audio Database. Proceedings of XXIV International Scientific and Practical Conference «Modern Scientific Challenges are the Driving Force of the Development of Scientific Research». May 22-24, 2024. Bruges, Belgium. International Scientific Unity. 2024. Pp. 91-96.
3. Shin H. General-use unsupervised keyword extraction model for keyword analysis / H. Shin, J. Lee, S. Cho. // Expert Systems with Applications. 2023. №233. С. 120889.

4. Chen X. Named Entity Recognition via Unified Information Extraction Framework / X. Chen, Z. Zhang, X. Lu. // 4th International Conference on Computer Communication and Artificial Intelligence. 2024. С. 308–313.
5. Мазурець О.В., Віт Р.В. Метод виявлення цільових об'єктів предметної області у текстовому контенті. Науковий журнал «Вісник Хмельницького національного університету» серія: Технічні науки. Хмельницький, 2024. №6, Т.1 (343). С. 152-157.
6. Віт Р.В., Мазурець О.В. Метод формування множин цільових об'єктів предметних областей у цифрових текстах засобами машинного навчання. Науковий журнал «Наука і техніка сьогодні». Київ, 2024. №13 (41). С. 926-937.
7. Віт Р.В., Мазурець О.В. Підхід до тематичної класифікації текстової інформації засобами обробки природної мови. Науковий журнал «Наукові праці Донецького національного технічного університету», серія «Проблеми моделювання та автоматизації проектування». 2025. №1 (21). С. 94-99.
8. Мазурець О., Віт Р. Інтелектуальний метод виявлення цільових об'єктів предметної області для класифікації текстової інформації. Матеріали XII Міжнародної науково-практичної конференції «Інформаційні управляючі системи та технології ІУСТ-ОДЕСА-2024». 23-25.09.2024. Одеса. 2024. С.205-208.
9. Віт Р.В., Мазурець О.В. Метод виявлення множин цільових об'єктів предметної області у текстовому контенті. Збірник наукових праць за матеріалами XVI Всеукраїнської науково-практичної конференції «Актуальні проблеми комп'ютерних наук АПКН-2024». 15-16 листопада 2024. Хмельницький, 2024. с. 78-82.
10. Мазурець О.В., Віт Р.В. Дослідження ефективності методу виявлення цільових об'єктів предметної області. Інформаційні технології і автоматизація. Матеріали XVII міжнародної науково-практичної конференції. 31 жовтня – 1 листопада 2024 р. Одеса, ОНТУ. 2024. С.650-653.
11. Віт Р.В., Мазурець О.В. Тематична класифікація текстової інформації засобами обробки природної мови. Збірник наукових праць XXIII Міжнародної наукової конференції «Нейромережні технології та їх застосування НМТіЗ-2024». 11-12 грудня 2024. Краматорськ-Тернопіль, ДДМА. 2024. с. 63-66.
12. Мазурець О.В., Віт Р.В. Інтелектуальний метод виявлення цільових об'єктів предметної області за показниками семантичної зв'язності для класифікації текстової інформації. Розвитки інформаційно-керуючих систем та технологій.: монографія. Львів-Торунь: Lina-Pres, 2024. С.223-244.