УДК 004.8

Віт Р.В., Мазурець О.В.

Хмельницький національний університет

МЕТОД ВИЯВЛЕННЯ МНОЖИН ЦІЛЬОВИХ ОБ'ЄКТІВ ПРЕДМЕТНОЇ ОБЛАСТІ У ТЕКСТОВОМУ КОНТЕНТІ

Запропоновано метод виявлення множин цільових об'єктів предметної області у текстовому контенті засобами машинного навчання, що призначений для автоматизації процесу ідентифікації ключових елементів у великих масивах текстових даних. Метод виявлення цільових об'єктів предметної області дозволяє одержувати вихідні дані у вигляді сформованої множини цільових об'єктів з досліджуваного тексту, яка є об'єднаною множиною ключових слів знайденими різними методами без повторів та множиною NER що згруповані шляхом лематизації. Запропонований метод виявлення цільових об'єктів відрізняється від існуючих урахуванням ключових слів та іменникових сутностей предметної області, що дало змогу підвищити точність виявлення цільових об'єктів предметної області внаслідок урахування іменникових сутностей.

Method for identifying target objects sets of subject area in text content by means of machine learning is proposed, which is designed to automate the process of identifying key elements in large arrays of textual data. The method of detecting target objects of the subject area allows you to receive output data in the form of a formed set of target objects from the researched text, which is a combined set of keywords found by various methods without repetitions and a set of NER grouped by lemmatization. The proposed method of detecting target objects differs from the existing ones by taking into account keywords and noun entities of the subject area, which made it possible to increase accuracy of detection of target objects of subject area as result of taking into account noun entities.

В умовах зростаючої складності даних, які охоплюють різноманітні предметні області, методи виявлення цільових об'єктів у предметній області є критично важливими для ефективного аналізу і обробки великих обсягів інформації [1, 2]. Відсутність надійних й ефективних методів виявлення цільових об'єктів може призвести до втрати важливої інформації, зниження точності прийняття рішень й збільшення витрат на аналіз даних. Враховуючи швидкий розвиток технологій і постійне зростання обсягів інформації, дослідження методів виявлення цільових об'єктів набуває особливої ваги [3].

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 [4]. У роботі об'єкти будуть шукатись в текстових даних, а під цим терміном буде матись на увазі сукупність множини ключових слів й множини NER із групуванням

шляхом лематизації. Виявлення цільових об'єктів в системах NLP, зокрема розпізнавання іменованих сутностей, відіграє важливу роль в багатьох завданнях аналізу тексту і обробки інформації. Одним із перспективних напрямків для задачі виявлення цільових об'єктів є використання методів машинного навчання, які дозволяють автоматично адаптуватися до особливостей даних [5].

Одним із перспективних напрямків для задачі виявлення цільових об'єктів є використання методів машинного навчання, які дозволяють авто-матично адаптуватися до особливостей даних та поліпшувати точність виявлення об'єктів з часом. Отже, варто 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.

Метою роботи є розробка методу виявлення множин цільових об'єктів предметної області, який дозволяє підвищити точність виявлення цільових об'єктів предметної області внаслідок врахування іменникових сутностей за рахунок урахування ключових слів та іменникових сутностей предметної області й дозволяє перетворювати вхідні дані у вигляді досліджуваного тексту і попередньо обробленого та збалансованого корпусу текстів досліджуваної предметної області в вихідні дані у вигляді сформованої множини цільових об'єктів з досліджуваного тексту, яка є об'єднаною множиною ключових слів знайденими різними методами без повторів та множиною NER що згруповані шляхом лематизації.

Метод виявлення множин цільових об'єктів предметної області у текстовому контенті призначений для автоматизації процесу ідентифікації ключових елементів у великих масивах даних, схема методу наведені на рисунку 1.

Вхідними даними методу є досліджуваний текст й попередньо оброблений збалансований корпус текстів досліджуваної предметної області.

Першим етапом є підготовка досліджуваного тексту для аналізу, який включає в себе токенізацію, лематизацію та видалення стоп-слів.

Наступним етапом є пошук ключових слів різними методами, такими як TF, TF-IDF, YAKE! та методом дисперсної оцінки. Кожним перерахованим методом відбувається формування множини ключових слів.

На третьому етапі здійснюється виявлення цільових об'єктів, що включає в себе декілька кроків. Цільові об'єкти є об'єднаною множиною ключових слів знайденими різними методами без повторів та множиною NER що згруповані шляхом лематизації.

Для валідації запропонованого методу для пошуку цільових об'єктів предметної області було розроблено програмний продукт мовою C# для перетворення текстового контенту файлів із тестової вибірки у множину цільових об'єктів предметної області. Для створення вектора значущих слів українською мовою було об'єднано кілька частотних словників [5], з відсіканням стоп-слів. Після об'єднання й фільтрації довжина вектора значущих слів склала 1500 елементів. Для цього було використано тексти з двох ортогональних множин. Такий

вибір ресурсів обумовлений необхідністю забезпечити достатній обсяг текстів, які мають понад 200 слів, для навчання та перевірки запропонованого підходу. Головне вікно розробленого застосунку зображено на рисунку 2.

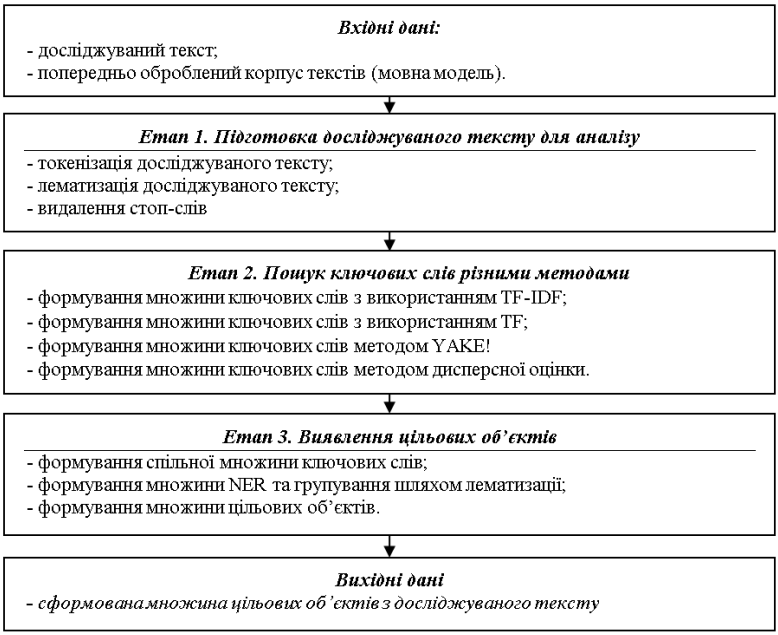

Рисунок 1 – Схема кроків методу виявлення множин цільових об'єктів предметної області у текстовому контенті засобами машинного навчання

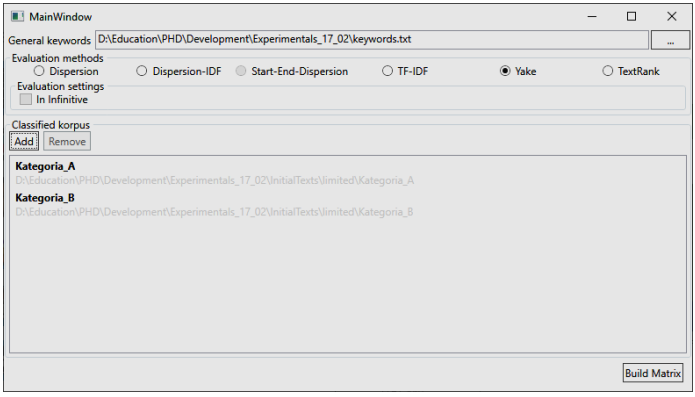

Рисунок 2 – Застосунок для виявлення цільових об'єктів предметної області у текстовому контенті

Для дослідження ефективності запропонованого підходу було перевірено Евклідові відстані між текстами одного спрямування, а також були обраховані Евклідові відстані між векторами протилежних категорій. Дані експерименту наведено на рисунку 3.

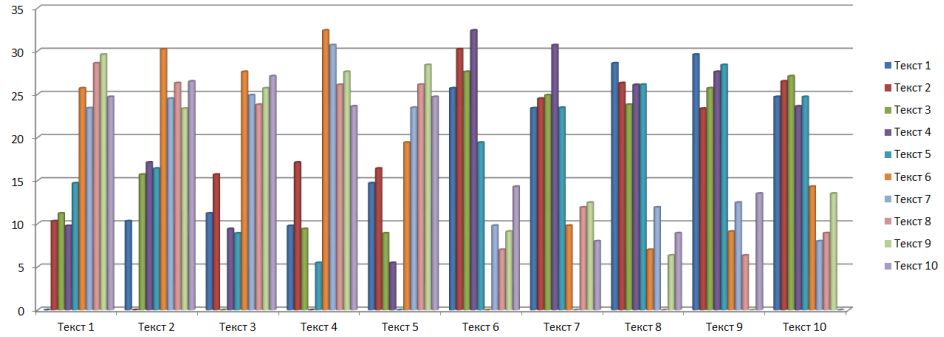

Рисунок 3 – Евклідові відстані між текстами двох протилежних категорій

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 із першої групи. Це свідчить про те, що ці групи належать до різних тематик. Тексти всередині кожної групи мають невеликі відстані, це свідчить про їх тематичну схожість.

Отже, було запропоновано метод виявлення множин цільових об'єктів предметної області. Метод виявлення цільових об'єктів предметної області дозволяє перетворювати вхідні дані у вигляді досліджуваного тексту і попередньо обробленого та збалансованого корпусу текстів досліджуваної предметної області в вихідні дані у вигляді сформованої множини цільових об'єктів з досліджуваного тексту, яка є об'єднаною множиною ключових слів знайденими різними методами без повторів та множиною NER що згруповані шляхом лематизації. 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.

Для дослідження ефективності розробленого методу виявлення множин цільових об'єктів предметної області було сформовано навчальний датасет побутовою українською мовою. Для валідації запропонованого методу було розроблено програмну систему для перетворення текстового контенту з тестової вибірки у множину цільових об'єктів предметної області. Також створено окреме консольне програмне забезпечення для використання отриманого списку цільових об'єктів для досліджуваних текстів та словників з предметних областей, обраних відповідно до датасету.

Виконане дослідження ефективності розробленого методу виявлення множин цільових об'єктів предметної області у текстовому контенті засобами машинного навчання виявило, що знайдені за методом цільові об'єкти предметних областей спроможні виконувати подальшу задачу класифікації, демонструючи на метриці Евклідових відстаней групування текстів однієї категорії та збільшення відстані ортогональної їй. Це визначає корисний ефект та область застосування розробленого методу [6, 7] в завданнях аналізу текстів і обробки інформації.

Перелік посилань

1. 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». Одеса. 2024. С.205-208.
2. 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.
3. Mazurets O., Uspenska K., Vit R., Tyschenko O. Intelligent System for Determining the Object Attributes Values by Neural Networks Means by Graphic Images in Databases. Current Trends in the Development of Scientific Research in Today's Conditions. Proceedings of XXV International scientific and practical conference. International Scientific Unity. Florence, Italy. 2024. Pp. 86-91.
4. Залуцька О.О., Молчанова М.О., Віт Р.В., Мазурець О.В. Конфігурування нейронної мережі для класифікації емоційної тональності текстової інформації за показниками семантичної зв'язності. Збірник наукових праць за матеріалами XV Всеукраїнської науково-практичної конференції «Актуальні проблеми комп'ютерних наук АПКН-2023». Хмельницький, 2023. с. 102-107.
5. Zalutska O., Molchanova M., Sobko O., Mazurets O., Pasichnyk O., Barmak O., Krak I. Method for Sentiment Analysis of Ukrainian-Language Reviews in E-Commerce Using RoBERTa Neural Network. CEUR Workshop Proceedings, 2023, vol. 3387, pp. 344–356.
6. 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». Bruges, Belgium. International Scientific Unity. 2024. Pp. 91-96.
7. Krak I., Molchanova M., Mazurets O., Sobko O., Zalutska O., Barmak O. Method for Neural Network Detecting Propaganda Techniques by Markers With Visual Analytic. CEUR Workshop Proceedings, 2024, vol. 3790, pp. 158-170.