УДК 004.8
Віт Р.В.[1], Мазурець О.В.[2]
[1] асп. кафедри комп'ютерних наук, Хмельницький національний університет
[2] канд. техн. наук, доц., доц. кафедри комп'ютерних наук, Хмельницький національний університет

ВИЯВЛЕННЯ ЦІЛЬОВИХ ОБ'ЄКТІВ ПРЕДМЕТНОЇ ОБЛАСТІ У ТЕКСТОВОМУ КОНТЕНТІ ЗАСОБАМИ МАШИННОГО НАВЧАННЯ

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]. 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].

Метою роботи є реалізація методу виявлення цільових об'єктів предметної області, який дає змогу підвищити точність виявлення цільових об'єктів предметної області внаслідок врахування іменникових сутностей за рахунок урахування ключових слів та іменникових сутностей предметної області.

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

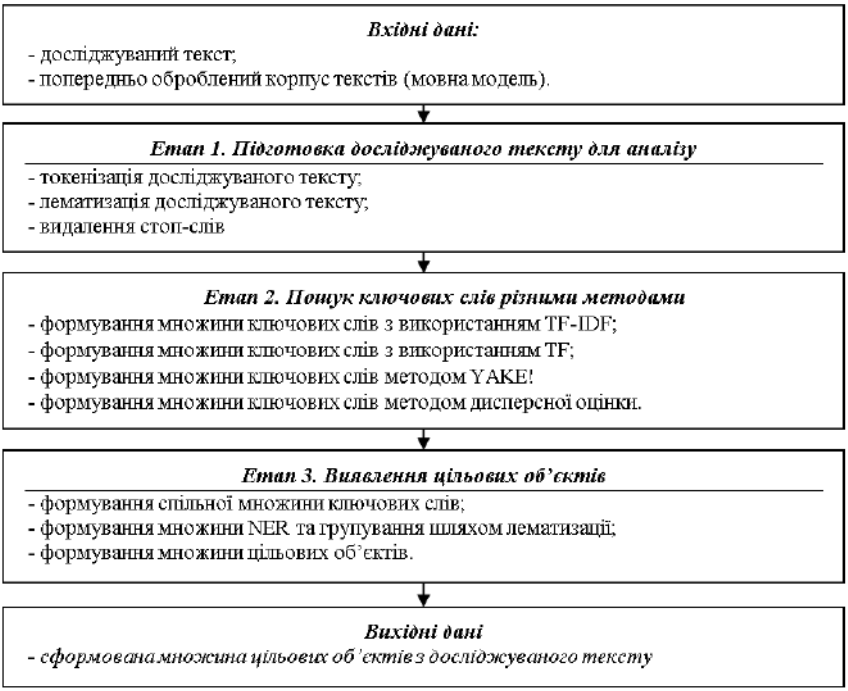

Рисунок 1 – Схема методу виявлення цільових об'єктів предметної області у текстовому контенті засобами машинного навчання

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. 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. Для створення вектора значущих слів українською мовою було вирішено об'єднати 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.

Отже, було запропоновано метод виявлення 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. 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.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

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. 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.