УДК 004.8

ІНТЕЛЕКТУАЛЬНИЙ МЕТОД ВИЯВЛЕННЯ ЦІЛЬОВИХ ОБ'ЄКТІВ ПРЕДМЕТНОЇ ОБЛАСТІ ЗА ПОКАЗНИКАМИ СЕМАНТИЧНОЇ ЗВ'ЯЗНОСТІ ДЛЯ КЛАСИФІКАЦІЇ ТЕКСТОВОЇ ІНФОРМАЦІЇ

Ph.D. О. Мазурець ORCID: 0000-0002-8900-0650
Хмельницький національний університет, Україна
E-mail: exe.chong@gmail.com
Р. Віт ORCID: 0009-0009-6958-4730
Хмельницький національний університет, Україна
E-mail: vit.roman.vit@gmail.com

Анотація. *Розглянуто інтелектуальний метод виявлення цільових об'єктів предметної області за показниками семантичної зв'язності, який призначений для автоматизації процесу ідентифікації ключових елементів у великих масивах текстових даних. Розроблений метод відрізняється від існуючих урахуванням ключових слів та іменникових сутностей предметної області, що дало змогу підвищити точність виявлення цільових об'єктів предметної області внаслідок врахування іменникових сутностей. Проведені дослідження встановили, що знайдені за створеним методом цільові об'єкти спроможні виконувати подальшу задачу класифікації, демонструючи на метриці евклідових відстаней групування текстів однієї категорії та збільшення відстані ортогональної їй. Для прикладної реалізації створеного інтелектуального методу виявлення цільових об'єктів предметної області, було сформовано модель сучасної української мови, що побудована шляхом об'єднання значимих для передачі сенсу відомих частотних словників. Для створення коректної моделі сучасної української мови було використано частотні словники існуючих корпусів української мови, які сукупно охоплюють як різні сфери діяльності та типи контенту, так і різні області спілкування, враховуючи спілкування в Інтернет. Шляхом об'єднання за частотними показниками словників цих корпусів української мови було одержано вектор слів, який для покращення спроможності класифікації був відфільтрований за іменниковою групою та обмежений кількісно. Проведені дослідження дали можливість стверджувати про можливість використання створеної моделі сучасної української мови для ефективного вирішення задач аналізу текстового контенту одиниць інтернет-спілкування та його класифікації за різними ознаками.*

Ключові слова: *цільові об'єкти предметної області, класифікація текстів, модель української мови, частотні словники, вектор слів, семантична зв'язність*

1. Вступ та постановка проблеми

За відсутності гарантій забезпечення вчасного й повноцінного процесу навчання, підвищення кваліфікації чи контролю рівня поточних знань фахівців сфери безпеки та медицини у зв'язку з складними епідеміологічними, екологічними й соціальними умовами, в сучасному світі зростає роль комп'ютерних засобів перевірки рівня знань [1, 2]. Одним із основних способів контролю рівня знань залишається комп'ютерне тестування [3]. Особливо перспективним є впровадження технологій адаптивного тестування, за яких складність чи інші властивості тестових завдань змінюються залежно від правильності попередніх відповідей [4]. 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 [2]. 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 [3]. 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 [4]. 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 [5].

З проведеного аналізу, запропоновано автоматизувати виявлення цільових об'єктів предметної області з використанням підходів машинного навчання. В даній праці наведено розроблений авторами комплексний підхід до побудови моделі сучасної української мови, у рамках якої стане можлива гіперплощинна класифікація контенту для задач автоматизації виявлення цільових об'єктів

предметної області, що сприятиме значному підвищенню ефективності та точності ідентифікації релевантних об'єктів у великих обсягах даних.

2. Аналіз останніх досліджень та публікацій

Проблему виявлення цільових об'єктів предметної області варто розглядати у контексті пошуку іменованих сутностей та пошуку ключових слів [6]. Даними задачами широко займаються науковці як по всьому світу, так і в Україні. Модель отримання ключових слів загального призначення, що призначена для роботи з групами документів різних розмірів і читабельності, а також наявності міток ключових слів запропонована у [7]. Для отримання кращого вибору ключових слів використано модель логістичної регресії з найменшим стисненням і регуляризацією оператора вибору. Заснована на класифікації структура такого підходу забезпечує виявлення слів, які чітко характеризують цю групу документів у порівнянні з групами порівняння, що підвищує репрезентативність вилучених ключових слів.

У роботі [8] були проведені експерименти на різних мовах, що показують, що зображення доповнюють моделі обробки природньої мови (включаючи BERT), навчені без зовнішнього попереднього навчання. Дослідження класифікації текстів були зосереджені на статтях з Вікіпедії, оскільки зображення зазвичай доповнюють текст і сторінка Вікіпедії може бути написана різними мовами. У статті [9] досліджують онлайн-розмови: їх перебіг, аргументи й як вони вирішуються. Експеримент проводився на основі функцій, використовуючи модель логістичної регресії від Scikit-learn. У роботі [10] було проведено виділення емоційних настроїв та класифікація їх полярності. У публікації були проведені експерименти з 8-и наборами даних англійською мовою. Результати показують, що продуктивність сучасних моделей для передбачення полярних зворотів мови погана, і це перешкоджає використанню цієї інформації на практиці.

Щодо задачі NER, у [11] запропоновано підхід до оптимізації завдання розпізнавання іменованих сутностей шляхом використання попередньо навчених мовних моделей для автоматичного дослідження слів, пов'язаних з віртуальними мітками, що представляють категорії сутностей. Метод передбачає розробку міток через встановлення зв'язків між початковими словами міток і відповідними словами сутності на основі розподілу даних, отриманих з попередньо навченої мовної моделі. Завдяки цьому покращується семантичне представлення слів міток, що в результаті підвищує точність моделі в ідентифікації конкретних сутностей. Крім того, завдання NER переналаштовується у формат text2text, що дозволяє краще використовувати знання мовної моделі та оптимізує процес вилучення інформації.

Отже, зважаючи на проведені дослідження, поставлена задача виявлення цільових об'єктів предметної області за показниками семантичної зв'язності є актуальною для сучасної української мови. Ця задача вимагає формування моделі української мови, для чого перспективним є використання векторної моделі, у якій ознаками будуть слугувати статистичні міри, які

використовуються для оцінки важливості слова в контексті повідомлення, яке в свою чергу є частиною колекції повідомлень або корпусу (TF-IDF, BM25, Yake, дисперсійна оцінка тощо).

3. Постановка задачі

Метою роботи є розробка моделі сучасної української мови, у рамках якої стане можлива гіперплощина класифікація текстів для виявлення цільових об'єктів предметної області за показниками семантичної зв'язності.

Основним результатом роботи є створений інтелектуальний метод виявлення цільових об'єктів предметної області за показниками семантичної зв'язності для класифікації текстової інформації з використанням розробленої моделі сучасної української мови, який відрізняється від існуючих урахуванням ключових слів та іменникових сутностей предметної області, що дало змогу підвищити точність виявлення цільових об'єктів предметної області внаслідок врахування іменникових сутностей.

4. Метод виявлення цільових об'єктів предметної області

Інтелектуальний метод виявлення цільових об'єктів предметної області за показниками семантичної зв'язності для класифікації текстової інформації призначений для автоматизації процесу ідентифікації ключових елементів у великих масивах текстових даних, спрямований на підвищення точності та ефективності аналізу текстової інформації. 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 упущення важливої інформації. Схема та кроки методу наведені на рис. 1.

Вхідними даними методу є досліджуваний текст та 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-IDF, TF, YAKE! та методом дисперсної оцінки. Кожним перерахованим методом відбувається формування множини ключових слів.

На третьому етапі здійснюється виявлення цільових об'єктів, яке включає в себе декілька кроків. Схематично виявлення цільових об'єктів зображено на рис. 2. Цільові об'єкти є об'єднаною множиною ключових слів знайденими різними методами без повторів та множиною NER що згруповані шляхом лематизації. Наведені етапи є основними у роботі роботи запропонованого інтелектуального методу виявлення цільових об'єктів предметної області за показниками семантичної зв'язності для класифікації текстової інформації з використанням моделі сучасної української мови.

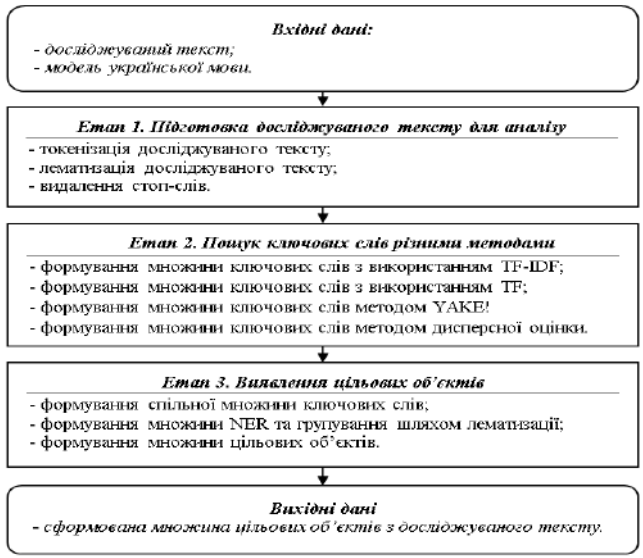

Рисунок 1. *Етапи роботи методу виявлення цільових об'єктів предметної області за показниками семантичної зв'язності*

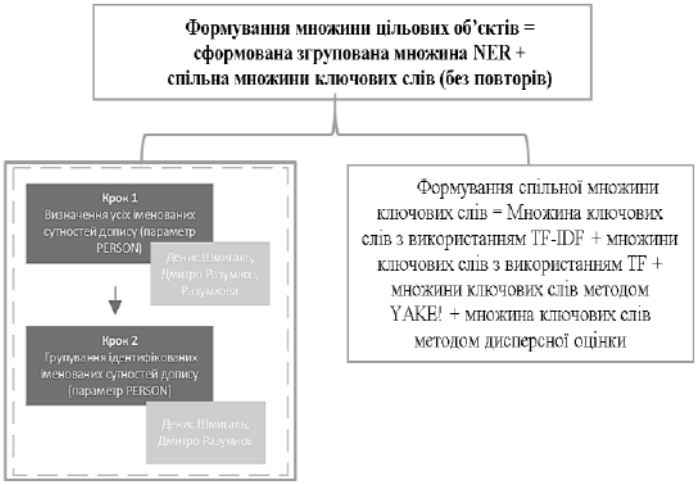

Рисунок 2. *Приклад виконання етапу формування цільових об'єктів методу виявлення цільових об'єктів предметної області*

Цей метод призначений для автоматизації процесу ідентифікації ключових елементів у великих масивах текстових даних й відрізняється від існуючих урахуванням ключових слів та іменникових сутностей предметної області, що

дало змогу підвищити точність виявлення цільових об'єктів предметної області внаслідок врахування іменникових сутностей.

5. Модель сучасної української мови

Модель сучасної української мови забезпечує гіперплощинну класифікацію текстів для її подальшого використання методом виявлення цільових об'єктів предметної області за показниками семантичної зв'язності. Схема запропонованого підходу до побудови моделі для контент-аналізу україномовного сегменту Інтернет-спілкування зображена на рисунку 3. Запропонований підхід до побудови моделі для контент-аналізу україномовного сегменту Інтернет-спілкування, яка формується для її використання в методі виявлення цільових об'єктів предметної області за показниками семантичної зв'язності, на першому кроці передбачає побудову вектора ключових слів, для чого послідовно виконуються вибір частотних словників української мови, об'єднання цих словників і подальше обмеження кількості ключових слів шляхом відкиданням стоп-слів та рідковживаних слів.

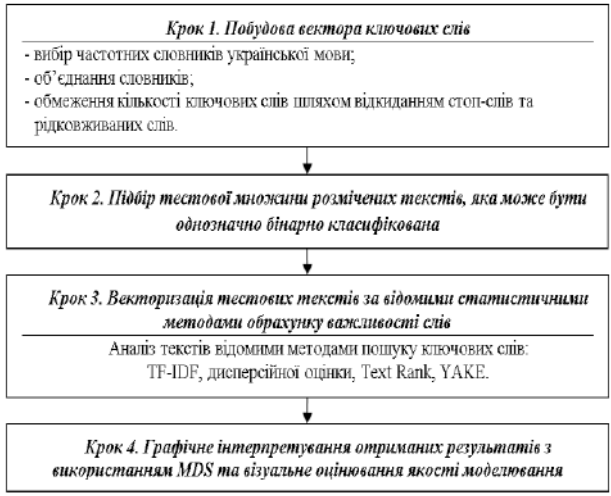

Рисунок 3. *Схема підходу до створення моделі сучасної української мови для методу виявлення цільових об'єктів предметної області*

Першим кроком підходу є побудова вектору ключових слів для передачі сенсу при комунікації у мережі Інтернет. Враховуючи зазначену особливість (використання для спілкування суржику, деформованих слів та ненормативної лексики), побудова вектору ключових слів є окремою підзадачею. Оскільки мова йде про специфічну флективну українську мову, це мають бути не просто ключові слова зі статей, а це повинен бути збалансований набір даних. Тому в роботі були використані частотні словники української мови [12-16]

($\overline{W_j} = \{\overline{w_1}, \overline{w_2}, ..., \overline{w_n}\}, j = \overline{1..n}$), де j – порядковий номер словника, n – кількість словників. Де кожен словник представляє собою відповідний набір слів $\overline{w} = \{слово1, слово2, ..., словоn\}, i = \overline{1..n}$, де n – кількість слів у словнику. У дослідженні з кожного частотного словника видалялися слова, які на думку авторів, не суттєво впливатимуть на модель (слова притаманні дуже великій кількості текстів (стоп-слова які мають службове значення та використовуються для зв'язування слів у тексті), надто рідковживані слова тощо). Тобто $\overline{W_j}$ містять тільки відібрані слова. Вектор ключових слів $\overline{Word}$ буде об'єднанням таких частотних словників:

$$\overline{Word} = \bigcup_{i=1}^{n} \overline{W_i} \quad . \tag{1}$$

На другому кроці при побудові моделі для контент-аналізу україномовного сегменту Інтернет-спілкування виконується підбір тестової множини розмічених текстів, яка може бути однозначно бінарно класифікована. Основним змістом кроку є вибір текстів, які підлягають ідеальній класифікації, коли однозначно можна зробити висновок щодо належності обраного тексту конкретній категорії. Формується множина текстів D, у якій кожен текст $d \in D$ може відповідати конкретній категорії $c \in C$, де C – множина категорій. У даному випадку буде використана бінарна класифікація, оскільки наша мета – виявлення текстів з негативною забарвленістю контенту як категорії.

Третій крок передбачає векторизацію тестових текстів за відомими статистичними методами обрахунку важливості слів, тобто аналіз текстів відомими методами пошуку ключових слів, до яких належать TF-IDF, дисперсійна оцінка, Text Rank, YAKE. Тобто, кожен текстовий документ векторизується відомими методами пошуку ключових слів.

На етапі препроцесінгу, кожен текстовий документ $d_i \in D$, де i – кількість документів у колекції, перетворюється у вектор слів. Після цього за допомогою кожного з методів пошуку ключових слів формується відповідний вектор оцінок входжень ключових слів $\overline{Wd}_i$, які є у векто $\overline{Word}$ рі.

Для формування оцінок пропонується використати відомі методи пошуку ключових слів TF-IDF [17], дисперсійної оцінки [18], Text Rank [19, 20, 21], YAKE [22]. Класично *TF-IDF* подається наступним чином:

$$TF - IDF(t,d,D) = TF(t,d) \times IDF(t,d,D) \tag{2}$$

та є вагою терміну t документа d корпусу D, а $TF(t,d)$ є значенням частоти терміну t в документі d.

Дисперсійна оцінка подається як оцінка важливості кожного слова в досліджуваному тексті, що проводиться з використанням методу дисперсійного оцінювання. Цей метод є оцінкою дискримінантної сили слів і дозволяє

відділити із загальної множини слів широкого вжитку в тексті слова, що розташовані рівномірно. Відповідно з [18], якщо деяке слово *A* в тексті, що складається з *N* слів, позначене як A_k^n, де індекс *k* – номер появи даного слова в тесті, а *n* – його позиція в тексті, то інтервалом між послідовними появами слова при таких позначеннях буде величина

$$\Delta A_k^m = A_{k+1}^m - A_k^n = m - n, \quad (3)$$

де на ітерації *m* і позиції *n* в тексті знаходиться слово *A*, яке зустрілось *k+1*-ий і *k*-ий рази. Таким чином, дисперсійна оцінка розраховується за формулою

$$\sigma = \sqrt{(\Delta A^2) - (\Delta A)^2} \Big/ (\Delta A), \quad (4)$$

де (*ΔA*) – середнє значення послідовності $\Delta A_1, \Delta A_2, .., \Delta A_k$. *K* – кількість появи слова *A* в тексті.

Метод Text Rank призначений для моделювання тексту як неорієнтованого зваженого графа *G=(V,E,W)*, у якому ключові слова-кандидати розглядаються як набір вузлів *V*, а взаємозв'язок між двома словами розглядається як ребро в *E*. *W* представляє частоту появ по відношенню до *E* [19, 20]. Для ітераційного обчислення ваг вузлів використовується:

$$WS(V_i) = (1-d) + d * \sum_{V_j \in \ln(V_i)} \frac{w_{i,j}}{\sum_{V_k \in Out(V_j)} w_{jk}} WS(V_j), \quad (5)$$

де *d* є коефіцієнтом амортизації ітераційного обчислення та може приймати значення 0,85 за замовчуванням [19]. $\ln(V_i)$ є набором вузлів, що вказують на V_i, $Out(V_j)$ є набором вузлів, на які вказує V_j. Формула (5) показує, що вага вузла V_i залежить від ваги ребра від V_j до V_i (on the edge weight from) та суми ваг ребер від вузла V_j до інших вузлів.

Алгоритм YAKE складається з 4 кроків: попередня обробка та генерація термінів-кандидатів; визначення особливості термінів; підрахунок балів за термін; асоціація подібних термінів [22]. На першому етапі виконується поділ на рівні речення, які в подальшому розбиваються на терміни. На етапі визначення особливості термінів кожен термін оцінюється завдяки використанню спеціальних функцій [22]. На етапі підрахунку балів за термін використовується нижченаведена формула:

$$S(t) = (Trel * Tposition) / Tcase + ((Tnorm / Trel) + (Tsentence / Trel)), \quad (6)$$

де *Tcase* – важливість використання великих літер і скорочень, *Tposition* – більше значення надається словам, які присутні на початку документа, *Tnorm* – частота слів, *Trel* – перевіряє різноманітність контексту, у якому зустрічається це слово, *Tsentence* – функція визначає, як часто слово-кандидат зустрічається з різними реченнями.

Вищий бал отримують слова, які часто зустрічаються в різних реченнях. Останнім етапом є об'єднання значень оцінок морфологічно подібних слів. Мінімальні бали одержують кращі ключові слова.

На *четвертому кроці* виконується графічне інтерпретування отриманих результатів з використанням MDS та візуальне оцінювання якості моделювання [23]. Важливим вмістом етапу інтерпретації отриманих результатів є вибір критерію якості моделювання за візуальною аналітикою [24, 25]. Для можливості оцінки якості отриманих моделей для задач класифікації пропонується використовувати метод Multidimensional scaling MDS [6, 19]. Це один з методів пониження розмірності векторного простору. Метою методу є пониження розмірності до такої яку можливо візуалізувати (3 чи 2-мірної). Критерієм для пониження розмірності виступає, наприклад, Евклідова відстань між векторами. Тобто розв'язуючи оптимізаційну задачу знаходять відображення $R^n \to R^2$ що дає можливим отримати двовимірний графік взаємного розташування точок-векторів та візуально оцінити якість моделі для задачі класифікації.

Запропоновано візуальні критерії для оцінки якості візуального моделювання (рисунки 4-6).

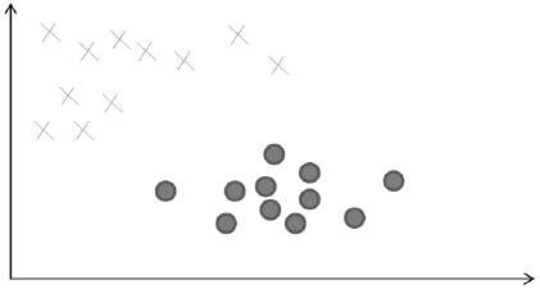

Рисунок 4. *Приклад високого рівня якості моделі для задачі класифікації*

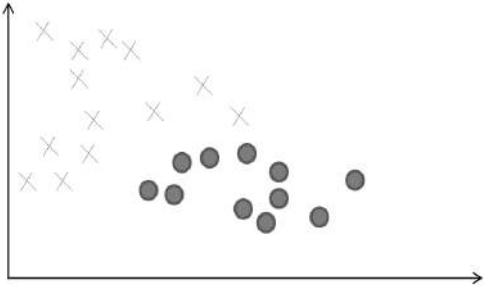

Рисунок 5. *Приклад прийнятного рівня якості моделі для задачі класифікації*

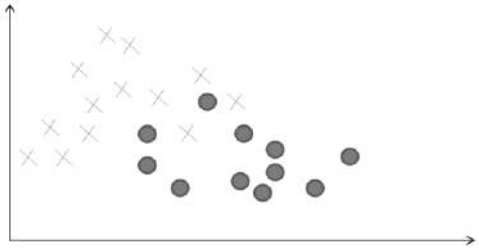

Рисунок 6. *Приклад незадовільного рівня якості моделі для задачі класифікації*

Критерій 1 – Високий рівень моделі для класифікації текстів. На рисунку 4 видно, що два класи чітко розділені між собою, що свідчить про коректність запропонованої моделі.

Критерій 2 – Прийнятний рівень моделі для класифікації текстів. На рисунку 5 видно, що два класи стикаються між собою. При такому показнику можна вважати модель працездатною, проте вона потребуватиме додаткового експертного висновка для підтвердження класифікації.

Критерій 3 – Незадовільний рівень моделі для класифікації текстів. На рисунку 6 видно, що два класи майже не розділені між собою, відстань між ними незначна, а місцями зустрічається перетин. При такому показнику модель не можна вважати працездатною, вона потребує доопрацювання.

Наведені критерії пропонується застосувати для перевірки якості запропонованої моделі сучасної української мови. Модель вважатимемо коректною, якщо значення результатів буде знаходитись в межах між першим та другим критеріями.

6. Підготовка навчальних вибірок даних

Підготовка навчальних вибірок даних для створення моделі сучасної української мови для методу виявлення цільових об’єктів предметної області є окремою складовою дослідження.

Для побудови моделі та її валідації (спроможності класифікувати тексти спілкування у Інтернеті) було використано такі корпуси української мови з частотними словниками та розміченими за категоріями текстами:

1. *БрУК* – збалансований корпус-мільйонник сучасної мови. Відкритий, збалансований за жанрами корпус сучасної української мови обсягом 1 млн слововживань. Корпус побудований на засадах, що були покладені в основу відомого корпусу англійської мови Brown. [12]. Також до складу цього корпусу входить словник VESUM (URL: https://r2u.org.ua/vesum/.), який містить слова-покручі, ненормативну лексика, суржик, що є невід’ємною частиною побутової української мови.

2. *Корпус української мови MOVA.info*. Призначений для пошуку лексем та словоформ в українських текстах певного стилю (для окремих частин корпусу також можливий пошук морфем, морфемних і синтаксичних структур) [13]. У даній роботі використовувався для побудови вектора ключових слів.

3. *UA-GEC*: перший анотований GEC-корпус української мови. Це колекція текстів, написаних звичайними людьми: есеї, дописи в блогах та соцмережах, відгуки, листи тощо. Ці тексти містять граматичні, стилістичні та орфографічні помилки, що максимально наближають їх до повсякденної мови [14].

4. *Ukrainian News Collection*. Українські новини – це колекція з понад 150 тисяч новинних статей, зібраних з понад 20 новинних ресурсів. Зразки наборів даних поділяються на 5 категорій: політика, спорт, новини, бізнес, технології. Набір даних надано некомерційною студентською організацією FIdo.ai (дослідницький відділ машинного навчання FIdo Національного університету «Києво-Могилянська академія») для дослідницьких цілей в області аналізу даних (класифікація, кластеризація, виділення ключових слів тощо) [15].

5. *Український веб-корпус Лейпцизького університету*. Корпус текстів української мови. Містить частотні словники. Словники побудовані на базі Вікіпедії, новинних сайтів, вебдокументів. Можна завантажити в різних обсягах токенів (слів): 10 000, 30 000, 100 000, 300 000, 1000 000 [16].

6. *Карпати буд каркас*. Набір статей де зібрані інформація про перебіг будівництва, новини сучасної архітектури, технології. Складається з понад 200 текстів в середньому по 500 слів (https://karpatybud.com.ua/statti/).

7. *Блог садівника*. Сайт, що містить понад 200 текстів розмірністю близько 500 слів кожен, присвячених тематиці садівництва (https://agro-market.net/ua/news/).

Така кількість джерел обумовлена тим, що сучасної українська мова охоплює багато сфер життєдіяльності.

Тому взявши тільки один з корпусів для дослідження не зможемо охопити весь лексичний запас мови. Кожен із взятих корпусів здатен передавати сенс сучасної української мови, тому для побудови вектору ключових слів $\overline{Word}$ було використано наведені джерела.

7. Прикладне застосування для дослідження ефективності методу виявлення цільових об’єктів предметної області

Для валідації запропонованого інтелектуального методу виявлення цільових об’єктів предметної області за показниками семантичної зв’язності для класифікації текстової інформації з використанням розробленої моделі сучасної української мови, було розроблено програмний застосунок мовою C# для перетворення текстового контенту файлів із тестової вибірки у множину цільових об’єктів предметної області [1]. Головне вікно розробленого застосунку зображено на рисунку 7.

Застосунок дозволяє задавати наступі параметри:

1) шлях до файлу, що містить ключові терміни;

2) обирати метод оцінки термінів в текстовому контенті;

3) обирати корпуси текстів які будуть оброблятися. В результаті роботи застосунку отримується файл який містить цифрове представлення кожного тексту із обраних корпусів.

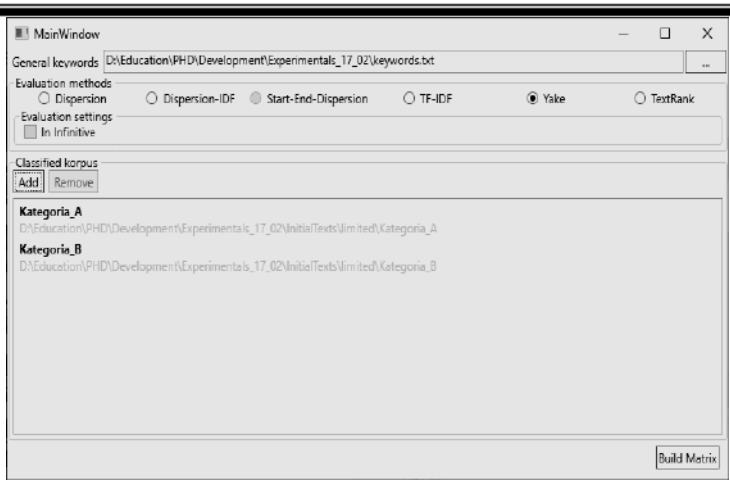

Рисунок 7. *Експериментальний застосунок для пошуку ключових слів альтернативними методами й перетворення текстового контенту в цифрове подання*

Параметри середовища аналізу текстових даних. Отриманий файл передається на обробку програмному застосунку розробленого на мові програмування Python із застосуванням бібліотеки Manifold. Даний застосунок зчитує дані із отриманого на вхід файлу із попереднього етапу та обробляє отримані дані за допомого методу MDS із бібліотеки Manifold. Результат отриманий з методу MDS візуалізується на графічному інтерфейсі.

8. Результати експерименту та дискусія

Для дослідження ефективності запропонованого методу було створено окреме консольне програмне забезпечення мовою Python, яке передбачає використання отриманого списку цільових об’єктів для досліджуваних текстів, та словників для окреслених тем «Карпати буд каркас» та «Блог садівника». Відповідно, знайдені цільові об’єкти були переведені у векторне представлення розміром 1500 (як розмір словника) методом One-Hot Encoding.

Надалі було перевірено Евклідові відстані між текстами одного спрямування (5 текстів категорії «Карпати буд каркас» та 5 текстів «Блог садівника»), а також були обраховані Евклідові відстані між векторами протилежних категорій. Дані експерименту наведено в таблиці 1.

Матриця відстаней таблиці 1 та рисунку 8 демонструють чітке розділення текстів на дві основні групи з різним змістом.

Перша група текстів (1–5), що належать категорії «Карпати буд каркас» має тісніші зв'язки між собою, аналогічно як друга група (6–10) також має менші внутрішні відстані (категорія «Блог садівника»), 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.

Таблиця 1

Евклідові відстані між текстами (№1-10) одного спрямування

	№1	*№2*	*№3*	*№4*	*№5*	*№6*	*№7*	*№8*	*№9*	*№10*
№1	0	10.3	11.2	9.75	14.7	25.7	23.4	28.6	29.6	24.7
№2	10.3	0	15.7	17.1	16.4	30.21	24.5	26.3	23.34	26.5
№3	11.2	15.7	0	9.4	8.89	27.6	24.9	23.8	25.7	27.1
№4	9.75	17.1	9.4	0	5.47	32.4	30.7	26.1	27.6	23.6
№5	14.7	16.4	8.89	5.47	0	19.4	23.45	26.12	28.4	24.7
№6	25.7	30.21	27.6	32.4	19.4	0	9.78	6.99	9.1	14.3
№7	23.4	24.5	24.9	30.7	23.45	9.78	0	11.9	12.45	7.98
№8	28.6	26.3	23.8	26.1	26.12	6.99	11.9	0	6.33	8.91
№9	29.6	23.34	25.7	27.6	28.4	9.1	12.45	6.33	0	13.5
№10	24.7	26.5	27.1	23.6	24.7	14.3	7.98	8.91	13.5	0

Результати отримані з таблиці проілюстровані графіком на рисунку 8.

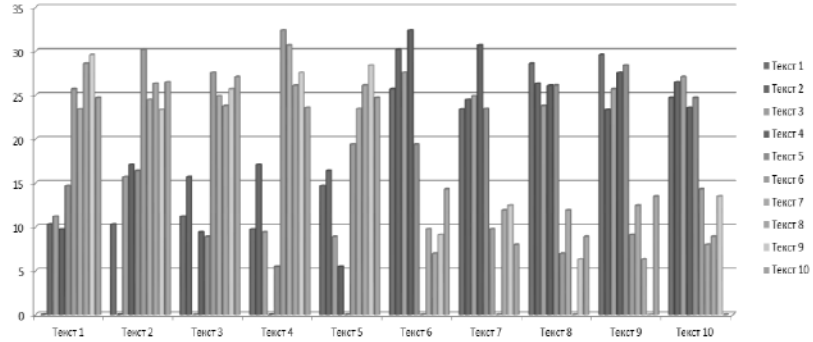

Рисунок 8. *Евклідові відстані між тестовими текстами двох категорій*

При побудові мовної моделі основою пропонованого узагальненого вектору $\overline{Word}$ є словник MOVA.info [13], оскільки він найближче підходить для задач класифікації інтернет-контенту з додаванням слів інших словників. Для досягнення поставленої задачі було також проведено фільтрацію за іменниками, оскільки отриманий словник мав різні частини мови та був неоднорідним. Отриманий таким чином вектор складається з 1500 слів.

Довжина вектору слів моделі для векторизації україномовного сегменту Інтернет у 1500 одиниць була встановлена за результатом досліджень. Наприклад, на рисунку 9 наведено візуалізацію бінарної класифікації описаної

множини текстів методом TF-IDF за довжини вектора у 3000 слів, а на рисунку 10 наведено візуалізацію бінарної класифікації описаної множини текстів методом TF-IDF за довжини вектору у 2000 слів. Дослідження встановили, що найкраща здатність до роздільності при класифікації текстів спостерігається при довжині вектора слів рівній 1500 одиниць.

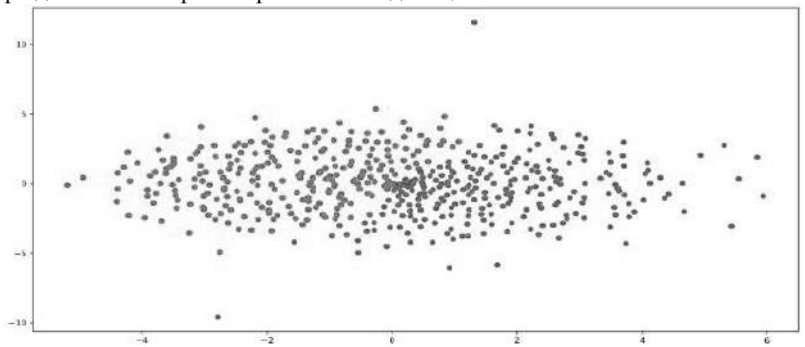

Рисунок 9. *Візуалізація бінарної класифікації текстів методом TF-IDF за довжини вектора у 3000 слів*

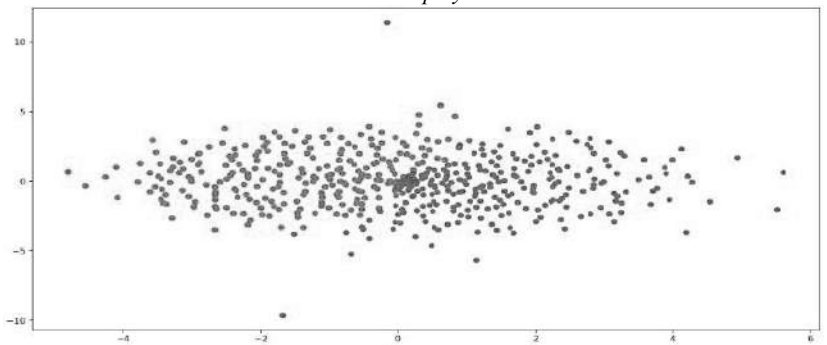

Рисунок 10. *Візуалізація бінарної класифікації текстів методом TF-IDF за довжини вектора у 2000 слів*

Для проведення експериментальних досліджень були зібрані тексти двох категорій: будівництво (https://karpatybud.com.ua/statti/) та садівництво (https://agro-market.net/ua/news/). Ці зібрання містять по 200 текстів кожен, в середньому довжиною в 500 слів кожен текст. Оскільки у завданнях дослідження метою є бінарна класифікація, то використати дві категорії достатньо для валідації пропонованої моделі та проведення експериментальних досліджень.

Результат 1. В результаті валідації якості моделі побутового лексикону Україномовного сегменту Інтернет за використання методу TF-IDF були отримані результати, зображені на рисунку 11. Результат можна оцінити як незадовільний, оскільки категорії деяких текстів було визначено помилково, а

область поділу/розмежування є нечіткою. Це пояснюється характерними рисами цього методу, які полягають у його високій чутливості до підбору текстів альтернативних категорій. Оскільки альтернативні категорії можуть бути поза межами використаних класів для класифікації, що особливо чутливо для текстів побутової тематики, то це може приводити до випадків неправильної класифікації текстів.

Результат 2. Отримані за використання методу дисперсійного оцінювання результати валідації якості моделі побутового лексикону проілюстровані на рисунку 12. Результат можна оцінити як прийнятний, оскільки є тексти, які містяться на межі класів. Метод дисперсійного оцінювання для ефективного обрахунку значень семантичної важливості слів потребує якомога більшої кількості появ значущих слів в окремих текстах. Побутове спілкування характеризується фрагментарним використанням значущих слів у невеликій кількості, тому в деяких випадках незадовільні статистичні показники унікальних слів тексту не забезпечують достатньої роздільності значень показників, актуальних для класифікації.

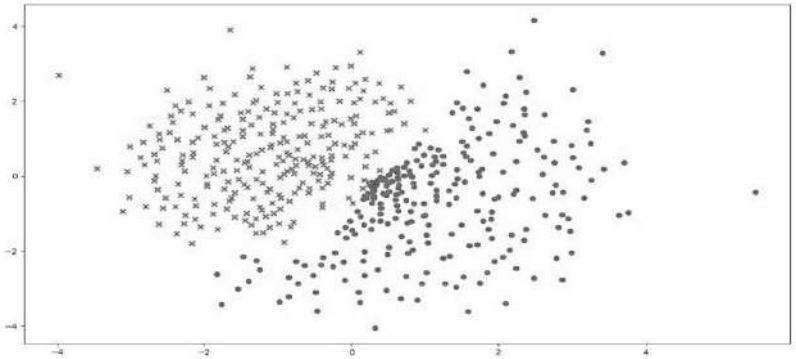

Рисунок 11. *Валідація якості моделі за методом TF-IDF*

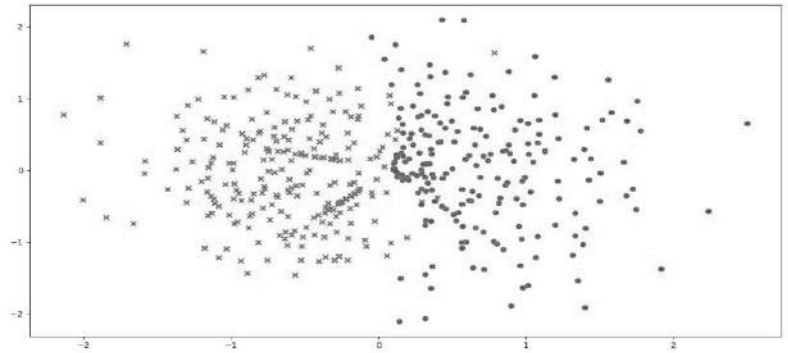

Рисунок 12. *Валідація якості моделі за методом дисперсійного оцінювання*

Результат 3. Отримані за використання методу Text Rank результати проілюстровані на рисунку 13. Якість результату можна оцінити як високу, оскільки обидві категорії мають чітке розділення. Метод Text Rank для обрахунку значень семантичної важливості слів використовує не тільки власне позиції слів у тексті, а й взаємозв'язки між словами та взаємозв'язок частот появ слів по відношенню до взаємозв'язків між словами.

Це дозволяє приймати до уваги більшу за інші методи кількість параметрів тексту, в той час як тематики побутового спілкування не дозволяють одержувати у великій кількості первинні показники для оцінювання. Тому даний метод виявив достатньо високу ефективність роздільності текстів за категоріями.

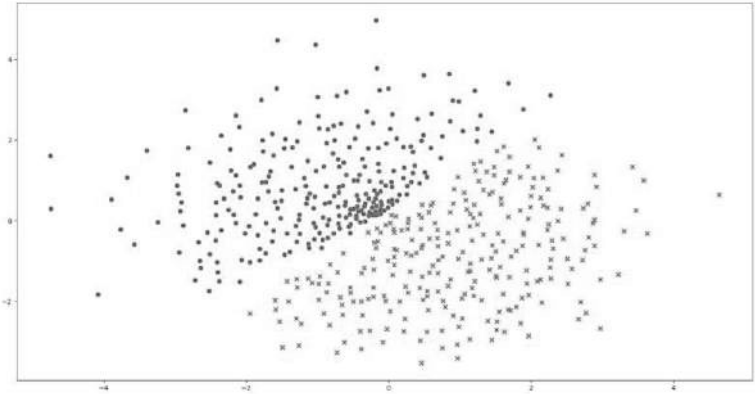

Рисунок 13. *Валідація якості моделі за методом Text Rank*

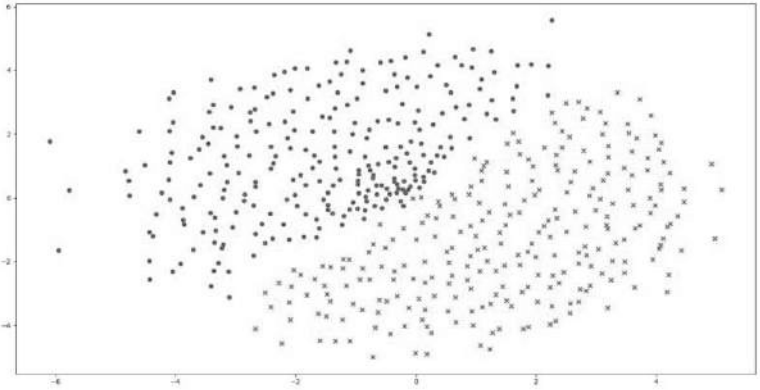

Рисунок 14. *Валідація якості моделі за методом YAKE*

Результат 4. За використання методу YAKE були отримані результати класифікації текстів, зображені на рисунку 14. Такі результати можна інтерпретувати як між високим та задовільним. Категорії розділені, проте чіткої межі немає. Цей метод враховує ряд важливих показників тексту, таких як частота слів, різноманітність контексту появ слова та частота появ слів у різних реченнях. Проте характерні особливості побутового спілкування не дозволяють методу використати його характерні переваги. До таких переваг належать врахування використання великих літер і скорочень, присутність слів на початку тексту. Також, при збільшенні числа кількості текстів можливості методу YAKE знижуються і його застосування може призвести до нечіткої класифікації. Тому цей метод, ефективний для текстів іншого характеру, таких як наукової статті, в випадку аналізу текстів з побутового спілкування виявив менш задовільний результат.

Результат 5. За отриманими різними методами числовими значеннями для вектору слів які відповідають важливості кожного з них для задачі класифікації, був проведений аналіз, шляхом їх перетину для різних методів. В результаті наведеного отримана множина слів яка є загальною для методів що розглядались. Об'єм множини склав біля 500 слів.

Отримані слова можна вважати достатніми для моделювання задачі класифікації за наборами текстів, що розглядались. В подальшому, отриманий таким чином набір можливо використати як базовий для формування векторної моделі. Формування моделі може проходити шляхом доповнення базової множини слів словами що притаманні конкретним задачам, що розглядатимуться: виявлення суїцидних настроїв [26], буллінгу [27], негативного емоційного забарвлення текстів [28], негативного контенту тощо.

З отриманих результатів валідації моделі побутового лексикону Україномовного сегменту Інтернет видно, що класифікація текстів побутового характеру є найбільш ефективною за використанням методу пошуку ключових слів Text Rank. Подальші дослідження направлені на вдосконалення та модифікацію вже описаних підходів шляхом перевірок припущень, використання для векторизації текстів композиції методів тощо. Також подальші дослідження будуть спрямовані на удосконалення загального вектору слів сучасної побутової української мови та розв'язання задач визначення негативного забарвлення текстових повідомлень в сегменті інтернет-спілкування.

9. Висновки

Було розглянуто поточний стан наукового напряму виявлення цільових об'єктів предметної області, та на основі опрацьованого матеріалу запропоновано власний інтелектуальний метод виявлення цільових об'єктів предметної області за показниками семантичної зв'язності для класифікації текстової інформації, який призначений для автоматизації процесу ідентифікації ключових елементів у великих масивах текстових даних,

спрямований на підвищення точності та ефективності аналізу текстової інформації.

Запропонований метод показав, що знайдені цільові об'єкти предметних областей спроможні виконувати подальшу задачу класифікації, демонструючи на метриці Евклідових відстаней групування текстів однієї категорії та збільшення відстані ортогональної їй. Супутнім результатом реалізації методу є розробка моделі сучасної української мови, у рамках якої можлива гіперплощина класифікація текстів для виявлення цільових об'єктів предметної області за показниками семантичної зв'язності. Тож було запропоновано модель сучасної української мови, що побудована шляхом об'єднання значимих для передачі сенсу відомих частотних словників.

Модель має використовуватись для аналізу контенту Україномовного сегменту Інтернет, який характеризується використанням граматично й синтаксично хибних, але в побутовому спілкування розповсюджених слів, враховуючи нецензурні; тому існуючі сучасні частотні словники не покривають у повній мірі вказаний сегмент. Для створення коректної моделі сучасної української мови було використано частотні словники існуючих корпусів української мови, які сукупно охоплюють як різні сфери діяльності та типи контенту, так і різні області спілкування, враховуючи спілкування в Інтернет. Шляхом об'єднання за частотними показниками словників цих корпусів української мови було одержано вектор слів, який для покращення спроможності класифікації був відфільтрований за іменниковою групою та обмежений кількісно.

Для валідації пропонованої моделі було взято дві ортогональних множини текстів розмірністю понад 200 у кожній. Кожен текст був векторизований одним із чотирьох пропонованих методів пошуку ключових слів (TF-IDF, дисперсійної оцінки, Text Rank, YAKE) за ключовими словами та був віднесений до заданої категорії. Для забезпечення можливості оцінки якості отриманих моделей був використаний метод MDS та запропоновані критерії для інтерпретації отриманих результатів за трьохрівневою шкалою. Це дозволило визначити метод пошуку ключових слів, який найкраще підходить для використання з запропонованою моделлю сучасної української побутової мови. Проведені дослідження визначили наступні результати:

1. Встановлено, що найкраща здатність до роздільності при класифікації текстів з використанням запропонованої моделі сучасної української побутової мови спостерігається при довжині вектора слів рівній 1500 одиниць, це визначає оптимальну розмірність моделі.

2. За результатом тестової класифікації понад 400 текстів, з використанням візуальної верифікації результатів класифікації методом MDS було визначено, що метод пошуку ключових слів Text Rank найкраще підходить для використання з запропонованою моделлю сучасної української побутової мови.

3. Підтверджено, що метод візуальної аналітики MDS є ефективним і достатнім інструментом для візуальної верифікації результатів класифікації цифрових текстів за різними категоріями, до яких належать як тематичні класи, так і категорії емоційного забарвлення тестів.

Наведене дає можливість стверджувати про можливість використання створеної моделі сучасної української мови для вирішення задач аналізу текстового контенту одиниць інтернет-спілкування та його класифікації за різними ознаками. У подальших дослідженнях планується вдосконалення наведених підходів до векторизації з перевіркою припущень використання композиції методів тощо. Також є потреба спрямувати дослідження на удосконалення загального вектору слів української мови та розв'язання задач визначення негативного забарвлення текстових повідомлень під час інтернет-спілкування.

Характерною рисою розглянутого підходу є його висока ефективність при роботі з сучасної українською мовою як характерного представника флективних мов. З огляду на особливості підходу, він виявиться ефективним також для аналітичних мов, зокрема англійської. Це підвищує цінність підходу для роботи з україномовним контентом побутового спілкування, оскільки в ньому часто присутні як запозичення англійські слова та власні назви. Ефективність підходу для аглютинативних мов таких як угорська вбачається нижчою, оскільки ідентифікація формантів та робота з ними дещо відрізняється від роботи з флексіями флективних мов і потребує окремих рішень. Наведене формує окремий напрямок подальших досліджень по роботі з текстами побутового спілкування, що мають змішаний багатомовний контент.

Основним результатом роботи є створений інтелектуальний метод виявлення цільових об'єктів предметної області за показниками семантичної зв'язності для класифікації текстової інформації з використанням розробленої моделі сучасної української мови, який відрізняється від існуючих урахуванням ключових слів та іменникових сутностей предметної області, що дало змогу підвищити точність виявлення цільових об'єктів предметної області внаслідок врахування іменникових сутностей. Подальші дослідження будуть спрямовані на розширення кількості категорій та експерименти із іншими метриками оцінки знайдених цільових об'єктів, у порівнянні їх із відомими великими мовними моделями, на кшталт GPT, Gemini тощо.

10. Література

[1] 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. 2024. Pp. 91-96.

[2] 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.

[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. May 29-31, 2024. Florence, Italy. 2024. Pp. 86-91.
[4] 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.
[5] Slobodzian V., Molchanova M., Kovalchuk O., Sobko O., Mazurets O., Barmak O., Krak I. An Approach Based on the Visualization Model for the Ukrainian Web Content Classification. 2022 12th International Conference on Advanced Computer Information Technologies, ACIT 2022. 2022. pp. 400-405.
[6] Krak Y., Barmak O., Mazurets O. The practice implementation of the information technology for automated definition of semantic terms sets in the content of educational materials. CEUR Workshop Proceedings. 2018. vol. 2139. pp. 245–254.
[7] Shin H., Lee H. J., Cho S. General-use unsupervised keyword extraction model for keyword analysis. Expert Systems with Applications. Volume 233, 2023.
[8] Ma Ch., Shen A., Yoshikawa H., Iwakura T., Beck D., Baldwin T. On the (In)Effectiveness of Images for Text Classification. Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics. Association for Computational Linguistics. 2021. pp. 42-48.
[9] Ghosh D., Shrivastava R., Muresan S. «Laughing at you or with you»: The Role of Sarcasm in Shaping the Disagreement Space. 2021.
[10] Barnes J., Øvrelid L., Velldal E. If you've got it, flaunt it: Making the most of fine-grained sentiment annotations. Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics. Association for Computational Linguistics. 2021. pp. 49–62.
[11] Chen X., Zhang, Z. Lu X. Named Entity Recognition via Unified Information Extraction Framework. 2024 4th International Conference on Computer Communication and Artificial Intelligence (CCAI). Xi'an, China. 2024. pp. 308-313.
[12] Corpus of Modern Ukrainian Language (BRUK). URL: https://r2u.org.ua/corpus.
[13] MOVA.info: about the Ukrainian language, linguistics and more. URL: http://www.mova.info/.
[14] UA-GEC: the first annotated GEC-corpus of the Ukrainian language. URL: https://ua-gec-dataset.grammarly.ai/.
[15] Ukrainian News is a collection. URL: https://github.com/fido-ai/ua-datasets/tree/main/ua_datasets/src/text_classification.
[16] Deutscher Wortschatz. Corpora Ukrainian. URL: https://wortschatz.uni-leipzig.de/en/download/Ukrainian#ukr_mixed_2014.
[17] Jiang Zh., Gao Bo, He Y., Han Y., Doyle P., Zhu Q. Text Classification Using Novel Term Weighting Scheme-Based Improved TF-IDF for Internet Media Reports. Mathematical Problems in Engineerin. 2021.

[18] Krak I., Barmak O., Mazurets O. The practice investigation of the information technology efficiency for automated definition of terms in the semantic content of educational materials. CEUR Workshop Proceedings. 2016. vol.1631. pp. 237–245.
[19] Kazemi A., P´erez-Rosas V., Mihalcea R. Biased TextRank: Unsupervised Graph-Based Content Extraction, Proceedings of the 28th International Conference on Computational Linguistics, Barcelona, Spain (Online), 2020. pp. 1642–1652.
[20] Huang Zh., Xie Zh. A patent keywords extraction method using TextRank model with prior public knowledge, Complex Intell. Syst. 2021.
[21] Zhang M., Li X., Yue Sh., Yang L. An Empirical Study of TextRank for Keyword Extraction, IEEE Access, Volume 8. 2020.
[22] Campos R., Mangaravite V., Pasquali A., Jorge A., Nunes C., Jatowt A. YAKE! Keyword extraction from single documents using multiple local features. Information Sciences. Volume 509. 2020. pp. 257–289.
[23] Manziuk E.A. Barmak O.V., Krak Iu.V., Pasichnyk O.A., Radiuk P.M., Mazurets O.V. Semantic alignment of ontologies meaningful categories with the generalization of descriptive structures. Problems in programming. 2023. Vol. 3-4. Pp. 355-363.
[24] Тимуш О.Ю. Шпичко А.В., Мазурець О.В. Дослідження ефективності інформаційної технології тематичного сортування текстових повідомлень. Збірник наукових праць за матеріалами XI всеукраїнської науково-практичної конференції «Актуальні проблеми комп'ютерних наук АПКН-2019». Хмельницький, 2019. Т.1. С.207-212.
[25] Molchanova M., Mazurets O., Sobko O., Boiarchuk I. Object-Oriented Approach for Ethnic Enmity Detection in Text Messages by NLP. Proceedings of XXI International Scientific and Practical Conference «Scientific Achievements and Innovations as a Way to Success». Vilnius, Lithuania. 2024. Pp. 73-77.
[26] Sobko O., Mazurets O., Didur V., Chervonchuk I. Recurrent Neural Network Model Architecture for Detecting a Tendency to Atypical Behavior Of Individuals by Text Posts. Theoretical and Practical Aspects of Modern Research. Proceedings of XXVI International scientific and practical conference. June 5-7, 2024. International Scientific Unity. Ottawa, Canada. 2024. Pp. 113-117.
[27] Залуцька О.О., Молчанова М.О., Мазурець О.В., Мельник О.І., Скрипник Т.К. Метод інтелектуального аналізу емоційної тональності текстової інформації для визначення поведінкових намірів нейромережевими засобами. Науковий журнал «Вісник Хмельницького національного університету» серія: Технічні науки. Хмельницький, 2023. №5 (325). Т.1. С. 67-73.
[28] Молчанова М.О., Мазурець О.В., Собко О.В., Кліменко В.І., Андрощук В.І. Метод нейромережевого виявлення кібербулінгу з використанням хмарних сервісів та об'єктно-орієнтованої моделі. Науковий журнал «Вісник Хмельницького національного університету» серія: Технічні науки. Хмельницький, 2024. №2 (333). С. 200-206.

INTELLIGENT METHOD FOR IDENTIFYING TARGET OBJECTS OF THE SUBJECT AREA BASED ON SEMANTIC CONNECTIVITY INDICATORS FOR TEXT INFORMATION CLASSIFICATION

Ph.D. **O. Mazurets** ORCID: 0000-0002-8900-0650
Khmelnytskyi National University, Ukraine
E-mail: exe.chong@gmail.com
R. Vit *ORCID: 0009-0009-6958-4730*
Khmelnytskyi National University, Ukraine
E-mail: vit.roman.vit@gmail.com

Abstract: The intelligent method for identifying target objects of subject area based on indicators of semantic connectivity was considered, which is designed to automate the process of identifying key elements in large arrays of text data. The developed method differs from the existing ones by taking into account keywords and noun entities of the subject area, which made it possible to increase the accuracy of detection of target objects of the subject area due to the consideration of noun entities. The conducted studies established that the target objects found by the created method are able to perform the further task of classification, demonstrating on the metric of Euclidean distances the grouping of texts of the same category and the increase of the distance orthogonal to it.

For the applied implementation of created intellectual method for identifying target objects of subject area, the model of modern Ukrainian language was formed, which was built by combining known frequency dictionaries that are significant for conveying meaning. To create the correct model of modern Ukrainian language, frequency dictionaries of existing corpora of the Ukrainian language were used, which collectively cover both different fields of activity and types of content, as well as different areas of communication, taking into account communication on the Internet. By combining the dictionaries of these corpora of the Ukrainian language according to frequency indicators, a vector of words was obtained, which was filtered by noun group and quantitatively limited in order to improve the ability of classification. The conducted research made it possible to assert the possibility of using the created model of modern Ukrainian language to effectively solve the problems of analyzing the text content of Internet communication units and classifying it according to various characteristics.

***Keywords*:** target objects of subject area, texts classification, Ukrainian language model, frequency dictionaries, words vector, semantic connectivity