Monograph

Advances in Information Control Systems and Technologies

Information control systems

Intelligent systems and data analysis

Modeling and software engineering

ICST-2024

Національний університет «Одеська політехніка»

РОЗВИТКИ ІНФОРМАЦІЙНО-КЕРУЮЧИХ СИСТЕМ ТА ТЕХНОЛОГІЙ

МОНОГРАФІЯ

1256 1233 1996 LIHA-PRES
Львів-Торунь
2024

УДК 004:37:001:62
Р64

Авторський колектив:
Д. Антонов, Н. Аксак, О. Базей, О. Беспалов, Ю. Бондар, Р. Віт, В. Вичужанін, О. Вичужанін, В. Власенко, О. Гузун, А. Давиденко, О. Задорожний, М. Корабльов, А. Король, Ю. Котух, М. Кушнарьов, Л. Ковальчук, С. Машталір, О. Мазурець, М. Молчанова, П. Михайловський, С. Науменко, Ю. Нетребін, О. Ніколенко, Г. Охотько, О. Орєхов, С. Орлов, Т. Отрадська, І. Петров, О. Поліщук, І. Розломій, М. Рудніченко, В. Саваневич, О. Собко, А. Татарніков, О. Ткачук, В. Троянський, Т. Трунова,Є. Хаджиєв, Г. Халімов, С. Хламов, Д. Шведов, Н. Шибаєва, І. Шпінарева, А. Ярмілко

Рекомендовано до друку вченою порадою
Національного університету «Одеська політехніка»
(протокол №1 від 25.09.2024 р.)

Рецензенти:
Професор **Є Чженмао**, Південний університет (США);
Професор **С. Овермайер**, Університет Південного Нью-Гемпширу (США);
Професор **А. Тріпаті**, Індійський технологічний інститут (BHU) (Індія)

Р64 **Розвитки** інформаційно-керуючих систем та технологій : монографія / Н. Аксак, Д. Антонов та ін.; під наук. ред. проф. В.Вичужаніна. – Львів-Торунь : Liha-Pres, 2024 – 380 с.
Укр., англ. мовами.

ISBN 978-966-397-422-4

DOI 10.36059/978-966-397-422-4

У монографії відображені результати наукових досліджень у галузі інформаційних, інтелектуальних систем та аналізу даних, моделювання. Матеріали монографії будуть корисними для аспірантів, магістрантів, викладачів вищих навчальних закладів, що спеціалізуються на галузі ІТ-технологій.

УДК 004:37:001:62

ISBN 978-966-397-422-4

CONTENTS

PREFACE

The monograph contains materials from articles received from the Council of articles sent to the XII International Scientific and Practical Conference "Information Control Systems & Technologies (ICST-Odesa-2024)"

The collective monograph presents the results of scientific research in the field of information systems and technologies, intelligent systems, data analysis, modeling and software development.

The monograph is compiled in the form of scientific articles- sections corresponding to thematic areas, which comprehensively reflect the results of research in the following areas: intelligent control technologies; intelligent systems and data analysis, modeling and software engineering, mathematical and simulation modeling.

In the monograph, the authors pay attention to solving problems in the field of information systems and technologies using improved encryption based on non-belian small groups ri, the method of assessment and forecasting the technical state of complex systems of critical application based on precedents, protection of multilayer network systems from sequential, group and system-wide targeted attacks, new statistical criteria for checking the independence of bit random variables and sequences. Considerable attention is paid to the agent-controlled approach to increasing the efficiency of e-learning, automated data extraction of individual astronomical objects from blurry ccd images, artificial intelligence in diagnostics, prognosis and treatment of diabetic neovascular glaucoma.

Thanks to the presented research results, readers will receive a useful amount of knowledge necessary for a better understanding problems and solutions in the field of information systems and technologies.

The articles presented in the monograph correspond to the original author's variants. The authors alone are responsible for the content of the articles.

The materials of the monograph will be useful for postgraduate students, master's students, and teachers of higher educational institutions specializing in the field of information systems and technologies.Scientists including 11 doctors of science, 15 candidates of science and 17 applicants took part in the work on the collective monograph.

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.

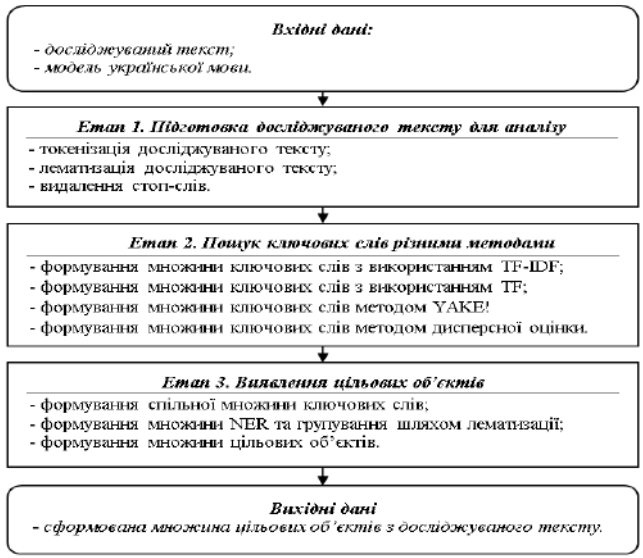

Рисунок 1. *Етапи роботи методу виявлення цільових об'єктів предметної області за показниками семантичної зв'язності*

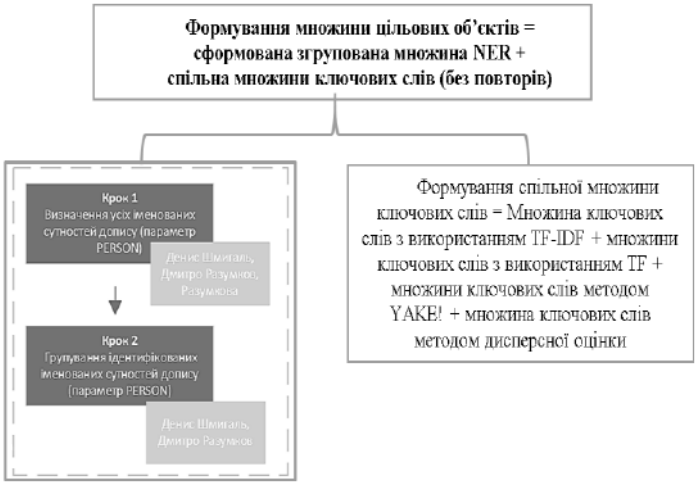

Рисунок 2. *Приклад виконання етапу формування цільових об'єктів методу виявлення цільових об'єктів предметної області*

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.

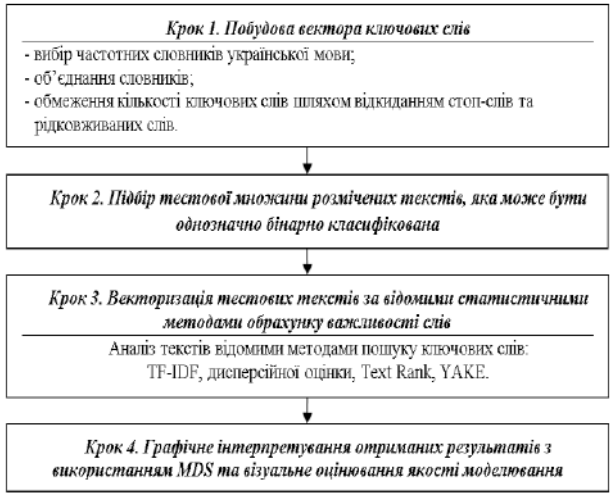

Рисунок 3. *Схема підходу до створення моделі сучасної української мови для методу виявлення цільових об'єктів предметної області*

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).

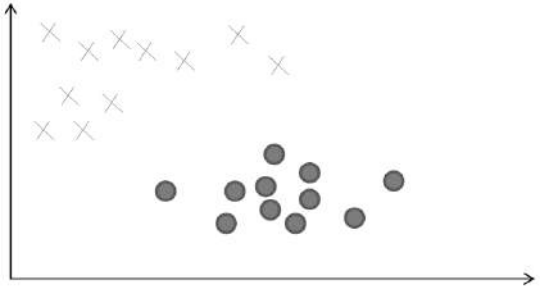

Рисунок 4. *Приклад високого рівня якості моделі для задачі класифікації*

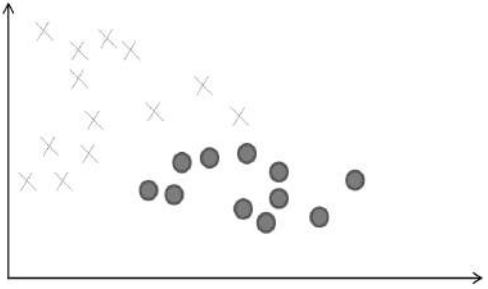

Рисунок 5. *Приклад прийнятного рівня якості моделі для задачі класифікації*

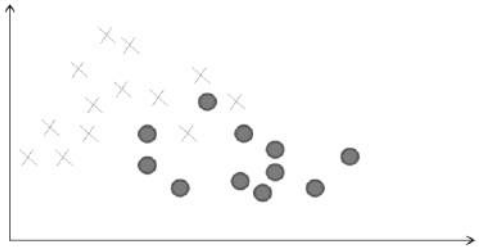

Рисунок 6. *Приклад незадовільного рівня якості моделі для задачі класифікації*

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.

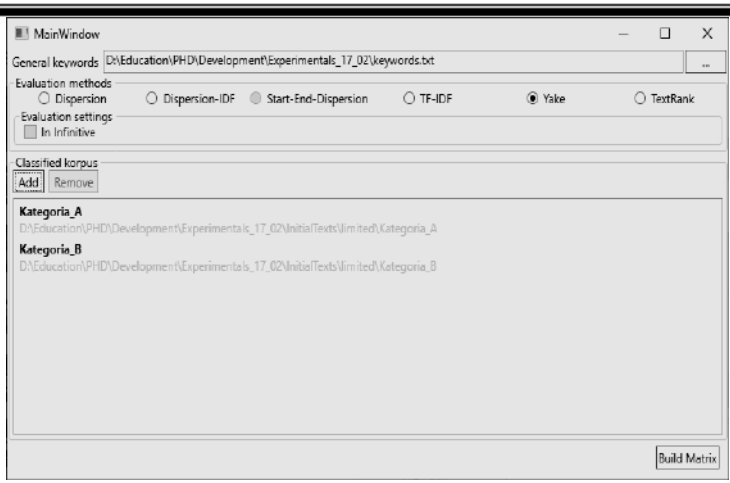

Рисунок 7. *Експериментальний застосунок для пошуку ключових слів альтернативними методами й перетворення текстового контенту в цифрове подання*

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.

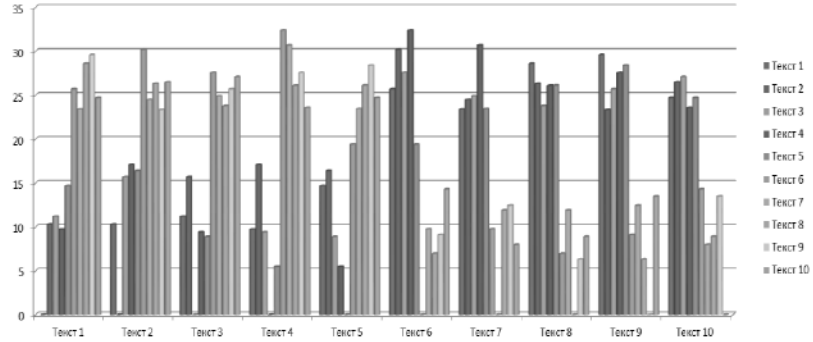

Рисунок 8. *Евклідові відстані між тестовими текстами двох категорій*

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.

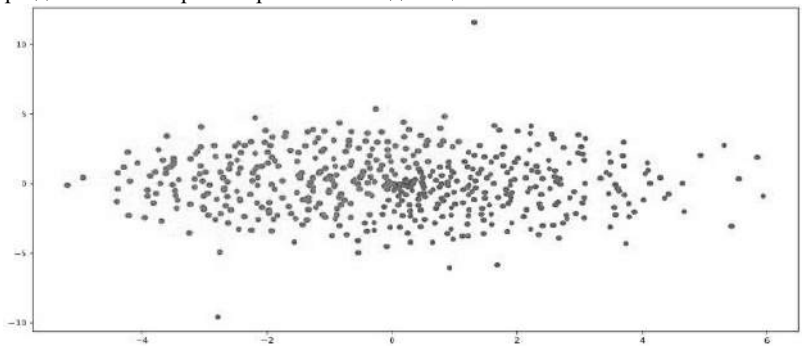

Рисунок 9. *Візуалізація бінарної класифікації текстів методом TF-IDF за довжини вектора у 3000 слів*

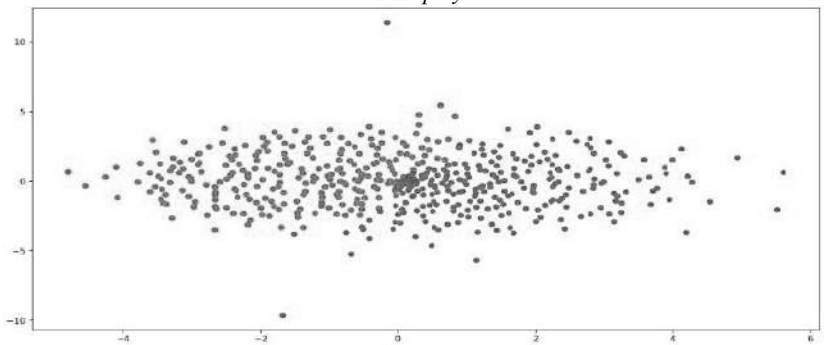

Рисунок 10. *Візуалізація бінарної класифікації текстів методом TF-IDF за довжини вектора у 2000 слів*

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.

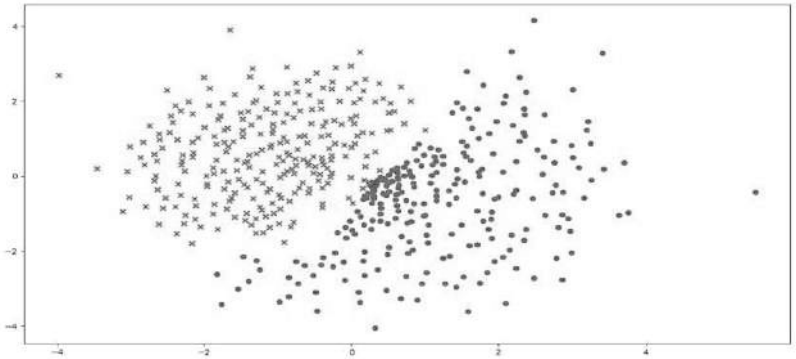

Рисунок 11. *Валідація якості моделі за методом TF-IDF*

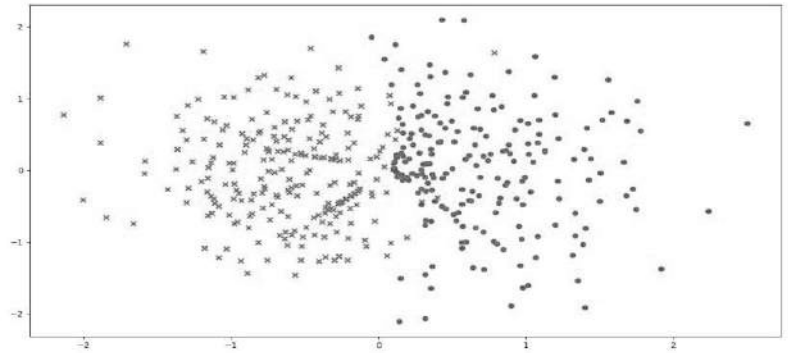

Рисунок 12. *Валідація якості моделі за методом дисперсійного оцінювання*

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.

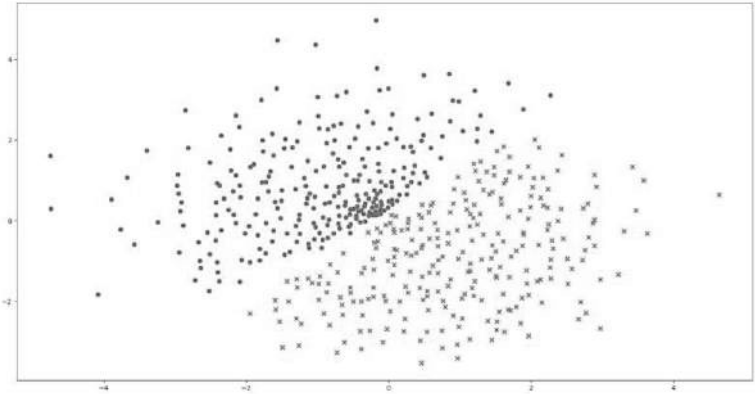

Рисунок 13. *Валідація якості моделі за методом Text Rank*

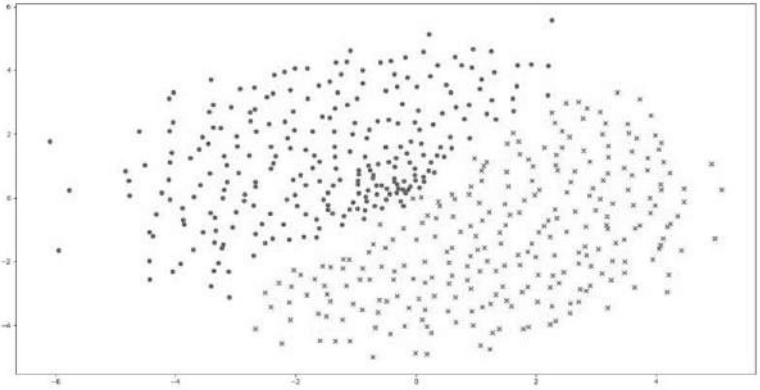

Рисунок 14. *Валідація якості моделі за методом YAKE*

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

УДК 004.8

МЕТОД ВИЯВЛЕННЯ ТА КЛАСИФІКАЦІЇ ТЕХНІК ПРОПАГАНДИ У ТЕКСТОВОМУ КОНТЕНТІ ЗАСОБАМИ ШТУЧНОГО ІНТЕЛЕКТУ

М. Молчанова ORCID: 0000-0001-9810-936X
Хмельницький національний університет, Україна
E-mail: m.o.molchanova@gmail.com

Анотація. *Робота присвячена створенню та апробації методу нейромережевого виявлення техніки пропаганди за маркерами з візуальною аналітикою, що дозволяє перетворювати вхідні дані у вигляді тексту для аналізу та моделей керованого машинного навчання у вихідні дані, що містять числові оцінки присутності кожного пропагандистського прийому та розміченого тексту з візуальною аналітикою присутності виявлених пропагандистських маркерів. Було проведено дослідження, яке дозволяє виявити 17 основних пропагандистських прийомів. У дослідженні порівнювалися 3 підходи, які найчастіше використовуються: традиційний підхід машинного навчання, підхід на основі рекурентних нейронних мереж і підхід на основі трансформаторних моделей. Найвищих результатів досяг підхід на основі моделі трансформатора, який використовує механізми самоуважності, які дозволяють кожному елементу послідовності безпосередньо взаємодіяти з усіма іншими елементами. Це забезпечує ефективне захоплення довготривалих залежностей, що характерно для пропагандистських технік. Такий підхід дозволив виявити методи пропаганди з точністю до 0,96.*

Ключові слова: *BERT, RNN, методи пропаганди, виявлення пропаганди, пропагандистські маркери, візуальна аналітика*

1.Вступ

Пропаганда, замаскована під звичайні новини, поширюється вже багато десятиліть, але сучасна цифрова ера додатково створює умови для її більш швидкого, масового та ефективного поширення [1].

Створюються нові методи генерації текстів, які дедалі частіше мало відрізняються від створених людиною [2], що призводить до стрімкого зростання кількості контенту. Тому це все підкреслює важливість створення автоматизованих методів для виявлення пропагандистських маніпуляцій, які допоможуть користувачам отримувати інформацію більш усвідомлено.

Метою дослідження є підвищення точності виявлення методів пропаганди шляхом розробки методу виявлення методів пропаганди за маркерами на основі набору моделей машинного навчання, окремо для кожного методу пропаганди, навчених на модифікованих маркованих даних.

Основні внески дослідження можна підсумувати так:

– Розроблено підхід до підготовки навчальних даних, що дозволяє проводити навчання моделей машинного навчання для окремих технік пропаганди;

– Запропоновано метод виявлення прийомів пропаганди, що дозволяє знаходити силу прояву кожного з 17-и прийомів пропаганди, а також візуально інтерпретовувати отриманий результат з використанням моделі LIME.

– Експериментально продемонстровано ефективність використання нейромережевих моделей-трансформерів в порівнянні з рекурентними моделями та Traditional machine learning approaches

В цій роботі, далі представлено огляд пов'язаних робіт у сфері виявлення прийомів пропаганди згідно двох складових дослідження, зокрема аналіз наявних шляхів до вирішення проблеми виявлення пропаганди та аналіз моделей машинного навчання для виявлення прийомів пропаганди. Третій розділ паперу містить схему та кроки методу нейромережевого виявлення технік пропаганди за маркерами. Четвертий розділ присвячений опису плану експерименту виявлення технік пропаганди за маркерами та підготовці датасету. П'ятий розділ містить результати експерименту, їх аналітику та обговорення.

2. Аналіз останніх досліджень та публікацій

Наявні шляхи вирішення проблеми виявлення пропаганди. Проблема виявлення пропаганди залишається актуальною, адже і досі з'являються нові способи впливу на користувачів для поширення пропагандиських повідомлень. З огляду на це, виникає потреба у постійному моніторингу нових способів створення пропагандиського вмісту та вдосконалення методів їх ідентифікації, що є важливою задачею забезпечення інформаційної безпеки та протидії дезінформації. Тому науковці працюють над виявленням нових маркерів та нових прийомів пропаганди, а також над покращенням існуючих підходів для її виявлення.

Досліджено основні методи аналізу газетних текстів для виявлення маніпулятивних технологій, що допомагає застерегти від дезінформації та пропаганди [1]. Представлено новий набір еталонних даних чеською мовою для навчання та оцінки сучасних і майбутніх методів розпізнавання 18 маніпулятивних прийомів, таких як нагнітання страху, релятивізація та навішування ярликів. Показано, що поєднання контент-аналізу з запропонованим стильовим аналізом підвищує точність виявлення 15 з 17 оцінених маніпулятивних технік від 0.05% до 1.46%. Метод перевірено на пропагандистській базі QCRI. Подальші дослідження будуть зосереджені на додаванні нових стилометричних характеристик, вдосконаленні існуючих методів та використанні методів доповнення даних для боротьби з дисбалансом етикеток. Також планується перейти до дрібнозернистої класифікації на рівні проміжків часу, а не на рівні документа в цілому.

У ще одному дослідженні представлено багатомовний набір даних про пропаганду та проведено експеримент для дослідження маркерів, за якими людські анотатори та алгоритми класифікації відрізняють пропагандистські статті від непропагандистських на певну тему [3]. Показано, що перебільшення, зменшення описовості та відсутність адекватних джерел часто зустрічаються у пропагандистській пресі. Аналізатор VAGO підтвердив, що використання невизначених маркерів значно корелює з цими особливостями. Виявлено, що моделі машинного навчання ефективні для виявлення пропаганди на певну тему, але потребують покращення щодо пояснюваності та узагальнення на інші теми. Подальші роботи зосередяться на вдосконаленні аналізу, розробці багатомовних моделей та покращенні інструментів пояснюваності. Також планується введення нових міток для уточнення анотацій та ідентифікації більшої кількості стилістичних особливостей.

Застосування моделі MVPROP, що використовує багатовимірні контекстні вбудовування, дозволяє покращити точність виявлення пропаганди. Експерименти показали, що модель може бути перенесена на новинні статті[4]. Для тестування представлено TWEETSPIN, набір даних із твітами, що містить слабкі анотації тонких пропагандистських технік, і модель MVPROP для їх виявлення. TWEETSPIN включає лише ідентифікатори твітів, що відповідає умовам використання Twitter, і містить потенційно образливі та ворожі висловлювання. Основним обмеженням є слабкі анотації через великий масштаб даних. У майбутньому планується дослідження виявлення пропаганди на рівні окремих фрагментів.

Дослідники застосували мовну модель RoBERTa для виявлення пропагандистських технік у новинних статтях [5]. Модель оцінювалась за допомогою референсного набору даних для завдання SemEval-2020 Task 11, демонструючи здатність виявляти складні техніки пропаганди і перевершуючи базову модель з показником F1-score у 60.2%. У той час, як [6] аналізувались можливості використання великих мовних моделей (LLMs), зокрема моделі GPT-3.5-Turbo від OpenAI, для виявлення ознак пропаганди в новинних статтях. Використовуючи технологію, що лежить в основі ChatGPT, дослідники аналізували тексти для визначення присутності різних технік пропаганди, визначених у попередній роботі [7]. Розроблено ретельно уточнений запит, який поєднується зі статтями з мережі Russia Today (RT) та датасету SemEval-2020 Task 11, щоб визначити наявність пропагандистських методик. Дослідження показало, що технологія LLM може давати розумні висновки про пропаганду, хоча точність виявлення складає всього 25.12% за датасетом SemEval-2022. Однак вона демонструє потенціал як інструмент для виявлення пропаганди для кінцевих користувачів, таких як медіа споживачі та журналісти.

Як підтверджено у роботах вище, пропаганда характеризується прийомами, за які відповідають певні маркери, які притаманні використовуваним прийомам. У роботі увага буде зосереджена на виявленні 17 known прийомів пропаганди, детально описаних в [8]. А саме: «Appeal to fear-prejudice», «Causal Oversimplification», «Doubt», «Exaggeration», «Flag-Waving», «Labeling»,

«Loaded Language», «Minimisation», «Name Calling», «Repetition», «Appeal to Authority», «Black and White Fallacy», «Reductio ad hitlerum», «Red Herring», «Slogans», «Thought terminating Cliches», «Whataboutism».

Моделі машинного навчання для виявлення прийомів пропаганди. У рамках дослідження будуть використані 3 підходи щодо моделей машинного навчання:

– Традиційний підхід на основі машинного навчання.

– Підхід на основі рекурентних рейромереж.

– Підхід на основі моделей-трансформерів.

Традиційний підхід на основі машинного навчання охоплює декілька методів і алгоритмів, призначених для розв'язання різноманітних завдань прогнозування, класифікації та кластеризації даних, в тому числі і для виявлення прийомів пропаганди. Лінійна регресія використовується для моделювання лінійних залежностей між вхідними функціями (ознаками) і цільовими значеннями, а також є одним з найпростіших методів регресійного аналізу і часто використовується для прогнозування числових значень. Метод опорних векторів (SVM) шукає оптимальну гіперплощину, яка найкращим чином розділяє два класи точок даних у просторі ознак. Він часто використовується для задач класифікації, особливо коли дані мають складну структуру [9]. Навчання на основі байєсових мереж ґрунтується на байєсівській ймовірнісній моделі, де кожна змінна розглядається як випадкова, і використовуються правила байєсівського висновку для побудови моделі. Також серед традиційні підходів машинного навчання для виявлення прийомів пропаганди використовується логістична регресія, k-NN, алгоритми кластеризації, наприклад, k-means. У роботі [10] автори наводять результати дослідження кілька моделей класифікації, включаючи мультиноміальний наївний метод Байєса, SVM, логістичну регресію та K-найближчих сусідів. У статті [11] автори представили два альтернативні методи (BERT та SVM) автоматичного визначення прокремлівської пропаганди в газетних статтях і дописах у Telegram.

Підхід до виявлення пропаганди на основі рекурентних нейронних мереж використовується для аналізу послідовних даних, зокрема текстів, що часто зустрічаються у соціальних мережах. Рекурентні нейронні мережі, Long Short-Term Memory (LSTM) і Gated Recurrent Unit (GRU) це різновиди архітектур рекурентних нейронних мереж, кожна з яких має свої особливості і застосування у вирішенні різних завдань у машинному навчанні, зокрема виявленні пропаганди в соціальних мережах. RNN є базовою архітектурою, яка здатна обробляти послідовні дані, зберігаючи інформацію у вигляді внутрішнього стану (пам'яті), який оновлюється при кожному новому вході [12]. LSTM – це розширена версія RNN, яка включає додаткові механізми, такі як ворота забування, ворота оновлення та ворота виходу. GRU являється спрощеною версією LSTM, яка має менше внутрішніх компонент. GRU вважається менш обчислювально витратною архітектурою порівняно з LSTM [13]. Результати дослідження ідентифікації пропаганди на платформі Twitter

під час пандемії COVID-19 авторів [14] показали, що запропонована пропагандистська ідентифікація на основі LSTM показала кращі результати, ніж інші розглянуті у роботі методи машинного навчання. За допомогою запропонованого підходу на основі LSTM досягається точність 77,15%. А у статті [15] автори використовують техніки глибокого навчання Bi-LSTM і Bi-GRU з методами SVM із слабким контролем забезпечили. Даний підхід забезпечив точність 90% у виявленні пропагандиських новин. Автори стверджують, що такий підхід є дуже ефективним і дієвим для нерозмічених даних.

Підхід на основі моделей-трансформерів передбачає використання таких архітектур нейромереж як BERT, RoBERTa, DistilBERT, GPT тощо [16]. BERT є однією з найвідоміших архітектур трансформерів, розроблена Google. BERT здатний досягати вражаючих результатів у завданнях обробки природної мови (NLP) завдяки здатності до контекстної обробки слів і здібності до підготовки звичайних моделей для багатьох NLP-завдань [17]. RoBERTa – це оптимізований підхід до BERT, який покращує навчання і результати моделі на різних NLP-завданнях шляхом застосування різних оптимізаційних стратегій [18]. DistilBERT вважається легковаговою версією BERT, яка зберігає сутність оригінальної моделі, зменшуючи кількість параметрів і зберігаючи високу продуктивність на різних NLP-завданнях. GPT – це родина моделей трансформерів, розроблених OpenAI. Даний підхід застосовують у дослідженні для класифікації пропаганди [19]. Автори використовують три моделі глибокого навчання, CNN, LSTM, Bi-LSTM і чотири моделі на основі трансформаторів, а саме багатомовні BERT, Distil-BERT, Hindi-BERT і Hindi-TPU-Electra. Експериментальні результати вказують на те, що багатомовні моделі BERT і Hindi-BERT забезпечують найкращу продуктивність із найвищим показником F1 84% за даними проведеного експерементального дослідження. Також у дослідженні [20] досліджується продуктивність BERT і RoBERTa, DeBERTa з комбінацією різних методів збільшення даних для виявлення пропагандистських текстів. Автори змогли досягти F1 micro 60% на тестовому наборі, використовуючи ансамбль моделей BERT, RoBERTa та DeBERTa.

Отже, наведені підходи знаходять своє застосування у завданні виявлення прийомів пропаганди.

3. Метод виявлення та класифікації технік пропаганди

Для реалізації методу виявлення та класифікації технік пропаганди за маркерами пропонується створити 17 моделей машинного навчання, кожна з яких буде відповідати за визначений прийом пропаганди. Такий підхід (рисунок 1) дозволить навчити моделі машинного навчання таким чином, щоб у них змогли вибудуватись залежності, притаманні конкретним видам пропаганди.

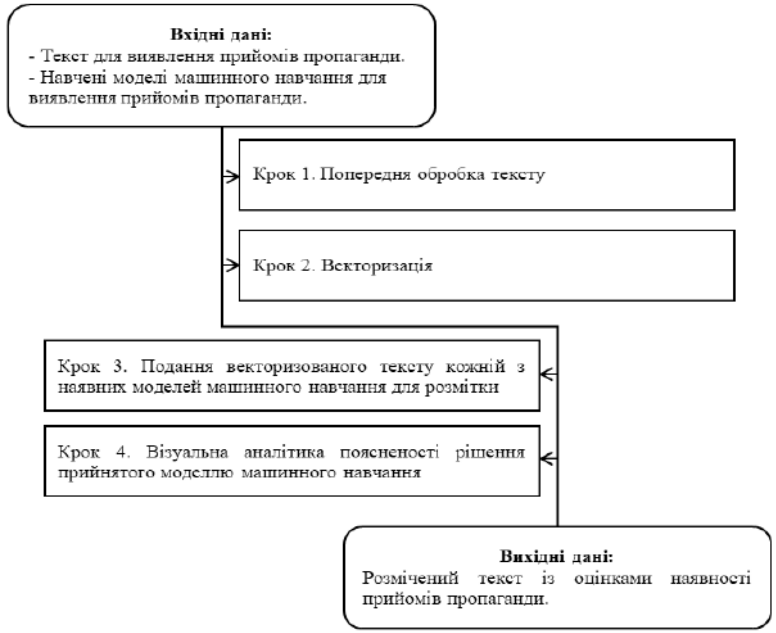

Рисунок 1. *Кроки методу виявлення та класифікації технік пропаганди*

В загальному, схема методу для виявлення прийомів пропаганди наведена на рисунку 1. Метод дозволяє перетворювати вхідні дані у вигляді тексту для аналізу та навчених моделей машинного навчання у вихідні дані, які містять числові оцінки наявності кожного з прийомів пропаганди та розмічений текст з візуальною аналітикою присутності детектованих маркерів пропаганди.

Вхідними даними методу виявлення прийомів пропаганди є текст для виявлення прийомів пропаганди та навчені моделі машинного навчання для виявлення прийомів пропаганди.

Попередня обробка тексту включає видалення розділових знаків та стоп-слів, хоча і розділові знаки, розміщені певним чином, також можуть впливати на наявність пропаганди [22]. Асоціація споріднених слів була виконана шляхом лематизації, яка показує кращі результати, ніж стемінг. Для лематизації була використана відповідна стандартна бібліотека Pyton. Однак, в рамках даного дослідження такий вплив досліджуватись не буде.

Наступним кроком є векторизація тексту після попередньої обробки. Векторизоване представлення подається на вхід кожній навченій моделі машинного навчання, яка виконує передбачення наявності для кожного прийому пропаганди та його сили прояву. Більш детально крок 3 наведено на рисунку 2.

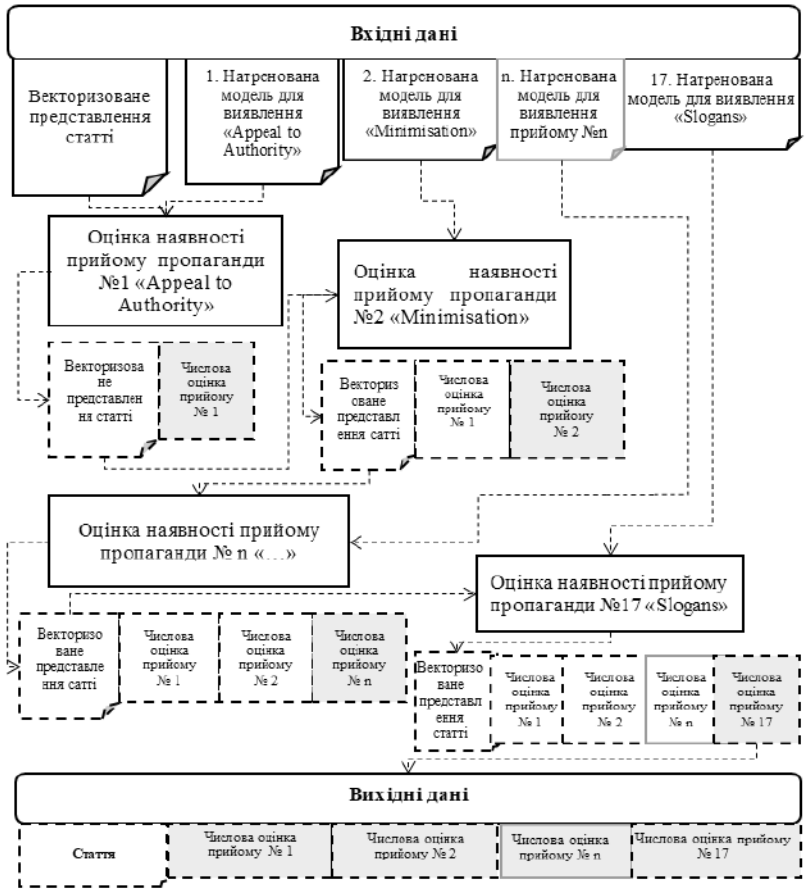

Рисунок 2. *Деталізація кроку подання векторизованого тексту кожній з моделей машинного навчання для розмітки*

Останнім етапом є проведення візуальної аналітики для поясненності рішення прийнятого кожною моделлю машинного навчання. Візуальна аналітика використовується із застосуванням методу Лайма, що є методом для інтерпретації прогнозів моделей машинного навчання, що розроблений для того, щоб пояснювати індивідуальні прогнози складних моделей [23]. LIME наближає будь-яку модель машинного навчання чорної скриньки до локальної, інтерпретованої моделі для пояснення кожного окремого прогнозу. LIME дозволяє зрозуміти, які частини вхідних даних вплинули на рішення моделі.

Вхідними даними кроку подання векторизованого тексту кожній з наявних моделей машинного навчання для розмітки є векторизоване представлення

статті та натреновані 17 моделей машинного навчання. Моделі по черзі оцінюють векторизоване представлення текстового контенту для аналізу на предмет наявності кожного з 17-и прийомів пропаганди. Вихідними даними є числові оцінки сили проявів прийомів пропаганди, що притаманні поданому векторному представленню тексту.

Таким чином, було створено метод нейромережевого виявлення технік пропаганди за маркерами, що дозволяє перетворювати вхідні дані у вигляді тексту для аналізу та навчених моделей машинного навчання у вихідні дані, які містять числові оцінки наявності кожного з прийомів пропаганди та розмічений текст з visual analytic присутності детектованих маркерів пропаганди.

4. Експеримент

У ході проведення експерименту досліджувалось використання 3-х підходів до навчання моделей для виявлення прийомів пропаганди:

– Традиційний підхід на основі машинного навчання.

– Підхід на основі рекурентних рейромереж.

– Підхід на основі моделей-трансформерів.

У рамках традиційного підходу на основі машинного навчання досліджувалось виявлення прийомів пропаганди з використанням моделей регресії, SVM, Random Forest та Naive Bayes. Для прийомів «Appeal to Authority», «Black and White Fallacy», «Reductio ad Hitlerum», «Red Herring», «Slogans», «Thought Terminating Cliches» та «Whataboutism» також буде проведено дослідження із застосуванням SMOTE балансування та без нього.

Підхід на основі рекурентних рейромереж включає в себе порівняння 3-х видів архітектур: RNN, LSTM та GRU.

Підхід на основі моделей-трансформерів включає в себе порівняння BERT-подібних моделей: RoBERTa, BERT, ELECTRA.

Для проведення експерименту було створено програмне забезпечення на мові програмування Python, з використанням бібліотек для машинного навчання Sklearn [24], Tensorflow [25], LimeTextExplainer [26], Numpy [27], Pandas [28]. Програмне забезпечення складається із консольного застосунку для навчання моделей машинного навчання, консольного застосунку для виявлення технік пропаганди за маркерами та веб-модуля для візуальної аналітики оцінювання прийнятих результатів обраною моделлю машинного навчання з її оцінками.

Для навчання моделей машинного навчання, що будуть виконувати функції виявлення прийомів пропаганди, буде використано набір даних «emnlp_trans_uk_dataset», що є перекладеним набором даних «emnlp_en_dataset» з відповідністю розмітки на українській мові, взятий з Kaggle-змагань «Disinformation Detection Challenge» [29] з посиланням на «Analysis Project».

Команда «Analysis Project» [30] провела аналіз текстів, виявивши всі фрагменти, які містять пропагандистські прийоми, а також їх тип. Розподіл статей за довжиною у символах наведено на рисунку 3.

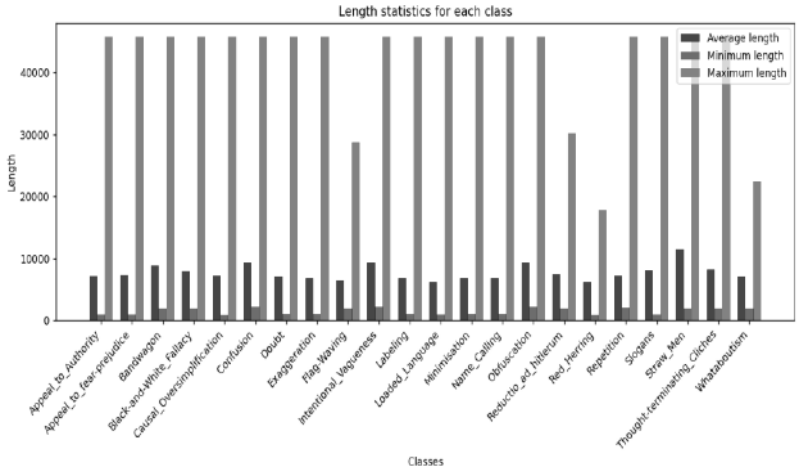

Рисунок 3.*Статистика за довжиною у символах по прийомам пропаганди*

Зокрема, ними створено корпус новинних статей, анотованих вручну на рівні фрагментів за допомогою вісімнадцяти пропагандистських прийомів. Набір даних налічує 788 статей.

Як видно з графіку на рисунку 3, для більшості прийомів пропаганди довжина текстів де вони представлені особливої ролі не грають. Однак, «Flag Waving», «Red Herring», «Reductio ad hitlerum» та «Whataboutism» все ж мають меншу максимальну довжину в текстах, де вони представлені.

Для тренування моделей машинного навчання даний датасет було модифіковано таким чином, щоб текст що містить кожен прийом пропаганди був розміщений в окремому каталозі. Після такого перерозподілу було виведено статистику наявних текстів, що репрезентують прийоми пропаганди. Статистика наведена на рисунку 4. Як видно з рисунку 4 деякі прийоми пропаганди, такі як «Bandwagon», «Confusion», «Intentional Vagueness», «Obfuscation» та «Straw Men» представлені у критично низькій кількості (менше 20 тестів), тому для них окремі класифікатори створені не будуть, ці дані будуть об’єднані у категорію «Інші прийоми пропаганди», однак таким чином, щоб у наявному наборі не були присутні інші прийоми, відмінні від п’яти перерахованих. До прийомів пропаганди, що представлені менш ніж у 100 документах, однак більше ніж 20 буде застосовано SMOTE-балансування під час навчання класифікаторів [31]. До таких категорій належать: «Appeal to

Authority», «Black and White Fallacy», «Reductio ad hitlerum», «Red Herring», «Slogans», «Thought terminating Cliches» та «Whataboutism».

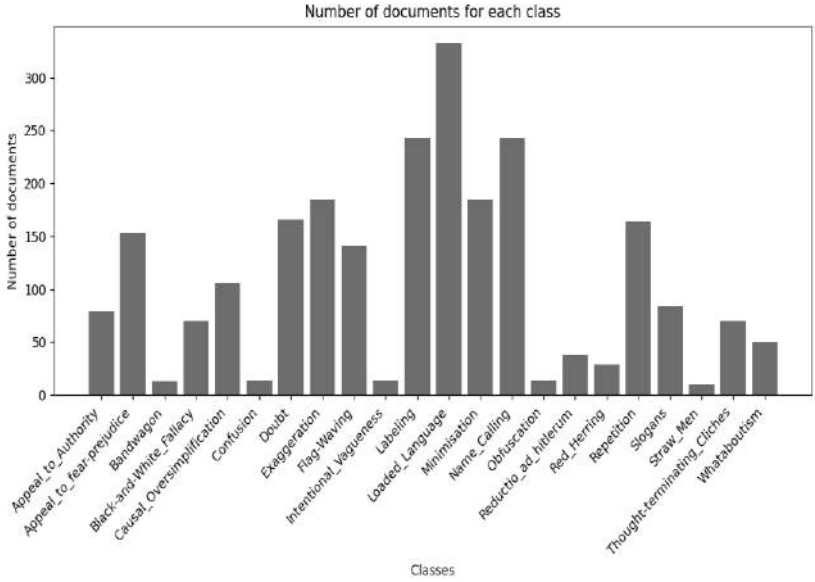

Рисунок 4. *Статистика по кількості текстів, що представляють прийоми пропаганди, шт.*

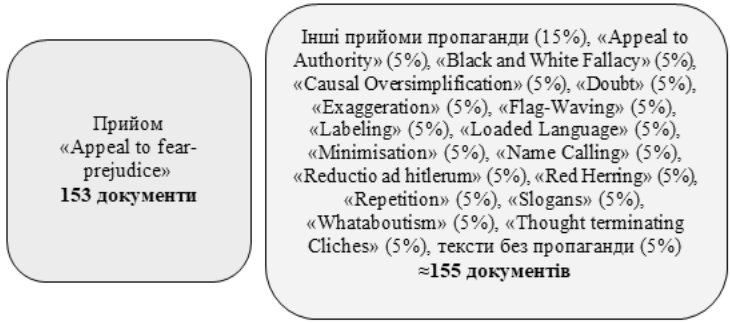

Рисунок 5. *Приклад балансування при формуванні набору даних для навчання та тестування моделі виявлення прийому «Appeal to fear-prejudice»*

Із розглянутого вище набору даних для кожної з 17 типових моделей машинного навчання буде сформовано власний дочірній набір текстів, що буде задовольняти такі вимоги:

- мати тексти з визначеним прийомом пропаганди;

- у противагу використовувати набір «Інші прийоми пропаганди» доповнений текстами без пропаганди та текстами, що представляють інші прийоми пропаганди, відмінні від цільового виду.

Приклад формування набору даних для виявлення прийому «Appeal to fear-prejudice» наведено на рисунку 5.

Отже, у дослідженні буде використано 18 класів: 17 цільових, що є репрезентативними по кількості та відповідають 17 визначеним прийомам пропаганди та 5 об'єднаних в категорію «Інші прийоми пропаганди».

5. Результати та дискусія

Результати дослідження для Traditional machine learning approaches для прийомів «Appeal to Authority», «Black and White Fallacy», «Reductio ad hitlerum», «Red Herring», «Slogans», «Thought terminating Cliches» та «Whataboutism» без використання SMOTE балансування за метрикою точності наведено в Таблиці 1.

Таблиця 1

Традиційний підхід на основі машинного навчання для виявлення прийомів пропаганди до SMOTE балансування за метрикою Accuracy

Techniques of propaganda	Regression	SVM	Random Forest	Naive Bayes
Appeal to Authority	0.57	0.63	0.55	**0.64**
Black and White Fallacy	0.55	**0.64**	0.51	0.58
Reductio ad hitlerum	**0.68**	0.56	0.61	0.59
Red Herring	**0.61**	**0.61**	0.59	**0.61**
Slogans	0.62	**0.63**	0.56	0.62
Thought terminating Cliches	0.59	0.58	**0.63**	0.58
Whataboutism	0.62	**0.65**	0.59	0.57

Як видно з таблиці 1, точність виявлення прийомів пропаганди коливається від 0.51 до 0.68, що є доволі низьким показником.

Наступним етапом для даних прийомів пропаганди було застосовано SMOTE балансування, збільшивши таким чином кількість навчальних зразків до рівня не менше 100. Результат експерименту наведено у таблиці 2.

Як видно з таблиці 2, застосування SMOTE балансування дало позитивні результати для більшості прийомів пропаганди, однак для «Slogans» результат покращення не дав.

Це пов'язано з тим, що кількість навчальних зразків близька до граничної і є достатньою для навчання запропонованих версій машинного навчання.

Таблиця 2

Традиційний підхід на основі машинного навчання для виявлення прийомів пропаганди зі SMOTE балансування за метрикою Accuracy

Techniques of propaganda	Regression	SVM	Random Forest	Naive Bayes
Appeal to Authority	0.59	**0.67**	0.54	0.60
Black and White Fallacy	0.61	0.62	0.55	**0.64**
Reductio ad hitlerum	**0.69**	0.63	0.62	0.58
Red Herring	**0.69**	0.64	0.58	0.61
Slogans	0.62	**0.63**	0.56	0.62
Thought terminating Cliches	0.62	**0.68**	0.62	0.61
Whataboutism	0.64	**0.66**	0.58	0.56

Для решти прийомів SMOTE балансування не застосовувалось, та результати застосування традиційного підходу машинного навчання наведено у таблиці 3.

Таблиця 3

Традиційний підхід на основі машинного навчання для виявлення прийомів пропаганди без SMOTE балансування за метрикою Accuracy

Techniques of propaganda	Regression	SVM	Random Forest	Naive Bayes
Appeal to fear-prejudice	**0.62**	0.61	0.55	0.59
Causal Oversimplification	0.58	**0.62**	0.59	0.55
Doubt	0.6	0.58	**0.63**	0.59
Exaggeration	0.61	**0.63**	0.61	0.53
Flag-Waving	0.62	0.61	**0.64**	0.6
Labeling	**0.67**	0.62	0.61	0.61
Loaded Language	**0.6**	0.59	**0.6**	0.58
Minimisation	**0.62**	0.61	0.61	0.60
Name Calling	0.55	0.59	0.57	**0.61**
Repetition	0.69	**0.7**	0.61	0.55

Результати з таблиці 3 також коливаються від 0.6 до 0.67, що має схожий результат із застосуванням SMOTE балансування (таблиця 2). Однак отримані результати також не є задовільними. Ілюстрація експерименту щодо застосування традиційного підходу машинного навчання наведена на рисунку 6.

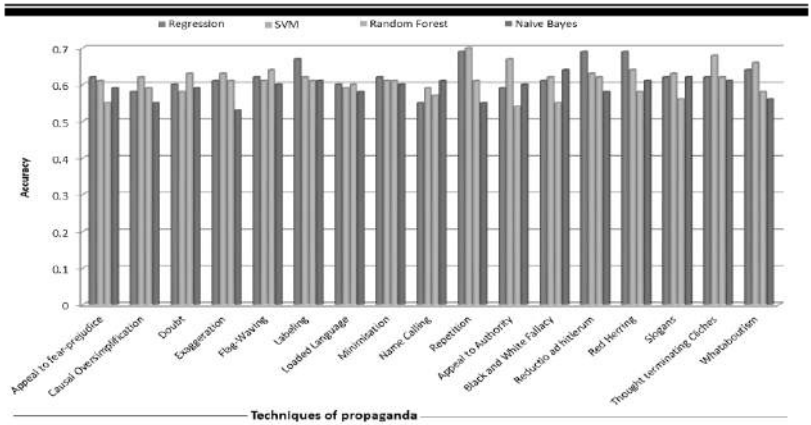

Рисунок 6. *Порівняння точності моделей для традиційного підходу машинного навчання*

Таблиця 4

Підхід на основі рекурентних нейромереж для виявлення прийомів пропаганди за метрикою Accuracy

Techniques of propaganda	RNN	LSTM	GRU
Appeal to fear-prejudice	0.69	0.69	**0.71**
Causal Oversimplification	0.73	0.71	**0.76**
Doubt	**0.75**	0.7	0.74
Exaggeration	0.64	0.72	**0.75**
Flag-Waving	0.69	0.7	**0.79**
Labeling	0.69	0.73	**0.8**
Loaded Language	**0.71**	0.7	0.68
Minimisation	**0.78**	**0.78**	0.74
Name Calling	**0.76**	0.74	**0.76**
Repetition	0.74	0.75	**0.76**
Appeal to Authority	0.71	0.72	**0.73**
Black and White Fallacy	0.7	0.68	**0.72**
Reductio ad hitlerum	**0.75**	0.68	0.71
Red Herring	0.65	**0.72**	0.70
Slogans	0.74	0.68	**0.75**
Thought terminating Cliches	0.63	**0.66**	0.65
Whataboutism	0.67	**0.69**	**0.69**

Наступним експериментом проводилось дослідження застосування An approach based on recurrent rail networks, що включало в себе порівняння використань 3-х видів архітектур: RNN, LSTM та GRU. Дані експерименту без використання SMOTE балансування наведено у таблиці 4.

Як видно з даних таблиці 4, результати для усіх прийомів пропаганди окрім Thought terminating Cliches, є вищими та знаходяться у діапазоні від 0.66 до 0.8. Однак до даного прийому пропаганди буде в подальшому застосовано SMOTE балансування, що можливо дозволить покращити показник. Наступним експериментом буде використання SMOTE балансування до навчання нейромережевих моделей для «Appeal to Authority», «Black and White Fallacy», «Reductio ad hitlerum», «Red Herring», «Slogans», «Thought terminating Cliches» та «Whataboutism». Із таблиці 5 видно, що SMOTE балансування дає позитивний ефект на точність виявлення прийомів пропаганди. Не вдалося покращити виявлення прийому «Reductio ad hitlerum», де результати до SMOTE балансування були на 0.01 вище, а також «Appeal to Authority» та «Black and White Fallacy» залишились на тому ж рівні, що і до виконання балансування. На рисунку 7 показано порівняння найвищих отриманих показників точності для рекурентних нейромережевих моделей.

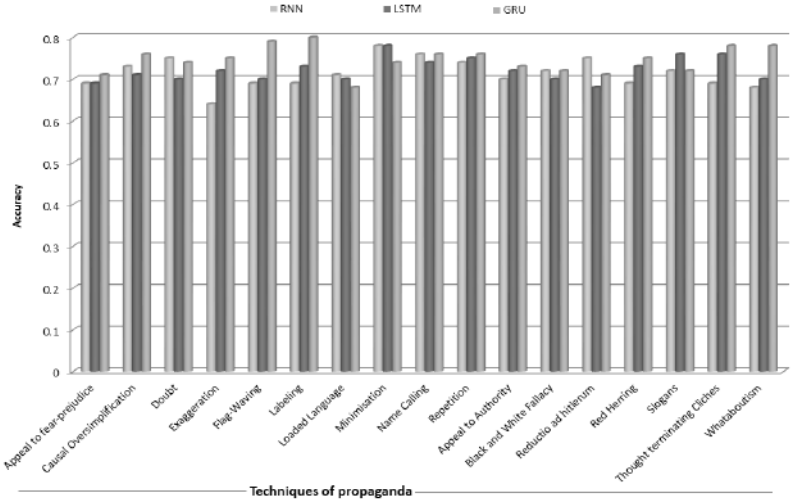

Рисунок 7. *Порівняння точності моделей для підходу на основі рекурентних нейронних мереж*

Останнім етапом дослідження є використання підходу на основі використання моделей-трансформерів, що включає в себе порівняння BERT-подібних моделей: RoBERTa, BERT, ELECTRA. Були використані попередньо натреновані моделі з ресурсу Hugging Face [32], які було донавчені вищеописаним способом протягом 3-х епох навчання. Отримані результати без використання SMOTE балансування наведені у таблиці 6.

Таблиця 5

Підхід на основі рекурентних нейромереж для виявлення прийомів пропаганди за метрикою Accuracy із SMOTE балансуванням

Techniques of propaganda	RNN	LSTM	GRU
Appeal to Authority	0.7	0.72	**0.73**
Black and White Fallacy	**0.72**	0.7	**0.72**
Reductio ad hitlerum	0.73	**0.74**	**0.74**
Red Herring	0.69	0.73	**0.75**
Slogans	0.72	**0.76**	0.72
Thought terminating Cliches	0.69	0.76	**0.78**
Whataboutism	0.68	0.7	**0.78**

Таблиця 6

Підхід на основі нейромереж-трансформерів для виявлення прийомів пропаганди за метрикою Accuracy

Techniques of propaganda	bert-base-multilingual-cased	roberta-base	ukr-electra-base
Appeal to fear-prejudice	0.81	0.8	**0.87**
Causal Oversimplification	0.78	0.79	**0.82**
Doubt	**0.93**	0.9	0.87
Exaggeration	**0.8**	**0.8**	**0.8**
Flag-Waving	**0.92**	0.9	0.89
Labeling	**0.96**	0.94	**0.96**
Loaded Language	0.93	**0.97**	0.94
Minimisation	0.89	0.86	**0.9**
Name Calling	**0.92**	**0.92**	0.91
Repetition	0.93	**0.94**	**0.94**
Appeal to Authority	0.87	**0.89**	0.88
Black and White Fallacy	0.89	**0.91**	0.88
Reductio ad hitlerum	0.85	**0.87**	0.86
Red Herring	0.67	**0.8**	0.78
Slogans	0.84	**0.86**	0.83
Thought terminating Cliches	**0.8**	0.73	0.79
Whataboutism	**0.79**	0.78	0.78

Із даних таблиці 6 видно, що BERT-подібні нейромережеві архітектури значно краще виявляють прийоми пропаганди, порівняно з рекурентними та традиційним підходами до навчання. Це пояснюється тим, що такі архітектури

є контекстно-орієнтованими, що є важливим аспектом для виявлення прийомів пропаганди. Застосування SMOTE балансування дозволило підвищити точність виявлення «Red Herring» нейромережевою моделлю ukr-electra-base до 0.89, а також «Whataboutism» до 0.83 з використанням bert-base-multilingual-cased. Порівняння найвищих оцінок за метрикою точності для 3-х розглянутих підходів наведено на рисунку 8.

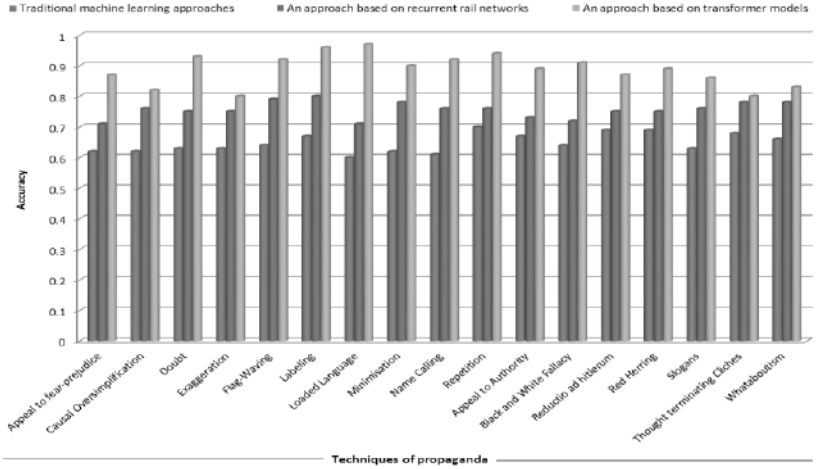

Рисунок 8. *Порівняння точності моделей альтернативних підходів для виявлення технік пропаганди*

Як видно з рисунку 8, традиційний підхід на основі машинного навчання очікувано показав гірші результати, оскільки він не вміє бачити контекст, що є важливим для виявлення технік пропаганди. Рекурентні нейромережеві моделі хоч і показали результати вищі від традиційного підходу, однак все ще мають проблеми з обробкою довгих залежностей. Найвищі результати з проведеного експерименту виявились у підході на основі моделей-трансформерів, це пояснюється використовуваними механізмами самоуваги, які дозволяє кожному елементу послідовності безпосередньо взаємодіяти з усіма іншими елементами. Це дозволяє ефективно захоплювати довготривалі залежності, що характерно для проявів технік пропаганди. Отримані результати забезпечили виявлення різних пропагандистських прийомів з мінімальною точністю 79,03% (мінімальні значення точності отримані для методики «Whataboutism»), що краще за відомі аналоги [8] щодо виявлення пропаганди незалежно від використовуваних методик. Порівняно з відомими аналогами [7] підвищилась точність виявлення різних пропагандистських прийомів:

– для техніки «Appeal to Authority», точність виявлення зросла на 9.81% (існуючий метод 77.27%, розроблений метод 87.08%);

– для техніки «Causal Oversimplification», точність виявлення зросла на 11.99% (існуючий метод 70.1%, розроблений метод 82.09%);

– для техніки «Doubt», точність виявлення зросла на 75.32% (існуючий метод 17.78%, розроблений метод 93.1%);

– для техніки «Exaggeration», точність виявлення зросла на 26.04% (існуючий метод 54.17%, розроблений метод 80.21%);

– для техніки «Flag-Waving», точність виявлення зросла на 27.65% (існуючий метод 64.52%, розроблений метод 92.17%);

– для техніки «Labeling», точність виявлення зросла на 48.57% (існуючий метод 47.43%, розроблений метод 96.0%);

– для техніки «Loaded Language», точність виявлення зросла на 42.9% (існуючий метод 54.17%, розроблений метод 97.07%);

– для техніки «Name Calling», точність виявлення зросла на 44.6% (існуючий метод 47.43%, розроблений метод 92.03%);

– для техніки «Repetition», точність виявлення зросла на 58.11% (існуючий метод 35.98%, розроблений метод 94.09%);

– для техніки «Appeal to Authority», точність виявлення зросла на 11.84% (існуючий метод 77.18%, розроблений метод 89.02%);

– для техніки «Black and White Fallacy», точність виявлення зросла на 36.68% (існуючий метод 54.55%, developed method 91.23%);

– для техніки «Reductio ad hitlerum», точність виявлення зросла на 62.31% (існуючий метод 25.0%, розроблений метод 87.31%);

– для техніки «Red Herring», точність виявлення зросла на 41.07% (існуючий метод 39.22%, розроблений метод 80.29%);

– для техніки «Slogans», точність виявлення зросла на 10.54% (існуючий метод 75.5%, розроблений метод 86.04%);

– для техніки «Thought terminating Cliches», точність виявлення зросла на 26.74% (існуючий метод 53.57%, розроблений метод 80.31%);

– для техніки «Whataboutism», точність виявлення зросла на 39.81% (існуючий метод 39.22%, розроблений метод 79.03%).

Приклад візуальної поясненності [33, 34] щодо виявлення прийому пропаганди «Repetition» наведено на рисунку 9 (використано оригінальний текст повсякденною українською мовою із збереженням орфографії та помилок). Як видно з рисунку 9, є багаторазові повторення фраз на кшталт «економічні мігранти», «мусульманський», «Орбан» того. З означення виду пропаганди «Repetition», це є «повторення того самого повідомлення знову і знову, щоб глядачі зрештою прийняти це». Отже, запропонований метод дозволяє ефективно виявляти прийоми пропаганди, та має перевагу у точності в порівнянні із запропонованими моделями що використовують підхід багатокласової класифікації.

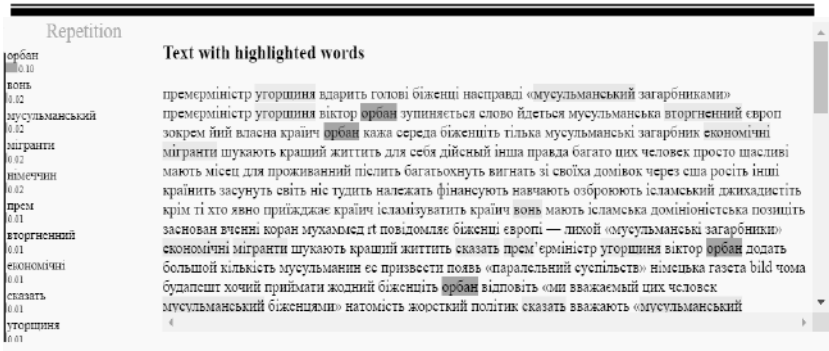

Рисунок 9. *Візуальна аналітика щодо виявлення прийому пропаганди «Repetition» розробленим програмним забезпеченням*

Експерименти, представлені в статті, проводилися з використанням різних можливостей бібліотеки SKLearn. У цій роботі представлені максимальні результати, яких вдалося досягти авторам емпіричним шляхом. Питання конфігурації та вибору гіперпараметрів є окремою проблемою, яка виходить за рамки питань, що розглядаються.

6. Висновки

Створено метод виявлення та класифікації технік пропаганди у текстовому контенті засобами штучного інтелекту, що дозволяє перетворювати вхідні дані у вигляді тексту для аналізу та навчених моделей машинного навчання у вихідні дані, які містять числові оцінки наявності кожного з прийомів пропаганди та розмічений текст з візуальною аналітикою присутності детектованих маркерів пропаганди. Було проведено дослідження, що дозволяє визначати 17 основних прийомів пропаганди, таких як: «Appeal to fear-prejudice», «Causal Oversimplification», «Doubt», «Exaggeration», «Flag-Waving», «Labeling», «Loaded Language», «Minimisation», «Name Calling», «Repetition», «Appeal to Authority», «Black and White Fallacy», «Reductio ad hitlerum», «Red Herring», «Slogans», «Thought terminating Cliches», «Whataboutism».

У рамках дослідження було порівняно 3 найчастіше застосовувані підходи: традиційний підхід машинного навчання, підхід на основі рекурентних нейронних мереж та підхід на основі трансформерних моделей. Традиційні підходи машинного навчання очікувано показали гірші результати, оскільки вони не здатні враховувати контекст, що є важливим для виявлення технік пропаганди. Досягнута точність для традиційного підходу становила від 0.60 до 0.67. Рекурентні нейромережі, хоча й перевершили традиційні підходи, все ж мають труднощі з обробкою довгих залежностей. Для цього підходу було досягнуто точності від 0.66 до 0.80. Найвищі результати були досягнуті підходом на основі трансформерних моделей, завдяки використанню механізмів самоуваги, які дозволяють кожному елементу послідовності безпосередньо взаємодіяти з усіма іншими елементами. Це забезпечує

ефективне захоплення довготривалих залежностей, що є характерним для технік пропаганди. Даний підхід дозволив виявляти техніки пропаганди з точністю до 0.96. Отримані результати забезпечили виявлення різних пропагандистських прийомів з мінімальною точністю 79,03% (мінімальні значення точності отримані для методики «Whataboutism»), що краще за відомі аналоги [8] щодо виявлення пропаганди незалежно від використовуваних методик. Порівняно з відомими аналогами [7] покращилась точність виявлення різноманітних прийомів пропаганди: за методикою «Appeal to Authority» точність виявлення зросла на 9,81%, за методикою «Причинне спрощення» — на 11,99%, за методикою «Causal Oversimplification» точність виявлення зросла на 75,32% для методики «Doubt», точність виявлення зросла на 26,04% для техніки «Exaggeration», точність виявлення зросла на 27,65% для техніки «Flag-Waving», точність виявлення зросла на 48,57% для «Labeling» методики, точність виявлення зросла на 42,9% для методики «Loaded Language», точність виявлення зросла на 44,6% для техніки «Name Calling», точність виявлення зросла на 58,11% для техніки «Repetition», точність виявлення зросла на 11,84% для техніки «Appeal to Authority», точність виявлення зросла на 36,68% для техніки «Black and White Fallacy», точність виявлення зросла на 62,31% для техніки «Reductio ad hitlerum», точність виявлення зросла на 41,07% для «Red Herring», точність виявлення зросла на 10,54% для техніки «Slogans», точність виявлення зросла на 26,74% для техніки «Thought terminating Cliches», точність виявлення зросла на 39,81% для техніки «Whataboutism».

Подальші дослідження будуть спрямовані на розширення датасету для навчання та пошуку додаткових міток у текстах, що характеризують прийоми пропаганди, таких як наявність булінгу, емоційна тональність тощо, що дозволить зробити більш поясненим рішення моделі машинного навчання та дозволить зробити більш точним виявлення прийомів.

7. Література

[1] A. Horak, R. Sabol, O. Herman, V. Baisa, Recognition of propaganda techniques in newspaper texts: Fusion of content and style analysis, Expert Systems with Applications, Volume 251, 202. doi:10.1016/j.eswa.2024.124085.

[2] A. Bhattacharjee, H. Liu, Fighting Fire with Fire: Can ChatGPT Detect AI-generated Text? SIGKDD Explor, Newsl, 25, pp. 14–21, 2023. doi:10.1145/3655103.3655106.

[3] G. Faye, B. Icard, M. Casanova, J. Chanson, F. Maine, F. Bancilhon, G. Gadek, G. Gravier, P. Egre, Exposing propaganda: an analysis of stylistic cues comparing human annotations and machine classification, in: Proceedings of the Third Workshop on Understanding Implicit and Underspecified Language, pp. 62–72, Malta, Association for Computational Linguistics.

[4] P. Vijayaraghavan, S. Vosoughi, TWEETSPIN: Fine-grained Propaganda Detection in Social Media Using Multi-View Representations, in: Proceedings of the 2022 Conference of the North American Chapter of the Association for

Computational Linguistics, pp. 3433–3448, Seattle, United States, Association for Computational Linguistics.

[5] M. Abdullah, O. Altiti, R. Obiedat, Detecting Propaganda Techniques in English News Articles using Pre-trained Transformers, in: 2022 13th International Conference on Information and Communication Systems (ICICS), Irbid, Jordan, 2022, pp. 301-308. doi: 10.1109/ICICS55353.2022.9811117.

[6] D.G. Jones, Detecting Propaganda in News Articles Using Large Language Models, Eng OA, 2(1), 2024, pp. 1-12. URL: https://www.opastpublishers.com/peer-review/detecting-propaganda-in-news-articles-using-large-language-models-6952.html.

[7] G. D. S. Martino, A. Barron-Cedeno, H. Wachsmuth, R. Petrov, P.Nakov, SemEval-2020 Task 11: Detection of Propaganda Techniques in News Articles, in: Proceedings of the Fourteenth Workshop on Semantic Evaluation, pp. 1377–1414, Barcelona (online), International Committee for Computational Linguistics.

[8] G. Martino, S. Yu, A. Barron-Cedeno, R. Petrov, P. Nakov, Fine-Grained Analysis of Propaganda in News Article, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp. 5640-5650, 2019. doi:10.18653/v1/D19-1565.

[9] Analytics Vidhya, Guide on Support Vector Machine (SVM) Algorithm, 2024. URL: https://www.analyticsvidhya.com/blog/2021/10/support-vector-machinessvm-a-complete-guide-for-beginners/.

[10] S. Mann, D. Yadav, D.Rathee, Identification of Racial Propaganda in Tweets Using Sentimental Analysis Models: A Comparative Study, in: Swaroop, A., Kansal, V., Fortino, G., Hassanien, A.E. (eds) Proceedings of Fourth Doctoral Symposium on Computational Intelligence. DoSCI 2023. Lecture Notes in Networks and Systems, vol 726, Springer, Singapore. doi: 0.1007/978-981-99-3716-5_28.

[11] L. Syed, A. Alsaeedi, L. A. Alhuri, H. R. Aljohani, Hybrid weakly supervised learning with deep learning technique for detection of fake news from cyber propaganda, Array, Volume 19, 2023. doi:10.1016/j.array.2023.100309.

[12] Krak I., Zalutska O., Molchanova M., Mazurets O., Bahrii R., Sobko O., Barmak O. Abusive Speech Detection Method for Ukrainian Language Used Recurrent Neural Network. CEUR Workshop Proceedings, 2023, vol. 3387, pp. 16-28. URL: https://ceur-ws.org/Vol-3688/paper2.pdf.

[13] Geeks for geeks, What is LSTM – Long Short Term Memory, 2024. URL: https://www.geeksforgeeks.org/deep-learning-introduction-to-long-short-term-memory/.

[14] A.M.U.D. Khanday, Q.R. Khan, S.T. Rabani, M.A. Wani, M. ELAffendi, Propaganda Identification on Twitter Platform During COVID-19 Pandemic Using LSTM, in: Abd El-Latif, A.A., Maleh, Y., Mazurczyk, W., ELAffendi, M., I. Alkanhal, M. (eds) Advances in Cybersecurity, Cybercrimes, and Smart Emerging

Technologies, CCSET 2022, Engineering Cyber-Physical Systems and Critical Infrastructures, vol 4. Springer, Cham. doi:10.1007/978-3-031-21101-0_24.

[15] Dev, Large Language Models: Comparing Gen 1 Models (GPT, BERT, T5 and More), 2024. URL: https://dev.to/admantium/large-language-models-comparing-gen-1-models-gpt-bert-t5-and-more-74h.

[16] Hugging Face, BERT, 2024. URL: https://huggingface.co/docs/transformers/model_doc/bert.

[17] 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. URL: https://ceur-ws.org/Vol-3387/paper26.pdf

[18] Hugging Face, DistilBERT, 2024. URL: https://huggingface.co/distilbert/distilbert-base-uncased.

[19] D. Chaudhari, A. V. Pawar, Empowering Propaganda Detection in Resource-Restraint Languages: A Transformer-Based Framework for Classifying Hindi News Articles, Big Data and Cognitive Computing, 2023, 7(4):175. doi:10.3390/bdcc7040175.

[20] A. Malak, D. Abujaber, A. Al-Qarqaz, R. Abbott, M. Hadzikadic, Combating propaganda texts using transfer learning, IAES International Journal of Artificial Intelligence (IJ-AI). Volume 12, pp. 956-965, 2023. doi: 10.11591/ijai.v12.i2.pp956-965

[21] A. Malak, D. Abujaber, A. Al-Qarqaz, R. Abbott, M. Hadzikadic, Combating propaganda texts using transfer learning, IAES International Journal of Artificial Intelligence (IJ-AI). Volume 12, pp. 956-965, 2023. doi: 10.11591/ijai.v12.i2.pp956-965.

[22] I. Krak, O. Barmak, O. Mazurets, 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. doi:10.15407/pp2016.02-03.237.

[23] C3.ai, What is Local Interpretable Model-Agnostic Explanations (LIME)?, 2024. URL: https://c3.ai/glossary/data-science/lime-local-interpretable-model-agnostic-explanations/.

[24] Scikit-learn, Machine Learning in Python, 2024. URL: https://scikit-learn.org/stable/.

[25] Tensorflow, An end-to-end platform for machine learning. Machine Learning in Python, 2024. URL: https://www.tensorflow.org.

[26] GitHub, Lime, 2024. URL: https://github.com/marcotcr/lime.

[27] Numpy, The fundamental package for scientific computing with Python, 2024. URL: https://numpy.org.

[28] Pandas, Pandas, 2024. URL: https://pandas.pydata.org.

[29] Kaggle, Disinformation Detection Challenge, 2024. URL: https://www.kaggle.com/competitions/disinformation-detection-challenge/data?select=emnlp_trans_uk_dataset.
[30] Propaganda, Propaganda Analysis Project, 2024. URL: https://propaganda.qcri.org/index.html.
[31] Y. Elor, H. Averbuch-Elor, To SMOTE, or not to SMOTE. URL: https://arxiv.org/abs/2201.08528.
[32] Hugging Face, The AI community building the future, 2024. URL: https://huggingface.co/.
[33] V. Slobodzian, M. Molchanova, O. Kovalchuk, O. Sobko, O. Mazurets, O. Barmak, I. Krak. 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. doi: 10.1109/ACIT54803.2022.9913162
[34] O. Kovalchuk, V. Slobodzian, O. Sobko, M. Molchanova, O. Mazurets, O. Barmak, I. Krak, N. Savina. Visual Analytics-Based Method for Sentiment Analysis of COVID-19 Ukrainian Tweets. Book Chapter. Lecture Notes on Data Engineering and Communications Technologies. 2023. Vol. 149. pp. 591–607. doi: 10.1007/978-3-031-16203-9_33

METHOD FOR DETECTING AND CLASSIFYING PROPAGANDA TECHNIQUES IN TEXTUAL CONTENT USING ARTIFICIAL INTELLIGENCE

M. Molchanova ORCID: 0000-0001-9810-936X
Khmelnytskyi National University, Ukraine
E-mail: m.o.molchanova@gmail.com

Abstract. *The paper is devoted to the creation and approbation of the method for neural network detecting propaganda techniques by markers with visual analytic, which allows converting input data in the form of text for analysis and supervised machine learning models into output data containing numerical estimates of the presence of each propaganda technique and marked-up text with visual analytical presence of detected propaganda markers. Research was conducted that allows us to detect 17 main propaganda techniques. The study compared the 3 most commonly used approaches: a traditional machine learning approach, an approach based on recurrent neural networks, and an approach based on transformer models. The highest results were achieved by the transformer model approach, which uses self-attention mechanisms that allow each element of the sequence to interact directly with all other elements. This ensures efficient capture of long-term dependencies, which is typical for propaganda techniques. This approach allowed us to detect propaganda techniques with an accuracy of 0.96.*

Keywords: *BERT, RNN, propaganda techniques, detecting propaganda, propaganda markers, visual analytics*

УДК 004.8

МЕТОД ІНТЕЛЕКТУАЛЬНОГО ВИЯВЛЕННЯ КІБЕРЗАЛЯКУВАНЬ У ТЕКСТОВОМУ КОНТЕНТІ

О. Собко ORCID: 0000-0001-5371-5788

Хмельницький національний університет, Україна
E-mail: olena.sobko.ua@gmail.com

Анотація. *Дослідження присвячене створенню та апробації методу інтелектуального виявлення кіберзалякувань в текстовому контенті. Метод здатний оцінювати за введеною текстовою інформацією рівень кіберзалякувань з використанням рекурентної нейронної мережі. Виявлення кіберзалякувань передбачає використання комбінованого підходу, що поєднує використання словника слів мови ворожнечі та нейромережевий підхід для визначення наявності кіберзалякувань. Запропонований метод дозволяє встановити наявність кіберзалякувань у тексті та визначити числову оцінку рівня прояву кіберзалякувань для подальшого моніторингу комунікаційного процесу, попередження шкідливого впливу та вжиття необхідних заходів з боку модераторів чи автоматизованої модерації контенту відповідними системами управління контентом.*

Було використано тришарову архітектуру з Embedding Layer, LSTM Layer та Dense Layer із сигмоїдною функцією активації. Навчена за розробленим методом модель RNN була в подальшому протестована й виявила показники ефективності: точність 0.96, Recall 0.959 та F1 0.957. Як метод, так і модель враховують контекст, тон висловлювань та наявність мови ворожнечі. Це свідчить про те, що результати аналізу відповідають сучасним підходам до виявлення кіберзалякування, підвищуючи їх достовірність.

Метод інтелектуального виявлення кіберзалякувань у текстовому контенті було протестовано на створеному програмному забезпеченні, й встановлено, що метод має високу ефективність виявлення кіберзалякувань у текстовому контенті. Згідно проведених прикладних досліджень, метод має оцінку точності ідентифікації кіберзалякувань понад 90%, однак оцінка наявності кіберзалякувань може бути суб'єктивною і сприйняття контенту може різнитися від особи до особи. Для покращення результату можна доповнювати словник виразів мови ворожнечі. Також метод має обмеження, працюючи з текстами довжиною від 3 до 500 слів.

Ключові слова: *кіберзалякування, інтелектуальний аналіз текстів, мова ворожнечі, нейронної мережі, класифікація кіберзалякувань.*

1. Вступ та постановка проблеми

Кіберзалякуваня, або інтернет-агресія, являє собою новітню форму насильства, що здійснюється у віртуальному просторі. Зловмисник використовує різні онлайн-комунікаційні канали, такі як соціальні мережі,

електронна пошта або месенджери, з метою психологічного тиску, нанесення шкоди або приниження особистості [1]. Це тривала і повторювана маніпулятивна дія, спрямована на окремих осіб або групи з метою створення почуття страху, гніву або приниження. Основним завданням кіберзалякування є завдання психічної або емоційної шкоди його жертвам [2].

Серед типових прикладів кіберзалякування можна навести наступні ситуації:

– поширення неправдивої інформації або публікація компрометуючих матеріалів щодо особи в соціальних мережах;

– надсилання образливих повідомлень чи погроз, спрямованих на приниження або завдання шкоди через сервіси обміну повідомленнями;

– видавання себе за іншу особу з метою введення в оману та надсилання повідомлень третім сторонам від її імені.

Значущою відмінністю кіберзалякування є наявність цифрових слідів, таких як повідомлення та публікації, які можуть слугувати доказами та сприяти припиненню агресії. Незважаючи на те, що кіберзалякування і традиційні форми залякувань можуть перетинатися, перший залишає конкретні електронні свідчення.

Законодавче регулювання кіберзалякувань є порівняно новим явищем і ще не прийняте в усіх країнах. У зв'язку з цим багато держав використовують для боротьби з кіберзалякуваннями інші правові норми, зокрема, закони про захист від домагань. На рисунку 1 подана інформація про кількість судових справ, пов'язаних з різними формами цькувань. Загалом було розглянуто 335 справ. Основна частина справ (271 або 80.9%) стосується традиційних залякувань. Інша частина (61 справа або 18.2%) пов'язана із застосуванням засобів електронних комунікацій. Лише 3 справи (0.9%) належать до категорії неназваних зловмисних дій або некласифікованих випадків залякувань [3].

Рисунок 1. *Статистика судових справ щодо цькувань*

На рисунку 2, більш детально подано типи зловживань, що включають використання цифрових технологій [3]. Найбільша кількість справ стосується публікації та поширення фото чи відеоматеріалів (16 справ). На другому місці розташовані справи, пов'язані з погрозами чи переслідуванням через електронну пошту чи месенджери (12 справ). Інші види зловживань включають створення фейкових сторінок (6 справ), злом акаунтів або блокування доступу (4 справи), поширення чуток (3 справи), шантаж (3 справи), а також погрози з боку невідомих осіб у соціальних мережах чи месенджерах (1 справа).

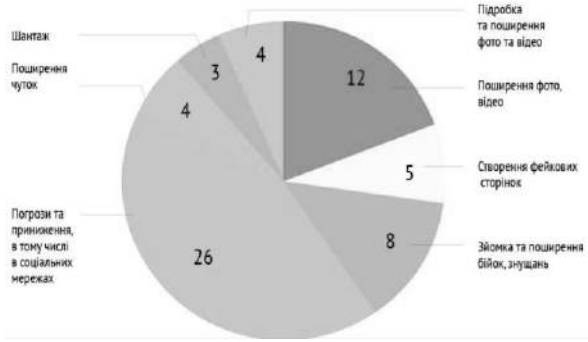

Рисунок 2. *Кількість судових справ за проявами залякувань з використанням засобів електронних комунікацій*

Останні дослідження [4] показують, що кіберзалякування часто включають мову ворожнечі (hate speech), яка є важливим компонентом цього явища. У кіберзалякуваннях агресивні та образливі висловлювання часто спрямовані на дискримінацію за ознаками раси, статі, релігії, сексуальної орієнтації тощо. Наприклад, дослідження за допомогою технології BERT та інших методів машинного навчання показують, що розпізнавання мови ворожнечі є ключовим завданням у виявленні кіберзалякувань у текстовому контенті, особливо на платформах соціальних мереж, таких як Twitter. Актуальність розробки методу інтелектуального виявлення кіберзалякувань у текстовому контенті зумовлена поширенням агресивної поведінки в онлайн-середовищі та її впливом на психологічний стан жертв [5]. Як свідчить статистика судових справ, значна частка інцидентів цькування вже переміщується у цифровий простір, де нападники використовують електронні засоби комунікації для переслідувань, погроз і приниження. Ці дані вказують на те, що майже п'ята частина випадків булінгу відбувається саме через використання таких технологій, як соціальні мережі, месенджери або електронна пошта. Незважаючи на те, що кількість справ про традиційний булінг залишається високою, випадки кіберзалякування, зокрема розповсюдження зловмисного контенту, створення фейкових акаунтів, шантаж та поширення неправдивої інформації, становлять серйозну проблему. Оскільки ця форма агресії залишає цифровий слід, розробка інтелектуальних

систем, здатних автоматично ідентифікувати кіберзалякування в текстовому контенті, є важливою для швидкого реагування і запобігання подальшій ескалації конфлікту. Такий метод дозволить ефективніше виявляти прояви агресії, мінімізувати емоційні наслідки для жертв і слугуватиме інструментом для забезпечення безпеки в онлайн-просторі.

2. Аналіз останніх досліджень та публікацій

Задача виявлення кіберзалякування у текстовому контенті залишається вкрай актуальною через зростаючу кількість шкідливих взаємодій на платформах соціальних мереж, особливо серед молоді. У дослідженнях останніх років науковці зробили значний прогрес у цій галузі, використовуючи алгоритми машинного навчання та обробки природної мови (NLP).

У статті [6] розглядається проблема кіберзалякувань, яка набула актуальності через зростання залежності суспільства від онлайн-платформ внаслідок пандемії. Автори акцентують увагу на тому, що існуючі алгоритми виявлення кіберзалякувань часто класифікують дружні жарти як образливі, оскільки використовують лише бінарну класифікацію на основі ключових слів, що містять образи. Відсутність аналізу контексту та недоступність відкритих навчальних даних ускладнює точне навчання моделей. Саме тому дослідження зосереджується на врахуванні контексту як важливого параметра для класифікації онлайн-повідомлень. Використаний набір даних був анотований на основі п’яти параметрів, що враховують контекст онлайн-розмов. У роботі застосовуються різні алгоритми машинного навчання, такі як SVM (метод опорних векторів), випадковий ліс (random forest), AdaBoost і багатошаровий персептрон (MLP), до вибірки даних про кіберзалякування, отриманих із Twitter. Найкращі результати показав алгоритм SVM, на якому було виконано рандомізований перезапис, що привело до значного підвищення середнього значення F1-міри, перевищивши базовий показник.Відомий підхід [7] до виявлення кіберзалякувань в соціальних мережах за допомогою ансамблевого навчання на основі стекування. Основна увага приділяється використанню глибоких нейронних мереж (DNNs) та модифікованої моделі BERT (BERT-M) для аналізу даних із Twitter. Зібраний набір даних було попередньо оброблено для видалення нерелевантної інформації. Основним засобом класифікації у представленому підході став стековий ансамбль моделей, який продемонстрував високі показники ефективності. Для валідації моделей були використані загальновизнані метрики оцінки, і запропонований стековий підхід досягнув F1-міри 0.964, точності 0.950, і повноти 0.92.

У статті [8] розглянуто методи виявлення кіберзалякувань. Використано шість наборів даних із Facebook, Twitter та Instagram, а також спеціально розроблений арабський лексикон кіберзалякувань. Перед класифікацією було здійснено попередню обробку тексту, зокрема очищення даних, і застосовано методи вбудовування слів для обробки природної мови. У дослідженні оцінювали різноманітні алгоритми машинного та глибинного навчання: наївний баєс, метод опорних векторів, нейронні мережі та інші. Проведено

детальний порівняльний аналіз, який показав, що гібридні моделі мають перевагу над окремими алгоритмами. Представлено гібридну глибинну модель зі стекуванням вбудовування слів, яка продемонструвала покращену здатність до вилучення ознак та точну класифікацію текстів, зокрема в різних мовних контекстах.

Незважаючи на значний прогрес у дослідженнях в області виявлення кіберзалякувань, ця тема залишається актуальною через постійно зростаючу кількість цифрового контенту та змінюваний характер онлайн-комунікацій. Нові форми комунікації, соціальні мережі та анонімність користувачів створюють сприятливі умови для кіберзалякувань, що ускладнює їх автоматичне виявлення.

3. Постановка задачі

Метою роботи є розробка методу інтелектуального виявлення кіберзалякувань у текстовому контенті, що поєднує використання словника слів, що притаманні мові ворожнечі, та нейромережевого підходу для визначення наявності кіберзалякування у текстовому контенті.

4. Метод інтелектуального виявлення кіберзалякувань у текстовому контенті

Для виявлення кіберзалякувань у текстовому контенті, а також з метою забезпечення надійності результату при визначення рівня кіберзалякувань буде використано нейронну мережу для визначення кіберзалякувань та підхід словника для перевірки наявності елементів мови ненависті.

Вхідними даними методу є: попередньо навчена модель виявлення кіберзалякування RNN, текстовий контент для аналізу, словник ключових виразів мови ненависті, яка притаманна кіберзалякуванням. Текстовий контент соціальних мереж для аналізу характеризуються деякими особливостями [9], основні з них:

– невелика довжина (часто мають обмеження на кількість символів, що спричиняє стислості тексту та вимагає ясного та лаконічного висловлення думок);

– інформальність (зазвичай, інформація подана у неформальному характері, і може містити скорочення, нестандартну лексику);

– використання мультимедіа (тексти можуть включати гіфки, емодзі тощо);

– діалогічний характер (соціальні мережі мають розгалужену структуру, що сприяє взаємодії між користувачами та створенню діалогів).

Це все робить обробку короткого контенту специфічною, відмінною від загальних методів для роботи з текстами. Узагальнена схема пропонованого методу проілюстрована на рисунку 3.

Кроком 1 є нейромережева оцінка оцінка кіберзалякуваннь в тексті, що відбувається на основі попередньо натренованої моделі RNN. Результатом виконання кроку є числова оцінка кіберзалякуваннь в текстовому контенті в

проміжку від 0 до 1, де 0 – текст не містить кіберзалякування, 1 – містить кіберзалякування.

Рисунок 3. *Кроки методу інтелектуального виявлення кіберзалякувань у текстовому контенті*

Крок 2 відбувається паралельно з першим, і полягає в оцінці частоти (TF) появи мови ворожнечі, яка буде визначатись при порівнянні вмісту текстового повідомлення із словником «Hatebase». Оцінка частоти появи мови ворожнечі буде розраховуватись за формулою:

$$TF = \frac{O_w}{TotalCount} \tag{1}$$

де O_w – кількість слів мови ворожнечі, що містяться у словнику, *TotalCount* – загальна кількість слів у досліджуваному текстовому контенті. Результатом виконання кроку є числова оцінка концентрації складових мови ворожнечі.

На кроці 3 здійснюється обрахунок оцінки кіберзалякування текстового контенту за виявленим кіберзалякуванням нейромережею та концентрацією виразів мови ворожнечі. Числова оцінка кіберзалякування текстового контенту буде розраховуватись за формулою:

$$CyberbullyingLevel = \frac{k \cdot TF + (1-k)(1-CybBullVal)}{2} \tag{2}$$

де *TF* – числова оцінка концентрації виразів мови ворожнечі, *CybBullVal* – нейромережева оцінка кіберзалякування контенту, *k* – коефіцієнт порогу

чутливості слів мови ворожнечі на загальний рівень виявлення мови ворожнечі. Підбирається згідно до політики досліджуваного соціального сервісу.

Відповідно, вихідними даними методу інтелектуального виявлення кіберзалякувань у текстовому контенті за допомогою рекурентної нейронної мережі є числова оцінка кіберзалякування текстового контенту.

5. Підготовка навчальних вибірок даних

У якості експериментальних даних для реалізації методу виявлення кіберзалякувань за допомогою рекурентної нейронної мережі буде використано «Cyberbullying Data for Multi-Label Classification», «Cyberbullying Tweets», «Cyberbullying Dataset» «A Comprehensive Dataset for Automated Cyberbullying Detection», (для навчання RNN), а також «Hatebase» (для ідентифікації мови ненависті).

Датасет «Cyberbullying Data for Multi-Label Classification» [10], доступний на Kaggle, призначений для багатоміткової класифікації текстів з соціальних мереж. Він включає 47 тисяч твітів, що належать до 6 різних категорій кіберзалякувань: стать, вік, етнічність, релігія, ознаки інвалідності та зовнішній вигляд. Цей датасет допомагає ідентифікувати різні форми кіберзалякувань на основі текстів і класифікувати їх за кількома параметрами одночасно. Дані попередньо оброблені, включаючи очистку текстів від зайвої інформації, що робить цей набір готовим для використання в моделях машинного навчання та глибинного навчання.

Датасет «Cyberbullying Tweets», доступний на Kaggle [11], містить твіттер-пости, які використовуються для дослідження та моделювання виявлення кіберзалякувань. Цей набір даних включає твіти, мічені за категоріями, пов'язаними з кіберзалякуваннями, що дозволяє проводити багатокласову класифікацію. Він призначений для тренування моделей машинного та глибинного навчання з метою ідентифікації образливих повідомлень або ворожих висловлювань.

Датасет «Cyberbullying Dataset», доступний на Kaggle [12], містить твіти, які використовуються для тренування моделей виявлення кіберзалякування. Цей датасет був створений для аналізу образливого контенту в соціальних мережах, дозволяючи проводити класифікацію текстів на основі того, чи містять вони ознаки кіберзалякувань.

Для автоматичного виявлення кіберзалякувань дослідники створюють набори даних, що включають слова та фрази, які часто використовуються для агресії в інтернеті. Наприклад, датасет від дослідників [13] містить не лише агресивні тексти, але й включає чотири аспекти кіберзалякувань, такі як повторюваність, намір заподіяння шкоди та агресивність.

«Hatebase» [14] – це платформа, яка спеціалізується на зборі та систематизації термінів, пов'язаних з мовою ненависті. Вона була створена для боротьби з дискримінацією, ненавистю та насильством, що виражаються через мову, в контексті як онлайн, так і офлайн. Платформа містить великий обсяг даних, включаючи терміни, фрази та сленг, що можуть використовуватися для

дискримінаційних висловлювань проти різних груп людей, включаючи расові, етнічні, релігійні та інші меншини. Платформа підтримує багатомовність, що дозволяє користувачам аналізувати ненависть у різних культурних контекстах. «Hatebase» надає можливість систематизувати дані за різними категоріями ненависті, такими як расова чи сексуальна, що полегшує дослідження. Інструменти аналітики на платформі допомагають розуміти, де і як використовуються ненависницькі висловлювання, що може бути корисним для розробки стратегій протидії цим явищам. Дані платформи «Hatebase» буде використаний у якості даних для числової оцінки рівня кіберзалякувань у текстовому контенті.

Складові дані для пропонованого методу наведено на рисунку 4.

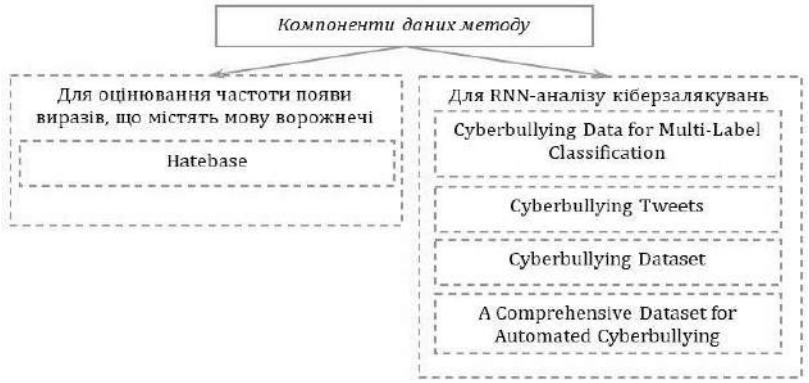

Рисунок 4. *Складові набору даних методу інтелектуального виявлення кіберзалякувань у текстовому контенті*

Вищеописаний набір даних буде використано з метою реалізації методу інтелектуального виявлення кіберзалякувань у текстовому контенті, що буде спроможний визначати за уведеною текстовою інформацією рівень наявності кіберзалякувань.

6. Формування навченої моделі RNN для виявлення кіберзалякувань у текстовому контенті

Оскільки одними з вхідних даних методу виявлення кіберзалякувань у текстовому контенті є навчена модель, виникає потреба її отримання. Схема виявлення кіберзалякування у текстовому контенті нейромережевою моделлю показана на рисунку 5.

Схема ілюструє процес виявлення кіберзалякування в текстових даних за допомогою RNN. Процес починається з визначення вхідних даних, які

складаються з векторизатора, навченої моделі, що спеціалізується на виявленні кіберзалякування, та текстового зразка, що підлягає аналізу.

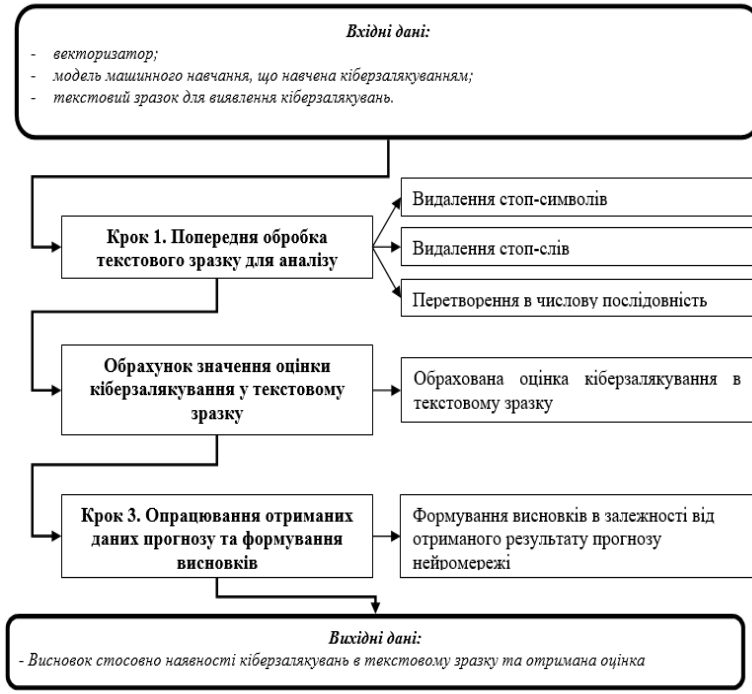

Рисунок 5. *Схема виявлення кіберзалякування у текстовому контенті нейромережевою моделлю*

Першим кроком цього процесу є попередня обробка текстового зразка для аналізу. Крок є важливим, оскільки забезпечує чистоту та правильне форматування вхідних даних, що робить їх придатними для подальшого аналізу. У процесі попередньої обробки видаляються стоп-символи, які є непотрібними символами, що не додають значення тексту, а також видаляються стоп-слова, які є часто вживаними словами, що можуть не нести суттєвої інформації, наприклад, «і», «те» чи «є». Після цього текст перетворюється на числове представлення, що необхідно для роботи алгоритмів машинного навчання. Такого роду перетворення дозволяє представити текстовий зміст у форматі, який кількісно описує інформацію та робить її підлягаючою статистичному аналізу. На наступному кроці модель оцінює значення індикаторів кіберзалякування в текстовому зразку. Це передбачає порівняння обробленого тексту з попередньо визначеними метриками, які характеризують випадки кіберзалякування. Вихідні даними моделі є оцінка кіберзалякування в тексті, який відображає ймовірність або присутність шкідливого контенту.

Останній крок зосереджується на обробці отриманих даних прогнозу та формуванні висновків. На основі результатів аналізу модель генерує висновки, які відображають ймовірність присутності кіберзалякування в текстовому зразку. Цей прогноз служить основою для подальшої інтерпретації та прийняття рішень стосовно аналізованого контенту.

В результаті всього цього процесу формується визначення наявності кіберзалякування в аналізованому тексті, а також оптимальна оцінка, що вказує на ступінь або інтенсивність виявленого кіберзалякування.

Тренування нейромережевої моделі відбувається за алгоритмом, зображеним на рисунку 6.

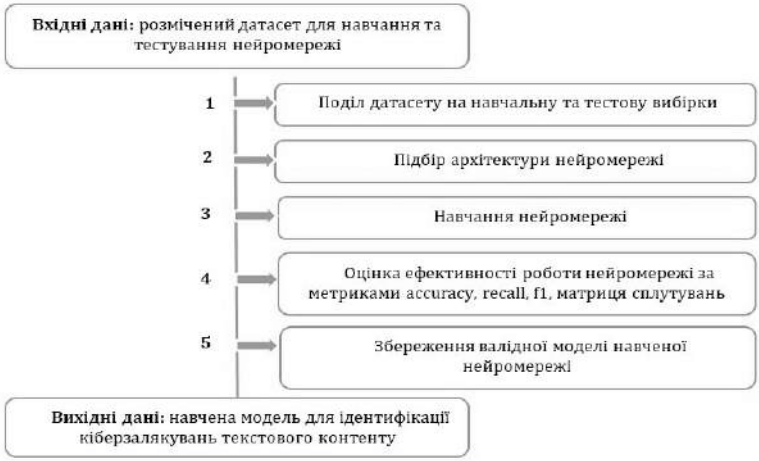

Рисунок 6. *Етапи формування навченої моделі аналізу тональності RNN для виявлення образливого мовлення*

Відповідно, вхідними даними для підбору моделі є розмічений датасет, що використовується для навчання нейромережі та оцінки його ефективності. Рівень присутності кіберзалякувань у текстовому контенті буде визначатись з числового проміжку від 0 до 1, де 0 – відсутність кіберзалякувань, 1 – присутність кіберзалякувань.

Першим етапом є поділ датасету на навчальну та тестову вибірки. Було прийнято рішення поділити датасет у пропорції 60 на 40, де 60 % це навчальна вибірка, а 40 % – тестова. Наступним етапом був підбір архітектури нейромережі. Було прийнято рішення використовувати тришарову архітектуру з Embedding Layer, LSTM Layer та Dense Layer із сигмоїдною функцією активації. Наступним етапом було навчання нейромережі з вищеописаною архітектурою. Етап навчання проводився спільно з етапом подальшого оцінювання моделі на основі таких метрик, як: accuracy, recall, f1 та матриці сплутування [15]. Accuracy визначається як відношення кількості правильно

класифікованих прикладів до загальної кількості прикладів [16]. Recall визначається як відношення кількості правильно класифікованих позитивних прикладів до загальної кількості позитивних прикладів. F1-score обчислюється як гармонічне середнє між точністю і повнотою [17].

На рисунках 7-10 наведено ілюстрацію сплутувань зразків нейромережевими моделями. Зелений колір – істино позитивні, червоний – істино негативні, жовтий – хибно позитивні і синій – хибно негативні.

Таблиця 1

Параметри навчання RNN та результати

Параметри навчання	**Моделі**						
	V1	**V2**	**V3**	**V4**	**V5**	**V6**	**V7**
Кількість епох навчання	20	20	20	20	10	10	10
Batch size	128	64	32	16	64	32	16
	Результати						
Час навчання (sec)	257	343	503	921	183	255	442
Accuracy	0.951	0.951	0.957	0.949	**0.96**	0.956	0.947
Recall	0.963	0.968	0.948	0.936	0.959	0.943	**0.978**
F_1	0.959	0.961	0.956	0.95	0.957	0.956	**0.960**
True positive	0.96	0.97	0.95	0.94	0.96	0.94	0.98
True negative	0.96	0.96	0.97	0.97	0.97	0.98	0.95
False positive	0.036	0.037	0.028	0.026	0.035	0.02	0.047
False negative	0.037	0.032	0.05	0.06	0.04	0.06	0.022

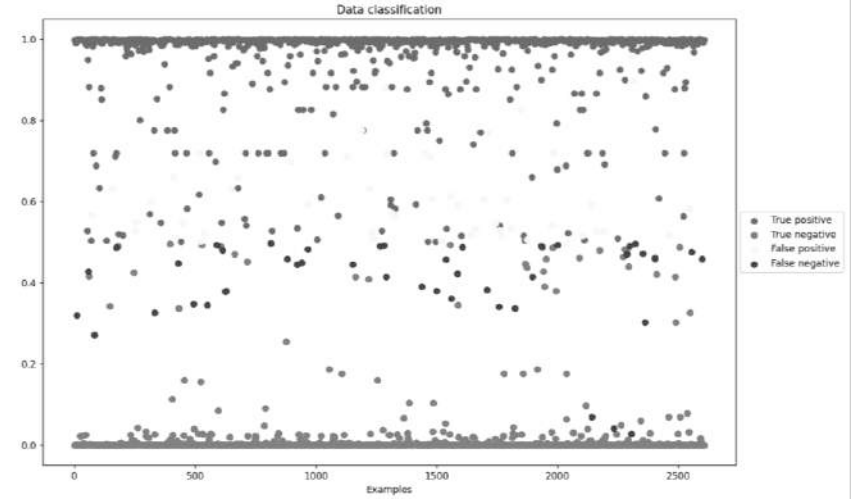

Рисунок 7. *Класифікація текстового контенту моделлю V1*

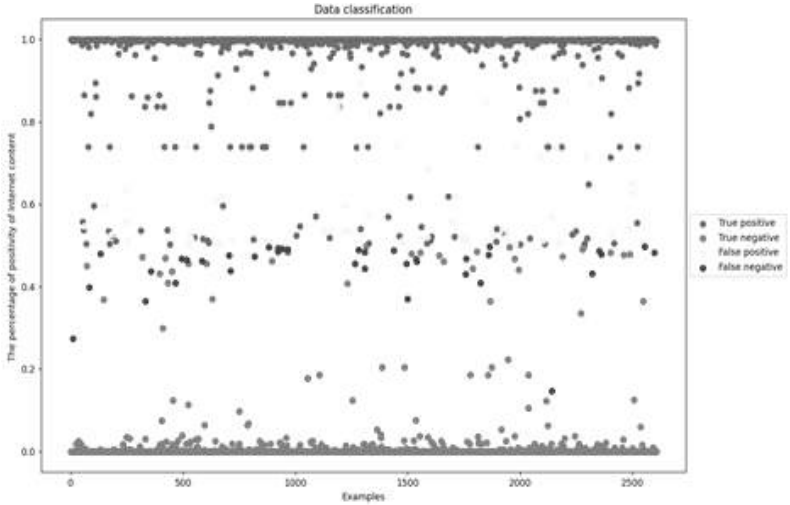

Рисунок 8. *Класифікація текстового контенту моделлю V2*

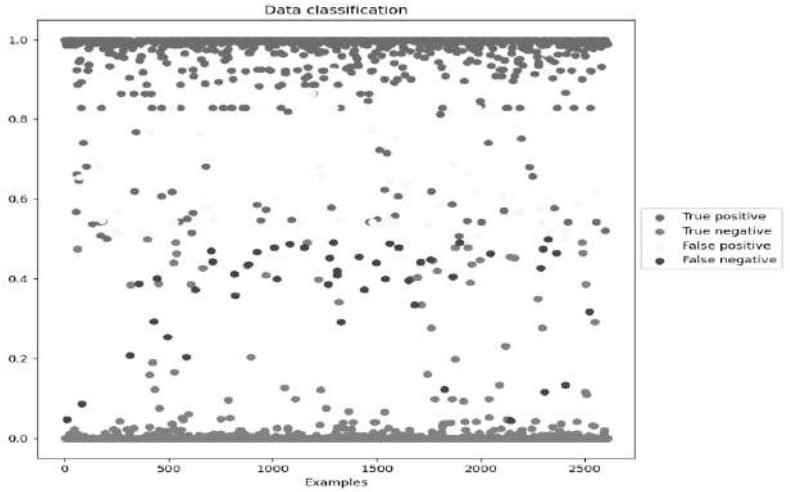

Рисунок 9. *Класифікація текстового контенту моделлю V5*

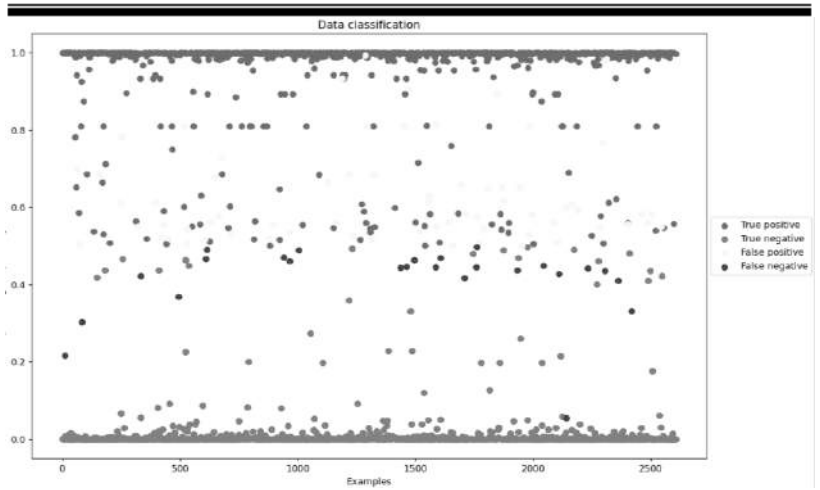

Рисунок 10. *Класифікація відгуків моделлю V7*

Для процесу навчання досліджувались такі вхідні параметри: Batch size, кількість епох. Кількість епох навчання показує, скільки разів модель підлягає навчанню. Batch size показує кількість навчальних прикладів, що використовуються в межах однієї ітерації навчання нейромережі. Дуже важко відразу визначити, який ідеальний розмір партії для потреб конкретної задачі [18], тому даний параметр буде підібрано експериментальним шляхом. Статистика метрик за проведеним навчанням наведена у таблиці 1.

Як видно з рисунків 7-10, в цілому, всі моделі справляються із поставленою задачею, але зважаючи на мету дослідження, було прийнято рішення у подальшому використовувати модель V5, яка має найвищий показник Accuracy. Хоча і модель V7 має досить високі показники Recall та F1, проте виходячи з мети, більш важливим є більш точна ідентифікація саме позитивних зразків.

7. Дослідження ефективності пропонованого методу

Для дослідження ефективності методу виявлення кіберзалякувань за допомогою рекурентної нейронної мережі було створено відповідну програмну реалізацію. Для розробки було використано засоби мови Python, для інтерфейсу користувача було використано бібліотеку «wx» [19]. Для навчання та подальшого використання нейронної мережі використовувалась бібліотека «Sklearn» [20]. Приклад ідентифікації контенту проілюстровано на рисунку 11. Для того, щоб проаналізувати текстовий контент на наявність кіберзалякувань необхідно у верхній частині вікна розташоване текстове поле з підписом «Текст для аналізу» ввести або вставити вміст, який потрібно оцінити на наявність ознак кіберзалякування.

Рисунок 11. *Приклад ідентифікації контенту як «контент з кіберзалякуванням»*

Після введення тексту треба натиснути кнопку «Розпізнавання», що активує алгоритм виявлення для обробки введеного тексту. Програма оцінить текст і надасть прогноз щодо наявності кіберзалякування, а також ймовірність, що вказує на ступінь присутності кіберзалякування у вхідному текстовому контенті. Якщо потрібно навчити модель на основі нових даних, слід натиснути кнопку «Навчити модель», що відкриє нове вікно, яке надасть можливість налаштувати параметри нейромережі. Функція вибору моделі дає змогу обрати конкретну модель для аналізу, що може бути корисним у різних сценаріях використання програми. Після виконання всіх необхідних дій результати з'являться в нижній частині інтерфейсу, де відобразяться клас прогнозу та відповідна ймовірність.

У проілюстрованих прикладах для тестування текстового контенту були взяті два текстових зразка, які слугують прикладами для виявлення кіберзалякувань. Перший зразок: «Original: *Erm do you think the nations football fans are dominated by veggie muslim trans people Your in for a shock Mr Lineker. You are misreading the public mood around this quite spectacularly. U still seem to think that the BBC and Sky somehow capture the mood they dont* / Переклад українською: *Емм, як ви думаєте, серед футбольних уболівальників домінують вегетаріанці-мусульмани-трансгендери? Ви в шоці, пане Лінекер? Ви досить вражаюче неправильно розумієте суспільні настрої навколо цього. Здається, ви все ще думаєте, що BBC і Sky якимось чином вловлюють настрій, яким вони не можуть*». Цей коментар був охарактеризований нейромережею як той, що містить кіберзалякування з оцінкою 0.8670.

Другий текстовий зразок: «Original: *In today's digital age, it's essential to foster a culture of respect and support online. Social media platforms can be powerful tools for connection, allowing us to share our thoughts, celebrate our achievements, and uplift one another.* / Переклад українською: *У сучасну епоху цифрових технологій дуже важливо розвивати культуру поваги та підтримки в Інтернеті. Платформи соціальних медіа можуть бути потужними інструментами для спілкування, дозволяючи нам ділитися своїми думками, відзначати наші досягнення та підбадьорювати один одного*». Нейромережа класифікувала текстовий зразок, як той, що не має кіберзалякування з показником 0.028.

Таким чином, ці два коментарі демонструють різні рівні кіберзалякування та мови ворожнечі, що важливо для ефективного виявлення кіберзалякувань в текстовому контенті.

8. Результати експерименту та дискусія

Було розроблено та практично реалізовано метод інтелектуального виявлення кіберзалякувань у текстовому контенті за допомогою рекурентної нейронної мережі, що містить у собі комбінований підхід: мережа RNN для визначення числової оцінки наявності кіберзалякувань у текстовому контенті та підхід числової оцінки концентрації мови ворожнечі на основі «Hatebase». Значення метрик для навчених версій нейромереж при 20-и епохах і різних розмірах батча наведено на рисунку 12. Значення метрик для навчених версій нейромереж при 10-и епохах і різних розмірах батча наведено на рисунку 13. Як видно з наведених діаграм, показники не опускаються метрик не опускаються нижче 94%, отже нейронна мережа показує на всіх моделях високі показники до визначення присутності кіберзалякувань у текстовому контенті.

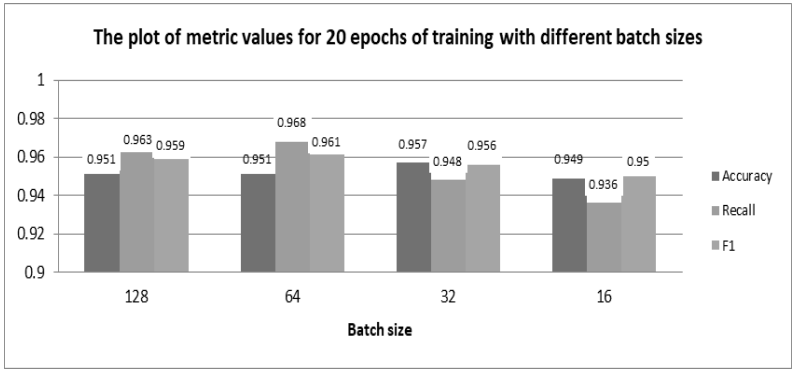

Рисунок 12. *Значення метрик для визначення кіберзалякувань у текстовому контенті RNN для 20-и епох*

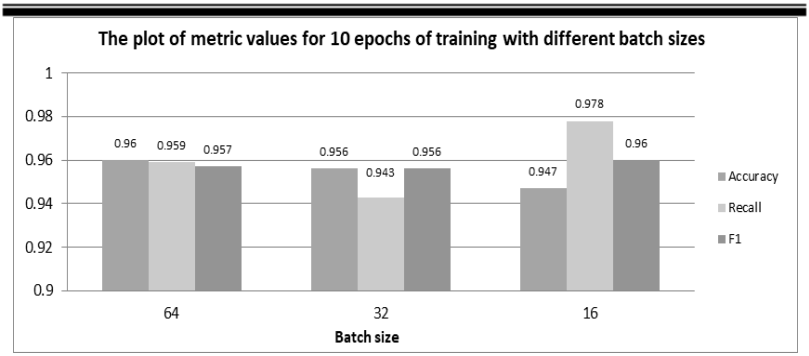

Рисунок 13. *Значення метрик для визначення кіберзалякувань у текстовому контенті RNN для 10-и епох*

У рамках проведення дослідження було зібрано 50 коментарів, які не входять у навчальні та тестові дані з соціальних мереж «Facebook» та «Instagram», які було оцінено експертами на 2 категорії: «з кіберзалякуванням» та «без кіберзалякувань». З них за оцінкою експертів, 32 коментарі мали маркер «без кіберзалякувань» та 18 «з кіберзалякуванням». З проведеного тестування розробленим методом, 30 текстових зразків є з кіберзалякуванням, а 20 – без кіберзалякування.

Однак, спірні текстові зразки, що були охарактеризовані попереднім експертом як «з кіберзалякуванням», а розробленим методом, як «без кіберзалякування» було запропоновано оцінити ще 3-м експертам, і у першому випадку 2-є з 3-х їх теж віднесли до «без кіберзалякування». Текст був наступним: «Original: *Honestly, I can't believe how clueless some people are. It's like they just don't get it. Maybe if you actually paid attention, you'd understand why everyone is frustrated with you. Do you even care about what others think? It seems like you just enjoy making things difficult for everyone around you. Just saying.* / Український переклад: *Чесно, я не можу повірити, як деякі люди можуть бути такими нездатними. Здається, вони просто не розуміють. Може, якби ти справді звертав увагу, ти б зрозумів, чому всі незадоволені тобою. Тобі взагалі цікаво, що думають інші? Схоже, ти просто отримуєш задоволення від того, що ускладнюєш життя всім навколо. Просто кажу*». Оцінка даного коментаря розробленим методом склала 0.37, при пороговому значення 0.4. Це повідомлення містить елементи критики та розчарування, але не містить явних образ або загроз, мови ворожнечі, що ускладнює його класифікацію як чітке кіберзалякування.

Порогові значення для визначення кіберзалякувань в текстовому контенті є важливими для аналізу даних, оскільки вони впливають на точність і чутливість моделей виявлення. Загальноприйнятим є те, що оптимальні порогові значення можуть варіюватися від 0.4 до 0.7, залежно від контексту та застосованих підходів до виявлення кіберзалякувань. Зокрема, оцінка 0.4 часто розглядається як мінімальний поріг для виявлення образливого контенту, що

значення дозволяє класифікувати коментарі, що містять потенційно небезпечні висловлювання, як такі, що потребують подальшого аналізу, водночас виключаючи деякі менш агресивні повідомлення. Важливо, щоб порогове значення забезпечувало достатній рівень чутливості, що дозволяє виявляти численні випадки кіберзалякувань, які в іншому випадку могли б залишитися непоміченими. Крім того, використання порогового значення 0.4 забезпечує можливість швидкого виявлення небезпечних висловлювань, що є критично важливим для оперативного реагування на ситуації кіберзалякувань. Нижчі порогові значення здатні виявляти значну частину випадків кіберзалякувань, що є важливим для модерації контенту [21].

Також результати з прикладами використання методу наведені на рисунку 14.

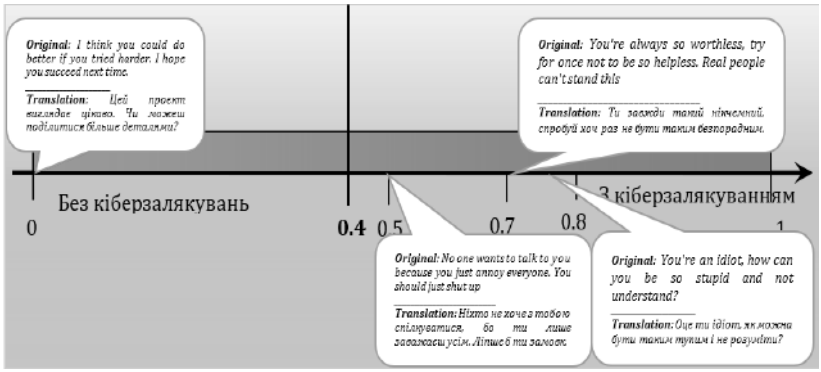

Рисунок 14. *Результат прикладного використання методу*

Для порівняння, ці ж коментарі були подані для оцінки чату GPT, де він дав оцінки, наведені на рисунку 15.

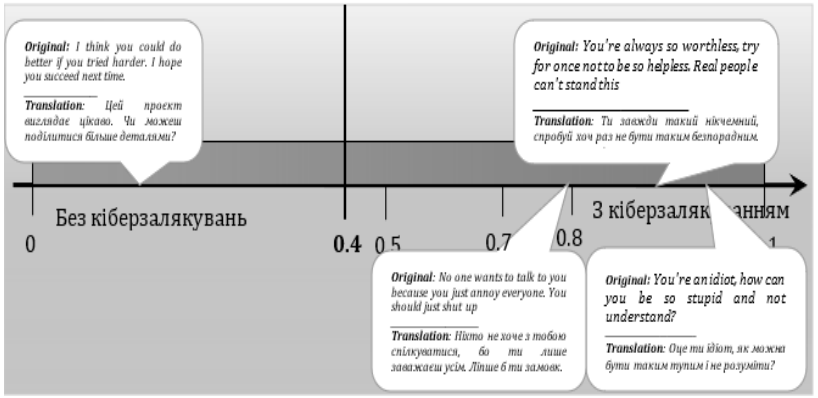

Рисунок 15. *Оцінка тексту мовною моделлю ChatGPT*

При цьому, ChatGPT визначив наступне обґрунтування своїм оцінкам:

Текст 1: *«I think you could do better if you tried harder. I hope you succeed next time.»* – оцінка 0.2. Цей коментар має позитивний підхід, пропонуючи підтримку, хоча може бути сприйнятий як критичний. Однак у ньому відсутнє пряме приниження чи образи, тому загальна оцінка кіберзалякування залишається низькою.

Текст 2: *«You're always so worthless, try for once not to be so helpless. Real people can't stand this.»* – оцінка 0.9. У цьому коментарі є явне приниження і негативне висловлювання про людину, що підвищує оцінку кіберзалякування. Слова "worthless" та "helpless" є образливими, що свідчить про ненависть.

Текст 3: *«No one wants to talk to you because you just annoy everyone. You should just shut up.»* – оцінка 0.8. Цей коментар також містить ненависть і приниження. Він не лише висловлює негативні почуття, а й закликає мовчати, що підкреслює деструктивний характер повідомлення.

Текст 4: «You're an idiot, how can you be so stupid and not understand?» – оцінка 0.95. Цей коментар містить дуже сильні образи і ненависть, безпосередньо принижуючи особу, що є явним прикладом кіберзалякування.

Оцінка наявності кіберзалякування, отримана шляхом застосування розробленого методу, відображає результати, що не суперечать оцінкам, отриманим за допомогою моделі, такої як ChatGPT. Обидва підходи використовують схожі критерії для аналізу коментарів, що свідчить про прийнятність отриманих результатів. Зокрема, як метод, так і модель враховують контекст, тон висловлювань та наявність мови ворожнечі. Це свідчить про те, що результати аналізу відповідають сучасним підходам до виявлення кіберзалякування, підвищуючи їх достовірність. Таким чином, можна стверджувати, що застосований метод надає об'єктивні та коректні оцінки, які можуть бути використані для подальших досліджень у цій галузі.

9. Висновки

Було розглянуто сучасний стан напряму виявлення кіберзалякувань у текстовому контенті, й відповідно до проведеного аналізу були виділені основні підходи до вирішення задачі виявлення кіберзалякувань, серед яких є: аналіз тональності, виявлення мови ворожнечі (підхід словника) та використання машинного навчання. Було прийнято рішення використати комбінований підхід на базі виявлення кіберзалякування та виявленні мови ворожнечі (підхід словника), яка підсилювала б оцінку кіберзалякування при його наявності. Також сформовано відповідний набір даних, що складався з датасетів «Cyberbullying Data for Multi-Label Classification», «Cyberbullying Tweets», «Cyberbullying Dataset» «A Comprehensive Dataset for Automated Cyberbullying Detection» (для навчання RNN), а також «Hatebase» (для ідентифікації мови ненависті). На твіти було накладено фільтрацію, було видалено твіти що складались менше ніж з 3-х слів. Для навчання RNN було прийнято рішення поділити датасет у пропорції 60/40, де 60 % це навчальна вибірка, а 40 % тестова. Навчена модель RNN, що була в подальшому

використана для аналізу тональності, мала точність 0.96, а Recall та F1 мали показники 0.959 та 0.957 відповідно. Метод інтелектуального виявлення кіберзалякувань у текстовому контенті було протестовано на створеному програмному забезпеченні, та з проведеного дослідження показано що метод має високу ефективність виявлення кіберзалякувань у текстовому контенті. Запропонований метод дозволяє оцінити рівень присутності кіберзалякувань у текстовому контенті для його подальшого моніторингу, попередження шкідливого впливу та вжиття необхідних заходів з боку модераторів чи автоматизованих систем управління контентом. Згідно проведених досліджень, метод має оцінку точності ідентифікації понад 90 %, однак, оцінка наявності кіберзалякувань може бути суб'єктивною, і сприйняття контенту може різнитися від особи до особи. Однак, для покращення результату необхідно доповнити словник виразів мови ворожнечі. Також пропонований метод має ряд обмежень: працює з текстовим контентом довжиною від 3 до 500 слів.

Подальші дослідження можуть бути спрямовані на прикладне застосування, що може бути корисним інструментом для оцінки рівню кіберзалякувань у текстовому контенті, який публікується в соціальних мережах, та для запобігання поширенню шкідливої чи образливої інформації.

10. Література

[1] 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.

[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] Департамент ГО «Докудейз». URL: https://cyber.bullyingstop.org.ua/storage/media-archives/cyberbuling_виправл15-10_compressed.pdf

[4] Nouas S., Boumahdi F., Madani A., Berrhail F. Cyberbullying: A BERT Bi-LSTM Solution for Hate Speech Detection. Proceedings of the International Conference on Applied Cybersecurity (ACS) 2023. ACS 2023. Lecture Notes in Networks and Systems, vol 760. Springer, Cham.

[5] 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.

[6] Dhingra N., Chawla S., Saini, O. An Improved Detection of Cyberbullying on Social Media Using Randomized Sampling. Int Journal of Bullying Prevention (2023).

[7] Muneer A., Alwadain, A., Ragab M.G. Alqushaibi A. Cyberbullying Detection on Social Media Using Stacking Ensemble Learning and Enhanced BERT. Information 2023, 14, 467.
[8] Daraghmi E.-Y., Qadan S., Daraghmi Y.-A., Yousuf R., Cheikhrouhou O., Baz M. From Text to Insight: An Integrated CNN-BiLSTM-GRU Model for Arabic Cyberbullying Detection. IEEE Access. 2024. vol. 12.pp. 103504-103519.
[9] 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.
[10] CyberBullying Detection Dataset. URL: https://www.kaggle.com/datasets/ sayankr007/cyber-bullying-data-for-multi-label-classification
[11] Cyberbullying Tweets. URL: https://www.kaggle.com/datasets/ soorajtomar/ cyberbullying-tweets
[12] Cyberbullying Dataset. URL: https://www.kaggle.com/datasets/saurabhshahane/ cyberbullying-dataset
[13] A Comprehensive Dataset for Automated Cyberbullying Detection. URL: https://data.mendeley.com/datasets/wmx9jj2htd/2dataset .
[14] Hatebase. URL: https://hatebase.org .
[15] Y. Krak Y., O. Barmak O., 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.
[16] Hartmann M., Heitmann Ch., Siebert Ch., Schamp, More than a Feeling: Accuracy and Application of Sentiment Analysis, International Journal of Research in Marketing, Volume 40, Issue 1, 2023, pp. 75-87.
[17] Scikit-learn, sklearn.metrics.f1_score. URL: https://scikit-learn.org/ stable/modules/generated/sklearn.metrics.f1_score.html
[18] Medium, How does Batch Size impact your model learning. URL: https://medium.com/geekculture/how-does-batch-size-impact-your-model-learning-2dd34d9fb1fa.
[19] wxPython. URL: https://wxpython.org/index.html.
[20] Scikit-learn. Machine Learning in Python. URL: https://scikit-learn.org/stable/.
[21] 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.

METHOD FOR INTELLIGENT DETECTION OF CYBERBULLYING IN THE TEXT CONTENT

O. Sobko ORCID: 0000-0001-5371-5788
Khmelnytskyi National University, Ukraine
E-mail: olenasobko.ua@gmail.com

Abstract. *Research is devoted to creation and testing of method for intelligent detection of the cyberbullying in text content. The method is able to estimate the level of cyberbullying based on the entered text information using a recurrent neural network. The detection of cyberbullying involves the use of combined approach, that combines the use of hate speech words dictionary and neural network approach to determine the cyberbullying presence. The proposed method allows establishing the presence of cyberbullying in the text and determining numerical assessment of the level of cyberbullying, for further monitoring of communication process, prevention of harmful effects and taking necessary measures by moderators or automated content moderation by relevant content management systems.*

Three-layer architecture with Embedding Layer, LSTM Layer and Dense Layer was used. The RNN model, trained according to the developed method, was further tested and found performance indicators: accuracy 0.96, recall 0.959 and F1 0.957. Both the method and the model take into account the context, the tone of statements and the presence of hate speech. This indicates that the results of the analysis correspond to modern approaches to the detection of cyberbullying, increasing their credibility.

The method for intelligent detection of the cyberbullying in text content was tested on the created software, and it was established that the method has high efficiency of detecting cyberbullying in text content. According to applied studies, the method has an accuracy rating of over 90% for identifying cyberbullying, however, the assessment of the presence of cyberbullying can be subjective and the perception of content can vary from person to person. Method has limitations: it working with texts from 3 to 500 words long.

Keywords: *cyberbullying, text mining, hate speech, neural network, classification of cyberbullying.*

Наукове видання

РОЗВИТКИ ІНФОРМАЦІЙНО-КЕРУЮЧИХ СИСТЕМ ТА ТЕХНОЛОГІЙ

Монографія

Під наук. ред. проф. В. Вичужаніна

Укр. та англ. мовами

Підписано до друку 27.09.2024. Формат 60×84/16.
Папір офсетний. Гарнітура Times New Roman. Цифровий друк.
Умовно-друк. арк. 22,2. Тираж 300. Замовлення № 1024-73.
Ціна договірна. Віддруковано з готового оригінал-макета.

Українсько-польське наукове видавництво «Liha-Pres»
79000, м. Львів, вул. Технічна, 1
87-100, м. Торунь, вул. Лубіцка, 44
Телефон: +38 (050) 658 08 23
E-mail: editor@liha-pres.eu
Свідоцтво суб'єкта видавничої справи
ДК № 6423 від 04.10.2018 р.


1256
1233
1996
LIHA-PRES
9 789663 974224