МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
Західноукраїнський національний університет
Навчально-науковий інститут новітніх освітніх технологій
Кафедра кібербезпеки

УНІЧЕНКО Світлана Максимівна

Метод ідентифікації Інтернет користувачів на основі стилістичних характеристик повідомлень / Method for Identifying Internet Users Based on Stylistic Characteristics of Messages

спеціальність: 125 – Кібербезпека та захист інформації

освітньо-професійна програма –Кібербезпека

Кваліфікаційна робота

Виконала студентка групи КБзм -21
С.М. Уніченко

Науковий керівник
д.т.н., професор М.М.Касянчук

Кваліфікаційну роботу допущено
до захисту:

«____» _____ ___2024 р.

Завідувач кафедри

_____ В.В.Яцків

ТЕРНОПІЛЬ – 2024

Навчально-науковий інститут новітніх освітніх технологій
Кафедра кібербезпеки
Освітній ступінь «магістр»
спеціальність: 125 – Кібербезпека та захист інформації
освітньо-професійна програма – Кібербезпека

ЗАТВЕРДЖУЮ
Завідувач кафедри
_____ В.В.Яцків
«____» _____ 2023 року

ЗАВДАННЯ
НА КВАЛІФІКАЦІЙНУ РОБОТУ СТУДЕНТУ
УНІЧЕНКО СВІТЛАНА МАКСИМІВНА

1. Тема кваліфікаційної роботи:
Метод ідентифікації Інтернет користувачів на основі стилістичних характеристик повідомлень / Method for Identifying Internet Users Based on Stylistic Characteristics of Messages

керівник роботи д.т.н., професор М.М. Касянчук
затверджені наказом по університету від «12» грудня 2023 року №753.
2. Строк подання студентом закінченої кваліфікаційної роботи 12 грудня 2024 р.

3. Вихідні дані до кваліфікаційної роботи: завдання на кваліфікаційну роботу студента, наукові статті, технічна література.

4. Основні питання, які потрібно розробити:
– проаналізувати сучасний стан проблеми ідентифікації Інтернет-користувачів при інформаційному обміні електронними повідомленнями;
– визначити основні особливості задачі ідентифікації на основі стилістичних характеристик текстів електронних повідомлень;
– розробити метод формування динамічного стилістичного профілю користувача;
– розробити методику ідентифікації Інтернет-користувача на основі стилістичних і лінгвістичних характеристик коротких електронних повідомлень;
– розробити метод порівняння ДСПК потенційних користувачів з еталонними на основі методу випадкового лісу;
– розробити рекомендації для використання результатів дослідження для підвищення безпеки інформаційних процесів в мережі Інтернет..

5. Перелік графічного матеріалу у роботі:
– дерево рішень із довільною кількістю нащадків;
– блок-схема дій першого кроку першого етапу методики ідентифікації;

– блок-схема послідовності кроків розробленого методу порівняння ДСПК з еталонними ДСПК потенційних користувачів;

– набори даних з різним рівнем незбалансованості;

– точність ідентифікації при нормальній кількості текстів навчальної вибірки;

– точність ідентифікації з використанням розробленої КБМПК та різних сучасних методів класифікації;

– порівняння точності ідентифікації за розробленою методикою та відомими методами.

6. Консультанти розділів кваліфікаційної роботи

	Прізвище, ініціали та посада консультанта	Підпис, дата	
		завдання видав	завдання прийняв

7. Дата видачі завдання 12 грудня 2023 р.

КАЛЕНДАРНИЙ ПЛАН

№ з/п	Назва етапів кваліфікаційної роботи	Строки виконання етапів кваліфікаційної роботи	Примітка
1	Аналіз стану проблеми ідентифікації інтернет-користувачів при інформаційному обміні електронними повідомленнями	12.2023 р. – 03.2024 р.	
2	Розробка методики ідентифікації інтернет-користувача на основі стилістичних та лінгвістичних характеристик коротких електронних повідомлень	03.2024 р. – 06.2024 р.	
3	Експериментальні дослідження та оцінка результатів	06.2024 р. – 11.2024 р.	

Студент _____ Уніченко С.М.
(підпис)

Керівник роботи _____ д.т.н., професор М.М.Касянчук

АНОТАЦІЯ

Випускна кваліфікаційна робота на тему „Метод ідентифікації Інтернет користувачів на основі стилістичних характеристик повідомлень” на здобуття освітнього ступеня «Магістр» зі спеціальності 125 „Кібербезпека та захист інформації” освітньо-професійної програми «Кібербезпека» написана обсягом 85 сторінок і містить 26 ілюстрації, 13 таблиць, 1 додаток та 40 джерел за переліком посилань.

Метою випускної кваліфікаційної роботи є розробка методів ідентифікації Інтернет користувачів шляхом використання лінгвістичних і стилістичних характеристик їхніх текстів.

Методи дослідження. Методи моделювання, методи ідентифікації, методи дослідження точності.

Результати дослідження. Здійснено аналіз сучасного стану проблеми ідентифікації Інтернет-користувачів при інформаційному обміні електронними повідомленнями, що дозволило обгрунтувати необхідність розробки методів ідентифікації Інтернет-користувачів шляхом використання лінгвістичних характеристик їхніх текстів. Розроблено модель загроз безпеки інформаційних процесів при обміні електронними повідомленнями, що дозволило встановити типи моделей ймовірних порушників інформаційної безпеки. Розроблено методику ідентифікації Інтернет-користувача на основі стилістичних характеристик коротких електронних повідомлень, що дозволило встановити основні етапи методики ідентифікації Інтернет-користувача. Розроблено методи визначення та підвищення точності ідентифікації з використанням різних методів порівняння динамічного стилістичного профілю користувача.

Результати роботи можуть успішно застосовуватися для ідентифікації Інтернет користувачів на основі стилістичних характеристик повідомлень.

КЛЮЧОВІ СЛОВА: ІДЕНТИФІКАЦІЯ, ІНТЕРНЕТ КОРИСТУВАЧ, СТИЛІСТИЧНА ХАРАКТЕРИСТИКА, КОРОТКЕ ПОВІДОРМЛЕННЯ, ЗАХИСТ ІНФОРМАЦІЇ.

ABSTRACT

The graduate work on the topic „Method for Identifying Internet Users Based on Stylistic Characteristics of Messages” for Master’s degree on speciality 125 "Cybersecurity and information protection " is written on 85 pages and contains 26 illustrations, 13 tables, 1 suplement and 40 references.

The aim of graduate work is to develop a methods for identifying Internet users by using the linguistic and stylistic characteristics of their texts.

Research methods. Modeling methods, identification methods, accuracy research methods.

Results of the study. An analysis of the current state of the problem of identifying Internet users in the information exchange of electronic messages was carried out, which allowed to substantiate the need to develop methods for identifying Internet users by using the linguistic characteristics of their texts. A model of security threats to information processes in the exchange of electronic messages was developed, which allowed to establish the types of models of probable information security violators. A method for identifying an Internet user based on the stylistic characteristics of short electronic messages was developed, which allowed to establish the main stages of the method for identifying an Internet user. Methods for determining and increasing the accuracy of identification using various methods for comparing the dynamic stylistic profile of the user were developed.

The results of the work can be successfully applied to identify Internet users based on the stylistic characteristics of messages.

Keywords: IDENTIFICATION, INTERNET USER, STYLISTIC CHARACTERISTICS, SHORT MESSAGE, INFORMATION PROTECTION.

ЗМІСТ

ВСТУП

Сучасні дослідження підтверджують зростання кількості кіберзлочинів, які здійснюються за допомогою Інтернету [1-4]. До таких злочинів належать переслідування, шантаж, анонімні погрози, корпоративне шпигунство, тероризм та екстремізм [5-6]. Ці дії часто виконуються через форуми, соціальні мережі, електронну пошту та месенджери, де злочинці можуть діяти приховано завдяки анонімності, яку надає Інтернет [7-8]. Це підвищує рівень злочинної активності в кіберпросторі та вимагає нових підходів до ідентифікації користувачів. Інтернет є унікальною системою з погляду можливості анонімізації джерела інформації - користувача, що розміщує електронне повідомлення у відкритий доступ. Під Інтернет-користувачем розуміється конкретна фізична особа, яка своїми діями з ресурсами порталу виявляє деякі ознаки (характеристики користувача) [9]. Електронне повідомлення - інформація, передана або отримана користувачем інформаційно-телекомунікаційної мережі

Отже, поширення Інтернету та засобів масової комунікації створює ситуацію, коли кількість електронної текстової інформації зростає, а разом із цим збільшується потреба в методах ідентифікації користувачів [10-11]. Існуючі методи ідентифікації, такі як ідентифікація за технічними характеристиками пристроїв або поведінковими ознаками користувачів на веб-ресурсах, не завжди дозволяють визначити особистість конкретної людини [12]. Це обмежує їхню ефективність, особливо у випадках, коли користувачі залишають короткі повідомлення або діють під вигаданими іменами.

Система розмежування доступу до сервісів публікації електронних повідомлень, що включає додаткову приховану ідентифікацію, в загальному випадку може бути представлена наступним чином. Користувач проходить основну процедуру ідентифікації/автентифікації, передбачену на Інтернет-ресурсі. Далі до розміщення повідомлення у відкритих джерелах проводиться процедура прихованої ідентифікації. Якщо дані процедури успішно пройдені і користувач раніше не був помічений у скоєнні протиправних дій, йому надається

доступ до розміщення повідомлення, інакше може бути зроблено відмітку про спробу порушення прав доступу. Таким чином, прихована ідентифікація є додатковим засобом захисту для забезпечення інформаційної безпеки інформаційних процесів, що протікають в Інтернет [13-15].

Одним із перспективних напрямів розвитку технологій є використання біометричної ідентифікації, зокрема методів, що базуються на аналізі лінгвістичних та стилістичних характеристик тексту [16-18]. Цей підхід дозволяє ідентифікувати користувача за унікальними рисами його письма, що є своєрідним «відбитком» автора.

Мета роботи. Метою даної роботи є розробка методів ідентифікації Інтернет користувачів шляхом використання лінгвістичних і стилістичних характеристик їхніх текстів.

Для вирішення поставленої мети вирішуються наступні **завдання**:

– проаналізувати сучасний стан проблеми ідентифікації Інтернет-користувачів при інформаційному обміні електронними повідомленнями;

– визначити основні особливості задачі ідентифікації на основі стилістичних характеристик текстів електронних повідомлень;

– розробити метод формування динамічного стилістичного профілю користувача;

– розробити методику ідентифікації Інтернет-користувача на основі стилістичних і лінгвістичних характеристик коротких електронних повідомлень;

– розробити метод порівняння ДСПК потенційних користувачів з еталонними на основі методу випадкового лісу;

– розробити рекомендації для використання результатів дослідження для підвищення безпеки інформаційних процесів в мережі Інтернет.

Об’єкт дослідження. Процес ідентифікації Інтернет користувачів на лінгвістичних і стилістичних характеристик їхніх текстів.

Предмет дослідження. Методи і засоби ідентифікації Інтернет користувачів на лінгвістичних і стилістичних характеристик їхніх текстів.

Методи дослідження. Методи моделювання, методи ідентифікації, методи дослідження точності.

Наукова новизна одержаних результатів.

1. Здійснено аналіз сучасного стану проблеми ідентифікації Інтернет-користувачів при інформаційному обміні електронними повідомленнями, що дозволило обгрунтувати необхідність розробки методів ідентифікації Інтернет-користувачів шляхом використання лінгвістичних і стилістичних характеристик їхніх текстів.

2. Розроблено модель загроз безпеки інформаційних процесів при обміні електронними повідомленнями з використанням інтернет-ресурсів, що дозволило встановити типи моделей ймовірних порушників інформаційної безпеки.

3. Розроблено методику ідентифікації Інтернет-користувача на основі стилістичних і лінгвістичних характеристик коротких електронних повідомлень, що дозволило встановити основні етапи методики ідентифікації Інтернет-користувача.

Практичне значення отриманих результатів. Розроблено методи визначення та підвищення точності ідентифікації з використанням різних методів порівняння динамічного стилістичного профілю користувача при різній кількості текстів і різному рівні незбалансованості навчальної вибірки.

Публікації та апробація КР.

1. Уніченко С., Лисик Г., Закревська М. Метод випадкового лісу для ідентифікації користувача на основі стилістичних характеристик електронних текстів. Збірник матеріалів науково - практичної конференції молодих вчених, аспірантів та студентів «Кібербезпека та комп’ютерно-інтегровані технології» (КБКІТ-2024), Тернопіль, 2024. С.158-160 [19].

2. Уніченко С., Зарихта О. Метод лінгвістичної ідентифікації користувачів на основі стилістичних характеристик текстів електронних повідомлень. Матеріали науково-практичного симпозіуму «Захист інформації». Тернопіль, 2024. С.111-113 [20].

1 АНАЛІЗ СТАНУ ПРОБЛЕМИ ІДЕНТИФІКАЦІЇ ІНТЕРНЕТ-КОРИСТУВАЧІВ ПРИ ІНФОРМАЦІЙНОМУ ОБМІНІ ЕЛЕКТРОННИМИ ПОВІДОМЛЕННЯМИ

1.1 Сучасний стан проблеми ідентифікації Інтернет-користувачів при інформаційному обміні електронними повідомленнями

Ідентифікація – це дії з присвоєння суб’єкту або об’єкту доступу ідентифікатора та (або) порівняння наданого ідентифікатора зі списком існуючих. У контексті Інтернету, ідентифікація користувача полягає у розпізнаванні його під час обміну короткими електронними повідомленнями, шляхом порівняння певних характеристик, які формують унікальний ідентифікатор, для визначення, чи є цей об’єкт тим, що шукають.

Аутентифікація користувача – це перевірка належності ідентифікатора суб’єкту доступу, підтвердження його дійсності.

Швидкий розвиток засобів інфокомунікацій призвів до глобалізації процесів інформаційного обміну, ліквідувавши межі між країнами. Масове поширення Інтернету та засобів масової комунікації (блоги, форуми, соціальні мережі, сервіси миттєвих повідомлень) сприяє стрімкому зростанню обсягу текстової інформації та залежності суспільства від неї.

Інформація може існувати у двох формах: повідомлення та відомості. Відомості відображають дії об'єктів реального світу через психічну діяльність людини, а повідомлення передають відомості від однієї особи до іншої у вигляді знакової системи. Поширення засобів масової комунікації спричиняє те, що електронні повідомлення можуть досягати значно більшої кількості адресатів без перевірки достовірності.

Існують дві ключові особливості поширення інформації в Інтернеті:

1) висока швидкість поширення;

2) можливість анонімності джерела інформації.

Завдання ідентифікації Інтернет-користувача стає дедалі актуальнішим через можливість анонімності в Інтернеті, що дозволяє публікувати

повідомлення під вигаданими іменами або навіть анонімно. Це створює ризики, пов'язані з поширенням неправдивої або маніпулятивної інформації, а також зловживанням обліковими записами легальних користувачів.

Анонімний користувач може створювати та поширювати інформацію без ідентифікації та аутентифікації, що сприяє дезінформації та маніпуляціям. Також можлива ситуація, коли зловмисник отримує доступ до облікового запису легального користувача і діє від його імені. Це робить ідентифікацію користувачів важливою при підготовці до злочинів або маніпуляцій у реальному світі.

Існуючі методи ідентифікації та аутентифікації користувачів Інтернету мають певні обмеження та не забезпечують високу точність. Збільшення точності ідентифікації стає важливим науковим завданням.

Інтернет-користувач – це фізична особа, яка виявляє певні ознаки під час взаємодії з ресурсами Інтернету, що можуть визначатися технічними засобами, які він використовує, або лінгвістичними та стилістичними особливостями його письмової мови.

1.2 Модель загроз безпеки інформаційних процесів при обміні електронними повідомленнями з використанням інтернет-ресурсів

Інформаційна безпека визначається як стан захищеності інформаційного середовища, яке забезпечує його розвиток та використання в інтересах громадян, організацій, держави.

Об'єкти захисту - це інформаційні процеси обробки та поширення інформації в Інтернеті, а також сама інформація, яку необхідно захистити від модифікації, щоб уникнути матеріальної або моральної шкоди для організацій або фізичних осіб.

Об'єкти доступу - це сервіси публікації електронних повідомлень або веб-портали, що знаходяться у відкритому доступі. До публікації інформації користувач має пройти основну та додаткову процедури ідентифікації та

аутентифікації. Проте зловмисники часто отримують доступ до облікових записів користувачів, що вимагає додаткових перевірок для запобігання несанкціонованого розміщення інформації.

Цілі загроз - отримання можливості анонімного розповсюдження інформації під вигаданим іменем або від імені іншої особи.

Потенційні загрози:

1) підробка ідентифікатора (маскарад);

2) анонімний доступ;

3) створення численних ідентифікаторів.

Ці загрози порушують цілісність та автентичність інформації, адже кожне електронне повідомлення складається з реквізитної (дані про автора, час та місце) та змістової частин. Якщо змінено хоча б одну з частин, джерело інформації вважається новим, а питання про автентичність даних стає невизначеним. Порушення цілісності інформації впливає на її достовірність.

Одним з видів спотворення інформації є публікація повідомлення від імені іншої особи. Це може призвести до маніпуляцій громадською думкою, економічних втрат і репутаційних збитків. Також поширені «інформаційні вкиди», коли в мережу масово завантажується неправдива інформація, що викликає масову реакцію.

Загрози можуть бути використані для пропаганди тероризму та екстремізму, завдаючи шкоди як інформаційній, так і національній безпеці.

Потенційний збиток від загроз:

- моральний або матеріальний збиток репутації організації через публікацію спотвореної інформації;
- збиток фізичній особі через публікацію дискредитуючої інформації від її імені;
- моральний збиток отримувачів інформації через поширення завідомо неправдивих даних;
- матеріальний збиток від розголошення конфіденційної інформації анонімно або під вигаданим ім'ям;

- збиток національній безпеці та інформаційній безпеці держави.

Джерела загроз - це фізичні особи, які безпосередньо створюють загрози, модифікуючи або поширюючи інформацію, зловмисно або через недбалість.

1.3 Модель ймовірного порушника інформаційної безпеки

Усі порушники в контексті Інтернет-ресурсу вважаються внутрішніми, оскільки вони мають певні права доступу. Можливості порушника залежать від прав допуску до розповсюдження електронних повідомлень та заходів, спрямованих на запобігання несанкціонованому доступу. Однак на багатьох ресурсах можливий анонімний доступ або доступ після простої парольної ідентифікації, яка часто недостатня.

До внутрішніх порушників можуть належати:

- користувачі з санкціонованим доступом (категорія I);
- анонімні користувачі (категорія II);
- користувачі, що маскуються під легальних (категорія III);
- користувачі, які створюють безліч ідентифікаторів (категорія IV).

Користувачі всіх категорій мають доступ до засобів Інтернет-ресурсу для поширення повідомлень та можуть бути обізнані з процедурами ідентифікації, аутентифікації та системами розмежування доступу. Вони потенційно здатні реалізовувати загрози інформаційної безпеки, використовуючи вразливості систем ідентифікації та доступу. Ймовірні порушники з категорій III і IV є основними загрозами.

Інформація, якою може володіти порушник:

- дані про інформаційні ресурси та правила їх створення, зберігання, передачі;
- відомості про вразливості систем, включаючи недекларовані можливості;
- дані про процедури ідентифікації та аутентифікації;
- паролі та інші дані, необхідні для реалізації загроз.

Основні способи реалізації загроз:

1) маскування під легального користувача через використання його ідентифікатора (методи соціальної інженерії, крадіжка обладнання, збережені дані тощо);

2) створення множинних ідентифікаторів або повторний анонімний доступ для розміщення нелегальної інформації.

Через неможливість контролю фізичної особи, яка здійснює доступ за наданою інформацією, необхідно впроваджувати додатковий контроль перед поширенням повідомлень.

Для протидії виявленим загрозам були розроблені додаткові методи ідентифікації користувачів Інтернету.

1.4 Існуючі методи ідентифікації Інтернет-користувачів при інформаційному обміні електронними повідомленнями

Ідентифікаційним набором є сукупність ознак, яка, після виявлення і детального вивчення, може стати основою для ототожнення об'єкта. Ідентифікаційні ознаки - це стійкі властивості або особливості користувача, які мають відмінні риси.

Сьогодні існує кілька підходів до ідентифікації користувача при розміщенні електронних повідомлень. Всі методи можна поділити на три основні групи на основі типів ідентифікаційних ознак користувача:

1) методи ідентифікації за технічними характеристиками робочої станції користувача;

2) методи ідентифікації на основі поведінкових характеристик користувача на порталі;

3) методи ідентифікації на основі лінгвістичних або стилістичних характеристик електронних повідомлень.

Далі розглянемо особливості застосування цих методів для ідентифікації Інтернет-користувачів при обміні короткими електронними повідомленнями.

Ця група методів є найбільш розробленою та поширеною. Ідентифікація проводиться за сукупністю характеристик апаратного та програмного забезпечення робочої станції користувача, з якої здійснюється доступ до Інтернет-порталу. В якості ідентифікаторів використовуються комбінації таких характеристик:

- версія операційної системи;
- часовий пояс;
- версія браузера;
- мова браузера за замовчуванням;
- дозвіл екрана;
- мережева інформація (IP, MAC-адреса).

Попри глобальну унікальність IP-адрес, їх точність знижується через динамічні адреси, проксі-сервери, анонімайзери та NAT. Крім того, кілька користувачів можуть використовувати одну IP або MAC-адресу.

Додаткові технології:

- ActiveX (використовується лише в Internet Explorer);
- Flash (дає інформацію про шрифти, ОС, параметри екрана);
- Java (отримання інформації вимагає підтвердження користувача);
- JavaScript (інформація про User Agent і Etag, хоча легко підробити).

Недоліки методу:

1) відсутність унікальності - ідентифікує лише пристрій, а не конкретного користувача;

2) необхідність додаткових компонентів - для отримання технічних характеристик потрібні серверні компоненти чи додатки до браузера;

3) легка підробка - ідентифікаційні ознаки можна легко підробити, особливо в агресивному середовищі;

4) не підходить для постінцидентного аналізу - технічні характеристики неможливо отримати після інциденту, якщо вони не були зібрані завчасно;

5) законодавчі обмеження - отримання деяких характеристик суперечить законодавству.

Другий підхід, на якому хотілося б зупинитися, — це підхід, заснований на аналізі «клавіатурного почерку» та аналізі поведінки користувача на веб-порталі.

Суть цих методів полягає в тому, що проводиться збір і аналіз даних про дії, які користувач виконує на веб-сторінці. Джерела цього методу лежать у персоналізації веб-простору - задачі створення сайтів, які адаптують свою структуру, способи навігації, контент, банери, рекламні пропозиції та інші можливості під конкретного користувача на основі зібраної та проаналізованої інформації про його уподобання. Це робиться для вивчення аудиторії сайту, оптимізації його структури, а також для прямого маркетингу. Наприклад, користувач робить певну покупку в інтернет-магазині і далі починає отримувати пропозиції про купівлю схожих товарів. Іншим прикладом може бути аналіз кошика замовлень: на основі аналізу кошиків замовлень інших користувачів відбувається підбір товарів, і користувачеві пропонуються товари, що містяться в кошиках інших користувачів інтернет-магазину, які мають схожу поведінку на веб-сторінці.

Як правило, проводиться збір і аналіз інформації, характерної для поведінки користувача: послідовностей кліків на веб-сторінці та веб-сайті в цілому, переходів користувача на інші ресурси, періодів і часу активності в мережі, покупок, здійснених користувачем, а також надсилання форм, прокрутки сторінок, додавання в обране та інше.

Існує досить велика кількість підходів до вирішення задачі виявлення знань із шаблонів навігації користувачів і розробки методів і алгоритмів інтелектуального аналізу даних для пошуку шаблонів поведінки користувача. Але в основному всі вони базуються на застосуванні методик кластеризації та аналізу послідовностей.

Цей підхід набагато ефективніший і дає набагато вищу точність ідентифікації. Його застосування дозволяє ідентифікувати окремого користувача, а не його робочу станцію. Проте методи цієї групи також мають кілька загальних недоліків або обмежень:

1) необхідне використання додаткових серверних компонентів або доповнень до веб-браузера;

2) не застосовні для постінцидентного аналізу.

Як ідентифікатор може бути використано набір виявлених характеристик (ідентифікаційних ознак) короткого електронного повідомлення, у яких проявляються характерні риси користувача. Кожен користувач має свій стиль письма, який складає своєрідний унікальний «відбиток пальців» — набір характеристик, що дозволяють його ідентифікувати. Основна ідея полягає в тому, що вимірювання певних характеристик текстів дозволяє розрізняти тексти, написані різними користувачами.

У даному випадку ідентифікаційними ознаками є стійкі властивості або особливості письмової мови конкретного користувача, які мають відмінну здатність та виокремлюються з текстів електронних повідомлень для ідентифікації користувача.

Ідентифікація на основі стилістичних характеристик письмової мови користувача дозволяє проводити ідентифікацію без використання додаткових компонентів веббраузера, не вимагає підключення до сервера додаткових програмних компонентів або втручання в програмний код для збору характеристик програмного чи апаратного середовища користувача.

Тут ідентифікація — це процес встановлення користувача, що є автором повідомлень, на основі сукупності загальних та специфічних ознак тексту, які складають авторський стиль.

Попередні дослідження з лінгвістичної ідентифікації належать до двох груп: методи для ідентифікації автора літературного твору та методи для ідентифікації користувачів за повідомленнями іноземними мовами.

Історія лінгвістичної ідентифікації починається майже одночасно з появою літератури. Навіть найперші роботи в цій галузі були спрямовані на вирішення досить широкого кола завдань: встановлення автора анонімного тексту або тексту, написаного під псевдонімом, підтвердження або спростування авторства, визначення авторства фрагментів тексту у колективних роботах. Подібні завдання стоять і перед сучасними дослідниками, однак ускладнились умови.

Усі дослідження з лінгвістичної ідентифікації інтернет-користувачів присвячені двом ключовим питанням: пошук ефективного алгоритму ідентифікації (класифікації) та вибір ідентифікаційних ознак, які мають найбільшу розпізнавальну здатність. Усі

роботи з лінгвістичної ідентифікації користувачів можна розділити на дві основні групи: так званий профільний підхід і підхід з навчанням на прикладах.

У профільному підході всі тексти одного користувача об'єднуються в один, і розраховуються значення різних ідентифікаційних ознак, тобто користувач представлений як один вектор значень. У підході з навчанням на прикладах користувач представлений як набір його текстів, де кожен текст є вектором з ідентифікаційними ознаками і навчальним прикладом.

Профільний підхід більш корисний, коли доступна невелика кількість текстів від одного користувача або коли тексти є дуже короткими. Підхід з навчанням на прикладах дозволяє комбінувати різні групи ознак, що робить його більш застосовним для великої кількості потенційних користувачів.

Більшість методів ідентифікації користувачів за характеристиками електронних повідомлень засновані на методі опорних векторів та нейронних мережах.

У роботі [21] вперше для ідентифікації користувачів було використано метод опорних векторів, який наразі є одним із найпопулярніших методів класифікації текстів. Пізніше цей метод був використаний у інших роботах [22-24].

Деякі роботи (наприклад, [25]) присвячені ідентифікації користувачів за допомогою ланцюгів Маркова, де використовували ймовірності появи певних послідовностей літер.

В роботах [26, 27] були зроблені перші спроби застосувати алгоритм випадкового лісу для вирішення завдання верифікації авторства тексту англійською мовою. Однак цей алгоритм не використовувався для ідентифікації інтернет-користувачів.

У більшості проаналізованих робіт використовувались такі алгоритми класифікації:

1) метод опорних векторів (Support Vector Machine, SVM) ([21]);

2) нейронні мережі (Multilayer Perceptron, MLP);

3) дерева рішень (Decision Trees, DR) ([23]);

4) регресійні моделі (Logistic Regression, LR);

5) наївний байєсівський класифікатор (Naive Bayes, NB) ([28]).

Існує велика кількість стилістичних або лінгвістичних характеристик, які можуть виступати як ідентифікаційні ознаки користувача. Це може варіюватися від найпростіших лінгвістичних ознак, таких як частоти слів певної довжини, до більш складних, що вимагають певного синтаксичного аналізу.

У проаналізованих роботах використовуються різні ознаки, але в більшості з них застосовуються n-грами символів і слів ([28-30]), частоти службових слів, частоти певних частин мови, знаків пунктуації, а також слова і речення певної довжини. Часто використовуються комбінації різних типів ознак [31].

Класифікація характеристик електронних повідомлень:

1) символьні:

а) характеристики символів;

б) N-грами літер;

2) лексичні:

а) характеристики слів, зокрема розподіл довжин слів;

б) функціональні слова;

3) синтаксичні:

а) N-грами слів;

б) N-грами частин мови;

в) структурні характеристики (форматування тексту);

4) семантичні;

5) контентно-специфічні мета-характеристики:

а) характеристики електронних листів;

б) характеристики записів на веб-сторінках.

У роботі [21] в якості ідентифікаційних ознак використовувалися частоти біграм частин мови (поєднання граматичних характеристик, таких як іменники чоловічого роду та прикметники). Усі іменники, прикметники та дієслова були замінені на їх граматичні мітки, з яких були отримані біграми та розраховані їх частоти. Точність класифікації була значно вищою, ніж при використанні словникових біграм, оскільки в цьому випадку класифікація не залежала від контексту.

Дослідження [32] вивчало, які характеристики тексту можуть найбільш ефективно використовуватися для ідентифікації користувача за лінгвістичними та стилістичними характеристиками. Дослідження показало, що біграми, складені шляхом заміни слів на відповідні їм частини мови (дієслово-іменник, іменник-прикметник, прикметник-іменник), можуть давати позитивні результати для класифікації, проте цей метод не враховував зв'язки між словами в реченні.

У роботах [22] і [33] досліджується можливість ідентифікації користувача за текстами електронних листів. Увага приділяється характеристикам з п'яти різних категорій: загальні характеристики тексту, частотні характеристики функціональних слів, щільність розподілу довжин слів та специфічні характеристики електронних листів. Якість класифікації суттєво зростала при одночасному використанні специфічних характеристик електронних листів з класичними характеристиками тексту.

Дослідження, присвячені грецькій мові, викликають особливий інтерес і цитуються у роботах [34-35]. У роботі [23] проводилися дослідження з ідентифікації користувачів за електронними повідомленнями англійською та китайською мовами. У роботі [23] доведено, що комбіноване використання характеристик різних типів підвищує точність ідентифікації.

У роботі [24] описується метод вибору *n*-грам, який найточніше характеризує користувача. Розмір навчальної вибірки складає 2500 текстів 50 користувачів. Для проведення експерименту використовується інший набір даних, що також складається з 2500 текстів 50 користувачів. Точність класифікації складає 73,08%.

У роботі [36] запропоновано використовувати метод Common N-Grams (CNG), який полягає у обчисленні близькості профілю невідомого користувача і профілів відомих користувачів, що містять частоти n-gram. У роботі [37] пропонується спростити обчислення відстані між двома профілями користувачів, названий Simplified Profile Intersection (SPI). Пропонується просто підраховувати кількість спільних n-gram з двох профілів, не звертаючи уваги на інші. Використання SPI для ідентифікації користувача дає кращі результати, ніж початкова відстань CNG.

У роботі [38] досліджується вплив кількості потенційних користувачів і розміру навчальної вибірки на точність класифікації. У роботі доведено, що з ростом кількості користувачів точність класифікації зменшується; при 145 потенційних користувачах точність класифікації становить всього 12%.

У роботі [39] наведені результати застосування методу ідентифікації автора вихідного коду (Source Code Authorship Profile) до набору повідомлень мікроблогів Twitter. Суть методу полягає у складанні профілю користувача і профілю повідомлення за p-gram, а потім у порівнянні профілів для пошуку найбільш підходящого користувача. Для набору з 50 користувачів застосування цього методу дає точність класифікації 55%.

У роботі [40] визначено, що текст довжиною менше 250 слів достатньо складно ідентифікувати, і ідентифікація користувача за повідомленням такої довжини практично неможлива.

Порівняння всіх проаналізованих робіт з ідентифікації Інтернет-користувача наведено нижче (таблиця 1.1). При цьому використовуються такі позначення:

- довжина повідомлення: мала (М) - до 5000 символів, середня (С) - до 20 000 символів, велика (В) - більше 20000 символів;
- g - кількість потенційних користувачів: мала (М) - 2-5, середня (С) - 5-9, велика (В) - більше 10;
- *l* - кількість повідомлень одного користувача: мала (М) - до 10, середня (С) - до 25, велика (В) - більше 25.

Проведений аналіз методів ідентифікації користувачів дозволяє зробити висновок, що найбільш перспективною є ідентифікація на основі лінгвістичних і стилістичних характеристик тексту.

Проаналізовані методи мають досить низьку точність ідентифікації Інтернет-користувача, що відображено на рисунку 1.1.

Таблиця 1.1 - Основні роботи з ідентифікації користувача.

Робота	Тип характеристик	Підхід	Алгоритм	*g*	Довжина повідомле	*l*	Точність ідентифіка
Іноземна мова							
[21]	Зміш.	Прим.	SVM	B	M	C	60-80
[22]	Зміш.	Прим.	SVM	M	M	C	92,2
[34]	Зміш.	Прим.	Др.	B	C	-	81
[35]	Лекс.	Прим.	Др.	C	C	C	73,8
[36]	Симв.	Проф.	Др.	B	B	C	85,7
[23]	Зміш.	Проф.	SVM	B	M	B	82
[24]	Синт.	Прим.	SVM	B	B	B	73,08
[36]	Симв.	Проф.	Др.	C	B	M	71,4
[25]	Леке.	Проф.	SVM NB	M	B	B	85,7
[39]	Симв.	Проф.	Др.	C	B	B	78,6
[40]	Синт.	Прим.	SVM	B	M	M	29,8

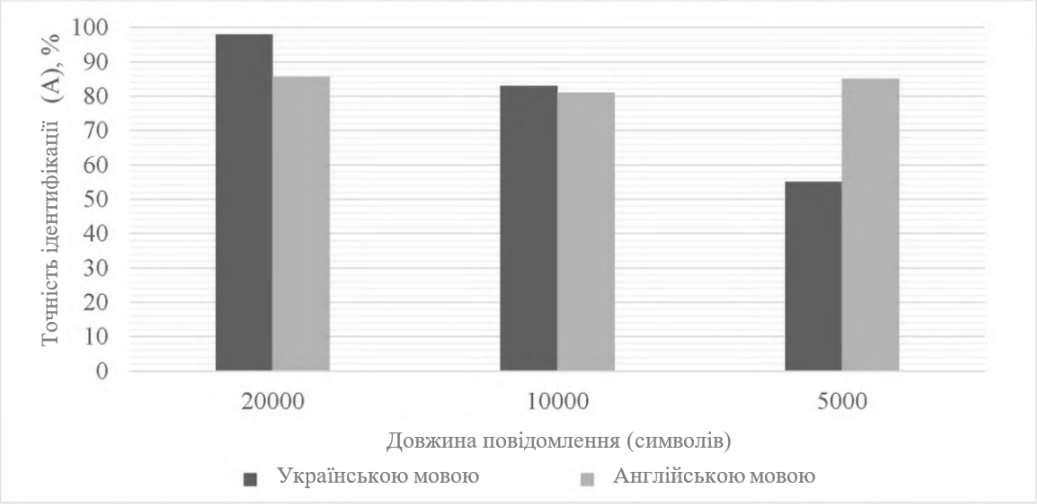

Рисунок 1.1 - Точність при використанні різних методів лінгвістичної ідентифікації Інтернет-користувача

Таким чином, не існує рішення, яке дозволяло б з достатньою точністю проводити ідентифікацію користувача на основі текстів електронних повідомлень українською мовою довжиною менше ніж 5000 символів. Необхідно провести дослідження, спрямоване на підвищення точності ідентифікації шляхом удосконалення існуючих або пошуку нових, раніше не застосовуваних для розв'язання цієї задачі методів ідентифікації, методів формування ідентифікаційного набору ознак, що має максимальну розпізнавальну здатність на текстах довжиною менше 5000 символів.

1.5 Основні особливості задачі ідентифікації на основі стилістичних характеристик текстів електронних повідомлень

Часто завдання ідентифікації користувача електронного повідомлення розглядається як одна з підзадач класифікації текстів, поряд з такими завданнями, як тематична ідентифікація, визначення мови, жанрова ідентифікація тощо. Однак існують деякі особливості, що виділяють ідентифікацію користувача як окрему самостійну задачу:

1) вибір характеристик тексту як ідентифікаційних ознак. Ідентифікація користувача на основі стилістичних і лінгвістичних характеристик текстів електронних повідомлень — це завдання категоризації текстів, що грунтується на стилістичних особливостях тексту. У таких задачах, зазвичай, найважливішими відмінними характеристиками є слова з найвищими частотними показниками або інші особливості стилю користувача, що найчастіше зустрічаються, тобто ті, що найбільше вживаються користувачем. У задачі категоризації текстів за тематикою, навпаки, найчастіше вживані слова зазвичай виключаються, оскільки вони є слабкими ознаками для класифікації, тому що не несуть семантичного (смислового) навантаження. У задачах класифікації, заснованих на стилістичних особливостях, такі службові або функціональні слова можуть бути вкрай корисними, оскільки вони

використовуються користувачем несвідомо, що певною мірою відображає індивідуальний стиль користувача;

2) недостатня кількість і стислість текстів навчальної вибірки. На практиці рішення типового завдання ідентифікації користувача ускладнюється тим, що зазвичай доступні лише декілька досить коротких текстів, які безсумнівно належать потенційним користувачам. Цей фактор є вирішальним при виборі методів ідентифікації, здатних коректно та ефективно обробляти випадки з невеликим розміром навчальної вибірки;

3) дисбаланс класів. Досить стандартною є ситуація, коли існує нерівномірний розподіл довжини текстів навчальної вибірки різних користувачів. Окрім проблеми, зазначеної в попередньому абзаці, різниця в загальній сумарній довжині текстів різних користувачів також може спричинити певний дисбаланс і призвести до некоректних результатів (наприклад, якщо у нас є короткі тексти одних користувачів та довгі тексти інших). Недостатня довжина текстів навчальної вибірки одного користувача порівняно з іншими не повинна знижувати ймовірність того, що цей користувач буде обраний як справжній автор тексту.

На якість вирішення завдання лінгвістичної ідентифікації впливають:

- довжина повідомлення,
- загальна кількість користувачів,
- кількість повідомлень одного користувача,
- дисбаланс між кількістю повідомлень користувачів.

1.6 Постановка задачі з ідентифікації Інтернет-користувачів за лінгвістичними та стилістичними характеристиками електронних повідомлень

Як ідентифікуюча інформація (ідентифікатор) у випадку лінгвістичної ідентифікації виступають стилістичні та лінгвістичні характеристики електронного повідомлення. Існують еталонні шаблони — електронні повідомлення користувачів.

Ідентифікувати Інтернет-користувачів під час інформаційного обміну електронними повідомленнями означає визначити користувача шляхом аналізу нового надісланого повідомлення та порівняння його з еталонними шаблонами. Визначення того, що певне електронне повідомлення належить певному користувачеві (класу) з числа можливих потенційних користувачів, здійснюється на основі аналізу значень ознак цього електронного повідомлення та повідомлень, які є в базі даних і належать потенційним користувачам.

Виходячи з вищесказаного, завдання ідентифікації Інтернет-користувачів за лінгвістичними та стилістичними характеристиками електронних повідомлень поділяється на дві підзадачі:

1) збір та формування бази характеристик потенційних користувачів, що містить еталонні шаблони користувачів U, включно з потенційними користувачами $U_{потенц}$;

2) ідентифікація та автентифікація користувача за новим надісланим електронним повідомленням t_j (здійснюється розбір повідомлення і порівняння з еталонними шаблонами користувачів). У разі успішного порівняння особистість користувача та її справжність можуть вважатися встановленими.

Задача ідентифікації користувача електронних повідомлень у загальному випадку зводиться до розв'язання задачі багатокласової класифікації.

Маємо деяке надіслане користувачем u_k повідомлення t_j та набір потенційних користувачів $U_{потенц}$.

Для кожного потенційного користувача необхідно визначити ймовірність того, що він є автором надісланого повідомлення, тобто потрібно класифікувати це повідомлення серед наявних користувачів.

Є множина повідомлень різних користувачів — еталонних шаблонів: $T=\{t_1, \dots, t_z\}$, де z - кількість повідомлень.

Множина користувачів $U = \{u_1, \dots, u_y\}$, де y — кількість користувачів.

Для $T' \in T$ повідомлень відома їхня приналежність до певного користувача, тобто існує деяка множина, що складається з пар відповідностей

користувачів і повідомлень. Кожен користувач представляється так: $u_k = T_k = \{t_1, \dots, t_l\}$, де l - кількість повідомлень користувача u_k.

Потрібно побудувати алгоритм a: $t_j \to U_{потенц}$, здатний визначити, з якими ймовірностями повідомлення t_j належить кожному користувачу з $U_{потенц}$, де t_j - повідомлення, за характеристиками якого необхідно провести ідентифікацію (пред'явлений ідентифікатор).

Необхідно здійснити навчання цього алгоритму на навчальній вибірці $T_{tr} \in T$ та його верифікацію на $T_{test} \in T$, де T_{tr} - навчальна вибірка, T_{test} - тестова вибірка.

Далі навчений алгоритм A повинен бути здатним класифікувати надіслане повідомлення t_j до користувачів з $U_{потенц}$.

Шуканим користувачем є користувач, для якого $P(t_j$ належить $u_{max})$ є максимальним.

При вирішенні задачі ідентифікації користувача за електронними повідомленнями на якість ідентифікації можуть впливати два основні фактори:

- використовувані характеристики електронного повідомлення;
- метод класифікації текстів за користувачами.

2 РОЗРОБКА МЕТОДИКИ ІДЕНТИФІКАЦІЇ ІНТЕРНЕТ-КОРИСТУВАЧА НА ОСНОВІ СТИЛІСТИЧНИХ ТА ЛІНГВІСТИЧНИХ ХАРАКТЕРИСТИК КОРОТКИХ ЕЛЕКТРОННИХ ПОВІДОМЛЕНЬ

2.1 Аналіз структури та характеристик електронних повідомлень Інтернет-порталів

З метою отримання характеристик текстів, які використовуються як ідентифікаційні ознаки, під час підготовки роботи було сформовано спеціальний корпус електронних текстів.

Для збору повідомлень користувачів порталу LiveJournal проводилася індексація веб-сторінок у автоматичному режимі з використанням розробленого пошукового робота («веб-павук», краулер), який послідовно здійснює обхід сторінок користувачів і збирає повідомлення, що на них містяться. Інформація про користувачів і повідомлення виділялася з вмісту веб-сторінок за певними критеріями — послідовність html-тегів, що однозначно визначає, що в певному місці html-коду міститься інформація про користувача або повідомлення. Під час збору повідомлень із них видалялася вся незначуща інформація (інформація про розмітку сторінки, коментарі розробників сторінки тощо), додатковий аналіз або обробка повідомлень не здійснювалися. Відбір повідомлень за тематикою не проводився. Усі повідомлення були різної довільної довжини. Більшість текстів мали довжину від 142 до 699 символів. Розподіл довжин повідомлень наведено на рисунку 2.1.

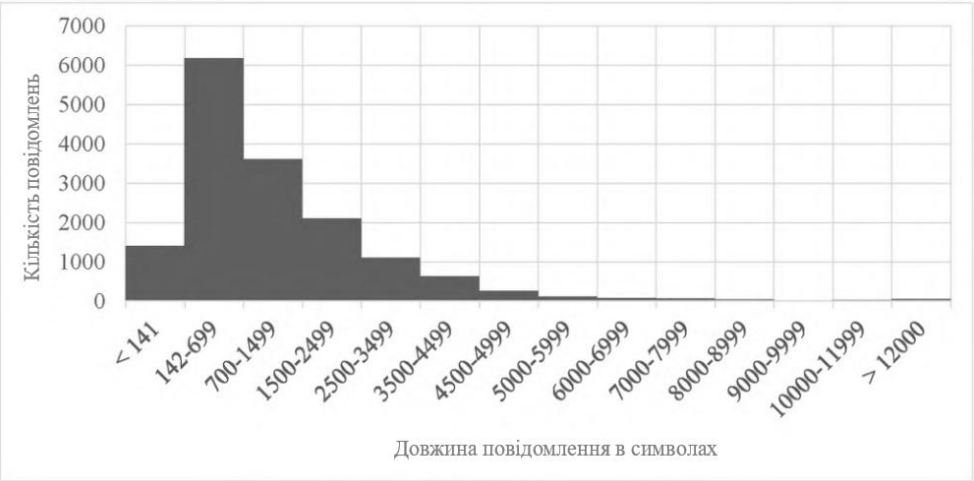

Рисунок 2.1 - Розподіл довжин повідомлень

Передбачається, що тексти були створені безпосередньо користувачем і не піддавалися редакторській правці чи іншій корекції.

2.2 Комплексна багаторівнева модель представлення Інтернет-користувача

Усі методи представлення користувача при вирішенні завдання лінгвістичної ідентифікації можна розділити на дві групи залежно від підходу до представлення користувача:

1) перша група: користувач представляється як один текст, отриманий шляхом конкатенації всіх його повідомлень. Обчислюються значення ідентифікаційних ознак, і формується профіль користувача. Цей підхід корисний, коли доступно мало повідомлень або вони дуже короткі;

2) друга група: користувач представляється як набір окремих повідомлень, кожне з яких є навчальним прикладом. У таких методах можна комбінувати різні типи ознак (бінарні, дискретні, неперервні), і вони демонструють вищу точність при великій кількості потенційних користувачів.

Аналіз попередніх досліджень привів до рішення використовувати саме другий підхід. Розроблена комплексна багаторівнева модель представлення Інтернет-користувача (КБМПІ) відрізняється застосуванням унікальних характеристик електронних текстових повідомлень і дозволяє ідентифікувати користувачів за повідомленнями довжиною менш ніж 5000 символів.

В векторній моделі електронне повідомлення розглядається як вектор в просторі ознак описів R^n, де n - це кількість частотних характеристик або ознак, що є однаковими для всіх повідомлень. Кожна з ознак характеризує певну особливість написання тексту.

Якщо ввести в розгляд n-вимірний простір ознак $\{Fi\}$, де $i = 1, \dots, n$, то кожне t_j повідомлення (об'єкт) в цьому просторі відображається точкою з координатами $t_j = F_j = (f_{j1}, \dots, f_{jn})$, а кожен клас об'єктів - користувач u_k - множиною таких точок.

Будь-яке повідомлення може бути представлене як набір деяких кількісних значень. Також раніше було доведено, що при розв'язанні задачі з ідентифікації користувача значення мають статистичні частотні характеристики. Одному

користувачеві в певний проміжок часу властиві певні характерні особливості. Вимірювання частот появи цих особливостей дозволяє нам представити повідомлення як набір частот характеристик і дозволяє перенести його в простір. Повідомленню може бути поставлена в відповідність точка в просторі, координатами якої є частоти його характеристик.

Припускаючи, що користувач має певні властиві йому стилістичні характеристики, можна зіставити йому, як конкретній фізичній особі, певні характеристики тексту на природній мові. Тобто можна представити текст у вигляді вектора довжиною *n*, де *n* - число характеристик або ідентифікаційних ознак.

Для представлення моделі Інтернет-користувача доцільно зазначити, що електронне повідомлення володіє характеристиками, властивими користувачу, який його написав. Можна визначити користувача u_k як множину повідомлень T_k, де кожне повідомлення t_{kj} представляється як кортеж, координатам якого відповідає множина характеристик, обумовлених особливостями використовуваної письмової та усної мови F_{kj}.

Отже, користувач u_k представляється як множина його повідомлень T_k або F_k — множина наборів векторних представлень повідомлень користувача u_k.:

$$u_k = T_k = \{t_{k1},...,t_{kl}\} = F_k = \{F_{k1},...,F_{kl}\}, \quad (2.1)$$

де $u_k \in U$ і l – кількість повідомлень u_k.

КБМПІ користувача u_k схематично представлена в таблиці 2.1.

2.3 Метод формування динамічного стилістичного профілю користувача, який має найбільшу розрізняючу здатність і дозволяє підвищити точність ідентифікації

Для формування динамічного стилістичного профілю автора і вибору оптимальної кількості інформативних ознак пропонується здійснювати відбір ознак на основі обчислення відстані за значенням ознаки до k-найближчих сусідів (алгоритм

Relief-f). Алгоритм Relief-f раніше не використовувався для відбору найбільш інформативних ознак користувачів, проте він був досить ефективним при вирішенні подібних завдань в інших галузях. Відбір ознак здійснюється на основі обчислення відстані за значенням ознаки до *k*-найближчих сусідів. У цій роботі приймається k=10. Кожній з ознак присвоюється певний коефіцієнт, розрахований шляхом оцінки відстані до найближчих повідомлень того ж користувача і до повідомлень інших користувачів.

Таблиця 2.1 - Комплексна багаторівнева модель представлення користувача

$u_k = T_k$	$T_k = \{t_{k1}, \dots, t_{kl}\}$	$t_{kj} = F_{kj} = (F_{kj}ls, F_{kj}lw, F_{kj}s, F_{kj}st, F_{kj}m)$	$t_{kj} = \{f_{kj1}, \dots, f_{kjn}\}$
Користувач u_k	Повідомлення t_{k1}	Лексичний рівень символів ($F_{k1}ls$)	f_1 Довжина повідомлення
		Лексичний рівень слів ($F_{k1}lw$)	…
		Синтаксичний рівень ($F_{k1}s$)	
		Структурний рівень ($F_{k1}st$)	
		Мета-рівень ($F_{k1}m$)	f_n День тижня публікації
	…	…	…
	Повідомлення t_{kl-1}	F_{kl-1}	$(f_1, f_2, \dots, f_n)$
	Повідомлення t_{kl}	F_{kl}	$(f_1, f_2, \dots, f_n)$

Вага ознаки тим більша, чим менша відстань за значенням цієї ознаки до повідомлень того ж користувача і чим більша відстань за значенням цієї ознаки до повідомлень інших користувачів. Тобто вага тим вища, чим краще ознака відокремлює повідомлення одного користувача від повідомлень інших

користувачів, і чим гірше розділяє повідомлення користувача між собою. Відбираються ознаки з позитивною інформативністю, чия вага $W[f_i] \geq 0$.

Алгоритм відбору ознак включає таку послідовність кроків:

1) вага всіх ознак $W[F] = 0$;

2) для всіх повідомлень набору $j = (1, l)$, де l - розмір вибірки:

а) відбір випадкового повідомлення $t_j \in T, j = (1, l)$ користувача u_x;

б) пошук k-найближчих повідомлень того ж користувача - H_c;

в) для всіх інших користувачів $u \neq u_x$ пошук k-найближчих повідомлень інших користувачів - $M_c(u), u \neq u_x$ (рис. 4);

г) обчислення інформативності для кожної ознаки f_i, $i = (1, n)$, де n - кількість ознак, які визначаються за формулами:

$$W[f_i] = W[f_i] - \sum_{c=1}^{k} \frac{diff(f_i, t_j, H_c)}{1 \times k} + \sum_{u \neq u_x} \sum_{c=1}^{k} [\frac{p(u)}{1 - p(u_k)} \times \frac{\mathrm{diff}(f_i, t_j, M_c(u))}{1 \times k}]; \quad (2.2)$$

$$diff(f, t_1, t_2) = \frac{|val(f, t_1) - val(f, t_2)|}{\max(f) - \min(f)}; \quad (2.3)$$

3) відбір ознак f, чия вага перевищує задане порогове значення $F'_{kj} = (f_{kj1}, \ldots, f_{kjm})$, де $W[f_i] \geq \tau, \tau = 0$.

Під час обчислення ваг у формулі (2.2) використовується нормалізована відстань. Також при обчисленні впливу ознаки на якість розділення повідомлень користувачів між собою у формулі (2.2) враховується апріорна ймовірність появи повідомлень даного користувача (рисунок 2.2). Це важливо, оскільки якщо повідомлень користувача небагато і ймовірності не враховуються, то вийде, що вага ознаки суттєво зростатиме через значення відстані для користувачів, у яких мало повідомлень.

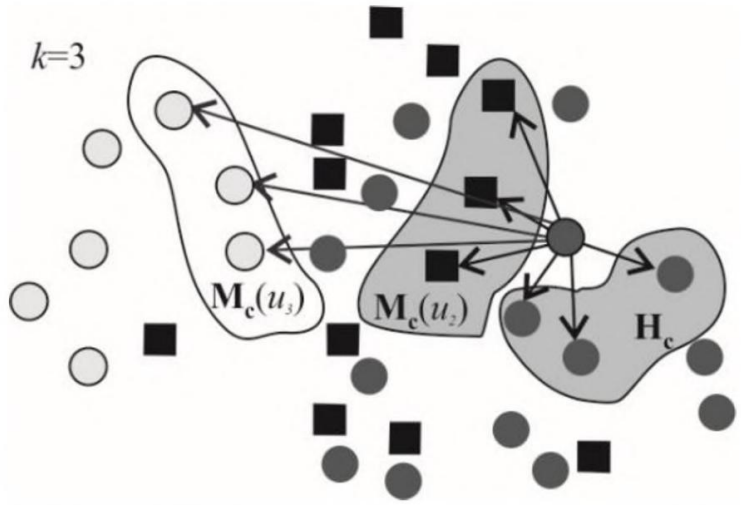

Рисунок 2.2 - Відбір ознак на основі розрахунку відстані до *k*-найближчих сусідів з урахуванням апріорних ймовірностей

2.4 Обгрунтування використання методу відбору на основі розрахунку відстані за значенням ознаки до *k*-найближчих сусідів для формування ДСПК

Проводилося дослідження можливості використання методів відбору ідентифікаційних ознак для формування динамічного стилістичного ідентифікаційного профілю користувача.

З метою виявлення найбільш ефективного алгоритму, для включення його в метод формування динамічного стилістичного профілю користувача досліджувалися два підходи:

1) на основі розрахунку відстані за значенням ознаки до *k*-найближчих сусідів (алгоритм Relief-f);

2) на основі розрахунку відношення приросту інформації (Gain Ratio, GR).

Розрахунок відношення приросту інформації проводиться за формулою:

$$GR(S, f_i) = \frac{H(S) - \sum_{x \in values(f_i)} \frac{|S_x|}{|S|} \times H(S_x)}{\sum_{x \in values(f_i)} \frac{|S_x|}{|S|} \times H(S_x)}, \quad (2.4)$$

де $H(S)$ — ентропія до розподілу за ознакою;

values(f_i) — усі можливі значення ознаки f_i;

S_x — підмножина набору даних;

$f_i = x$;

|*S*| - кількість елементів у множині.

Для обгрунтування вибору найкращого алгоритму відбору ознак, який використовується як функція перетворення, необхідно провести серію експериментів.

Було проведено дослідження точності ідентифікації за допомогою ДСПК, сформованих за використання зазначених алгоритмів. У переважній більшості випадків (у 8 з 10) найбільша точність була досягнута за допомогою запропонованого методу відбору, що дозволило зробити висновок про доцільність його використання (рисунок 2.3).

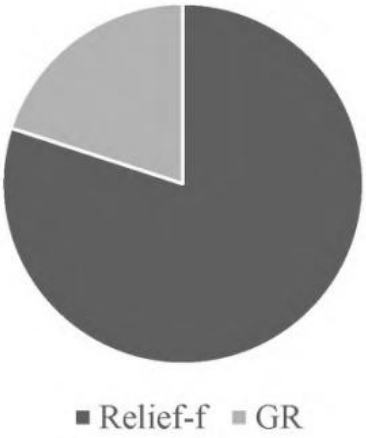

Рисунок 2.3 - Співвідношення кількості експериментів, у яких максимальна точність була досягнута з використанням різних підходів до відбору ознак

2.5 Метод порівняння ДСПК потенційних користувачів з еталонними на основі методу випадкового лісу

Для підтвердження можливості використання зазначених вище алгоритмів і їх відповідності відповідним вимогам проводився ряд додаткових експериментів. Було проведено оцінювання швидкості роботи та точності, також досліджувалася залежність часу роботи від кількості ідентифікаційних ознак.

Були складені п'ять наборів даних, в яких кількість ідентифікаційних ознак у ДСПК варіювалася від 30 до 90, та один набір даних, що містив повні КБМПІ.

Оцінювався час роботи кожного з зазначених алгоритмів на даних наборах ознак. Рисунок 2.4 містить результати проведених експериментів.

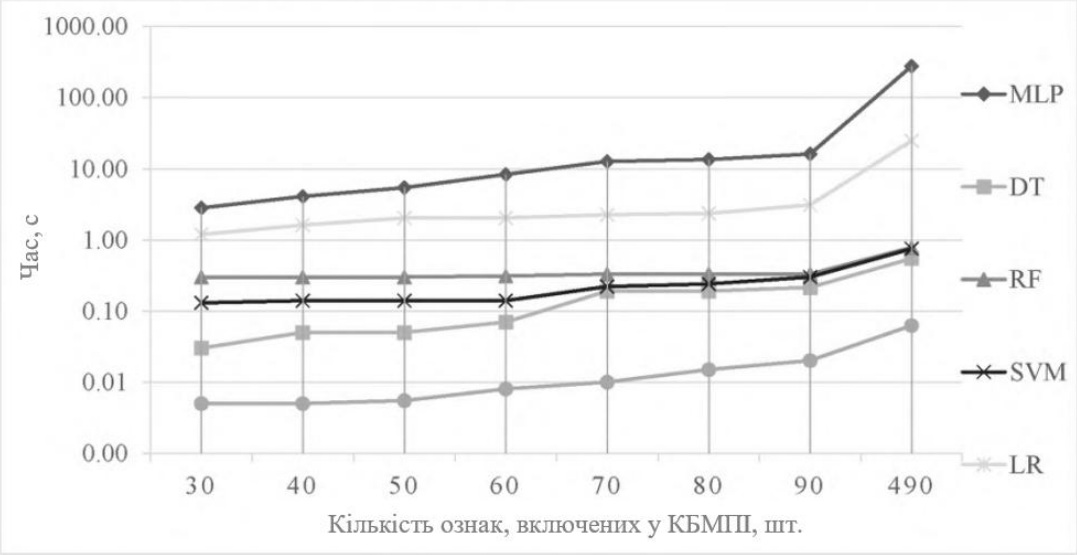

Рисунок 2.4 - Порівняння швидкості роботи методів порівняння КБМПІ

При порівнянні відібраних методів за часом їх роботи суттєво найгірший результат продемонстрували нейронні мережі та логістична регресія; зростання кількості ознак призводить до значного збільшення часу роботи. Цей результат збігається з результатами, отриманими раніше, і також підтверджується іншими дослідниками.

Далі проводилася оцінка точності класифікації з використанням зазначених вище алгоритмів на повній КБМПІ. Результати проведених експериментів дозволили виключити нейронні мережі та логістичну регресію з подальшого дослідження, оскільки час їх роботи є суттєво великим, а точність класифікації є порівнянною з точністю, отриманою при використанні інших методів, що явно видно в таблиці 2.2.

Алгоритм випадкового лісу (Random Forest, RF) раніше не використовувався для розв'язання задачі ідентифікації англомовних Інтернет-користувачів, хоча під час вирішення схожих задач в інших предметних областях було доведено, що алгоритм має ряд переваг.

Таблиця 2.2 - Порівняння точності ідентифікації при використанні різних методів класифікації на повній КБМПІ

Алгоритм класифікації	Точність ідентифікації (A) по КБМПІ (n=498), %	Час роботи, с.
Нейронна мережа	63,9344	277,22
Дерева рішень	61,45251	0,546904
Випадковий ліс	77,65363	0,781289
Метод опорних векторов	65,92179	0,750045
Логістична регресія	63,68715	24,82222
Наївний байєсівський класифікатор	54,18994	0,062503

Серед переваг використання дерев рішень досить часто виділяють простоту алгоритму для розуміння, наочність результатів побудови, відсутність необхідності вибору вхідних ознак (під час побудови дерева вибираються найбільш значущі ознаки, які й використовуються для побудови), швидкість навчання та високу точність класифікації, ефективну роботу з великими обсягами даних, а також здатність обробляти велику кількість ознак, можливість роботи з усіма типами ознак (дискретні, неперервні, бінарні, символічні тощо) та можливість роботи з даними з пропущеними значеннями.

Основна ідея алгоритму RF полягає в побудові ансамблю (лісу) випадкових дерев ухвалення рішень. Класифікація проводиться за принципом голосування: кожне дерево лісу класифікує об'єкт до одного з класів, тобто голосує за певний клас. Далі об'єкт належить до того класу, за який проголосувало найбільше дерев.

Структура дерева представляє собою деревоподібний граф, що має в своєму складі ребра (гілки) та вузли двох типів: листя та внутрішні вузли (рисунок 2.5). У листях дерева містяться значення цільової функції, у нашому випадку клас — користувач, на ребрах записані значення ознак, від яких залежить значення цільової функції, у вузлах — самі ознаки. Щоб класифікувати

повідомлення, необхідно спуститися від кореня дерева до листа та отримати значення класу, що в ньому міститься.

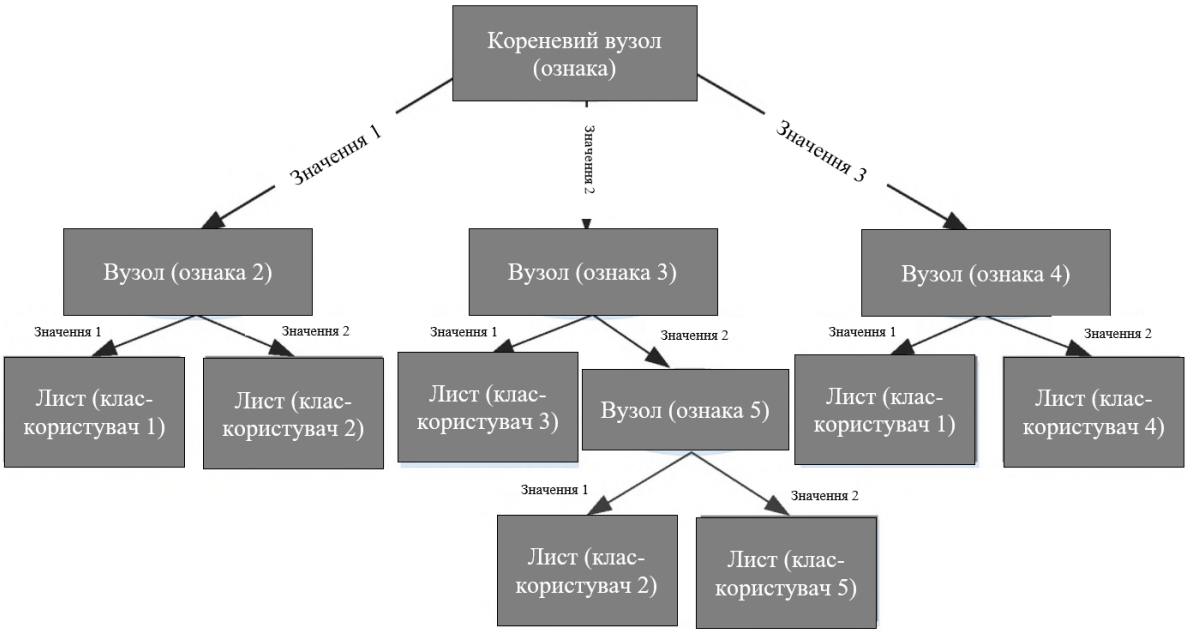

Рисунок 2.5 - Дерево рішень із довільною кількістю нащадків

Дано навчальну вибірку $T_{tr} \in T$, що складається з q повідомлень (рисунок 2.6), розмірність ознакового простору - n, n' - параметр, який визначає кількість випадково відібраних ознак із числа n (використовується $n' = \sqrt{n}$); $U_{потенц}$ — множина потенційних користувачів (g - кількість потенційних користувачів). Для кожного повідомлення $t_{us} \in T_{tr}$ відома його належність до певного користувача із множини $U_{потенц} = \{U_1, \dots, U_g\}$. Навчальну вибірку T_{tr} доцільно представити у вигляді матриці:

$$T_{tr} = \begin{Bmatrix} t_{us1} \\ \dots \\ t_{usq} \end{Bmatrix} = \begin{bmatrix} f_{11} & \dots & f_{1n} \\ \dots & \dots & \dots \\ f_{q1} & \dots & f_{qn} \end{bmatrix}. \quad (2.5)$$

Дерева ансамблю будуються за наступним принципом:

1) генеруються випадкові підвибірки з повторенням розміром q' з навчальної вибірки - $T_{tr} = \{T'_1, T'_2, \dots, T'_h\}$, де h - кількість згенерованих підвибірок. Деякі тексти

потрапляють у підвибірку кілька разів, а деякі не потрапляють у неї взагалі. Також випадковим чином відбираються n' ознак (рисунок 2.6).

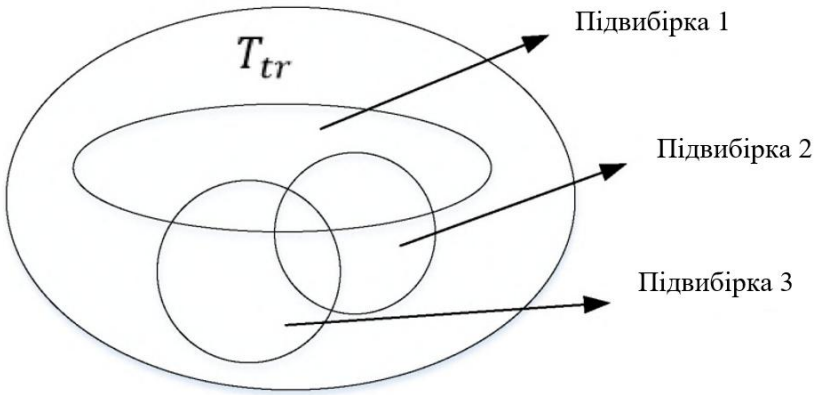

Рисунок 2.6 - Генерація підвибірок для побудови дерев рішень

2) на основі кожної такої підвибірки будується дерево рішень:

а) на першому кроці роботи алгоритму є корінь, множина повідомлень T', яку необхідно розбити на підмножини, та множина ознак $F' = \{f_1, ..., f_{m'}\}$ довжиною n'. Розбиття на підмножини здійснюється за значенням однієї із ознак. Для цього під час побудови обирається одна з ознак в якості критерію, що забезпечує найкраще розділення за n'. Для вибору найкращої з цих n' ознак використовується підхід на основі мінімізації ентропії:

$$H(S) = -(\sum_{u \in U} P(u) \times \log_2 P(u)) \tag{2.6}$$

та критерію приросту інформації:

$$IG(S, f_i) = H(S) - \sum_{x \in values(f_i)} \frac{|S_x|}{|S|} \times H(S_x), \tag{2.7}$$

де $H(S)$ - ентропія до розділення за ознакою;

$values(f_i)$ - всі можливі значення ознаки f_i;

S_x - підмножина набору даних, де $f_i = x$, $|S|$ - кількість елементів у множині;

б) обрана ознака f_i має k значень, що дає розбиття на k підмножин. Далі створюються k нащадків вузла, кожному з яких відповідає підмножина, отримана при розбитті початкової множини T'. Вибір ознаки та розбиття за нею на підмножини рекурсивно застосовуються до всіх k нащадків. Умовою виходу з рекурсії є випадки, коли після розгалуження у вузлі опиняються тексти одного користувача (тоді вузол стає листом), або вузол виявився асоційованим з порожньою множиною (тоді вузол також стає листом, але в якості користувача обирається користувач, який найчастіше зустрічається у безпосередніх предків цього вузла);

в) кожне дерево будується до вичерпання підвибірки, відсікання гілок не проводиться.

2) класифікація здійснюється за принципом голосування: кожне дерево лісу класифікує текст до одного з користувачів, тобто голосує за певного користувача. Далі текст відноситься до того користувача, за якого проголосувало найбільше дерев (рисунок 2.7).

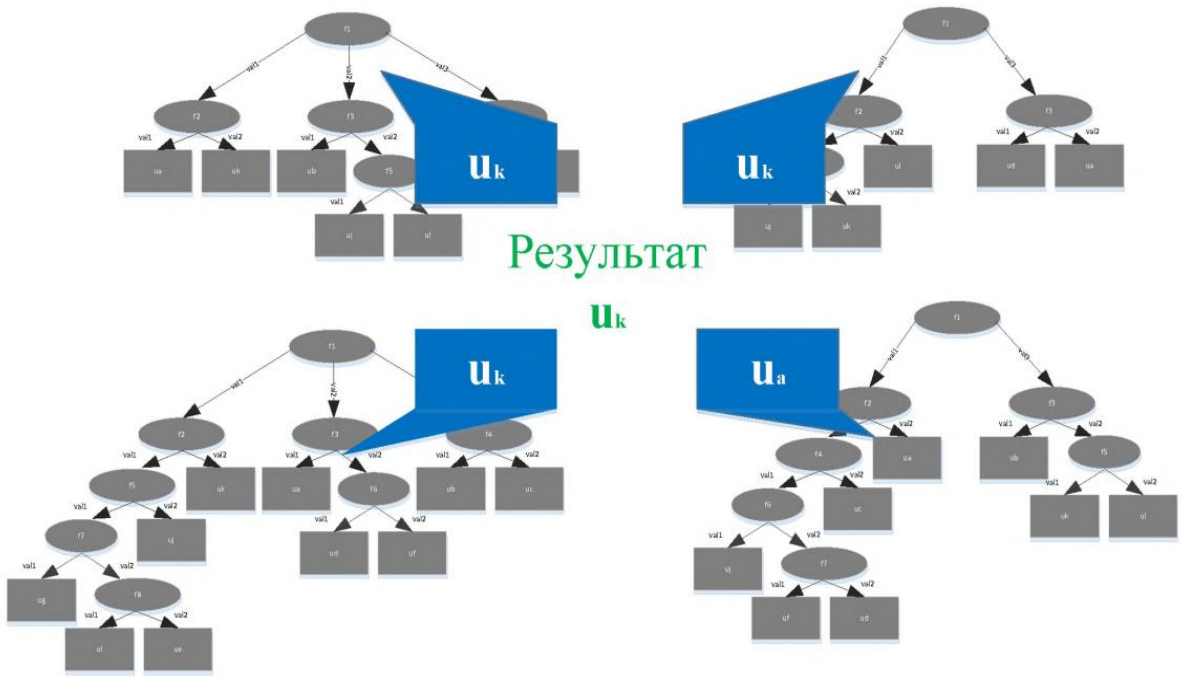

Рисунок 2.7 - Класифікація повідомлення за принципом голосування ансамблю дерев рішень

Передбачається, що більшість згенерованих дерев самі по собі правильно передбачають користувача, і що дерева, які помиляються, видають різні результуючі класи.

Для підвищення точності ідентифікації Інтернет-користувача шляхом віднесення електронного повідомлення до певного користувача, на основі аналізу значень його ідентифікаційних ознак, розроблений метод порівняння ДСПК з еталонними ДСПК потенційних користувачів на основі методу RF, що включає попередню дискретизацію ідентифікаційних ознак.

2.6 Методика ідентифікації Інтернет-користувача на основі стилістичних і лінгвістичних характеристик коротких електронних повідомлень

Ідентифікація користувача є можливою лише за наявності певної бази (БД), що містить еталонні шаблони U та $U_{потенц}$, оскільки ідентифікація здійснюється шляхом зіставлення набору виявлених характеристик електронного повідомлення з характеристиками повідомлень, зібраними раніше та вже наявними в базі даних, для яких відома їх належність до певного користувача. Еталонні шаблони користувачів у випадку лінгвістичної ідентифікації — це набір характеристик електронних повідомлень потенційних користувачів.

На цьому етапі формуються множини U та T, щоб потім було можливо віднести шуканий текст t_j до когось із користувачів підмножини $U_{потенц} \in U$.

На цьому етапі здійснюється:

1) збір повідомлень користувачів з певного Інтернет-порталу;

2) розбір та аналіз повідомлень для формування еталонних КБМПІ;

3) формування БД еталонних шаблонів користувачів КБМПІ еталони, яка також містить самих користувачів U та їх повідомлення T.

Першим кроком підготовчого етапу є збір, попередня обробка, структуризація та збереження електронних текстових повідомлень. Здійснюється формування бази даних користувачів, зареєстрованих на Інтернет-ресурсах (U). Далі відбувається збір і

попередня обробка повідомлень (*T*) користувачів, зареєстрованих на цих Інтернет-ресурсах, та подальше збереження повідомлень користувачів у базу даних.

Пропонується наступний алгоритм збору, обробки, структуризації та збереження повідомлень користувачів:

1) формування бази даних користувачів, зареєстрованих на Інтернет-ресурсах (*U*). Для початку необхідно сформувати список усіх користувачів, зареєстрованих на Інтернет-ресурсі *U*.

Для формування цього списку потрібно здійснити індексацію веб-сторінок в автоматичному режимі з використанням пошукового робота («веб-павук», краулер). Пошуковий робот аналізує та індексує вміст сторінки (html-код), виділяючи з нього користувачів та посилання на сторінки користувачів. Порядок обходу сторінок та критерії виділення значущої інформації визначаються розробленим пошуковим алгоритмом.

Інформація про користувачів виділяється з вмісту веб-сторінки за певними критеріями — послідовністю html-тегів, яка однозначно визначає, що в певному місці html-коду міститься інформація про користувача.

Якщо зустрічається новий користувач, запис про якого не міститься в БД, то він додається в БД. Якщо користувач зустрічається повторно, а запис про нього вже є в БД, то він ігнорується;

2) збір і попередня обробка повідомлень користувачів, зареєстрованих на певному Інтернет-ресурсі (*T*);

3) збереження повідомлень *T* у БД. Далі послідовно повідомлення кожного користувача з БД повинні бути отримані та додані до БД повідомлень користувачів. Для збору повідомлень користувачів також пропонується використовувати пошуковий робот, призначений для послідовного обходу сторінок користувачів і збору повідомлень, що містяться на них.

З БД витягується список користувачів. Далі паралельно запускається робота кількох пошукових роботів, які послідовно, відповідно до переданого їм списку, здійснюють обхід сторінок користувачів і збір повідомлень, що на них містяться.

Усі повідомлення проходять попередню обробку; з них видаляється вся незначна інформація (інформація про розмітку сторінки, коментарі розробників сторінки тощо). Повідомлення виділяються з html-коду сторінки і можуть містити певну кількість службових html-тегів, які не представляють інтересу в рамках цього дослідження.

Якщо раніше тексти цього користувача не збиралися, то в БД додаються всі повідомлення користувача. При оновленні інформації про повідомлення користувача враховуються лише нові повідомлення.

Після того, як повідомлення потрапляють у базу даних, необхідно здійснити індексацію кожного повідомлення з множини T з метою витягнення ідентифікаційних ознак (F), що входять до КБМПІ. Ці дії виконуються на наступному етапі;

4) аналіз електронних повідомлень множини T з метою витягнення та накопичення ідентифікаційних ознак (F), що входять до КБМПІ. Для кожного користувача, що міститься в БД та має повідомлення, необхідно провести обробку, витягнення та збереження ідентифікаційних ознак.

Для початку потрібно виконати попередню обробку повідомлення, замінити в тексті деякі фрагменти (гіперпосилання), символи або html-сутності (наприклад, «"»), які потім можуть заважати витягненню ідентифікаційних ознак. Далі проводиться підрахунок усіх кількісних ідентифікаційних ознак. До них належать, наприклад, кількість слів, речень, числових або великих символів, визначених html-тегів та багато інших. На основі отриманих даних обчислюються значення частотних ознак, що входять до КБМПІ;

5) збереження отриманих КБМПІ еталонів заноситься в БД.

2.7 Основні етапи методики ідентифікації Інтернет-користувача

На першому етапі починається процедура ідентифікації за лінгвістичними та стилістичними характеристиками електронного повідомлення t_j для користувача u_k. Виконується формування КБМПІ та ДСПК для u_k u_k та $U_{\text{потенц}}$.

На першому кроці (рисунок 2.8) здійснюється аналіз електронного текстового повідомлення t_j з метою вилучення характеристик F_j, що входять до КБМПІ, та формування КБМПІuk для користувача u_k, який розміщує повідомлення t_j. Також виконується вилучення КБМПІ потенц для $U_{\text{потенц}}$ з БД.

Рисунок 2.8 - Блок-схема дій першого кроку першого етапу методики ідентифікації

На другому кроці (рисунок 2.9) формуються динамічні лінгвістичні профілі потенційних користувачів (ДСПКпотенц), які мають максимальну роздільну здатність, та формується динамічний стилістичний профіль користувача ДСПКuk, для якого необхідно здійснити ідентифікацію, за розробленим методом формування динамічних стилістичних профілів з ідентифікаційних ознак, що входять до КБМПІ. Здійснюється безпосередня ідентифікація користувача u_k за електронним повідомленням t_j. Етап включає такі кроки:

1) дискретизація неперервних ознак з ДСПКпотенц і ДСПКuk, формування ДСПК'uk з еталонними ДСПК'потенц (рисунок 2.10);

2) порівняння ДСПК'uk з еталонними ДСПК'потенц;

3) прийняття остаточного рішення про автентичність наданого ідентифікатора.

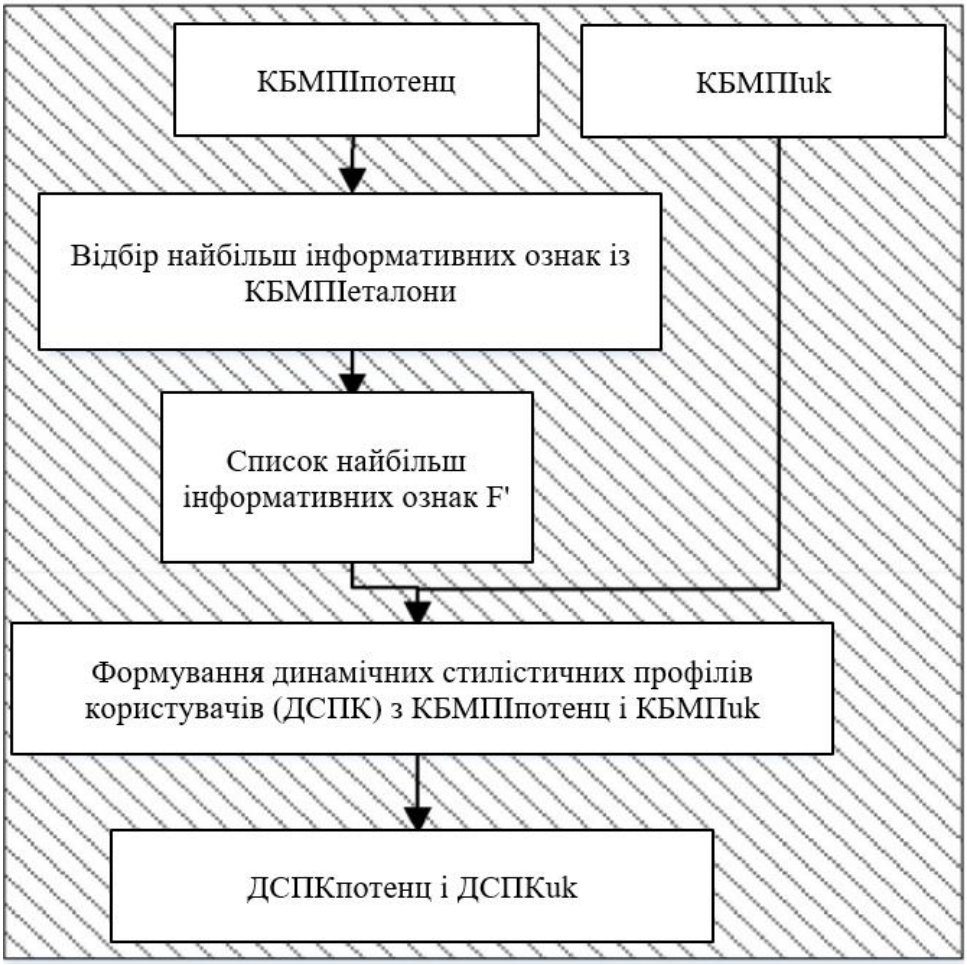

Рисунок 2.9 - Блок-схема дій другого кроку першого етапу методики ідентифікації

Рисунок 2.10 - Блок-схема дій першого кроку другого етапу методики.

Перший і другий кроки здійснюються за розробленим методом порівняння ДСПК'uk з еталонними ДСПК'потенц потенційних користувачів на основі методу випадкового лісу (Random Forest), який включає попередню дискретизацію ідентифікаційних ознак з ДСПКпотенц і ДСПКuk (рисунок 2.12).

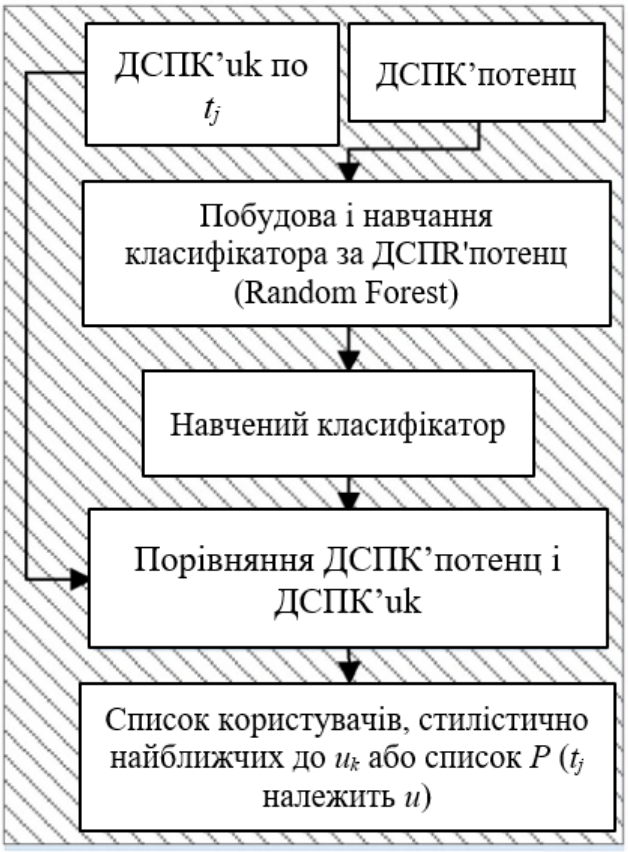

Рисунок 2.12 - Блок-схема послідовності кроків розробленого методу порівняння ДСПК з еталонними ДСПК потенційних користувачів

На третьому кроці визначається користувач, якому з найбільшою ймовірністю належить отримане електронне повідомлення, з числа потенційних користувачів шляхом порівняння ДСПК'uk з ДСПК'потенц. Шуканим користувачем є користувач, для якого $P(t_j$ належить $u_{max})$ є максимальним.

Визначається ідентифікатор користувача з найбільшою ймовірністю, що є автором повідомлення. І проводиться порівняння пред'явленого ідентифікатора з визначеним ідентифікатором; якщо ідентифікатори співпали, то надається доступ до ресурсу, тобто до розміщення повідомлення. В іншому випадку

робиться позначка про спробу НСД, після чого адміністратор системи обробляє дані цієї позначки.

Слід зазначити, що ідентифікація Інтернет-користувачів під час інформаційного обміну електронними повідомленнями може виконуватись у формі верифікації, аутентифікації або розпізнавання.

При верифікації підтверджується ідентичність ДСПКuk, сформованого за повідомленням t_j, та еталонного шаблона користувача u_k, що зберігається в базі даних.

Аутентифікація підтверджує відповідність ДСПКuk, сформованого за повідомленням t_j, одному з шаблонів, що зберігаються в базі даних.

При розпізнаванні, якщо отриманий ДСПКuk і один із збережених шаблонів виявляються однаковими, то система ідентифікує користувача з відповідним шаблоном.

3 ЕКСПЕРИМЕНТАЛЬНІ ДОСЛІДЖЕННЯ ТА ОЦІНКА РЕЗУЛЬТАТІВ

3.1 Вхідні дані експериментів

Для перевірки показників точності ідентифікації, отриманих при використанні розробленої методики, методів і КБМПІ, була проведена серія обчислювальних експериментів.

Окрім того, що Інтернет-тексти є досить короткими, існує друга проблема - нерівномірний розподіл кількості повідомлень по користувачах. Розроблена методика ідентифікації повинна дозволяти проводити ідентифікацію користувачів в умовах, коли існує дисбаланс між кількістю повідомлень різних користувачів.

Для того щоб наблизити експерименти до реальної ситуації, були складені набори тестових даних з різним рівнем незбалансованості (дисбалансом класів) та різною кількістю повідомлень на одного користувача. У ці набори були включені користувачі та повідомлення, відібрані випадковим чином. Було складено вісім груп ($Gtest_{num}$, $num = 8$) наборів тестових даних з різним рівнем незбалансованості та кількістю повідомлень. Таким чином, були сформовані дві групи наборів збалансованих даних та шість груп наборів незбалансованих даних, як показано в таблиці 3.1.

Таблиця 3.1 - Набори даних з різним рівнем незбалансованості

Рівень незбалансованості	Кількість повідомлень на користувача (/), мін.: макс.			
	Нормальна		Мала	
	Номер групи	*l*	Номер групи	*l*
Низький	$Gtest_1$	20:25	$Gtest_5$	8:10
Середній	$Gtest_2$	10:20	$Gtest_6$	5:10
Високий	$Gtest_3$	5:25	$Gtest_7$	2:10
Збалансований	$Gtest_4$	24:25	$Gtest_8$	10:10

Були сформовані два типи наборів даних. Перший з них моделює реальну ситуацію, коли є лише кілька повідомлень одного користувача. Другий моделює ситуацію більш оптимістично.

Набори даних з малою кількістю повідомлень:

1) збалансований (10 повідомлень на користувача);

2) низький рівень незбалансованості (мінімум 8 і максимум 10 повідомлень одного користувача);

3) середній рівень незбалансованості (мінімум 5 і максимум 10 повідомлень одного користувача);

4) високий рівень незбалансованості (мінімум 2 і максимум 10 повідомлень одного користувача).

Набори даних з нормальною кількістю повідомлень:

1) збалансований (25 повідомлень на користувача);

2) низький рівень незбалансованості (мінімум 20 і максимум 25 повідомлень одного користувача);

3) середній рівень незбалансованості (мінімум 10 і максимум 20 повідомлень одного користувача);

4) високий рівень незбалансованості (мінімум 5 і максимум 25 повідомлень одного користувача).

В незбалансованих наборах даних кількість повідомлень на користувача має нормальний розподіл (рисунки 3.1-3.3).

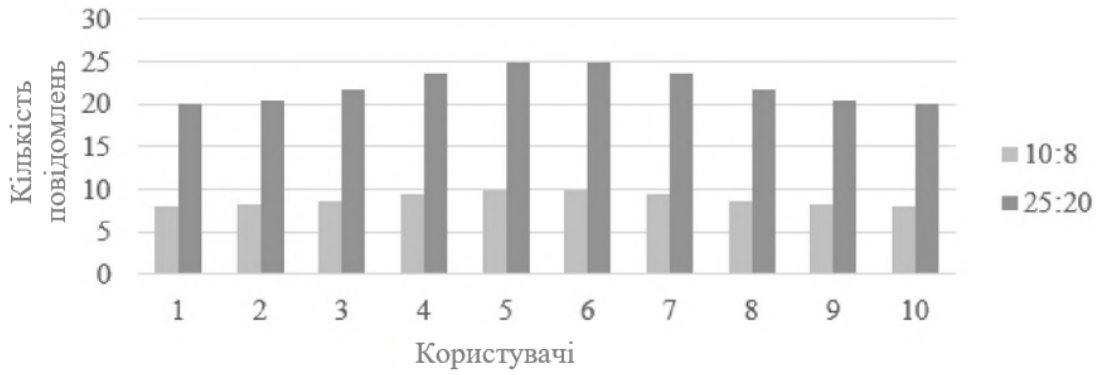

Рисунок 3.1 - Розподіл кількості повідомлень по користувачах у наборах даних з низьким рівнем незбалансованості для різної кількості навчальних прикладів

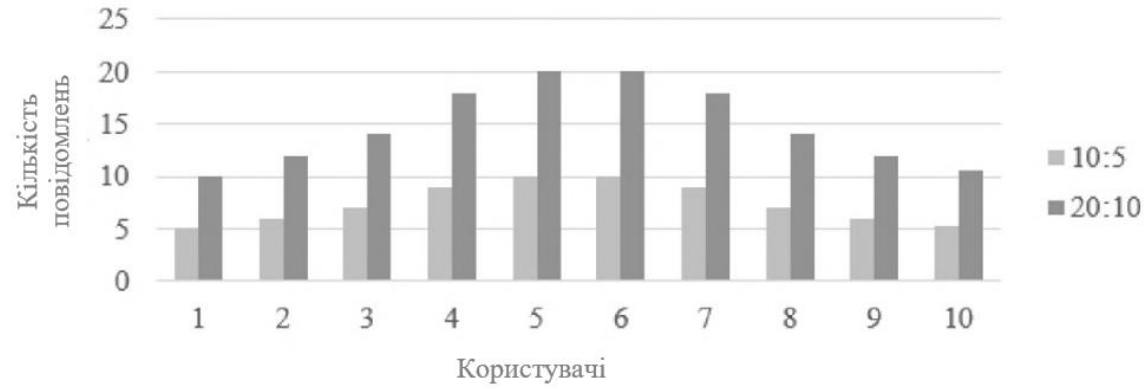

Рисунок 3.2 - Розподіл кількості повідомлень по користувачах у наборах даних із середнім рівнем незбалансованості для різної кількості навчальних прикладів

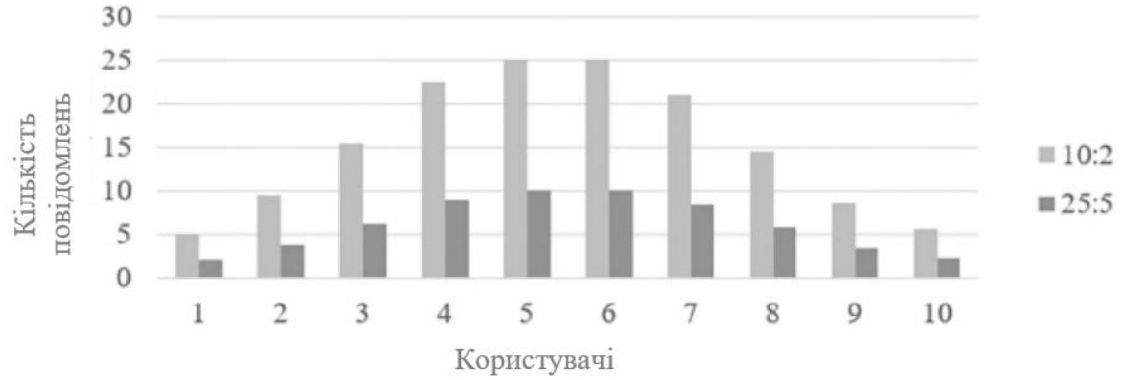

Рисунок 3.3 - Розподіл кількості повідомлень по користувачах у наборах даних з високим рівнем незбалансованості для різної кількості навчальних прикладів

Кожна з груп містить по 20 наборів тестових даних $Gtest_{num} = \{DS_1, \dots, DS_{dsnum}\}$, де $dsnum = 20$. Таким чином, загальна кількість наборів тестових даних становить 160. У свою чергу, кожен з наборів тестових даних DS_{dsnum} містить 10 потенційних користувачів $U_{потенц}$ та їх повідомлення T, де кожен із користувачів $U_{потенц} = \{u_1, \dots, u_k\}$, де $u_k = T_k = \{t_{k1}, \dots, t_{kl}\}$, l - кількість повідомлень користувача.

Всі повідомлення були випадкової довжини. Для того, щоб експерименти були максимально наближеними до реальної життєвої ситуації, відбір за тематикою також не проводився, хоча це дозволило б мінімізувати вплив тематики на класифікацію. Набори тестових даних були сформовані шляхом випадкової вибірки з бази даних.

У наведених нижче експериментах оцінка точності проводилась методом крос-валідації по 10 блоках, співвідношення навчальної та тестової вибірки - 90% та 10%.

3.2 Точність ідентифікації при використанні розробленої комплексної багаторівневої моделі представлення користувача

Мета експерименту - провести порівняльний аналіз точності ідентифікації з використанням розробленої КБМПК та різних сучасних методів класифікації, а також з точністю ідентифікації, отриманою попередніми дослідниками для повідомлень довжиною менше 5000 символів.

Проводилось дослідження точності ідентифікації при використанні розробленої КБМПК та різних методів класифікації, відібраних у ході теоретичного дослідження.

Оцінювалася точність ідентифікації (класифікації) для кожного з наборів даних, після чого розраховувалася середня точність для групи наборів (збалансований корпус, мала кількість текстів тощо).

Були отримані результати, представлені в таблицях 3.2-3.3 та на рисунках 3.4-3.5. На нормальній кількості навчальних прикладів випадковий ліс показав більш високу точність ідентифікації в середньому в порівнянні з методом опорних векторів на 10,35%, з деревами рішень на 14,42%, з наївним байєсівським класифікатором на 19,37%. Середня точність по всіх експериментах групи склала: SVM - 62,76%; DT - 58,7%; RF - 73,12%; NB - 53,75%.

На малій кількості навчальних текстів алгоритм RF показав найкращі показники точності в середньому в порівнянні з методом опорних векторів на 13,71%, з деревами рішень на 14,53%, з наївним байєсівським класифікатором на 17,08%. Середня точність по всіх експериментах склала: SVM - 51,04%; DT - 50,23%; RF - 64,75%; NB - 47,68%.

Таблиця 3.2 - Точність ідентифікації з використанням різних методів порівняння КБМПК при нормальній кількості текстів і різному рівні незбалансованості навчальної вибірки

Кількість текстів одного користувача (мін: макс)	Точність ідентифікації (A), %			
	SVM	DT	RF	NB
Збалансовані дані				
24:25	64,65	59,24	75,42	54,82
Слабо незбалансовані дані				
20:25	64,21	59,90	75,33	53,79
Середньо незбалансовані дані				
10:20	60,45	56,98	71,50	52,44
Сильно незбалансовані дані				
5:25	61,74	58,67	70,21	53,94

Таблиця 3.3 - Точність ідентифікації з використанням різних методів порівняння КБМПК при малій кількості текстів і різному рівні незбалансованості навчальної вибірки

Кількість текстів одного користувача (мін: макс)	Точність ідентифікації (A), %			
	SVM	DT	RF	NB
Збалансовані дані				
10:10	55,21	53,55	69,56	50,09
Слабо незбалансовані дані				
8:10	52,01	51,29	66,72	48,18
Середньо незбалансовані дані				
5:10	48,48	47,90	62,21	45,46
Сильно незбалансовані дані				
2:10	48,47	48,17	60,52	46,97

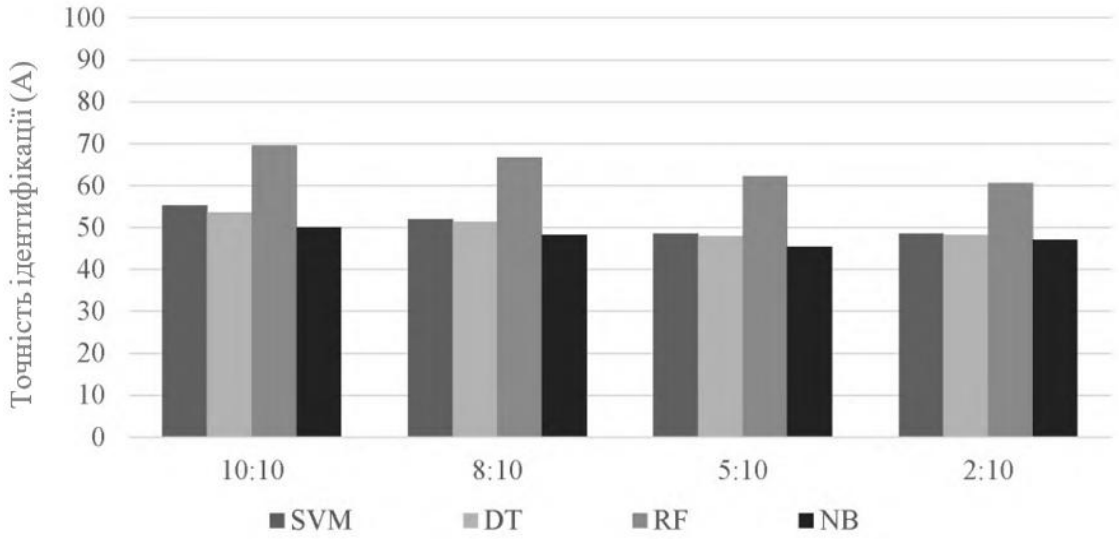

Рисунок 3.4 - Точність ідентифікації при малій кількості текстів навчальної вибірки

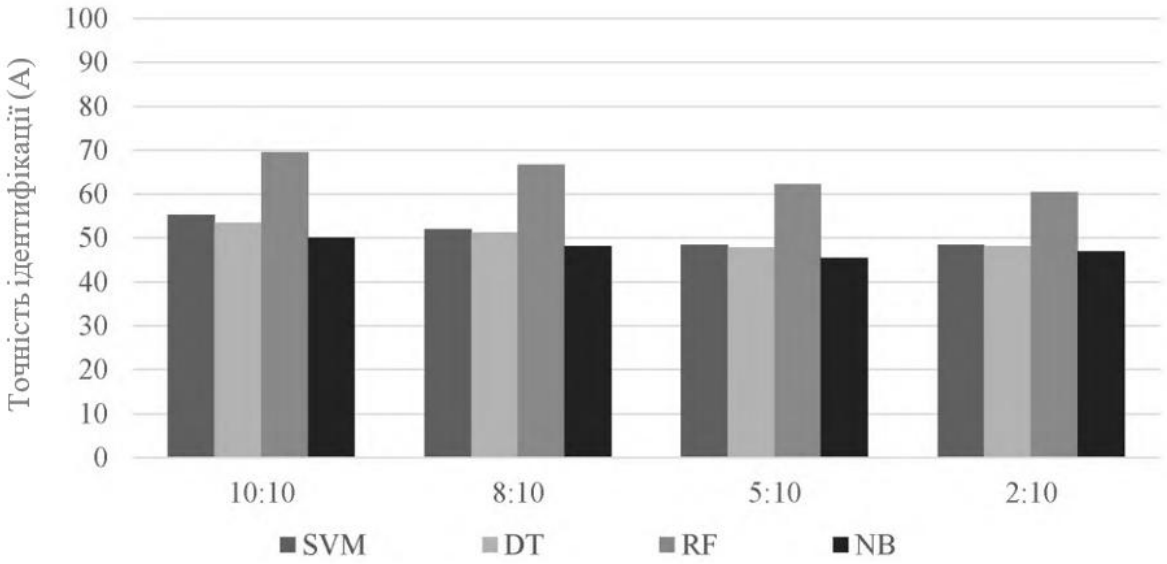

Рисунок 3.5 - Точність ідентифікації при нормальній кількості текстів навчальної вибірки

Оцінка точності ідентифікації з використанням розробленої КБМПК дозволяє зробити висновок, що розроблена КБМПК забезпечує вищу точність при використанні всіх розглянутих методів класифікації порівняно з точністю ідентифікації, отриманою іншими дослідниками для повідомлень довжиною менше 5000 символів (таблиця 3.4 і рисунок 3.6).

Таблиця 3.4 - Точність ідентифікації з використанням розробленої КБМПК та різних сучасних методів класифікації

Метод ідентифікації та простір ознак	Точність ідентифікації (А), %
Метод опорних векторів + найбільш уживані слова	47
Метод опорних векторів + Змішані ознаки	55
Метод опорних векторів + КБМПК	64,65
Наївний байєсівський класифікатор +	54,82
Випадковий ліс + КБМПК	Середнє 68,93

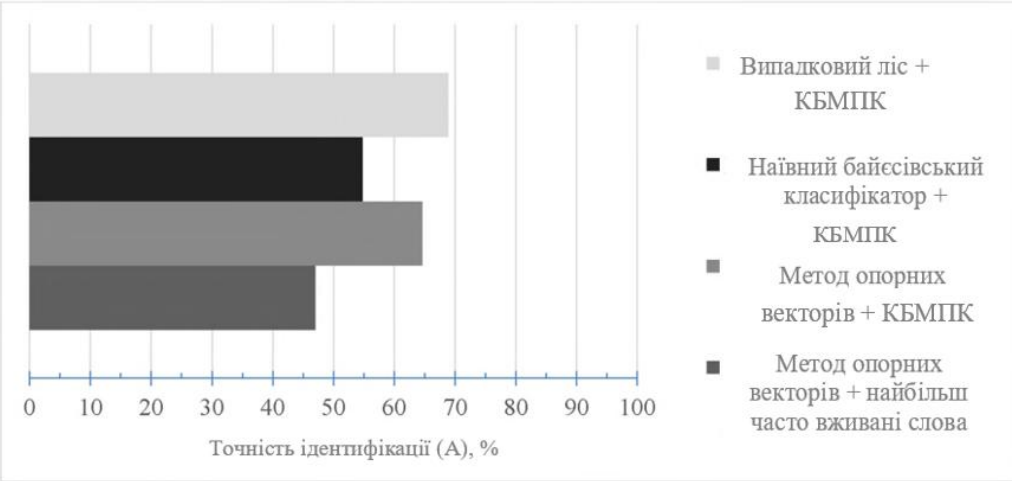

Рисунок 3.6 - Точність ідентифікації з використанням розробленої КБМПК та різних сучасних методів класифікації

Отримані результати не суперечать результатам, отриманим іншими дослідниками, але є значно кращими, що підтверджує якість розробленої КБМПК.

Точність ідентифікації з використанням алгоритму випадкового лісу для малого обсягу текстів становить у середньому 64,75%, для нормального обсягу текстів — 73,12%, а для всіх експериментів — 68,93%.

Розроблена КБМПК дозволяє підвищити точність ідентифікації в середньому на 21,93% порівняно з точністю, отриманою при використанні методу опорних векторів і найбільш уживаних слів.

3.3 Точність ідентифікації при використанні методу формування динамічних стилістичних профілів користувачів

Мета експерименту — оцінка точності ідентифікації при використанні методу формування динамічних стилістичних профілів користувачів у порівнянні з базовою КБМПК та відбором найбільш інформативних ознак для всіх користувачів.

Для перевірки ефективності розробленої КБМПК та методу відбору інформативних ознак при формуванні динамічних стилістичних профілів користувачів (ДСПК) було проведено серію експериментів. Оцінювалися точність класифікації при використанні динамічного стилістичного профілю користувача і при використанні статичного набору ідентифікаційних ознак (m=100), отриманого на основі аналізу ознак повідомлень 166 користувачів. Збільшення кількості ідентифікаційних ознак не має позитивного впливу на точність ідентифікації. Результати експериментів наведені в таблиці 3.5 та на рисунку 3.7.

Використання динамічної кількості ознак, що розраховується для кожного набору користувачів, підвищує точність класифікації на 0,93% порівняно зі статичним набором ознак, а порівняно з КБМПК – на 4,08%. Найбільший вплив динамічний відбір має в умовах, коли доступна лише мала кількість текстів (приріст точності порівняно з базовою КБМПК – 5,04%, порівняно зі статичним набором – 1,3%).

Найбільш доцільним для використання є запропонований у роботі метод формування динамічних стилістичних профілів користувачів, оскільки він забезпечив найвищу точність для переважної більшості рівнів незбалансованості та кількості повідомлень.

Таблиця 3.5 - Точність ідентифікації з використанням розробленого методу формування динамічних стилістичних профілів користувачів у порівнянні з базовою КБМПК та відбором найбільш інформативних ознак для всіх користувачів

Кількість текстів	Точність ідентифікації (А), %		
	Повна КБМПК	Динамічний відбір інформативних ознак	Відбір інформативних ознак для всіх
Збалансовані дані			
10:10	69,56	74	72,59
24:25	75,42	78,17	77,09
Слабо незбалансовані дані			
8:10	66,72	71,72	70,32
20:25	75,33	78,03	76,63
Середньо незбалансовані дані			
5:10	62,21	67,66	66,88
10:20	71,5	75,01	74,65
Сильно незбалансовані дані			
2:10	60,52	65,79	64,2
5:25	70,21	73,76	74,35

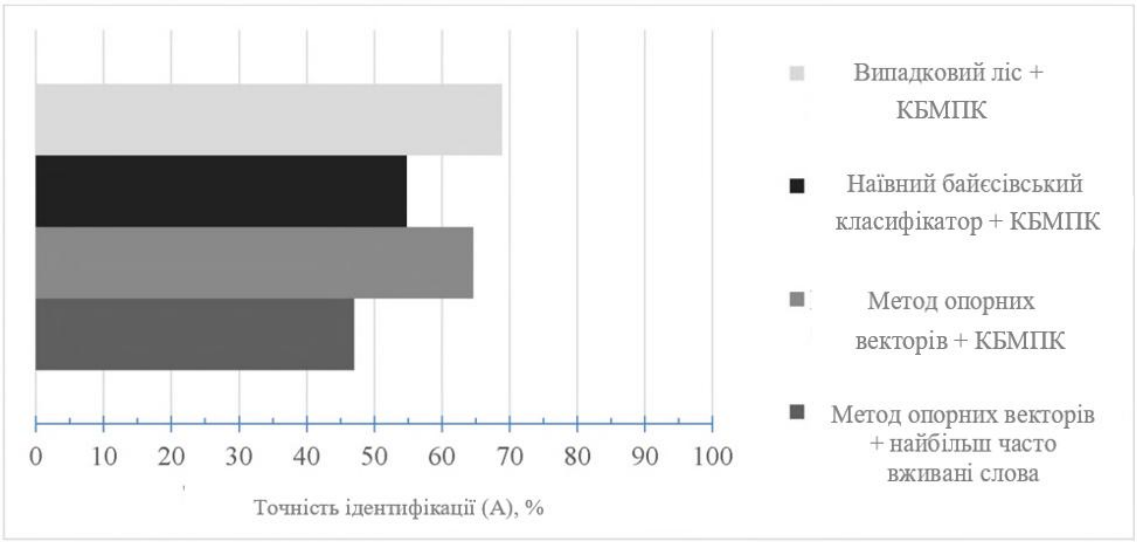

Рисунок 3.7 - Точність ідентифікації з використанням розробленого методу формування динамічних стилістичних профілів користувачів у порівнянні з базовою КБМПК та відбором найбільш інформативних ознак для всіх користувачів

Результати проведених експериментів підтверджують теоретичне припущення, що метод формування ДСПК (використання динамічної кількості та складу ідентифікаційних ознак) дозволяє підвищити точність ідентифікації та використовувати як ідентифікаційні ознаки значно меншу кількість ознак, ніж у початковому наборі.

3.4 Точність ідентифікації з використанням різних методів порівняння ДСПК при різній кількості текстів і різному рівні незбалансованості навчальної вибірки

Мета експерименту — оцінка точності ідентифікації з використанням різних методів порівняння ДСПК при різній кількості текстів та рівні незбалансованості навчальної вибірки, а також обгрунтування вибору алгоритму випадковий ліс як методу порівняння ДСПК з еталонними ДСПК.

Під час теоретичного дослідження було висунуто гіпотезу, що потребує експериментального підтвердження, про те, що для задачі класифікації текстів за користувачами може застосовуватися алгоритм випадковий ліс.

Було проведено серію експериментів з метою точнішої оцінки ефективності (точності) роботи відібраних методів. Результати наведено нижче (таблиці 3.6-3.7 і рисунок 3.8).

Найвищу точність ідентифікації для всіх рівнів незбалансованості та кількості повідомлень було досягнуто з використанням методу порівняння ДСПК на основі алгоритму випадковий ліс. Його було обрано для включення до методу ідентифікації користувачів. Середній виграш у точності порівняно з методом опорних векторів становить 7,92%, з методом дерев рішень — 17,41%, з наївним байєсівським класифікатором — 14,81%. Цей метод також показав себе досить ефективним за часом роботи.

Аналіз проведених експериментів дозволив підтвердити гіпотезу і обгрунтовано вибрати метод випадкового лісу для розв'язання задачі порівняння ДСПК.

Таблиця 3.6 - Точність ідентифікації з використанням різних методів порівняння ДСПК за нормальної кількості текстів та різної незбалансованості навчальної вибірки

Кількість текстів одного користувача (мін: макс)	Точність ідентифікації (A), %			
	SVM	DT	RF	NB
Збалансовані				
24:25	69,17	60,05	78,17	58,21
Слабко незбалансовані				
20:25	68,76	60,92	78,03	58,22
Середньо незбалансовані				
10:20	66,15	58,19	75,01	57,71
Сильно незбалансовані				
5:25	66,93	59,86	73,76	58,35

Таблиця 3.7 - Точність ідентифікації з використанням різних методів порівняння ДСПК за невеликої кількості текстів і різної незбалансованості навчальної вибірки

Кількість текстів одного користувача (мін: макс)	Точність ідентифікації (A), %			
	SVM	DT	RF	NB
Збалансовані дані				
10:10	66,01	54,61	74	59,17
Слабко незбалансовані дані				
8:10	63,38	52,62	71,72	59,55
Середньо незбалансовані дані				
5:10	59,1	49,26	67,66	56,69
Сильно незбалансовані дані				
2:10	6	4	6	5

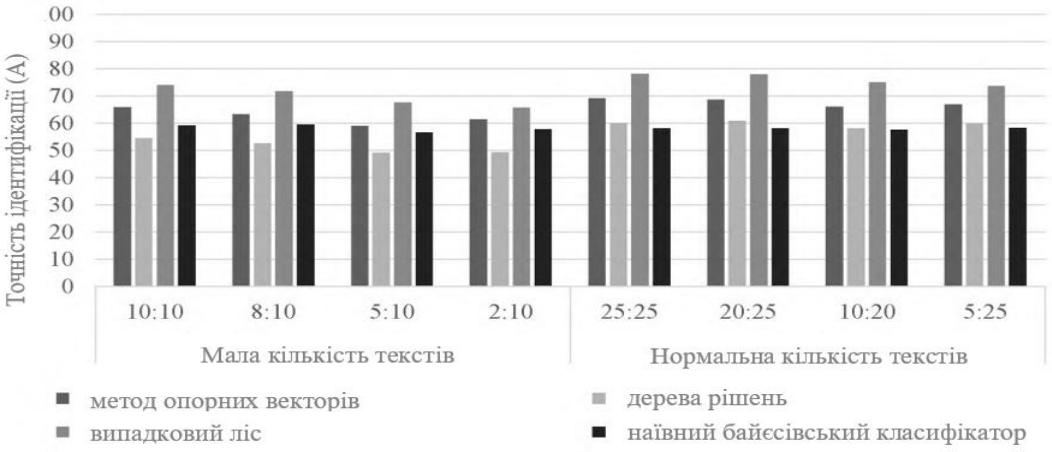

Рисунок 3.8 - Точність ідентифікації з використанням різних методів порівняння ДСПК за різної кількості текстів і різного рівня незбалансованості навчальної вибірки

3.5 Підвищення точності ідентифікації шляхом попередньої дискретизації ідентифікаційних ознак з ДСПК

Була проведена серія експериментів для оцінки впливу дискретизації ідентифікаційних ознак на точність порівняння ДСПК з використанням алгоритму випадковий ліс для ідентифікації веб-користувача. Перша серія експериментів була виконана на наборах даних із середньою кількістю текстів навчальної вибірки.

У всіх експериментах точність після дискретизації значно вища; максимальне зростання точності досягається на сильно незбалансованих наборах даних — 5,31%. Вплив дискретизації ознак на різних групах наборів даних наведено в таблиці 3.8 і на рисунку 3.9.

Другая серія експериментів була проведена для оцінки впливу дискретизації ознак на точність ідентифікації в ситуації, коли доступна мала кількість навчальних текстів на одного користувача. Як і в попередніх експериментах, найбільше підвищення точності спостерігалося на сильно незбалансованих даних; точність збільшилася на 14,58%. В середньому точність після дискретизації склала 79,7%, що на 9,95% вище в експериментах на даних до дискретизації (середня точність 69,8%) (таблиця 3.9 і рисунок 3.10).

Таблиця 3.8 - Порівняння точності ідентифікації до і після дискретизації ознак в умовах нормальної кількості текстів навчальної вибірки

	Сильно незбалансовані дані	Середньо незбалансовані дані	Слабо незбалансовані дані	Збалансовані дані
До дискретизації ознак	73,76	75,01	78,03	78,17
Після дискретизації ознак	81,62	82,2	81,15	81,23

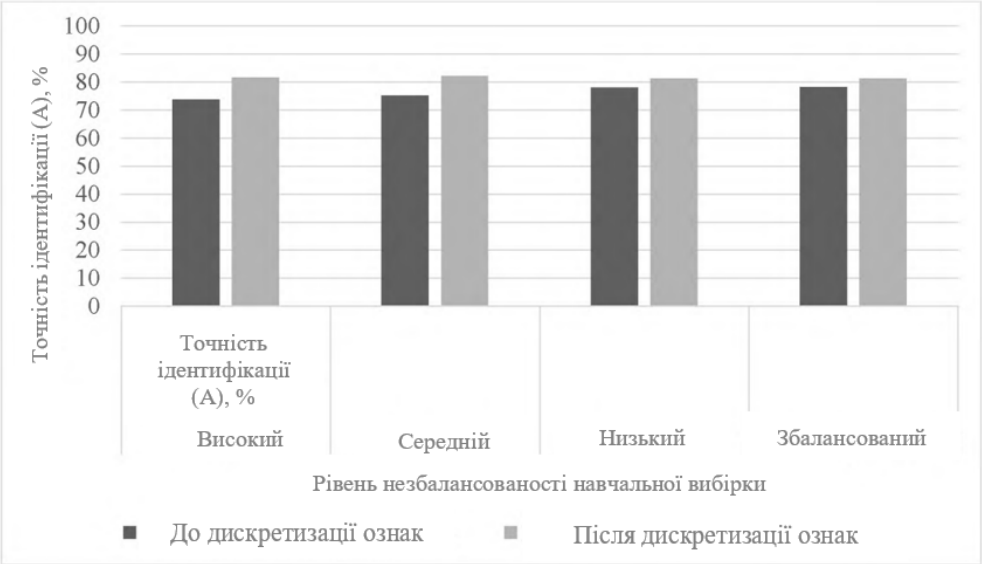

Рисунок 3.9 - Точність ідентифікації до і після дискретизації ознак в умовах нормальної кількості текстів навчальної вибірки

Проведені експерименти показали, що застосування запропонованого підходу — дискретизації ідентифікаційних ознак з ДСПК, дозволяє підвищити точність ідентифікації в середньому на 7,63%. Найбільший вплив дискретизація має при малому обсязі текстів та найвищому рівні їх незбалансованості, що має високу практичну значущість — 14,58% (рисунок 3.11).

Таблиця 3.9 - Порівняння точності ідентифікації до і після дискретизації ознак в умовах малого обсягу текстів навчальної вибірки

	Сильно незбалансовані дані	Середньо незбалансовані дані	Слабо незбалансовані дані	Збалансовані дані
До дискретизації ознак	65,79	67,66	71,72	74
Після дискретизації ознак	80,37	79,01	80,07	79,52

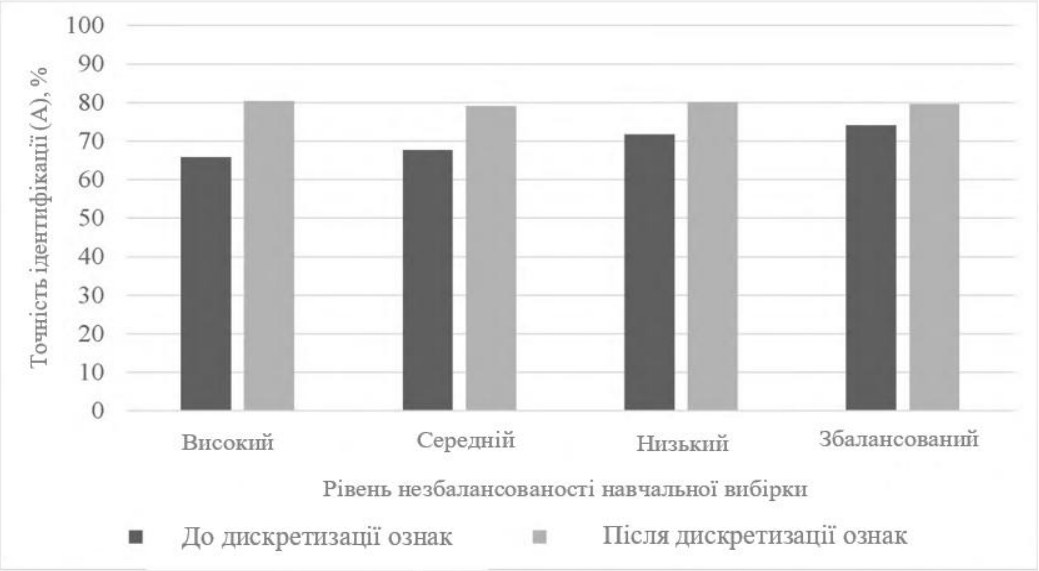

Рисунок 3.10 - Точність ідентифікації до і після дискретизації ознак в умовах малого обсягу текстів навчальної вибірки

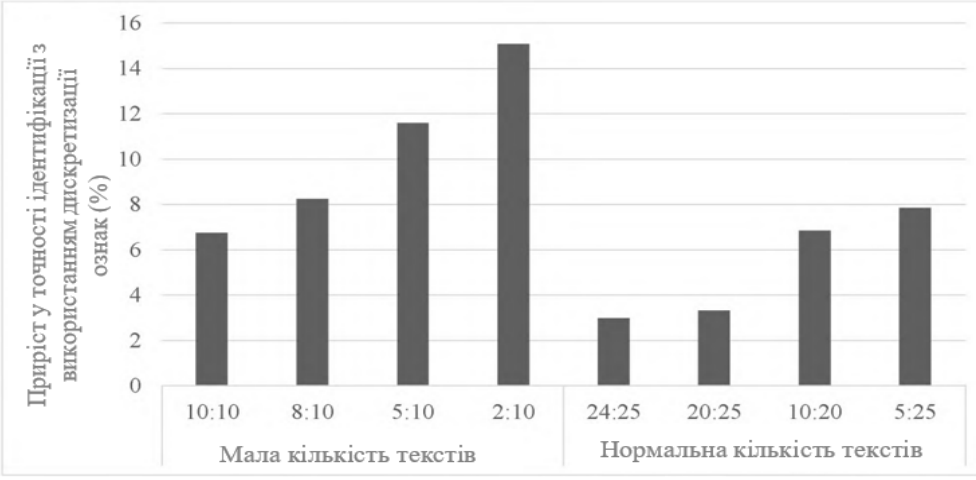

Рисунок 3.11 - Приріст у точності ідентифікації за рахунок попередньої дискретизації ідентифікаційних ознак з ДСПК

3.6 Визначення підсумкової точності ідентифікації на основі запропонованої методики

Проводився порівняльний аналіз точності ідентифікації з використанням розробленої методики на повідомленнях різної максимальної довжини. Аналіз показав, що методика дозволяє виконувати ідентифікацію Інтернет-користувача за електронними повідомленнями довжиною менше 500 символів з точністю 66,73%, а на текстах довжиною менше 1000 символів точність склала 71,7%. З рисунка 3.12 явно видно, що з підвищенням максимальної довжини повідомлення точність також зростає.

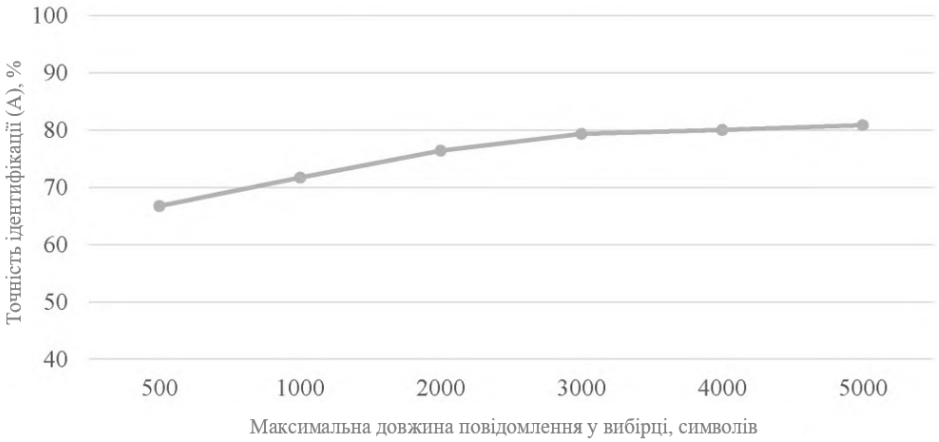

Рисунок 3.12 - Точність ідентифікації при використанні повідомлень різної довжини

Порівняльний аналіз підсумкової точності ідентифікації Інтернет-користувача за характеристиками електронних повідомлень з використанням розробленої методики проводився на текстах випадкової довжини. Було відібрано 3700 повідомлень від 166 користувачів з максимальною довжиною 4986 символів. Усі повідомлення мали довільну довжину. Більшість текстів були довжиною до 498 символів. Розподіл довжин повідомлень, використаних при оцінці точності ідентифікації за розробленою методикою, наведено на рисунку 3.13. В умовах обмеженої довжини текстів, їх малої кількості та нерівномірного розподілу по користувачам розроблена методика забезпечує більш високу точність ідентифікації. Точність ідентифікації за базовою КБМПК на малій

кількості текстів суттєво нижча за точність на нормальній кількості, тоді як після застосування всіх розроблених методів точність ідентифікації є однаково високою (таблиця 3.10 і рисунок 3.14).

Рисунок 3.13 - Розподіл довжин повідомлень, використаних при оцінці точності ідентифікації за розробленою методикою

Таблиця 3.10 - Точності ідентифікації з використанням розробленої методики

Кількість текстів одного користувача (мін: макс)	Точність ідентифікації (A), %
Збалансовані дані	
24:25	80,85
10:10	78,18
Слабко незбалансовані дані	
20:25	80,63
8:10	78,19
Середньо незбалансовані дані	
10:20	80,9
5:10	78,71
Сильно незбалансовані дані	
5:25	79,68
2:10	80,44

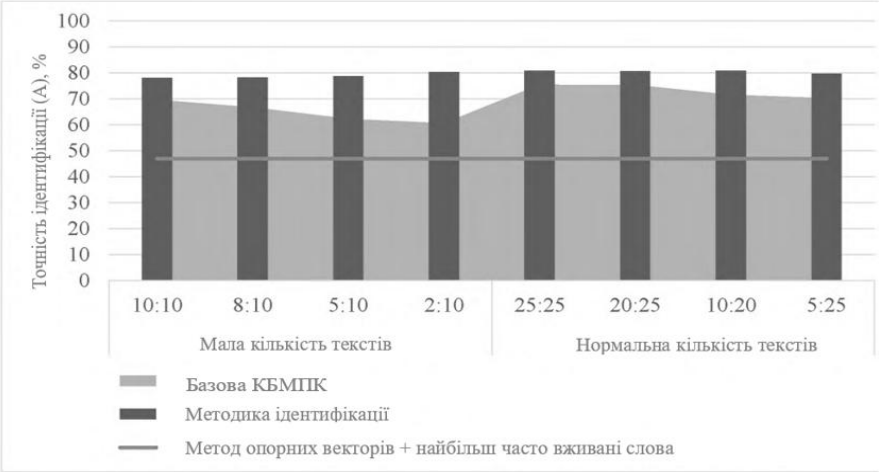

Рисунок 3.14 - Порівняння точності ідентифікації за розробленою методикою та відомими методами

Запропонована методика дозволяє проводити ідентифікацію Інтернет-користувача (в умовах обмеженої довжини текстів і незбалансованості навчальної вибірки) із середньою точністю на малому обсязі текстів - 78,88%, на нормальному обсязі текстів - 80,52% для електронних текстових повідомлень довжиною до 5000 символів. Остаточна середня точність 79,7% приблизно на 32,7% вища, ніж у існуючих статистичних методів.

3.7 Використання результатів дослідження для підвищення безпеки інформаційних процесів в мережі Інтернет

Методика ідентифікації Інтернет-користувача може бути використана для підвищення безпеки інформаційних процесів в мережі Інтернет. Методика дозволяє здійснювати ідентифікацію та аутентифікацію Інтернет-користувачів. Розроблені методи ідентифікації за стилістичними або лінгвістичними ознаками можуть бути застосовані як частина системи розмежування доступу. Лінгвістична ідентифікація є спорідненою біометричній ідентифікації.

Об'єктом доступу в даному випадку є сервіс розміщення електронного повідомлення на певному веб-порталі, що знаходиться в відкритому доступі в Інтернеті. Перед публікацією інформації у відкритому доступі повинна бути пройдена процедура додаткової прихованої ідентифікації та аутентифікації

користувача. Ця процедура необхідна для визначення повноважень користувача на розміщення електронного повідомлення. Допускається розміщення електронного повідомлення у разі успішної ідентифікації/аутентифікації, а також якщо раніше користувачем не було вчинено протиправних дій з ресурсами веб-порталу.

Розроблена методика може використовуватися для запобігання доступу зловмисників до створення та поширення електронних текстових повідомлень, зокрема:

- анонімного доступу без проходження процедури ідентифікації та аутентифікації;
- доступу під вигаданими іменами;
- неправомірних дій від імені легального користувача.

До застосування методики основні дії СРД включають:

1) введення ідентифікатора користувачів (логін, ім'я користувача) та ідентифікуючої інформації (зазвичай, пароль). Етап може бути доповнений двофакторною ідентифікацією;

2) перевірка введених даних та встановлення автентичності пред'явленого ідентифікатора;

3) у разі успішної аутентифікації - надання доступу до Інтернет-ресурсу та доступу до сервісів публікації електронних повідомлень. У цьому випадку СРД не може розрізнити, хто саме пред'явив ідентифікатор та ідентифікуючу інформацію. Оперуючи лише наданими ідентифікаційними даними, СРД не розрізняє легальних користувачів та зловмисників.

Застосування розробленої методики в СРД до сервісів публікації електронних повідомлень дозволяє доповнити стандартні механізми ідентифікації/автентифікації процедурою додаткової прихованої ідентифікації за такою схемою:

1) введення ідентифікатора користувачів (логін, ім'я користувача) та ідентифікуючої інформації (зазвичай, пароль). Етап може бути доповнений двофакторною ідентифікацією;

2) перевірка введених даних та встановлення автентичності пред'явленого ідентифікатора;

3) у разі успішної аутентифікації - надання доступу до Інтернет-ресурсу та доступу до сервісів публікації електронних повідомлень;

4) введення електронного повідомлення з метою його публікації в відкритий доступ;

5) додаткова прихована ідентифікація за надісланим повідомленням.

У разі успішного проходження додаткової прихованої ідентифікації користувач отримує доступ до розміщення повідомлення. Інакше, якщо користувач раніше був помічений у вчиненні протиправних дій, доступ до розміщення повідомлення не надається, або може бути зроблено позначку про спробу порушення прав доступу.

Таким чином, прихована ідентифікація є додатковим засобом захисту для забезпечення інформаційної безпеки інформаційних процесів, що відбуваються в Інтернеті.

Розроблені методи та КБМПК можуть використовуватися для виявлення факту порушення цілісності електронного повідомлення (підміни авторства).

Можна виділити суміжні завдання, в яких може бути застосована розроблена методика:

1) встановлення та перевірка авторства, підтвердження або спростування авторства певної особи;

2) перевірка того факту, що людина, яка опублікувала текст, є його справжнім автором.

Також пропонована методика ідентифікації Інтернет-користувача ефективно працює та показує стабільно високі результати, коли інші способи ідентифікації неможливі. Наприклад, у випадку, коли інцидент уже стався, але на Інтернет-порталі не велося обліку характеристик користувача. Також можлива ідентифікація користувачів, якщо відсутня можливість впровадження додаткових компонентів обліку та реєстрації на сервері Інтернет-порталу.

Методику можна використовувати як частину системи протидії інформаційно-психологічному впливу для ідентифікації джерела впливу.

Методика дозволяє проводити ідентифікацію користувача, а не програмного та апаратного оточення, і може бути затребувана спеціальними службами, які проводять експертизи в комп'ютерній криміналістиці.

ВИСНОВКИ

1. Здійснено аналіз сучасного стану проблеми ідентифікації Інтернет-користувачів при інформаційному обміні електронними повідомленнями, що дозволило обгрунтувати необхідність розробки методів ідентифікації Інтернет-користувачів шляхом використання лінгвістичних і стилістичних характеристик їхніх текстів.

2. Розроблено модель загроз безпеки інформаційних процесів при обміні електронними повідомленнями з використанням інтернет-ресурсів, що дозволило встановити типи моделей ймовірних порушників інформаційної безпеки.

3. Розроблено методику ідентифікації Інтернет-користувача на основі стилістичних і лінгвістичних характеристик коротких електронних повідомлень, що дозволило встановити основні етапи методики ідентифікації Інтернет-користувача.

4. Розроблено методи визначення та підвищення точності ідентифікації з використанням різних методів порівняння динамічного стилістичного профілю користувача при різній кількості текстів і різному рівні незбалансованості навчальної вибірки.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Бутузов В. Протидія комп'ютерній злочинності в Україні (системно-структурний аналіз): моногр. К.: КИТ, 2010. 148 с.
2. Лісовська Ю. Кібербезпека. Ризики та заходи. К.: Кондор, 2019. 272 с.
3. Дубов Д.В. Кіберпростір як новий вимір геополітичного суперництва: монографія. К.: НІСД, 2014. С. 36.
4. Бельський Ю. Щодо визначення поняття кіберзлочину. Юридичний вісник. 2014. № 6. С. 414–418.
5. Погорецький М. Кіберзлочини: до визначення поняття. Вісник прокуратури. 2012. № 8. С. 89–96.
6. Петров В.В. Щодо формування національної системи кібербезпеки України. Стратегічні пріоритети: [наук.-аналіт. щокварт. зб.]. Нац. ін-т стратег. дослідж. К.: НІСД, 2013. № 4 (29). С. 127–130.
7. Лужецький В.А. Інформаційна безпека: навч. посіб. Вінниця: УНІВЕРСУМ-Вінниця, 2009. 240 с.
8. Мельник А.О., Мохун С.В., Волошин К.В. Моделювання систем виявлення шкідливих додатків для Android. Збірник матеріалів проблемної наукової міжгалузевої конференції «Автоматизація та комп'ютерно-інтегровані технології» (АКІТ-2022). Тернопіль, 2022. С.59-61.
9. Бурячок В.Л., Толубко В.Б., Хорошко В. О., Толюпа С.В. Інформаційна і кібербезпека: соціотехнічний аспект: Підручник. К.: ДУТ, 2015. 288 с.
10. Eckersley P. How unique is your web browser? International Symposium on Privacy Enhancing Technologies Symposium. - Springer Berlin Heidelberg, 2010. С. 1-18.
11. Grcar M. User profiling: Web usage mining. Proc. 7th International Multiconference Information Society IS. 2004. С 79-82.
12. Ivancsy R. Analysis of Web User Identification Methods. International Journal of Computer Science. 2007. Vol. 2, №. 3. P. 212.

13. Бурячок В.Л. Кібернетична безпека – головний фактор сталого розвитку сучасного інформаційного суспільства. Сучас. спец. техніка. 2011. № 3 (26). С. 104–114.

14. Бурячок В.Л., Гулак Г.М., Толубко В.Б. Інформаційний та кіберпростори: проблеми безпеки, методи та засоби боротьби: Підручник. К.: ДУТ, 2015. 449 с.

15. Волокітін А.В., Маношкин А.П., Солдатенков А.В., Савченко С.А., Петров Ю.А. Інформаційна безпека державних організацій і комерційних фірм. К.: Юніор, 2012. 303 с.

16. Juola P. Authorship attribution. Foundations and Trends in Information Retrieval. 2006. Vol. 1(3). P. 233-334.

17. Rosenblum N., Zhu X., Miller B.P. Who Wrote This Code? Identifying the Authors of Program Binaries. Computer Security - ESORICS 2011: Lecture №tes in Computer Science. 2011. Vol. 6879. P. 172-189.

18. Iqbal F., Binsalleeh H., Fung B.C.M., Debbabi M. A unified data mining solution for authorship analysis in anonymous textual communications. Information Sciences. 2013. Vol. 231. P. 98-112.

19. Уніченко С., Лисик Г., Закревська М. Метод випадкового лісу для ідентифікації користувача на основі стилістичних характеристик електронних текстів. Збірник матеріалів науково - практичної конференції молодих вчених, аспірантів та студентів «Кібербезпека та комп’ютерно-інтегровані технології» (КБКІТ-2024), Тернопіль, 2024. С.158-160.

20. Уніченко С., Зарихта О. Метод лінгвістичної ідентифікації користувачів на основі стилістичних характеристик текстів електронних повідомлень. Матеріали науково-практичного симпозіуму «Захист інформації». Тернопіль, 2024. С.111-113.

21. Diederich J., Kindermann J., Leopold E., Paass G. Authorship Attribution with SuPort Vector Machines. Applied Intelligence. 2003. Vol. 19, №. l. P. 109-123.

22. de Vel O., Anderson A., Corney M., Mohay G. Mining e-mail content for author identification forensics. Newsletter ACM SIGMOD Record. 2001. Vol. 30. №4. P. 55-64.

23. Zheng R., Li J., Huang Z., Chen H. A Framework for Authorship Identification of Online Messages: Writing Style Features and Classification Techniques. Journal of the American Society for Information Science and Technology (JASIST). 2006. Vol. 57, №. 3. P. 378-393.

24. Houvardas J., Stamatatos E. N-gram feature selection for authorship identification. Lecture №tes in Computer Science. Artificial Intelligence: Methodology, Systems, and Applications. 2006. Vol. 4183. P. 77-86.

25. Sanderson C, Guenter S. Short text authorship attribution via sequence kernels, Markov chains and author unmasking: An investigation. Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing. 2006. P. 482-491.

26. Maitra P., Ghosh S., Das D. Authorship Verification - An Approach based on Random Forest. Notebook for PAN at CLEF 2015. 2015.

27. Pacheco M.L., Fernandes K., Porco A. Random Forest with Increased Generalization: A Universal Background Approach for Authorship Verification. Notebook for PAN at CLEF 2015. 2015.

28. Peng F., Schuurmans D., Keselj V., Wang S. Language independent authorship attribution using character level language models. Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics. 2003. Vol. 1. P. 267-274.

29. Qian T., Liu B., Chen L., Peng Z. Tri-Training for Authorship Attribution with Limited Training Data. ACL. 2014. Vol. 2. P. 345-351.

30. Sapkota U., Bethard S., Montes-y-G6mez M., Solorio T. №t All Character N-grams Are Created Equal: A Study in Authorship Attribution. HLT-NAACL. 2015. P. 93-102.

31. Abbasi A., Chen H. Writeprints: A stylometric aProach to identity-level identification and similarity detection in cyberspace. ACM Transactions on Information Systems. 2008. Vol. 26(2), №. 7.

32. McCombe N. Methods of author identification. BA (Mod) CSLL Final Year Project, TCD. 2002.

33. Corney M., Anderson A., Mohay G., de Vel. O. Identifying the authors of

suspect email [Електронний ресурс]. 2001. Режим доступу: http://eprints.qut.edu.au/8021/1/CompSecurityPaper.pdf.

34. Stamatatos E. Authorship attribution based on feature set subspacing ensembles // International Journal on Artificial Intelligence Tools. 2006. Vol. 15, №5. P. 823-838.

35. Stamatatos E. Author identification using imbalanced and limited training texts. 18th International Workshop on Database and Expert Systems APlications. 2007. P. 237-241.

36. Keselj V., Thomas C, Peng F., Cercone N. N-gram-based author profiles for authorship attribution. Pacific Association for Computational Linguistics, PACLING'03. 2003. P. 255-264.

37. Frantzeskou G., Stamatatos E., Gritzalis S. Identifying Authorship by Byte-Level N-Grams:The Source Code Author Profile (SCAP) Method. International Journal of Digital Evidence. 2007. Vol. 6, №. 1. P. 1-18.

38. Luyckx K., Daelemans W. Personae: a Corpus for Author and Personality Prediction from Text. LREC. 2008.

39. Layton R., Watters P., Dazeley R. Authorship attribution for twitter in 140 characters or less. Second Cybercrime and Trustworthy Computing Workshop(IEEE). 2010. P. 1-8.

40. Forsyth R., Holmes D. Feature-finding for text classification. 1996. Vol. 11, №. 4. P. 163-174.