Хмельницький національний університет
Факультет інформаційних технологій
Кафедра комп'ютерних наук

КВАЛІФІКАЦІЙНА РОБОТА БАКАЛАВРА

на тему *Метод інтелектуального визначення авторства текстів за стилем написання*

Галузь знань *12 – Інформаційні технології*
Шифр і назва галузі знань

Спеціальність *122 – Комп'ютерні науки*
Шифр і назва спеціальності

Освітня програма *Комп'ютерні науки*
Назва освітньої програми

Виконала: *студент групи КН-20-2* (Група виконавця) _____ (Підпис) *Богдан ШПОРТ* (Ім'я, ПРІЗВИЩЕ)

Керівник: *ст. викладач каф. КН* (Науковий ступінь, посада) _____ (Підпис) *Тетяна СКРИПНИК* (Ім'я, ПРІЗВИЩЕ)

Нормоконтроль: *к.т.н., доц. каф. КН* (Науковий ступінь, посада) _____ (Підпис) *Руслан БАГРІЙ* (Ім'я, ПРІЗВИЩЕ)

До захисту допускаю:
зав. кафедри КН, д.т.н., професор _____ (Підпис) *Олександр БАРМАК* (Ім'я, ПРІЗВИЩЕ)
24 06 2024 р.

Хмельницький 2024

ХМЕЛЬНИЦЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ

Факультет *інформаційних технологій*
Кафедра *комп'ютерних наук*
Освітній ступінь *бакалавр*
Галузь знань *12 – Інформаційні технології*
Спеціальність *122 – Комп'ютерні науки*

ЗАТВЕРДЖУЮ
Завідувач кафедри *комп'ютерних наук*

_____
(підпис)
д.т.н., професор Олександр БАРМАК
«*16*» *02* *2024* року

ЗАВДАННЯ
НА КВАЛІФІКАЦІЙНУ РОБОТУ БАКАЛАВРА

1. Тема кваліфікаційної роботи бакалавра: *«Метод інтелектуального визначення авторства текстів за стилем написання»*

2. Завдання видано студенту *Богдану ШПОРТУ*
(Ім'я, прізвище)

3. Керівник роботи *ст. викладач кафедри КН Тетяна СКРИПНИК*
(посада, ім'я, прізвище)

4. Затверджено наказом університету від «*15*» *02* *2024* р. № *8*

5. Дата видачі завдання студенту: «*16*» *02* *2024* р.

6. Зміст пояснювальної записки (перелік задач) та вихідні дані:

Мета роботи – спрощення роботи систем експертизи за рахунок автоматизованого визначення авторства текстів за стилем написання. Також слід вирішити такі завдання: виконати аналіз інформаційних моделей області інтелектуального визначення авторства текстів за стилем написання; створити метод інтелектуального визначення авторства текстів за стилем написання; описати інформаційну структуру системи для інтелектуального визначення авторства текстів за стилем написання; обрати набір даних для інтелектуального визначення авторства текстів; створити відповідну програмну реалізацію на основі створеного методу; виконати тестування створеного програмного забезпечення; виконати дослідження ефективності створеного методу з використанням розробленого ПЗ.

7. Календарний план виконання кваліфікаційної роботи бакалавра:

№	Назва етапів (розділів) кваліфікаційної роботи бакалавра	Термін виконання	Примітка
1	*Вибір напряму дослідження та узгодження тематики кваліфікаційної роботи бакалавра з керівником, складання календарного графіка виконання роботи*	*січень 2024*	виконано
2	*Ознайомлення з предметною областю, формулювання мети та задач дослідження, визначення об'єкта та предмета дослідження*	*лютий 2024*	виконано
3	*Проектування та розробка загальної архітектури програмного забезпечення, інтерфейсу користувача, вибір засобів реалізації програмного забезпечення*	*березень 2024*	виконано
4	*Створення та тестування програмного забезпечення*	*квітень 2024*	виконано
5	*Написання пояснювальної записки, урахування зауважень керівника, оформлення згідно вимог*	*травень 2024*	виконано
6	*Розробка презентаційних матеріалів та попередній захист кваліфікаційної роботи*	*травень 2024*	виконано
7	*Отримання відгуку керівника, рецензії, перевірка на плагіат, нормоконтроль*	*червень 2024*	виконано
8	*Підготовка до захисту та захист кваліфікаційної роботи бакалавра*	*червень 2024*	виконано

Виконавець: *студент групи КН-20-2* (Група виконавця) _____ (Підпис) *Богдан ШПОРТ* (Ім'я, ПРІЗВИЩЕ)

Керівник: *ст. викладач каф. КН* (Науковий ступінь, посада) _____ (Підпис) *Тетяна СКРИПНИК* (Ім'я, ПРІЗВИЩЕ)

Анотація

Тема кваліфікаційної роботи бакалавра: *«Метод інтелектуального визначення авторства текстів за стилем написання»*

Виконавець кваліфікаційної роботи бакалавра: *студент групи КН-20-2 Богдан ШПОРТ*

Керівник кваліфікаційної роботи бакалавра: *ст. викладач кафедри КН Тетяна СКРИПНИК*

Кваліфікаційна робота бакалавра містить:

Пояснювальна записка				Кількість додатків
Сторінок	Рисунків	Таблиць	Джерел інформації	
66	25	9	35	4

Метою кваліфікаційної роботи бакалавра є спрощення роботи систем експертизи за рахунок автоматизованого визначення авторства текстів за стилем написання. Для розробки інформаційної системи було використано мову програмування C# та систему керування базами даних MS SQL Server, також використано бібліотеку машинного навчання ML.NET.

Розроблена система призначена для адміністраторів соціальних мереж, а також може використовуватись в наукових цілях. Реалізована автоматизація визначення авторства текстів за стилем написання дозволяє відслідковування фактів несанкціонованих текстових запозичень, а також інтелектуальне відстеження зміни поведінки взламаних акаунтів користувачів.

Напрямками практичного використання розробленої інформаційної системи визначено автоматизоване визначення авторства текстів за стилем написання.

Ключові слова: визначення авторства, машинне навчання, інтелектуальне визначення авторства текстів.

Виконавець: *студент групи КН-20-2* _____ *Богдан ШПОРТ*

Група виконавця | Підпис | Ім'я, ПРІЗВИЩЕ

Зміст

Перелік скорочень

Скорочення, термін, позначення	Пояснення
NLP	Natural Language Processing
ПЗ	Програмне забезпечення
ШІ	Штучний інтелект
TF-IDF	Term frequency - Inverse document frequency
SVM	Support vector machines
RNN	Рекурентні нейронні мережі
LSTM	Long short-term memory
BiLSTM	Bidirectional LSTM
GRU	Gated recurrent units
МН	Машинне навчання
БД	База даних
L-BFGS	Limited-memory BFGS
ПІБ	Прізвище, ім’я, по-батькові
ML	Machine Learning
.NET	Network Enabled Technology
SQL	Structured Query Language
СКБД	Система керування базами даних
КРБ	Кваліфікаційна робота бакалавра
КН	Комп’ютерні науки

Вступ

Кваліфікаційна робота бакалавра присвячена спрощенню роботи систем експертизи за рахунок автоматизованого визначення авторства текстів за стилем написання.

Актуальність. Актуальність застосування методів інтелектуального визначення авторства текстів за стилем написання зростає в умовах сучасного інформаційного суспільства, де цифрова комунікація є важливою складовою багатьох сфер життя. Зокрема, одним із ключових завдань є інтелектуальне відстеження зміни поведінки взламаних акаунтів користувачів. Виявлення таких змін дозволяє своєчасно ідентифікувати випадки компрометації акаунтів та запобігти потенційним негативним наслідкам, зокрема поширенню дезінформації або шахрайству.

Пошук першоджерел та каналів розповсюдження пропаганди є ще однією важливою областю застосування методів інтелектуального визначення авторства текстів. У сучасному світі, де інформаційні війни та пропаганда стали звичними явищами, ідентифікація джерел пропагандистських матеріалів є критично важливою для забезпечення інформаційної безпеки та формування об'єктивної суспільної думки. Технології аналізу стилю написання текстів можуть суттєво сприяти виявленню та блокуванню пропагандистських ресурсів.

Відслідковування фактів несанкціонованих текстових запозичень також є важливою проблемою, особливо в академічному середовищі, де дотримання принципів академічної доброчесності є фундаментальним. Використання методів інтелектуального аналізу текстів для виявлення плагіату дозволяє не тільки захистити авторські права, але й забезпечити справедливість і об'єктивність у наукових дослідженнях. Таким чином, розробка та вдосконалення методів інтелектуального визначення авторства текстів є актуальною та важливою задачею в сучасному цифровому світі.

Об’єкт дослідження – процес визначення авторства текстів за стилем написання NLP-засобами.

Предмет дослідження – методи та засоби машинного навчання для роботи з текстовою інформацією.

Мета кваліфікаційної роботи бакалавра – спрощення роботи систем експертизи за рахунок автоматизованого визначення авторства текстів за стилем написання.

Завдання кваліфікаційної роботи бакалавра – Провести огляд теоретичних підходів, а також обрати підхід для інтелектуального визначення авторства текстів за стилем написання. Виконати аналіз існуючих публікацій за напрямком дослідження. Провести аналіз існуючого програмного забезпечення області інтелектуального визначення авторства текстів за стилем написання; створити метод інтелектуального визначення авторства текстів за стилем написання. Описати інформаційну структуру системи для інтелектуального визначення авторства текстів за стилем написання. Обрати набір даних для інтелектуального визначення авторства текстів. Створити відповідну програмну реалізацію на основі створеного методу. Виконати тестування створеного програмного забезпечення. Виконати дослідження ефективності створеного методу інтелектуального визначення авторства текстів за стилем написання з використанням розробленого ПЗ.

Розділ 1 Характеристика предметної області визначення авторства текстів за стилем написання

1.1 Аналіз інформаційних моделей

Визначення авторства текстів за стилем написання є важливою задачею в багатьох галузях людської діяльності, від літературознавства до юриспруденції та кібербезпеки. Розвиток методів інтелектуального аналізу текстів, зокрема за допомогою штучного інтелекту та машинного навчання, відкриває нові можливості для точного та ефективного визначення авторства. У наукових дослідженнях визначення авторства є ключовим для забезпечення академічної доброчесності та належної атрибуції ідей. Плагіат є серйозною проблемою, яка може призвести до втрати довіри до наукової спільноти. Тому точне визначення авторства допомагає підтримувати високу якість публікацій та захищати інтелектуальну власність [1].

У літературі та журналістиці визначення авторства також має велике значення. У випадках анонімних або псевдонімних публікацій, знання про справжнього автора може змінити інтерпретацію тексту або навіть його значення. Ідентифікація автора дозволяє розкрити можливі упередження або особисті інтереси, що впливають на зміст тексту.

У юридичному контексті визначення авторства є необхідним для вирішення спорів щодо авторських прав [2]. Правильна атрибуція дозволяє встановити, хто має юридичні права на текст і, відповідно, хто може вимагати компенсації у випадку порушення цих прав. Крім того, аналіз стилю письма може бути корисним у криміналістиці, де такі методи допомагають у розслідуванні злочинів, пов’язаних з анонімними погрозами або наклепами [3].

В епоху цифрових технологій питання кібербезпеки набувають особливої актуальності. Визначення авторства текстів в інтернеті може допомогти виявити та зупинити дезінформацію, фейкові новини та інші види шкідливого контенту. У цьому контексті, методи аналізу стилю письма можуть бути ефективним

інструментом для визначення джерела шкідливої інформації та запобігання її поширенню.

У соціально-політичному контексті ідентифікація авторства текстів може мати значний вплив на громадську думку та політичні процеси. Анонімні блоги, коментарі у соціальних мережах та інші форми онлайн-комунікації часто використовуються для маніпуляції масовою свідомістю [4]. Застосування інтелектуальних методів для визначення справжніх авторів таких текстів може сприяти підвищенню прозорості та відповідальності у суспільстві.

Авторські права є важливою частиною інтелектуальної власності, яка захищає авторів оригінальних творів, включаючи літературні тексти, музичні композиції, художні роботи та інші форми вираження. Основною метою авторських прав є надання виключного права на використання та поширення їхніх творів [5]. На міжнародному рівні захист авторських прав регулюється декількома основними договорами та конвенціями:

– Бернська конвенція [6] про охорону літературних і художніх творів: один із найважливіших міжнародних договорів, що забезпечує автоматичний захист авторських прав у країнах-учасницях без необхідності реєстрації.

– Всесвітня організація інтелектуальної власності [7]: ця організація підтримує міжнародну співпрацю у сфері інтелектуальної власності та адмініструє декілька важливих договорів, включаючи Бернську конвенцію.

В Україні авторське право регулюється законом «Про авторське право і суміжні права» [8], який був прийнятий у 1993 році і оновлювався кілька разів для узгодження з міжнародними стандартами. Нижче наведено основні положення українського законодавства про авторські права.

Авторське право виникає автоматично в момент створення твору і не потребує реєстрації. Проте, для підтвердження права автор може зареєструвати твір в Державному агентстві з питань інтелектуальної власності. Автор має виключне право на використання твору, включаючи право на його відтворення, поширення, публічне виконання та адаптацію. В Україні авторське право діє протягом життя автора та 70 років після його смерті. Також закон передбачає

відповідальність за порушення авторських прав, включаючи відшкодування збитків та штрафи.

В епоху цифрових технологій питання кібербезпеки набувають особливої актуальності. Визначення авторства текстів в інтернеті може допомогти виявити та зупинити дезінформацію, фейкові новини та інші види шкідливого контенту. У цьому контексті, методи аналізу стилю письма можуть бути ефективним інструментом для визначення джерела шкідливої інформації та запобігання її поширенню.

Визначення авторства текстів за стилем написання є важливою та багатогранною задачею, яка має значний вплив у різних сферах. Інтелектуальні методи аналізу стилю письма дозволяють підвищити точність та ефективність цього процесу, сприяючи забезпеченню академічної чесності, захисту авторських прав, підвищенню кібербезпеки та підтриманню соціальної справедливості. Це робить дослідження та розвиток методів визначення авторства надзвичайно актуальним напрямом сучасної науки і технологій.

1.2 Огляд теоретичних підходів щодо задачі визначення авторства текстів за стилем написання

Визначення авторства текстів за їх стилем написання є актуальною темою у різноманітних сферах, включаючи літературознавство, цифровий гуманітарний аналіз, правоохоронну діяльність та інформаційну безпеку. Ця проблематика спирається на теорії лінгвістики, статистики та машинного навчання для розробки методів, що дозволяють визначити авторство тексту на основі його унікального стилю. Відмінності в стилі письма між різними авторами можуть виявитися в різноманітних лінгвістичних аспектах, таких як вживання лексики, синтаксичні структури, а також унікальні манери вираження та побудови тексту [9].

Пошук ефективних методів ідентифікації авторства тексту вимагає інтеграції різних міждисциплінарних підходів та використання різноманітних

аналітичних інструментів. Це включає аналіз частоти вживання слів, структурні особливості тексту, семантичні зв’язки, а також використання технік машинного навчання для побудови моделей класифікації [10]. Дослідження в цій області спрямоване на розробку алгоритмів, які були б не лише ефективними, а й стійкими до різних стилістичних варіацій та здатними до застосування в реальних умовах, що відповідають вимогам наукової об’єктивності та практичної застосовності.

Штучний інтелект у контексті визначення авторства текстів за стилем написання являє собою технологію, що здатна аналізувати та інтерпретувати лінгвістичні особливості текстів для ідентифікації їх автора. ШІ використовує різні алгоритми та моделі для розпізнавання патернів у текстах, які характерні для певного автора. Це включає аналіз частоти використання слів, синтаксичних конструкцій, стилістичних особливостей та інших лінгвістичних характеристик [11].

Одним із ключових компонентів ШІ у цьому контексті є машинне навчання, яке дозволяє системам ШІ навчатися на основі великих обсягів текстових даних. За допомогою методів машинного навчання, таких як класифікація, кластеризація та аналіз тексту, системи можуть автоматично розпізнавати характерні особливості письма різних авторів [12]. Це включає як традиційні статистичні методи, так і сучасні нейронні мережі, які здатні обробляти складні лінгвістичні структури.

Машинне навчання – це підгалузь штучного інтелекту, яка зосереджується на розробці алгоритмів і статистичних моделей, що дозволяють комп'ютерам виконувати завдання без явного програмування [13]. У контексті визначення авторства текстів за стилем написання, машинне навчання використовує різні методи для аналізу текстових даних та побудови моделей, які можуть передбачати авторство на основі лінгвістичних особливостей (рисунок 1.).

Навчальні дані – це набір текстів, для яких відоме авторство. Ці дані використовуються для навчання моделей машинного навчання [15]. Великий та

якісний навчальний набір даних є критично важливим для успішного навчання моделей, оскільки вони дозволяють алгоритмам вивчати характерні особливості стилю кожного автора.

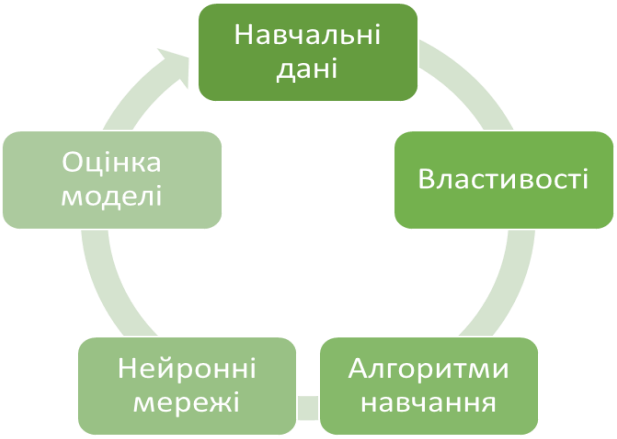


Рисунок 1.1 – Основні компоненти машинного навчання [14]

Властивості – це характеристики тексту, які використовуються для аналізу та класифікації [16]. У контексті визначення авторства, такими особливостями можуть бути частота вживання певних слів, середня довжина речень, синтаксичні структури, стильові маркери тощо. Ці особливості можуть бути визначені вручну (інженерія ознак) або автоматично виявлені моделями машинного навчання.

Існує багато алгоритмів, які можуть бути використані для навчання моделей. Логістична регресія є статистичним методом для бінарної класифікації, який використовується для передбачення ймовірності настання однієї з двох можливих подій. Це робиться шляхом застосування логістичної функції (сигмоїдної функції) до лінійної комбінації вхідних змінних. Логістична регресія добре працює з текстовими даними завдяки можливості обробляти великі обсяги високорозмірних даних, зокрема текстів, шляхом перетворення тексту на числові вектори, наприклад, за допомогою методів TF-IDF [17].

– Наївний баєсівський класифікатор є простим та ефективним алгоритмом для текстової класифікації, який базується на теоремі Баєса [18]. Він працює на припущенні, що всі особливості є незалежними одна від одної, що

нечасто буває в реальному житті, але на практиці цей підхід часто показує хороші результати.

SVM – це алгоритм класифікації, який знаходить гіперплощину в багатовимірному просторі, що найкраще розділяє дані на різні класи [19]. Основною ідеєю є знайти таку гіперплощину, яка має максимальну відстань до найближчих точок навчального набору, що забезпечує високу здатність до узагальнення.

Сучасні моделі машинного навчання часто використовують нейронні мережі, особливо для складних завдань. Деякі з них включають:

– рекурентні нейронні мережі, що використовуються для роботи з послідовними даними, такими як текст. Вони можуть зберігати контекст, що є важливим для аналізу стилю письма [20];

– LSTM, вдосконалення RNN, які можуть зберігати інформацію на довших часових відрізках, що є корисним для врахування довготривалих залежностей у тексті [21];

– багатоспрямовані LSTM, або BiLSTM. Розширення LSTM, які враховують контекст як з минулого, так і з майбутнього, що робить їх особливо ефективними для обробки тексту [22].

Після навчання моделі її необхідно оцінити для визначення точності та ефективності. Це включає такі метрики, як точність, повнота та F1-міра. Для цього використовуються тестові дані, які не входили до навчального набору.

Машинне навчання – це ітеративний процес. Моделі потребують постійного вдосконалення та перевірки з використанням нових даних для підтримання високої точності. Це включає додавання нових ознак, збільшення навчального набору та налаштування параметрів алгоритмів.

Таким чином, ШІ у сфері визначення авторства текстів за стилем написання є потужним інструментом, що поєднує в собі аналітичні можливості сучасних алгоритмів з обробкою великих обсягів даних. Це дозволяє не лише автоматизувати процес ідентифікації авторства, але й робить його більш точним

та надійним, що є важливим для застосувань у літературознавстві, журналістиці, правоохоронних органах та багатьох інших галузях.

Автори [23] пропонують використання GRU LSTM для вирішення задачі ідентифікації авторства текстів, написаних природною мовою. Для навчання моделей автори використовують наступну конфігурацію: вхідними даними для моделей є корпус авторів української літератури. Методами вбудовування є BoF та TFIDF. Найкращі результати показали наступні моделі: Decision Tree - 95,80 % точності, Multilayer Perceptron - 88,58 % точності. Що стосується моделей глибокого навчання, то вони показали досить схожі результати. За однаковий період навчання їхня точність коливається в межах 95%. GRU 94,63% LSTM 94,59%.

Автори [24] дослідили різні моделі глибокого навчання для ідентифікації авторства. В результаті було розроблено три моделі для ідентифікації авторства, а також по одній моделі для перевірки авторства на наборах даних C50 та Gutenberg. Найкраще з ідентифікацією авторства в статті впорався GRU, показавши точність 69,1% на базі даних C50 і 89,2% на базі даних Гутенберга. Крім того, мережа досягла точності верифікації 99,8% як для набору даних C50, так і для набору даних Gutenberg.

Штучний інтелект та машинне навчання становлять революційні інструменти для визначення авторства текстів, значно підвищуючи точність і ефективність цього процесу. Традиційні методи лінгвістичного аналізу, засновані на ручній ідентифікації стилістичних особливостей, мають обмеження, пов'язані з суб'єктивністю та трудомісткістю. У той же час, сучасні алгоритми ШІ та МН дозволяють автоматизувати та вдосконалити цей процес за рахунок глибокого аналізу текстових даних.

ШІ та МН дозволяють автоматично обробляти великі обсяги текстових даних, що значно скорочує час і ресурси, необхідні для аналізу. Завдяки можливості обробляти мільйони текстів, ці технології забезпечують високу масштабованість, що є особливо важливим у випадках великомасштабних досліджень або аналізу масових даних.

Моделі машинного навчання, такі як логістична регресія, наївний баєсівський класифікатор, дерева рішень і лісові моделі, а також SVM, демонструють високу точність у визначенні авторства текстів. Вони здатні враховувати складні лінгвістичні особливості та стилістичні патерни, що дозволяє більш точно ідентифікувати автора тексту.

Алгоритми ШІ можуть виявляти приховані стилістичні патерни, які можуть бути непомітні для людського ока. Це включає аналіз синтаксичних структур, морфологічних особливостей, частотного використання певних слів або фраз та інших лінгвістичних характеристик. Виявлення таких патернів є важливим для точного визначення авторства, особливо у випадках, коли автор намагається приховати свій стиль.

Моделі МН мають здатність навчатися на нових даних, що дозволяє їм адаптуватися до змін у стилі письма авторів або до появи нових авторів. Це забезпечує постійне вдосконалення моделей та їх відповідність сучасним вимогам.

Таким чином, ШІ та МН відкривають нові горизонти у сфері визначення авторства текстів, забезпечуючи високу автоматизацію, точність та адаптивність процесу. Ці технології значно підвищують ефективність наукових досліджень у лінгвістиці, літературознавстві, криміналістиці та інших галузях, де необхідно точно ідентифікувати авторство текстових матеріалів. Для інтелектуального визначення авторства текстів за стилем написання буде використано підхід машинного навчання.

1.3 Аналіз існуючих програмних засобів та наукових рішень до визначення авторства текстів за стилем написання

Для визначення авторства текстів за стилем написання існує кілька програм та онлайн-сервісів, які використовують інтелектуальні методи для визначення авторства текстів за стилем написання. Вони застосовують

різноманітні алгоритми машинного навчання та обробки природної мови для аналізу стилістичних особливостей текстів.

Authorship Attribution Tool від NeoNeuro [25] – це онлайн-інструмент, створений для визначення авторства текстів шляхом аналізу стилістичних ознак письма. Він використовує різні алгоритми машинного навчання, щоб порівнювати текст, що аналізується, з текстами відомих авторів і визначати ймовірного автора на основі стилістичної схожості. Цей інструмент є потужним засобом для дослідження авторства у випадках, коли ідентифікація автора має вирішальне значення, наприклад, у літературних студіях, криміналістиці або в контексті плагіату.

Рисунок 1.2 – Логотип компанії «NeoNeuro» [26]

Authorship Attribution Tool аналізує тексти на різних рівнях: від лексичних і синтаксичних ознак до більш складних стилістичних характеристик. Лексичний аналіз включає частоту вживання певних слів, словниковий запас автора, використання рідковживаних слів тощо. Синтаксичний аналіз охоплює структуру речень, середню довжину речень, частоту використання різних типів речень (наприклад, складних або складнопідрядних). Стилістичний аналіз включає визначення унікальних мовних зворотів, фразеологізмів та інших стилістичних особливостей, властивих конкретному автору.

Цей інструмент також може використовувати різні методи класифікації текстів, такі як наївний баєсівський класифікатор, підтримкові векторні машини та інші алгоритми машинного навчання, які дозволяють аналізувати великі обсяги текстових даних і робити висновки про авторство з високою точністю.

Крім того, Authorship Attribution Tool надає користувачам інтуїтивно зрозумілий інтерфейс для завантаження текстів і отримання результатів аналізу.

Однак, інструмент має і свої недоліки, наприклад, якщо база даних не охоплює всіх можливих авторів або стилів письма, точність визначення авторства може знизитися. Сервіс може бути менш точним при аналізі текстів, які містять велику кількість граматичних або орфографічних помилок, що може ускладнити ідентифікацію стилістичних ознак. Authorship Attribution Tool зосереджується переважно на стилістичних характеристиках тексту, не завжди враховуючи глибокі семантичні зв’язки, що можуть бути важливими для точного визначення авторства. Іноді інструмент може мати труднощі з адаптацією до нових стилів письма або авторів, які не були включені в початкову базу даних для навчання алгоритмів.

Наступна система для визначення авторства текстів – Writeprint. Це спеціалізована система для визначення авторства текстів, яка використовує передові методи машинного навчання для аналізу стилістичних особливостей письма [27]. Writeprint був розроблений з урахуванням необхідності аналізу текстів у складних випадках, таких як криміналістичні розслідування або літературні експертизи, де точність визначення авторства має вирішальне значення.

Writeprint аналізує тексти на декількох рівнях: лексичному, синтаксичному та стилістичному. На лексичному рівні система досліджує частоту використання слів, багатство словникового запасу та специфічні лексичні вибори, характерні для конкретного автора. На синтаксичному рівні аналізуються структура речень, використання граматичних конструкцій та патернів, що є типовими для певного стилю письма. Стилістичний рівень включає аналіз унікальних стилістичних ознак, таких як метафори, ідіоми, ритм та інші елементи, які визначають індивідуальний стиль автора.

Writeprint (рисунок 1.3) також використовує методи класифікації текстів, такі як класифікатори на основі дерев рішень, випадкових лісів та SVM. Ці

методи дозволяють точно ідентифікувати автора тексту шляхом виявлення і порівняння стилістичних ознак з базою даних відомих авторів.

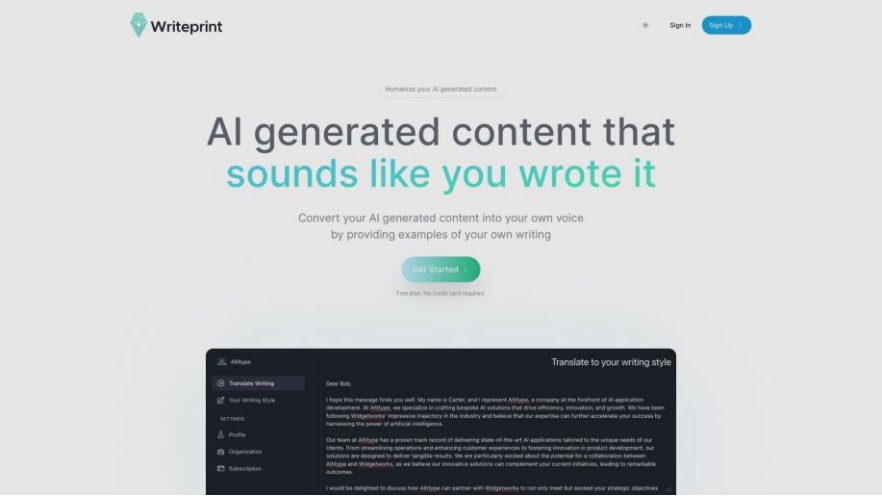


Рисунок 1.3 – Головна сторінка сайту «Writeprint»

Серед недоліків системи є складність налаштування, Writeprint може вимагати значних зусиль для налаштування та навчання моделей, особливо для користувачів, які не мають глибоких знань у галузі машинного навчання і стилометрії. Аналіз великих обсягів текстових даних за допомогою Writeprint може бути дороговартісним, вимагаючи потужного апаратного забезпечення для ефективної роботи. Writeprint може бути менш доступним для загального користування, оскільки часто використовується в специфічних наукових і криміналістичних контекстах. Інструмент має обмежену документацію та підтримку, що може створювати труднощі для нових користувачів у процесі налаштування та використання системи.

Таким чином, у результаті дослідження предметної області було визначено, що реалізація методу інтелектуального визначення авторства текстів за стилем написання є надзвичайно актуальним питанням. Це зумовлено зростаючою потребою у встановленні достовірності авторства в умовах цифрової епохи, де тексти поширюються швидко і часто анонімно. Такі методи дозволяють ефективно ідентифікувати авторів текстів, що є важливим для забезпечення захисту інтелектуальної власності, розслідування випадків

плагіату, підтвердження авторства в літературних та наукових творах, а також для аналізу комунікацій у соціальних мережах та інших онлайн-платформах.

Отож, із дослідження теоретичних підходів до розв'язку задачі інтелектуального визначення авторства текстів за стилем написання було встановлено, що моделі машинного навчання, такі як логістична регресія, наївний баєсівський класифікатор, дерева рішень, лісові моделі та SVM, демонструють високу точність у визначенні авторства текстів завдяки здатності враховувати складні лінгвістичні особливості та стилістичні патерни. Алгоритми ШІ ефективно виявляють приховані стилістичні патерни, включаючи аналіз синтаксичних структур, морфологічних особливостей і частотного використання певних слів, що важливо для точного визначення авторства, особливо коли автор намагається приховати свій стиль. Моделі машинного навчання здатні навчатися на нових даних, що дозволяє їм адаптуватися до змін у стилі письма або до появи нових авторів, забезпечуючи постійне вдосконалення та відповідність сучасним вимогам.

Актуальність і необхідність подальших розробок програмного забезпечення для інтелектуального визначення авторства текстів за стилем написання були продемонстровані на основі аналізу існуючих програмних продуктів. Описані інструменти надають потужні можливості для аналізу текстів та визначення авторства, кожен з них має свої унікальні особливості та недоліки, що дозволяє вибирати найбільш підходящий інструмент залежно від конкретних потреб і завдань.

Незважаючи на переваги цих інструментів, необхідно розробити продукт, який буде універсальним та адаптованим до специфічних вимог української мови та контексту. Отже, реалізація методу інтелектуального визначення авторства текстів за стилем написання є надзвичайно актуальною в сучасному світі, де питання авторства та достовірності інформації набувають все більшого значення у наукових дослідженнях, юридичних розслідуваннях, освітніх процесах та багатьох інших сферах.

1.4 Мета, задачі та вимоги до реалізації інформаційної системи

Метою кваліфікаційної роботи бакалавра є спрощення роботи систем експертизи за рахунок автоматизованого визначення авторства текстів за стилем написання.

Для досягнення поставленої мети слід вирішити такі завдання:

– виконати аналіз інформаційних моделей області інтелектуального визначення авторства текстів за стилем написання;

– виконати огляд теоретичних підходів, а також обрати підхід для інтелектуального визначення авторства текстів за стилем написання;

– виконати аналіз існуючих публікацій за напрямком дослідження;

– провести аналіз існуючого програмного забезпечення області інтелектуального визначення авторства текстів за стилем написання;

– створити метод інтелектуального визначення авторства текстів за стилем написання;

– описати інформаційну структуру системи для інтелектуального визначення авторства текстів за стилем написання;

– обрати набір даних для інтелектуального визначення авторства текстів;

– створити відповідну програмну реалізацію на основі створеного методу;

– виконати тестування створеного програмного забезпечення;

– виконати дослідження ефективності створеного методу інтелектуального визначення авторства текстів за стилем написання з використанням розробленого ПЗ.

Розділ 2 Створення методу інтелектуального визначення авторства текстів за стилем написання

2.1 Кроки методу інтелектуального визначення авторства текстів

Метод інтелектуального визначення авторства текстів за стилем написання призначений для інтелектуального відстеження зміни поведінки взламаних акаунтів користувачів, а також може бути використаний для відслідковування фактів несанкціонованих текстових запозичень. Кроки методу наведені на рисунку 2.1.

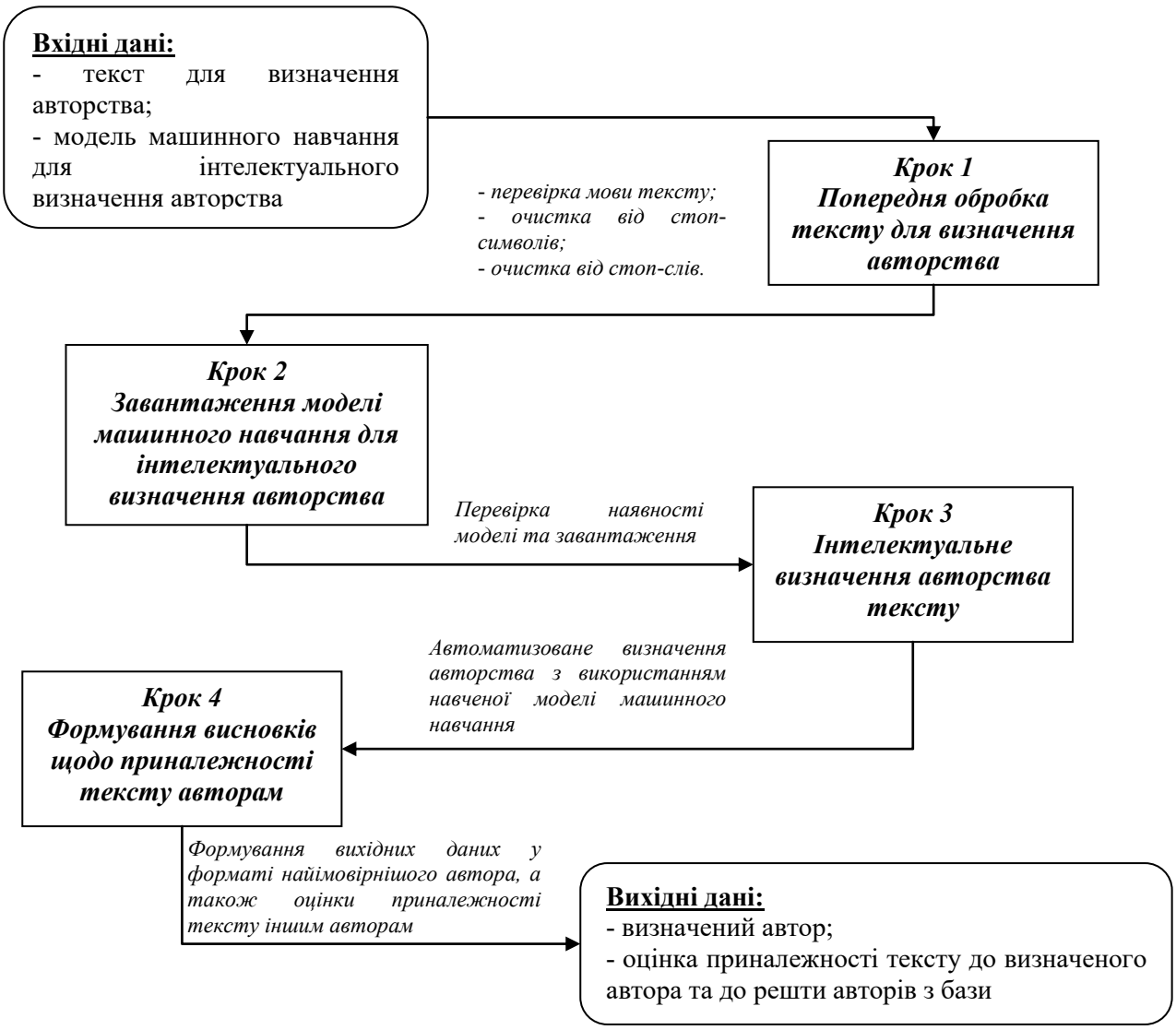


Рисунок 2.1 – Кроки методу інтелектуального визначення авторства текстів

Вхідними даними методу є текст для визначення авторства та попередньо натренована модель машинного навчання для інтелектуального визначення авторства.

На першому кроці здійснюється попередня обробка тексту для визначення авторства, яка включає в себе перевірку мови тексту, а також очистку від стоп-символів та від стоп-слів. Метод передбачає роботу з англомовними текстами.

Другим кроком є завантаження моделі машинного навчання для інтелектуального визначення авторства, яка є вхідними даними методу. При завантаженні також здійснюється перевірка, чи модель завантажена коректно.

На кроці інтелектуального визначення авторства тексту відбувається автоматизоване визначення авторства з використанням завантаженої навченої моделі машинного навчання.

І останнім кроком є крок формування висновків щодо приналежності тексту авторам, який включає формування вихідних даних у форматі найімовірнішого автора, а також оцінки приналежності тексту іншим авторам.

Відповідно, вихідними даними методу є визначений автор та оцінка приналежності тексту до визначеного автора, а також до решти авторів з бази.

Оскільки вхідними даними методу є попередньо натренована модель машинного навчання для інтелектуального визначення авторства за текстом, то дану модель попередньо потрібно навчити та оцінити її спроможність інтелектуального визначення авторства за текстом. На рисунку 2.2 наведено кроки отримання навченої моделі машинного навчання для інтелектуального визначення авторства.

Вхідними даними є набір текстів для навчання, поділений на авторів. Відповідно, набір даних повинен бути таким, щоб на кожного автора припадало щонайменше по 50 текстів.

На першому кроці здійснюється попередня обробка текстів, яка включає в себе перевірку мови тексту, а також очистку від стоп-символів та від стоп-слів. Якщо мова поточного тексту не англійська, такий текст видаляється.

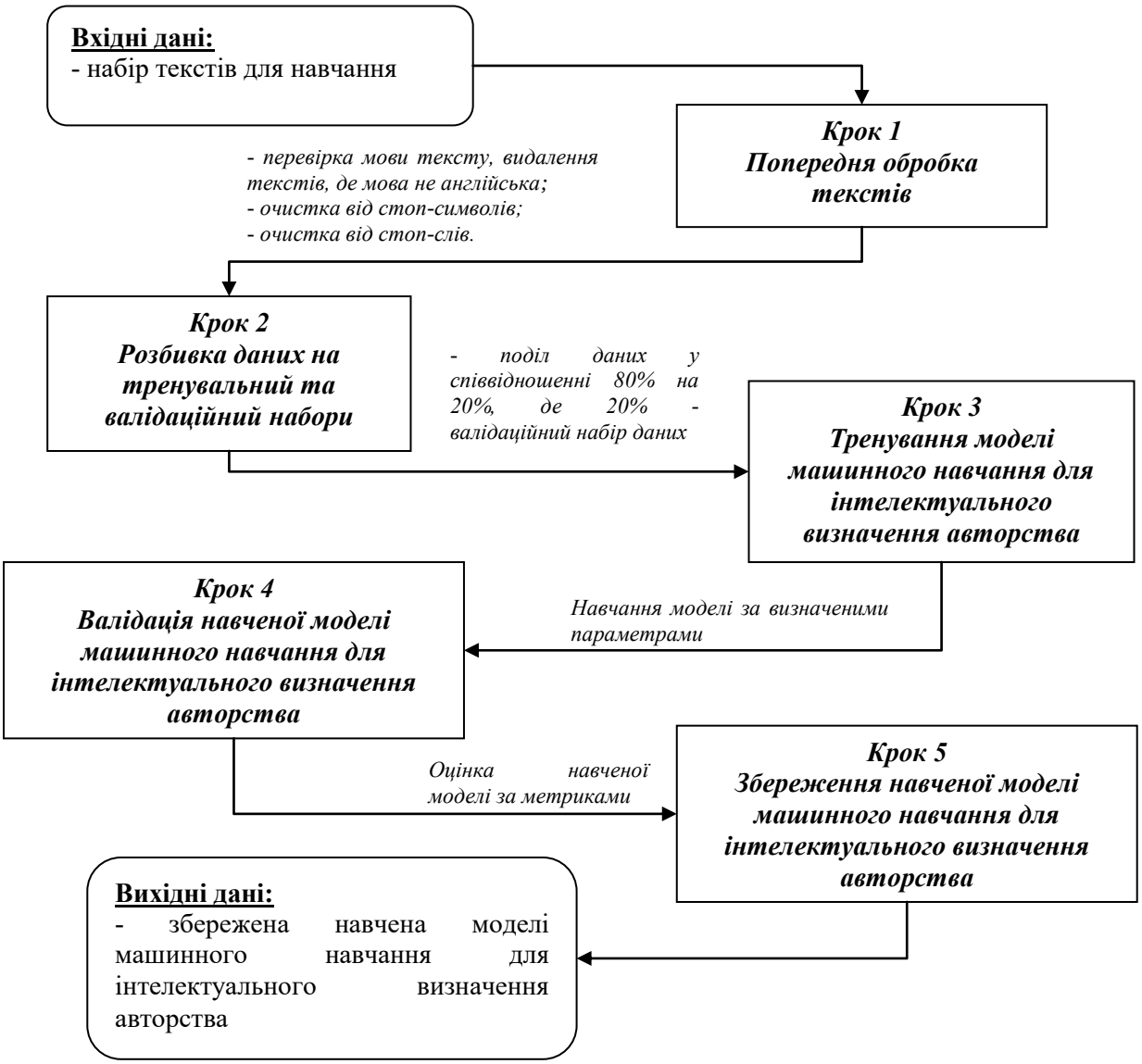


Рисунок 2.2 – Кроки отримання навченої моделі машинного навчання для інтелектуального визначення авторства

Наступним кроком відбувається розбивка даних на тренувальний та валідаційний набори. Діляться дані співвідношенні 80% на 20%, де 20% – валідаційний набір даних, а 80%, відповідно – навчальний.

Наступним кроком здійснюється тренування моделі машинного навчання для інтелектуального визначення авторства. Тренування залежить від гіпепараметрів, які підбираються емпіричним шляхом.

Після завершення тренування, навчену модель необхідно провалідувати. Валідність моделі відбувається шляхом оцінок навченої моделі за метриками точності, влучності та повноти.

Останнім кроком є збереження навченої моделі машинного навчання для інтелектуального визначення авторства за текстом, відповідно, вихідними даними є збережена навчена моделі машинного навчання для інтелектуального визначення авторства.

Отже, запропонований метод інтелектуального визначення авторства текстів за стилем написання призначений для інтелектуального відстеження зміни поведінки взламаних акаунтів користувачів, а також може бути використаний для відслідковування фактів несанкціонованих текстових запозичень. Метод працює шляхом перетворення вхідних даних у форматі тексту для визначення авторства та попередньо натренованої моделі машинного навчання для інтелектуального визначення авторства у вихідні дані у форматі визначеного автора та оцінки приналежності тексту до визначеного автора, а також оцінки приналежності тексту до решти авторів з бази.

2.2 Аналіз та автоматизація обробки потоків даних інтелектуальної системи визначення авторства за стилем написання

Автоматизація обробки потоків даних інтелектуальної системи визначення авторства за стилем написання наведена схематично на рисунку 2.2.

При виклику події завантаження та ініціалізації компонентів відбувається завантаження даних з БД для підсистеми роботи з текстами та авторами. Таким чином здійснюється стартове заповнення контролу зі списком авторів з БД, після вибору одного з них їх тексти повинні відобразитись переліком у відповідному контролі. Обравши текст для дослідження, його можна переглянути, а також можна модифікувати та зберегти зміни у БД, при натисканні на кнопку «Зберегти зміни».

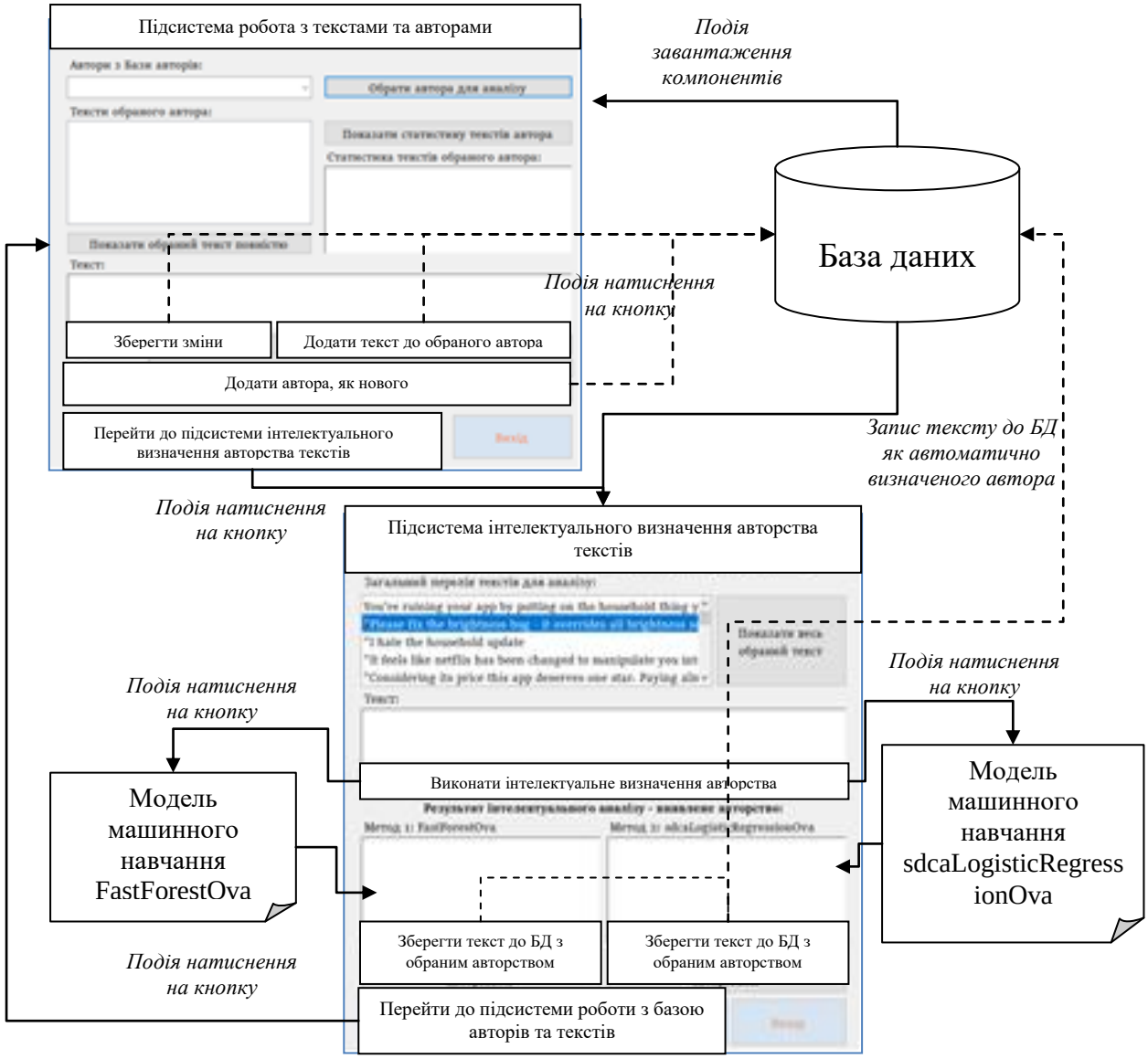


Рисунок 2.2 – Автоматизація обробки потоків даних інтелектуальної системи визначення авторства

Також можна вписати новий текст, що притаманний обраному автору, та теж зберегти в БД, як новий текст обраного автора. Також передбачено додавання нового автора, для якого є спеціально призначена кнопка, при натиснення на неї на неї породжує подію для збереження нового автора у БД.

Дана підсистема ще має кнопку для переходу на підсистему інтелектуального визначення авторства текстів, яка є основною та взаємодіє з навченими моделями машинного навчання та БД.

При натисканні на кнопку «Виконати інтелектуальне визначення авторства» породжується подія, що використовує для уведеного тексту дві

альтернативні моделі машинного навчання та виводить відповідні результати у текстове поле. Кнопки «Зберегти текст до БД з обраним авторством» для відповідних моделей дозволяють виконати збереження авторського тексту до БД як такого, що автоматизовано віднесений до визначеного автора. Також з даної підсистеми можна здійснити перехід по події натиснення на кнопку «Перейти до підсистеми роботи з базою авторів та текстів» до попередньо розглянутої підсистеми.

Отже, таким чином виконано аналіз та здійснено автоматизацію обробки потоків даних інтелектуальної системи визначення авторства за стилем написання.

2.3 Проєктування пайплайну для інтелектуального визначення авторства текстів за стилем написання

Для тренування та подальшого використання моделей машинного навчання необхідно спершу виконати проєктування пайплайну типової моделі машинного навчання. На рисунку 2.3 наведено пайплайн життєвого циклу моделі машинного навчання на прикладі FastForestOva.

Для моделі sdcaLogisticRegressionOva пайплайн будується аналогічно, адже містить всі ті самі етапи попередньої обробки та навчання здійснюється за таким самим алгоритмом L-BFGS Maximum Entropy.

Навчальна множина розмічених авторських текстів є набором текстів для навчання, що поділені на авторів. Відповідно, така множина повинна бути такою, щоб на кожного автора припадало щонайменше по 50 текстів, оскільки у рамках виконання кваліфікаційної роботи бакалавра робота буде здійснюватись саме з короткими текстами. Навчальна множина авторських текстів розбивається на множину міток та множину текстів. Множина текстів перетворюється у числові вектори ознак.

Далі здійснюється об'єднання векторів ознак у єдиний вектор Features, після чого мітки та ознаки перетворюються на числові ключі.

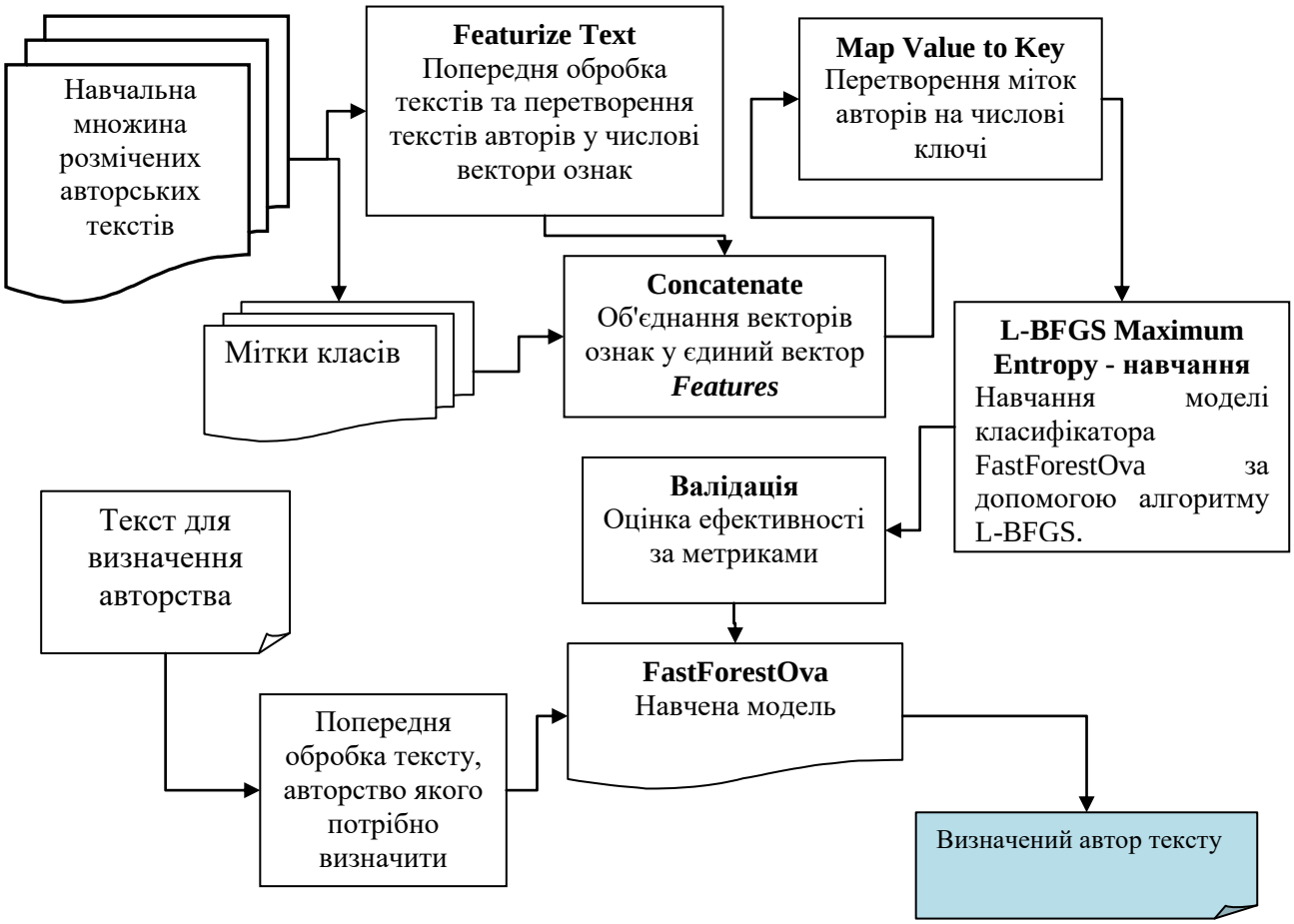


Рисунок 2.3 – Пайплайн типової моделі машинного навчання на прикладі FastForestOva

Наступним етапом є навчання типової моделі класифікатора на прикладі моделі «FastForestOva» за допомогою алгоритму L-BFGS. Навчена модель проходить етап оцінки продуктивності за метриками, та якщо результати задовільні, навчена модель зберігається для подальшого використання.

Збережена модель для визначення авторства текстів за стилем написання може надалі прогнозувати автора за текстом. Однак, текст для цього спершу проходить попередню обробку, після чого подається на вхід навченій моделі. Результат роботи типової моделі машинного навчання це визначений автор тексту.

Отже, спроєктовано пайплайн для інтелектуального визначення авторства текстів за стилем написання, що дозволяє створювати типові моделі машинного навчання.

2.4 Функціональна структура інтелектуальної системи визначення авторства текстів та взаємозв'язок компонентів

Відповідно до методу інтелектуального визначення авторства текстів за стилем написання, було спроєктовано відповідну структуру інформаційної системи, що зображена на Рисунку 2.4.

Рисунок 2.4 – Функціональна структура інтелектуальної системи визначення авторства текстів

Інтелектуальна системи визначення авторства текстів складається із бази даних та 4-х підсистем: «Підсистема робота з текстами та авторами», «Підсистема інтелектуального визначення авторства текстів», «Підсистема

навчання моделей машинного навчання для визначення авторства» та «Підсистема попередньої обробки текстів».

Підсистема робота з текстами та авторами призначена для взаємодії з базою даних, та дозволяє виконувати перегляд та редагування текстів авторів. Підсистема відповідає за такі основні функції, як: вибір автора з БД для аналізу, додавання нового автора в БД, перегляд текстів обраного автора, деталізація обраного тексту обраного автора, додавання нового тексту до обраного автора, редагування тексту обраного автора, перегляд статистики з наявних текстів обраного автора, перехід до підсистеми інтелектуального визначення авторства текстів. Підсистема має графічний інтерфейс користувача та взаємодіє з підсистемою інтелектуального визначення авторства текстів та базою даних.

Підсистема інтелектуального визначення авторства текстів призначена для визначення авторства текстів за стилем написання, та виконує такі основні функції: виведення загального переліку текстів з БД для аналізу, вибір тексту з БД для визначення авторства, написання нового тексту для визначення авторства, інтелектуальне визначення авторства альтернативними моделями машинного навчання, збереження тексту до БД з автоматично визначеним авторством, перехід до підсистеми роботи з базою текстів та авторів. Взаємодіє з підсистемою попередньої обробки текстів, підсистемою навчання моделей машинного навчання для визначення авторства та базою даних, є головною підсистемою інтелектуальної системи визначення авторства за стилем написання та має графічний інтерфейс.

Підсистема попередньої обробки текстів призначена для препроцесингу вхідного тексту автора, та виконує такі функції, як: видалення стоп-символів та видалення стоп-слів.

Підсистема навчання моделей машинного навчання для визначення авторства призначена для тренування та валідації моделей машинного навчання та збереження натренованих валідних моделей. Дана підсистема взаємодіє з підсистема попередньої обробки текстів, використовуючи її, а також є джерелом даних для підсистеми інтелектуального визначення авторства текстів. Не має

графічного інтерфейсу користувача, однак є ключовою підсистемою інтелектуальної системи.

Отже, наведено функціональну структура інтелектуальної системи визначення авторства текстів та проілюстровано взаємозв'язок компонентів. Інтелектуальна система складається із «Підсистеми робота з текстами та авторами», «Підсистеми інтелектуального визначення авторства текстів», «Підсистеми навчання моделей машинного навчання для визначення авторства» та «Підсистеми попередньої обробки текстів» та бази даних. У свою чергу, підсистема інтелектуального визначення авторства текстів є головною підсистемою, для якої передбачено графічний інтерфейс користувача.

2.5 Проєктування бази даних інтелектуальної системи визначення авторства текстів

Створення бази даних є важливим процесом, оскільки вона дозволяє ефективно зберігати, організовувати та обробляти великі обсяги інформації. Це забезпечує швидкий доступ до даних, підвищує точність і надійність збереженої інформації. Отже, для реалізації методу інтелектуального визначення авторства текстів за стилем написання необхідно спроєктувати БД та відповідну даталогічну модель. На рисунку 2.5 наведено даталогічну модель бази даних для методу інтелектуального визначення авторства текстів за стилем написання.

Таблиця «Genres» (таблиця 2.1) призначена для збереження назв жанрів, в яких написані тексти для подальшого дослідження (роман, повість, новела, тощо).

Таблиця 2.1 – Атрибути таблиці «Genres»

№	Назва	Тип даних	Опис
1.	ID	Integer	Первинний ключ, унікальний ідентифікатор запису таблиці
2.	name	String	Назва літературного жанру

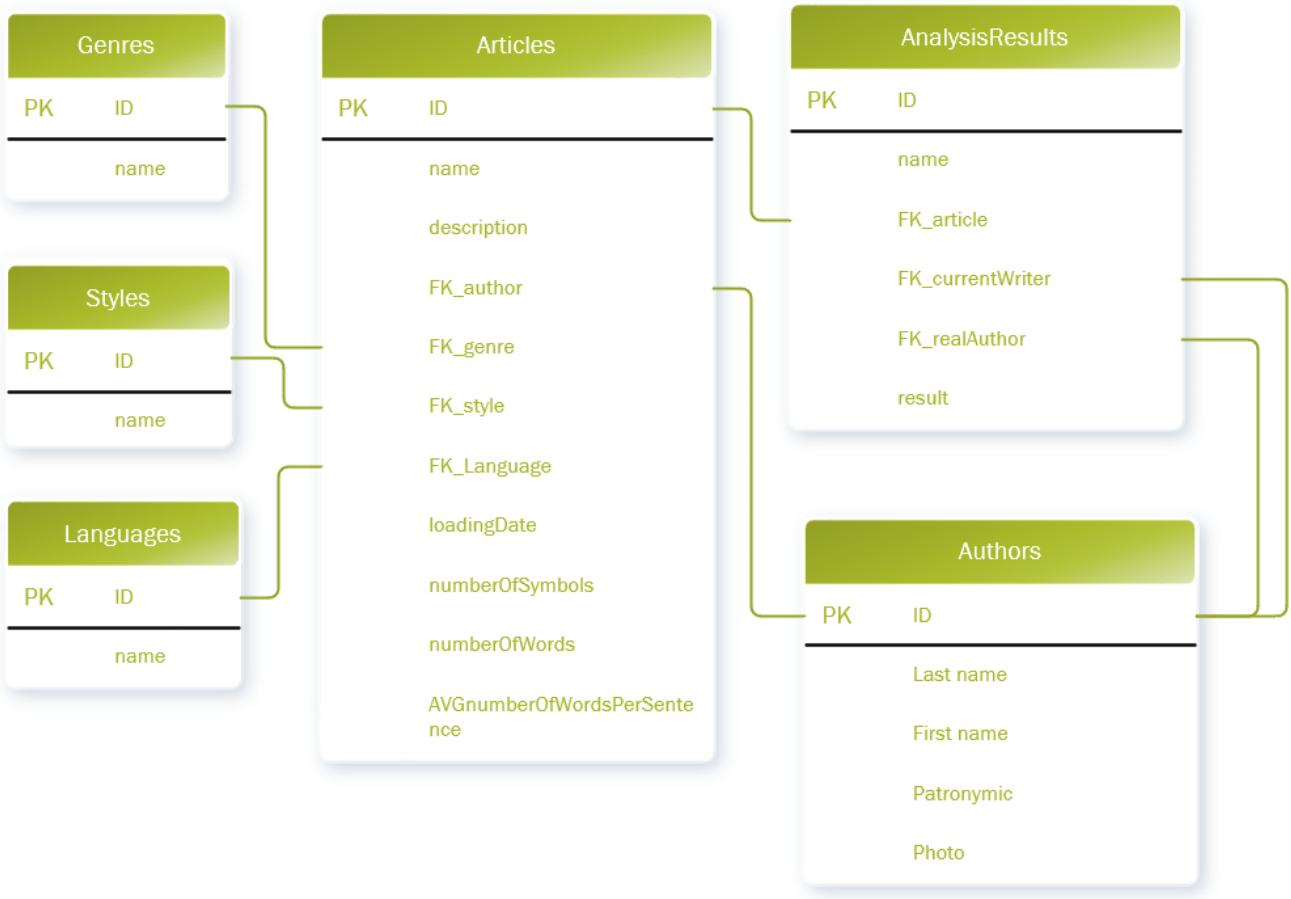


Рисунок 2.5 – Даталогічна модель бази даних для методу інтелектуального визначення авторства текстів за стилем написання

Таблиця «Styles» (таблиця 2.2) призначена для збереження назв літературних стилів в яких написані тексти для подальшого дослідження (науково-популярний, публіцистичний, епістолярний, тощо).

Таблиця 2.2 – Атрибути таблиці «Styles»

№	Назва	Тип даних	Опис
1.	ID	Integer	Первинний ключ, унікальний ідентифікатор запису таблиці
2.	name	String	Назва літературного стилю

Таблиця «Articles» (таблиця 2.3) призначена для збереження даних про тексти, що міститимуться в БД, такі як назви, описи, повний текст, автори,

жанри, стилі, дати написання та завантаження, доступність, кількість слів і символів, а також середня кількість слів у реченні.

Таблиця 2.3 – Атрибути таблиці «Articles»

№	Назва	Тип даних	Опис
1.	ID	Integer	Первинний ключ, унікальний ідентифікатор кожного запису у таблиці, який використовується для однозначної ідентифікації кожного твору
2.	name	String	Назва художнього твору
3.	description	Text	Вміст твору
4.	FK_Author	Integer	Вторинний ключ, який забезпечує зв'язок з відповідним записом у таблиці "Authors" і вказує на автора твору
5.	FK_Genre	Integer	Вторинний ключ, який вказує на жанр твору і посилається на відповідний запис у таблиці "Genres"
6.	FK_Style	Integer	Вторинний ключ, який вказує на стиль твору і співвідноситься з відповідним записом у таблиці "Styles"
7.	FK_Language	Integer	Вторинний ключ, який вказує на мову твору і посилається на відповідний запис у таблиці "Languages"
8.	LoadingDate	Date	Дата завантаження твору в базу даних,
9.	numberOfSymbols	Integer	Кількість символів у творі
10.	numberOfWords	Integer	Кількість слів у творі
11.	AVGnumberOf	Integer	Середня кількість слів у творі

Таблиця «Languages» (таблиця 2.4) призначена для збереження назв мов, якими написано текст для дослідження.

Таблиця 2.4 – Атрибути таблиці «Languages»

№	Назва	Тип даних	Опис
1.	ID	Integer	Первинний ключ, унікальний ідентифікатор запису таблиці
2.	name	String	Назва мови

Таблиця «AnalysisResults» (таблиця 2.5) призначена для збереження результатів дослідження, в таблиці створено поля для збереження посилання на твір, реального та зазначеного авторів.

Таблиця 2.5 – Атрибути таблиці «AnalysisResults»

№	Назва	Тип даних	Опис
1.	ID	Integer	Первинний ключ, унікальний ідентифікатор кожного запису у таблиці, який використовується для однозначної ідентифікації дослідження
2.	FK_article	Integer	Вторинний ключ, який вказує на твір і посилається на відповідний запис у таблиці "Articles"
3.	FK_currentAuthor	Integer	Вторинний ключ, який вказує на зазначеного автора і посилається на відповідний запис у таблиці "Authors"
4.	FK_realAuthor	Integer	Вторинний ключ, який вказує на дійсного автора і посилається на відповідний запис у таблиці "Authors"
5.	result	String	Результат дослідження

Таблиця «Authors» (таблиця 2.6) призначена для збереження даних про авторів творів, що будуть аналізуватись. Таблиця містить поля для збереження головної інформації про автора, зокрема ПІБ та фото.

Таблиця 2.6 – Атрибути таблиці «Authors»

№	Назва	Тип даних	Опис
1.	ID	Integer	Первинний ключ, унікальний ідентифікатор кожного запису у таблиці, який використовується для однозначної ідентифікації автора
2.	LastName	String	Прізвище автора
3.	FirstName	Text	Ім’я автора
4.	Patronymic	Integer	По батькові автора
5.	Photo	Integer	Шлях до файлу фотозображення автора

Таким чином, було спроектовано базу даних для методу інтелектуального визначення авторства текстів за стилем написання. Було створено та описано таблиці, зв’язки між ними, а також представлено даталогічну модель бази даних інтелектуальної системи визначення авторства текстів.

2.6 Підготовка робочих вхідних даних для інтелектуальної системи визначення авторства текстів

Для навчання моделей машинного навчання що спроможні виявляти авторство текстів за стилем написання було використано два набори даних: «Netflix Reviews» та «Spotify Reviews» [28].

Набір даних «Netflix Reviews» містить інформацію про відгуки користувачів netflix у Google Play Store. Окрім відгуків, він також містить інформацію стосовно рейтингу та дату перегляду, оцінки «подобається» для кожного огляду та ідентифікатор користувача. Набір даних містить

напівструктурований формат для оглядів і оцінок кожного користувача, та в цілому складається 111 836 відгуків.

Набір даних «Spotify Reviews» містить відгуки користувачів про програму Spotify Music у магазині Google Play. Також містить інформацію про рейтинги та дату переглядів. Файл слугує записом відгуків і оцінок користувачів у реальному часі, містить додаткові відомості про релевантність оглядів і дати їх публікації, а також ідентифікатор користувача, що залишив відгук. Кількість відгуків у даному наборі становить 84 165.

Два описані набори даних пройшли програмну обробку, та були груповані по користувачам. До робочого набору даних були включені лише ті записи, кількість авторських коментарів яких була більше 50. Таким чином було отримано дані 8-ми авторів, які були поділені на навчальні (по 50 зразків) та тестові (зразки, що залишились).

Відповідно, розподіл текстів по отриманих авторах, що задовольнили кількісну умову наведено на рисунку 2.6.

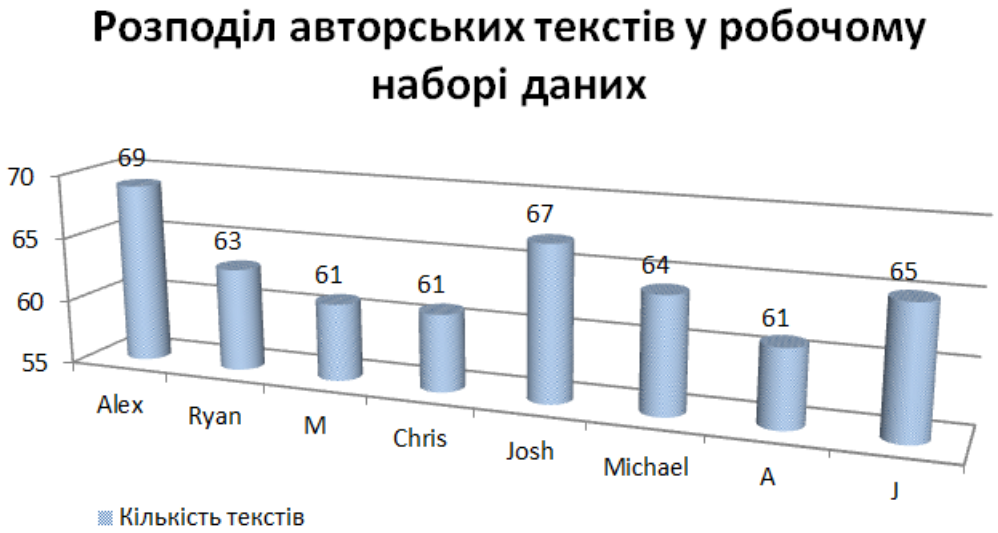


Рисунок 2.6 – Розподіл авторських текстів у робочому наборі даних

Отже, для навчання та тестування моделей машинного навчання для виявлення авторства текстів буде використано модифікований набір даних, величиною у 511 текстів відгуків, що належать 8-ми авторам.

2.7 Особливості використання спеціалізованих програмних компонентів

Під час програмної реалізації інтелектуальної системи визначення авторства текстів за стилем написання буде використано бібліотеку машинного навчання ML.NET.

ML.NET є потужною бібліотекою для машинного навчання на платформі .NET, яка значно спрощує реалізацію задачі інтелектуального визначення авторства текстів за стилем написання. Завдяки ML.NET є можливість створювати, тренувати та використовувати моделі машинного навчання безпосередньо в .NET-застосуваннях, використовуючи знайомі інструменти та мови програмування, такі як C# [29]. Це дозволяє інтегрувати процеси обробки тексту та аналізу стилістичних ознак у різні типи застосунків, від веб-застосунків до настільних і серверних рішень. Зокрема, ML.NET підтримує роботу з великими обсягами текстових даних, надаючи засоби для передобробки тексту, екстракції ознак, таких як частота слів, біграми та синтаксичні структури, і застосування алгоритмів класифікації, які можуть точно визначити авторство тексту.

ML.NET також забезпечує зручні інтерфейси для налаштування та оптимізації моделей машинного навчання, включаючи можливості для автоматичного налаштування гіперпараметрів та оцінки продуктивності моделей за допомогою метрик точності, влучності та повноти. Це дозволяє ефективно експериментувати з різними підходами та обирати найбільш підходящі моделі для конкретних задач. Інтеграція з іншими компонентами екосистеми .NET, такими як SQL Server для зберігання даних і Azure для масштабування обчислювальних ресурсів, робить ML.NET потужним інструментом для створення високопродуктивних рішень з аналізу тексту та визначення авторства.

System.Windows.Forms є бібліотекою, яка входить до складу платформи .NET [30]. Вона використовується для створення графічних інтерфейсів

користувача для настільних застосувань у Windows. Ця бібліотека надає інструменти для створення вікон, кнопок, текстових полів, меню та інших елементів інтерфейсу, дозволяючи створювати інтерактивні застосування. System.Windows.Forms є частиною Windows Forms (WinForms), яка забезпечує простий і швидкий спосіб розробки настільних додатків з графічним інтерфейсом у середовищі .NET. Буде використано для створення інтерфейсу користувача застосунку визначення авторства текстів за стилем написання.

Отже, буде використано набір бібліотек у вигляді використання ML.NET для навчання, валідації та збереження моделей машинного навчання, а також бібліотеку System.Windows.Forms для реалізації графічного інтерфейсу користувача для реалізації методу визначення авторства текстів за стилем написання.

2.8 Висновки до розділу 2

У рамках виконання другого розділу запропоновано метод інтелектуального визначення авторства текстів за стилем написання, що призначений для інтелектуального відстеження зміни поведінки взламаних акаунтів користувачів, а також може бути використаний для відслідковування фактів несанкціонованих текстових запозичень. Запропонований метод працює шляхом перетворення вхідних даних у вигляді тексту для визначення авторства та попередньо натренованої моделі машинного навчання для інтелектуального визначення авторства у вихідні дані у форматі визначеного автора та оцінки приналежності тексту до визначеного автора, а також оцінки приналежності тексту до решти авторів з бази.

Виконано аналіз та здійснено проєктування автоматизації обробки потоків даних інтелектуальної системи визначення авторства за стилем написання. Планується реалізувати дві інтерфейсні форми, які зможуть задовольнити всі потреби користувача для реалізації задачі інтелектуального визначення авторства.

Виконано проєктування пайплайну для інтелектуального визначення авторства текстів за стилем написання, що дозволяє створювати типові моделі машинного навчання.

Наведено функціональну структура інтелектуальної системи визначення авторства текстів та проілюстровано взаємозв'язок її компонентів. Запропонована інтелектуальна система складається із 4-х підсистем: «Підсистеми робота з текстами та авторами», «Підсистеми інтелектуального визначення авторства текстів», «Підсистеми навчання моделей машинного навчання для визначення авторства», «Підсистеми попередньої обробки текстів» та бази даних. Підсистема інтелектуального визначення авторства текстів є головною підсистемою, для якої передбачено графічний інтерфейс користувача.

Спроєктовано базу даних, створено та описано таблиці, зв'язки між ними, а також представлено даталогічну модель БД.

Виконано підготовку робочих вхідних даних, для навчання та тестування моделей машинного навчання для виявлення авторства текстів буде використано модифікований набір даних, величиною у 511 текстів відгуків, що належать 8-ми авторам.

Для розробки інтелектуальної системи визначення авторства обрано до використання набір бібліотек у вигляді використання ML.NET для навчання, валідації та збереження моделей машинного навчання, а також бібліотеку System.Windows.Forms для реалізації графічного інтерфейсу користувача для реалізації методу визначення авторства текстів за стилем написання.

У подальшому всі спроєктовані елементи інтелектуальної системи необхідно програмно реалізувати, а також дослідити коректність виконання функцій. За допомогою розробленої програмної реалізації також необхідно дослідити ефективність методу інтелектуального визначення авторства текстів за стилем написання.

Розділ 3 Експериментальне дослідження методу інтелектуального визначення авторства текстів

3.1 Визначення шляхів дослідження та засобів створення інформаційної системи інтелектуального визначення авторства текстів

Для дослідження методу інтелектуального визначення авторства текстів необхідно обрати засоби розробки, а також виконати саму програмну реалізацію. Створену програмну реалізацію необхідно протестувати на предмет коректності роботи заявлених функцій, основними з яких є:

– інтелектуальне визначення авторства текстів за стилем написання;

– робота з авторами (перегляд, додавання нового тексту автора);

– попередня обробка тексту;

– виведення статистики для авторів за наявними текстами (кількість текстів, довжина тощо).

Для дослідження ефективності інтелектуального визначення авторства текстів за стилем написання буде використано дві альтернативних моделі машинного навчання, між якими буде здійснено порівняння. Для порівняння авторських текстів буде використано розмічений попередньо набір даних, на якому буде обраховано основні метрики, такі як точність, влучність та повнота.

Метрики точності, влучності та повноти є ключовими для оцінки продуктивності моделей машинного навчання. Точність визначає частку правильно передбачених випадків серед усіх передбачень моделі, що відображає загальну коректність [31] Влучність вимірює частку правильних позитивних передбачень серед усіх позитивних передбачень, що вказує на здатність моделі уникати хибно-позитивних результатів. Повнота оцінює частку правильно передбачених позитивних випадків серед усіх реальних позитивних випадків, що показує здатність моделі виявляти всі істинні позитивні випадки. Ці метрики допомагають зрозуміти різні аспекти ефективності моделей та їхню здатність коректно класифікувати дані.

3.2 Вибір засобів розробки інформаційної системи інтелектуального визначення авторства текстів за стилем написання

Для розробки інформаційної системи інтелектуального визначення авторства текстів за стилем написання буде використано платформу .NET, середовища програмування Visual Studio 2022, мову програмування C#, СКБД SQLserver.

Платформа .NET надає потужні засоби для реалізації задачі інтелектуального визначення авторства текстів за стилем написання, зокрема за допомогою машинного навчання. Завдяки ML.NET, можна створювати, навчати та впроваджувати моделі машинного навчання безпосередньо у своїх .NET-застосуваннях [32]. Це включає обробку та аналіз тексту, екстракцію стилістичних ознак і застосування алгоритмів класифікації для визначення авторства. Тому у якості платформи буде використано .NET.

У якості середовища програмування буде використано Visual Studio 2022, що є потужним інструментом для реалізації задачі інтелектуального визначення авторства текстів за стилем написання. Visual Studio 2022 надає інтегроване середовище для розробки та тестування складних програмних рішень, а завдяки підтримці різних мов програмування, таких як Python та C#, є можливість використання бібліотек для обробки природної мови та машинного навчання [33]. Це дозволяє створювати та впроваджувати алгоритми, які аналізують текстові дані, визначаючи характерні стилістичні ознаки, притаманні певним авторам.

Крім того, Visual Studio забезпечує зручні інструменти для налагодження та оптимізації коду, що є критично важливим для ефективної роботи з великими обсягами текстових даних, а також має засоби для швидкої побудови інтерфейсів користувача. Вбудовані можливості аналізу коду, тестування та інтеграції з системами контролю версій сприяють створенню стабільного та надійного програмного забезпечення. Таким чином, використання Visual Studio 2022

дозволяє ефективно реалізувати проекти, пов'язані з визначенням авторства текстів, забезпечуючи високу продуктивність і точність аналізу.

Мова програмування C# у контексті задачі інтелектуального визначення авторства текстів за стилем написання на платформі .NET була обрана завдяки своїй потужності, зручності та інтеграції з ML.NET. Використання C# дозволяє скористатися перевагами об'єктно-орієнтованого програмування для структуризації та модульності коду [34]. Завдяки бібліотеці ML.NET процес реалізації алгоритмів машинного навчання для аналізу тексту, екстракції стилістичних ознак та класифікації авторства є значно легшим. Крім того, багатий набір вбудованих функцій C# та потужні інструменти розробки Visual Studio сприяють ефективній розробці, тестуванню та налагодженню застосувань, забезпечуючи високу продуктивність і надійність рішень для обробки великих обсягів текстових даних.

Система керування базами даних SQL Server є ключовим компонентом для реалізації задачі інтелектуального визначення авторства текстів за стилем написання на платформі .NET, оскільки однією із складових програмного комплексу є БД. Використовуючи SQL Server, можна ефективно зберігати, керувати та обробляти великі обсяги текстових даних, необхідних для аналізу та навчання моделей машинного навчання. SQL Server забезпечує високу продуктивність, надійність і масштабованість, що є критичним для застосунків, які працюють з великими даними [35]. Інтеграція з .NET та підтримка таких функцій, як індексація, транзакції, запити T-SQL та аналітичні сервіси, дозволяє здійснювати складні операції з даними, прискорюючи процес екстракції текстових ознак і підвищуючи точність моделей класифікації. Завдяки потужним можливостям аналітики та бізнес-інтелекту, SQL Server сприяє глибокому розумінню текстових даних і забезпечує надійну основу для побудови інтелектуальних додатків на базі машинного навчання.

Отже, для розробки інтелектуального виявлення авторства текстів за стилем написання обрано такий набір засобів: платформа .NET, середовище програмування Visual Studio 2022, мова програмування C#, СКБД SQLserver.

3.3 Структура та функціональне призначення програмних складових інформаційної системи інтелектуального визначення авторства текстів

Програмний застосунок для інтелектуального визначення авторства текстів за стилем написання складається із таких ключових компонентів, як дві альтернативних моделі машинного навчання для інтелектуального визначення авторства, модулю попередньої обробки, а також модулю роботи з авторами та модулю інтелектуального визначення авторства.

Діаграма класів розроблюваного застосунку наведена на рисунку 3.1.

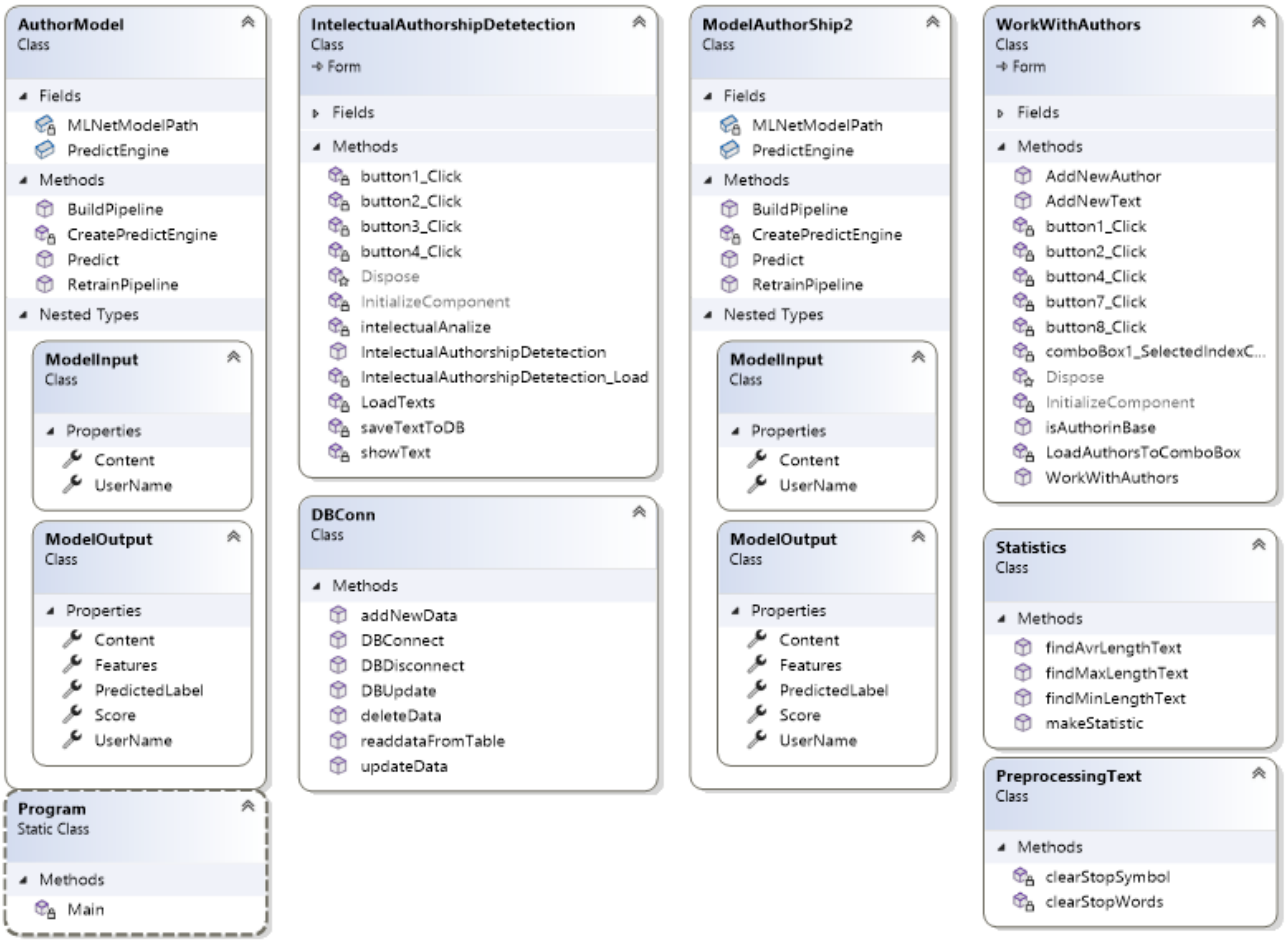


Рисунок 3.1 – Діаграма класів інтелектуальної системи визначення авторства текстів за стилем написання

Класи «AuthorModel» та «ModelAuthorShip2» мають схожу структуру, та є класами, що відповідають за створення, навчання та збереження навчених альтернативних моделей машинного навчання. У рамках роботи використовується дві альтернативних моделі машинного навчання – FastForestOva та sdcaLogisticRegressionOva відповідно.

Клас «DBConn» відповідає за з'єднання та взаємодію з базою даних, та має для цього відповідні методи.

Клас «PreprocessingText» відповідає за попередню обробку досліджуваного тексту.

Клас «Statistics» відповідає за знаходження статистики за обраним автором по довжині текстів та по кількості.

Клас «WorkWithAuthors» відповідає за роботу з авторами та текстами, взаємодіє з класом «Statistics» та «DBConn», а також взаємодіє з інтерфейсом користувача.

Клас «IntelectualAuthorshipDetetection» призначений для безпосередньої взаємодії з досліджуваними текстами з метою інтелектуального виявлення авторства за стилем написання. Даний клас взаємодіє з «DBConn», а також із моделями машинного навчання «AuthorModel» та «ModelAuthorShip2».

Отже, наведено діаграму класів інтелектуальної системи визначення авторства текстів за стилем написання та описано функціональне призначення програмних складових.

3.4 Особливості реалізації програмних складових інформаційної системи інтелектуального визначення авторства текстів

Наступним етапом після проєктування є реалізація програмних складових інформаційної системи інтелектуального визначення авторства текстів за стилем написання. Моделі машинного навчання перед використанням тренуються, та натреновані екземпляри використовуються відповідною підсистемою інтелектуального визначення авторства текстів. Оскільки алгоритми навчання

дуже схожі та відрізняються лише архітектурою моделі, нижче розглянуто реалізацію sdcaLogisticRegressionOva.

Програмна реалізація класу «ModelAuthorShip2» включає два основні методи: RetrainPipeline і BuildPipeline. Метод RetrainPipeline використовується для повторного навчання моделі, застосовуючи конвеєр, створений у процесі навчання. Він приймає об'єкт MLContext, що забезпечує контекст виконання для операцій машинного навчання, та дані для навчання у форматі IDataView. Спочатку цей метод викликає функцію BuildPipeline для побудови конвеєра обробки даних та навчання моделі, а потім навчає модель, використовуючи метод Fit, застосовуючи отриманий конвеєр до навчальних даних. Результатом є тренована модель типу ITransformer.

Метод BuildPipeline відповідає за створення конвеєра обробки даних та навчання моделі. Конвеєр починається з перетворення текстових даних у числові вектори ознак за допомогою методу FeaturizeText, де вхідні дані з колонки content перетворюються на вихідні ознаки тієї ж колонки. Далі ці ознаки об'єднуються в єдиний вектор під назвою Features. Наступним кроком є перетворення значень міток (імен авторів) на числові ключі за допомогою MapValueToKey. Для нормалізації ознак використовується метод NormalizeMinMax, який масштабує значення ознак до діапазону [0, 1]. Потім до конвеєра додається навчальний алгоритм LbfgsMaximumEntropy, що використовує регуляризацію L1 і L2 для покращення узагальнювальної здатності моделі. Нарешті, конвеєр завершується перетворенням ключів передбачених міток назад у вихідні значення за допомогою MapKeyToValue. Цей конвеєр дозволяє ефективно обробляти дані та навчати модель класифікації тексту за авторством.

Цей клас ModelAuthorShip2 реалізує функціональність для завантаження, тренування та використання моделі машинного навчання для визначення авторства текстів за стилем написання. Він включає внутрішні класи для представлення вхідних та вихідних даних моделі, методи для передбачення результатів та ініціалізації конвеєра передбачення.

Також клас «ModelAuthorShip2» містить два внутрішні класи: «ModelInput» та «ModelOutput». Клас «ModelInput» представляє структуру вхідних даних з двома полями: UserName (ім'я користувача) та Content (вміст тексту), які позначені атрибутом ColumnName для відповідності назвам колонок у наборі даних. «ModelOutput» містить результати передбачення моделі і включає кілька полів: UserName, Content, Features, PredictedLabel та Score, де кожне поле також позначено атрибутом ColumnName для відповідності результатам конвеєра.

Основний метод для передбачення, Predict, приймає об'єкт «ModelInput» та повертає об'єкт «ModelOutput». Він використовує об'єкт PredictionEngine, який створюється лінивим способом через Lazy<PredictionEngine<ModelInput, ModelOutput>>. Це забезпечує одноразове створення та ініціалізацію об'єкта PredictionEngine, який відповідає за здійснення передбачень. Метод CreatePredictEngine ініціалізує MLContext, завантажує збережену модель з файлу ModelAuthorShip2.zip і створює об'єкт PredictionEngine, використовуючи завантажену модель.

Приклад використання вже навченої моделі наведено на рисунку 3.2.

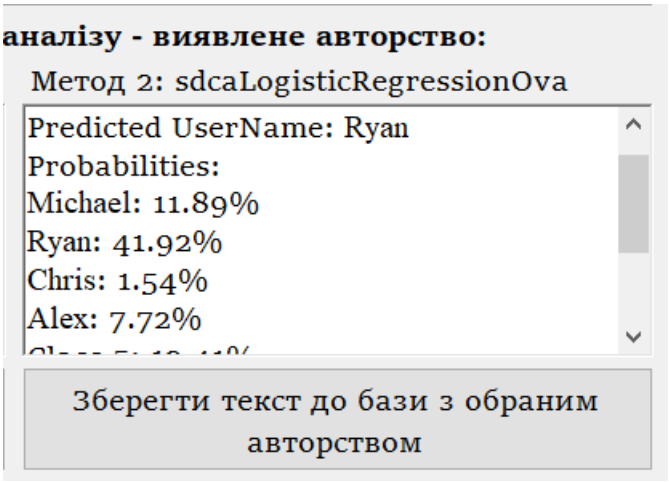


Рисунок 3.2 – Приклад використання навченої моделі sdcaLogisticRegressionOva

Для використання моделі машинного навчання для передбачення авторства тексту і відображення результатів у графічному інтерфейсі користувача (GUI), створеному за допомогою Windows Forms спочатку

створюється екземпляр класу «AuthorModel.ModelInput», який представляє вхідні дані для моделі. Поле Content заповнюється текстом з елемента інтерфейсу richTextBox2. Цей текст потім передається як вхідні дані для моделі.

Викликається метод Predict, який використовує модель для передбачення авторства на основі переданого тексту. Результат передбачення зберігається в змінній predictionResult, яка є екземпляром класу «ModelOutput». Це об'єкт, що містить результати передбачення, включаючи передбачений мітку (автора) та ймовірності для кожного класу (автора).

Ілюстрація роботи двох альтернативних моделей машинного навчання наведена на рисунку 3.3.

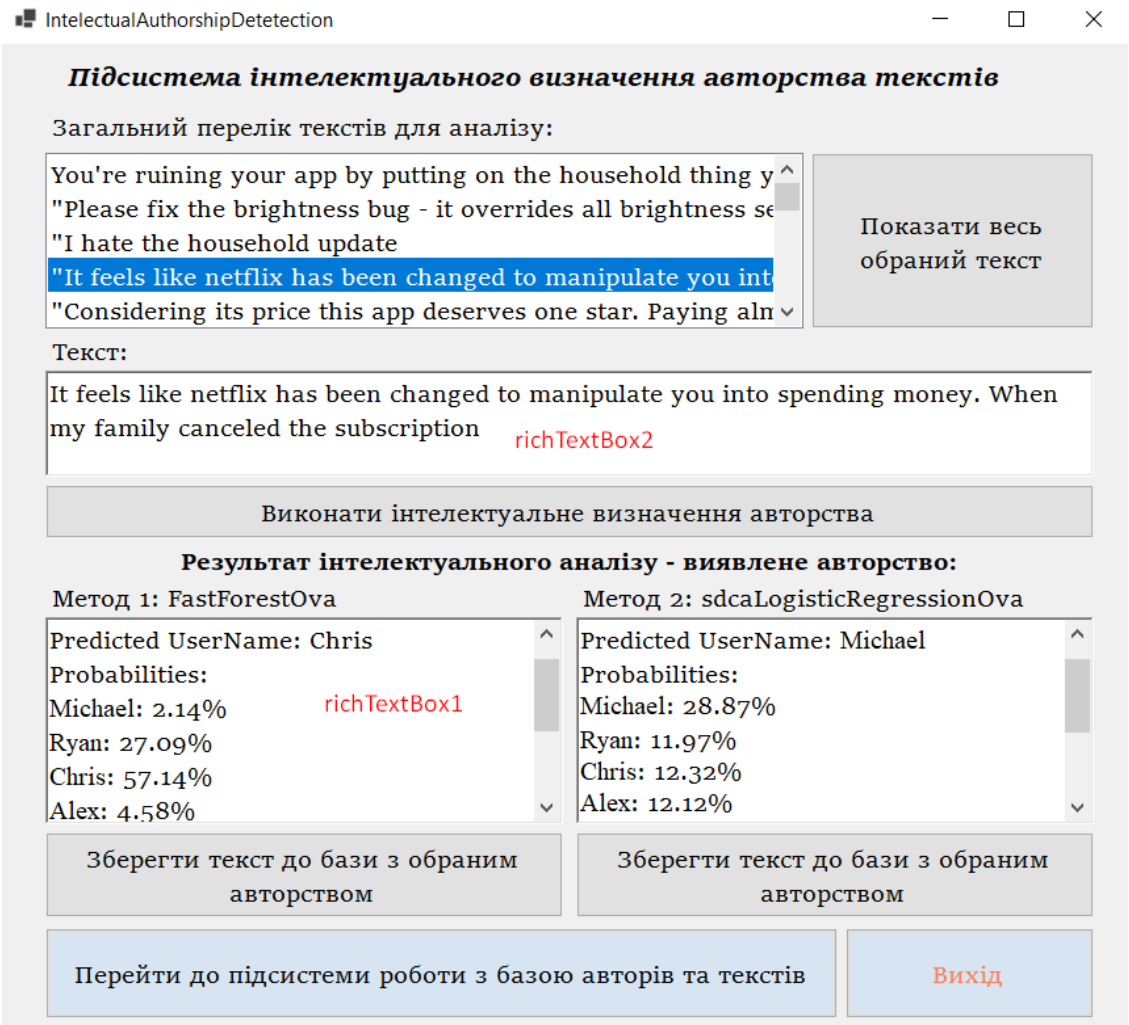


Рисунок 3.3 – Приклад роботи двох альтернативних моделей машинного навчання

Результат передбачення відображається в елементі інтерфейсу richTextBox1. Виводиться передбачений автор тексту, що зберігається в полі PredictedLabel об'єкта predictionResult. Якщо об'єкт predictionResult також містить ймовірності (Score), вони виводяться нижче. Для цього використовується цикл, який проходить через всі елементи масиву ймовірностей і додає їх до текстового поля. Ймовірності відображаються у форматі відсотків з двома знаками після коми.

Отже, виконано програмну реалізацію складових інтелектуальної системи визначення авторства текстів за стилем написання та описано особливості з програмування та використання підсистеми навчання моделей машинного навчання.

3.5 Тестування інформаційної системи інтелектуального визначення авторства текстів за стилем написання

Метою виконання даного підрозділу є оцінити, чи здатний програмний продукт ефективно виконувати завдання визначення авторства текстів за стилем написання, як це задано в постановці задачі. Визначити ступінь відповідності програмного продукту для його застосування за призначенням. Для цього буде використано підхід юніт-тестування та тескт-кейсів.

Для проведення юніт-тестування було створено тестовий клас для перевірки коректності роботи методів RetrainPipeline і Predict в контексті завдання визначення авторства текстів за стилем написання. Він використовує фреймворк MSTest для організації та виконання тестів.

У класі «ModelAuthorShip2Tests» створено кілька тестових методів. Перший тест, RetrainPipeline_ShouldReturnTrainedModel, перевіряє, чи метод RetrainPipeline повертає навчену модель, яка не є null. Для цього спочатку створюється контекст ML (MLContext) і підготовлюються зразки навчальних даних, які потім завантажуються у форматі IDataView. Після виклику методу RetrainPipeline перевіряється, чи повернута модель не є null.

Метод ClassInitialize використовується для ініціалізації контексту ML та навчальної моделі, що забезпечує підготовку перед запуском інших тестів. Це включає збереження моделі на диск, що дозволяє використовувати її для передбачень в інших тестах.

Другий тест, Predict_ShouldReturnNotNull, перевіряє, чи метод Predict повертає результат, який не є null. Для цього створюється зразок вхідних даних (ModelInput), після чого викликається метод Predict. Результат передбачення перевіряється на те, щоб він не був null.

Третій тест, Predict_ShouldReturnCorrectPrediction, перевіряє коректність передбачення для конкретного тексту. Для цього створюється зразок вхідних даних з відомим текстом, що належить певному користувачеві, і викликається метод Predict. Результат передбачення перевіряється на відповідність очікуваному користувачу.

Таким чином, ці тести забезпечують перевірку якості роботи методів навчання та передбачення моделі машинного навчання в задачі визначення авторства текстів, а результат їх успішного виконання наведено на рисунку 3.4.

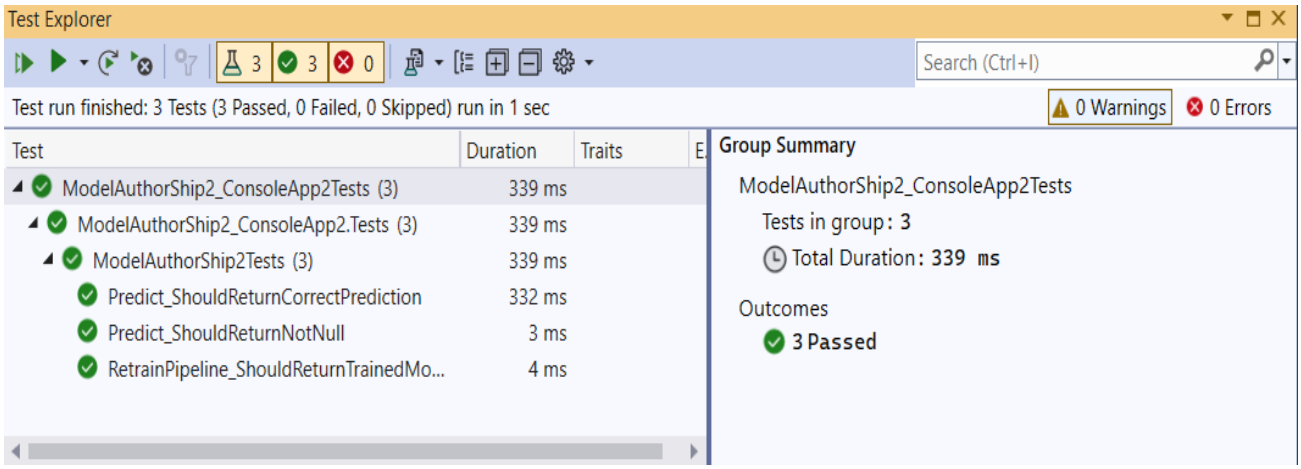


Рисунок 3.4 – Усішне виконання юніт-тестів

Отже, використовуючи модульне тестування було виконано перевірку коректності виконання програмних складових інформаційної системи інтелектуального визначення авторства текстів за стилем написання, що відповідають за процес навчання та використання моделей машинного навчання

для інтелектуального визначення авторства текстів за стилем написання. Подальший функціонал інтелектуальної системи планується перевіряти з використанням тест-кейсів.

Наступною буде виконана перевірка функціональності підсистеми роботи з авторами та текстами. Першим тест-кейсом буде перевірка виведення статистики для обраного автора. Кроки тест-кейсу наведено в таблиці 3.1.

Таблиця 3.1 – Тест-кейс 00001

Тест-кейс ID: 00001	**Приорітет:** 1	**Створено:** 10.05.2024, Богдан ШПОРТ
Назва: Перевірка виведення статистики для обраного автора		
Кроки		**Очікуваний результат**
1. Відкрити застосунок. 2. З випадаючого списку авторів обрати автора «Alex». 3. Натиснути кнопку «Показати статистику текстів автора». 4. Порівняти очікуваний результат з реальним.		Відкрився застосунок. Обрано автора «Alex». В полі «Тексти обраного автора» відображено наявні в БД тексти для аналізу. Виведено статистику текстів автора «Alex». Очікуваний результат відповідає отриманому
Результат виконання тест-кейсу: перевірку пройдено успішно.		

Результат успішного виконання тест-кейсу наведено на рисунку 3.5.

Наступним тестовим випадком буде перевірка додавання нового тексту, як тексту обраного автора. Кроки тест-кейсу наведено в таблиці 3.2.

Після виконання кроків, описаних у таблиці 3.2, виведено повідомлення про успішне додавання тексту до БД, а також текст додався до переліку авторських. На рисунку 3.6 наведено успішне виконання тест-кейсу 00002.

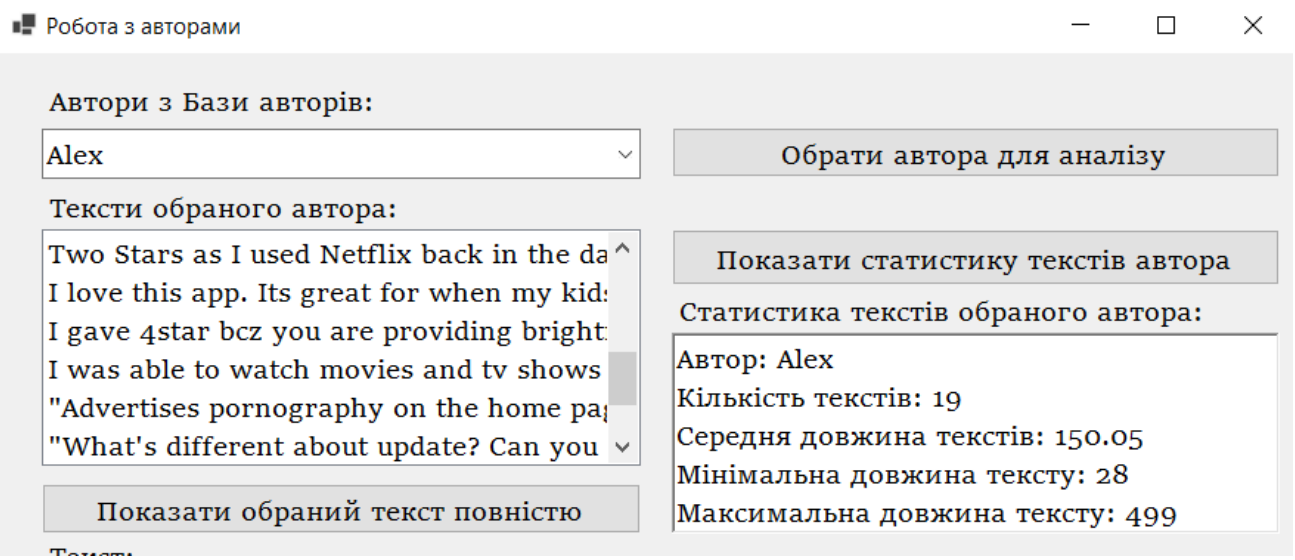


Рисунок 3.5 – Виведення статистики для обраного автора

Таблиця 3.2 – Тест-кейс 00002

Тест-кейс ID: 00002	**Приорітет:** 1	**Створено:** 15.05.2024, Богдан ШПОРТ
Назва: Перевірка додавання нового тексту, як тексту обраного автора		
Кроки		**Очікуваний результат**
1. Відкрити застосунок. 2. З випадаючого списку авторів обрати автора «Alex». 3. У текстове поле «Текст» ввести текст, який потрібно зберегти до бази даних. 4. Натиснути кнопку «Додати текст до обраного автора» 5. Перевірити наявність тексту у базі даних та у інтерфейсі користувача.		Відкрився застосунок. Обрано автора «Alex». В полі «Тексти обраного автора» відображено наявні в БД тексти для аналізу. Уведено текст для додавання у БД. Повідомлення про успішне додавання тексту до БД. Текст додано до БД та відображено у інтерфейсі користувача.
Результат виконання тест-кейсу: перевірку пройдено успішно.		

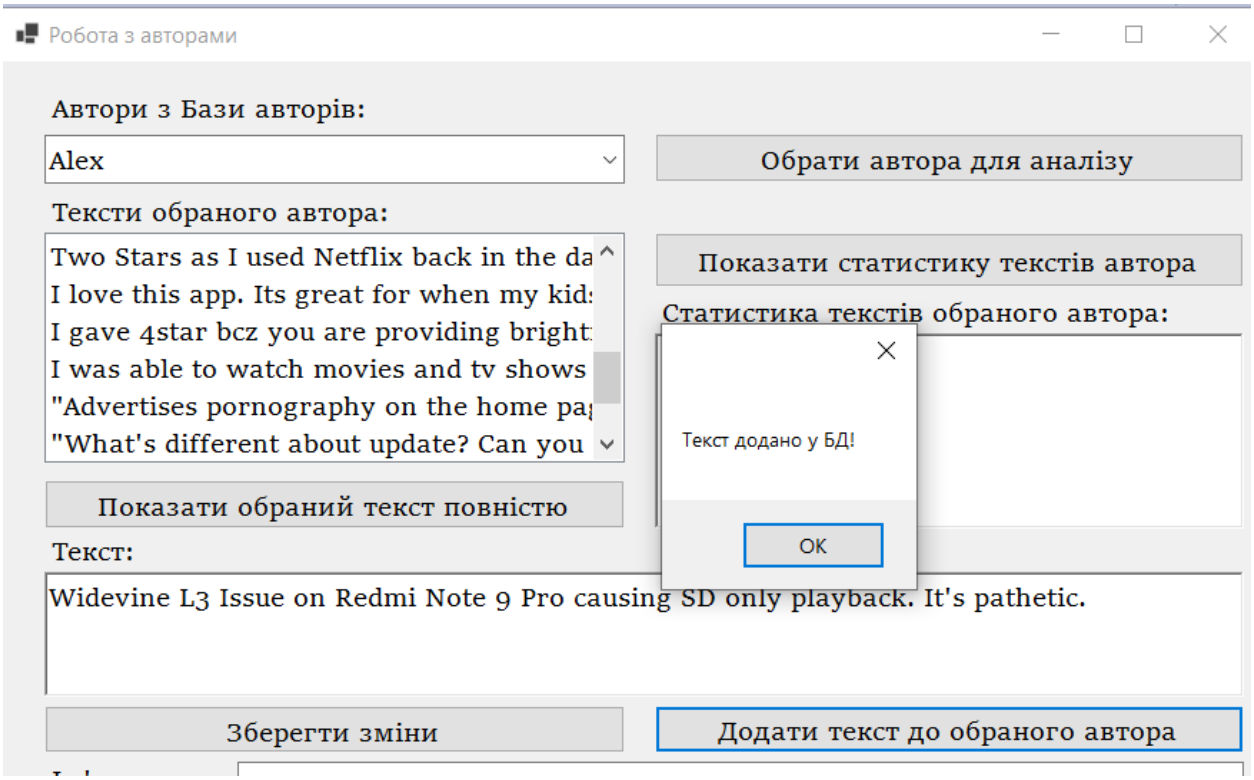


Рисунок 3.6 – Додавання тексту до БД інтелектуальної системи визначення авторства текстів за стилем написання

Отже, було проведено тестування основних функцій програмної реалізації інтелектуальної системи визначення авторства текстів за стилем написання, яке показало що весь заявлений функціонал працює коректно. Тестування проводилось з використанням юніт-тестів та тест-кейсів. Наведено підтвердження успішного проходження тестів у вигляді скріншотів роботи програми.

3.6 Аналіз функціональності інформаційної системи інтелектуального визначення авторства текстів за стилем написання

Окрім проведення тестування також є окрема необхідність проведення аналізу функціональності інтелектуальної системи визначення авторства за текстом. Після запуску програми, відкриється підсистема роботи з текстами та авторами (рисунок 3.7).

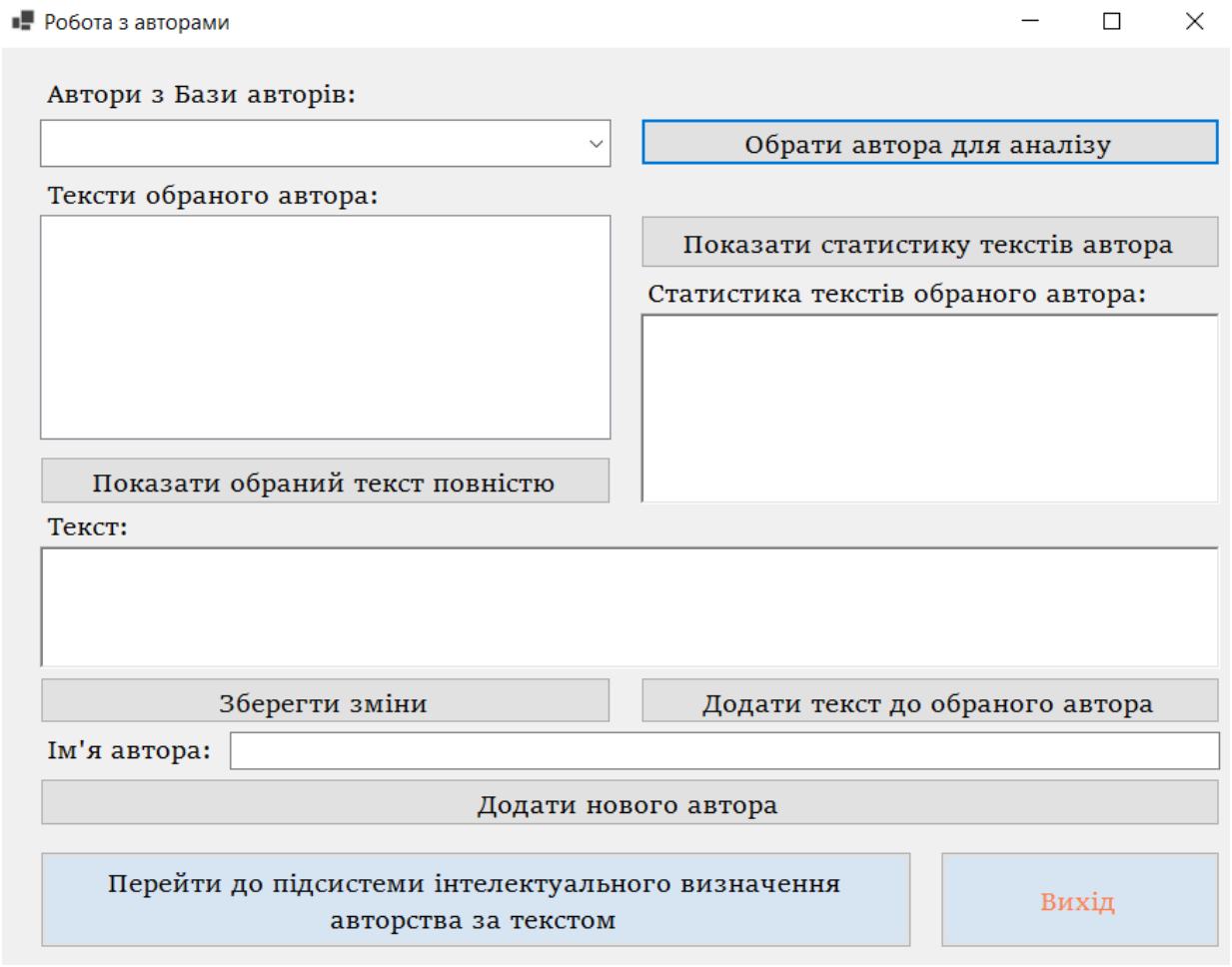


Рисунок 3.7 – Підсистема роботи з текстами та авторами

Для вибору автора з бази авторів необхідно натиснути на випадаючий перелік, та обрати автора для дослідження. Для закріплення вибору необхідно натиснути кнопку «Обрати автора для аналізу» (рисунок 3.8).

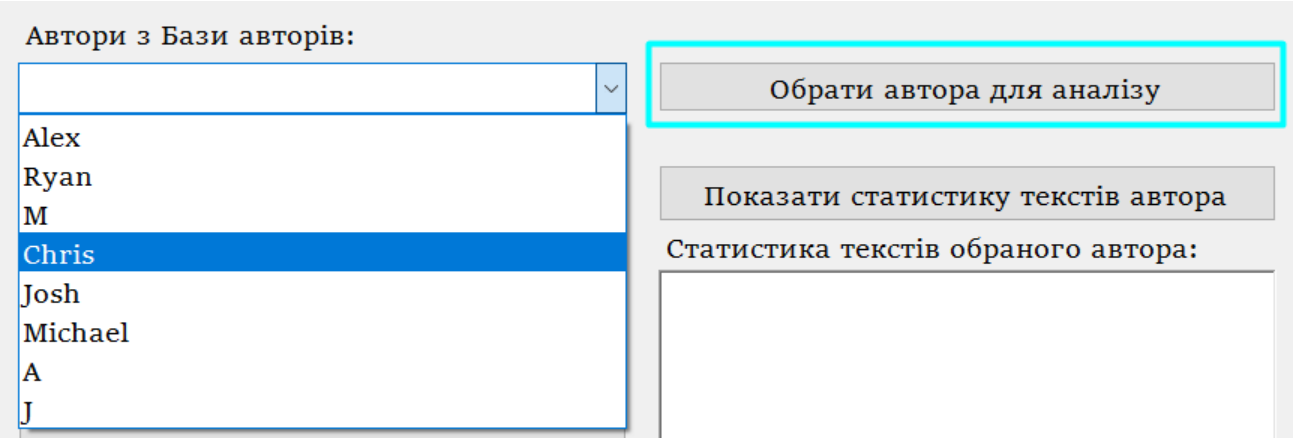


Рисунок 3.8 – Вибір ідентифікатора автора для дослідження

Після вибору автора, у текстове поле «Тексти обраного автора» буде виведено перелік його текстів для аналізу. Виводяться лише ті тексти, які не використовувались у тренуванні моделей машинного навчання.

Для перегляду статистики необхідно натиснути на кнопку «Показати статистику текстів автора» (рисунок 3.9).

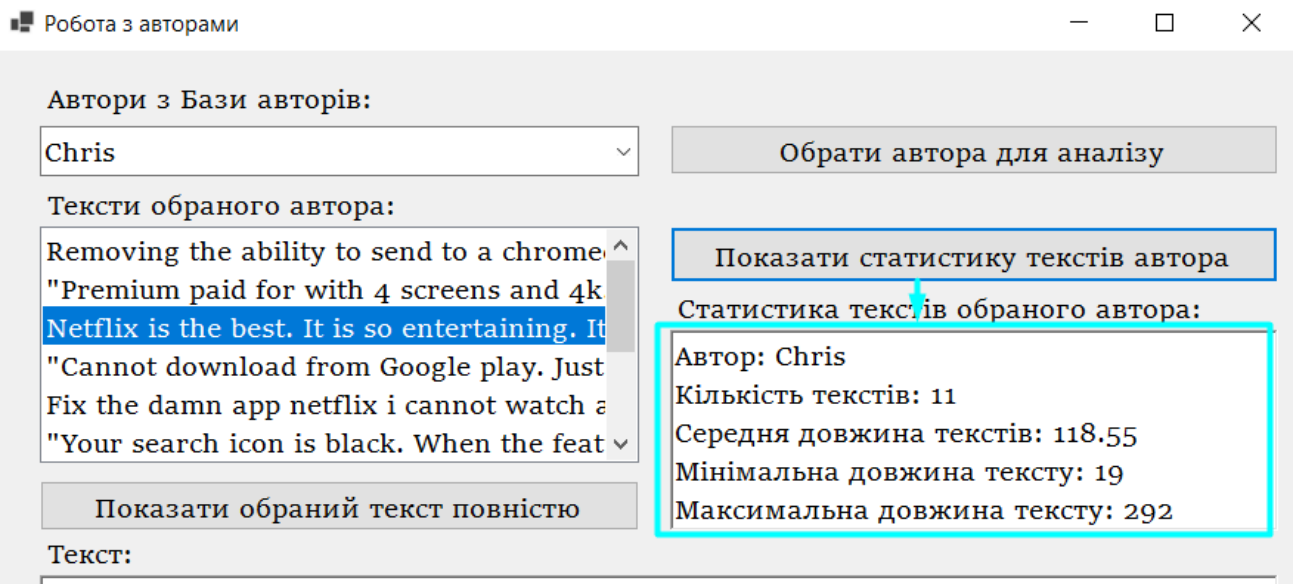


Рисунок 3.9 – Виведення статистики за обраним автором

Для перегляду конкретного тексту, що цікавить користувача, необхідно обрати текст з переліку, та натиснути кнопку «Показати обраний текст повністю» (рисунок 3.10).

Можна модифікувати текст, та зберегти зміни. Для збереження змін необхідно натиснути кнопку «Зберегти зміни». Якщо потрібно додати новий текст, то у текстове поле «Текст:» необхідно ввести авторський текст та натиснути кнопку «Додати текст до обраного автора».

Також можна додати нового автора до переліку авторів, для додавання нового автора необхідно увести ідентифікатор автора у полі «Ім’я автора», а також натиснути кнопку «Додати нового автора».

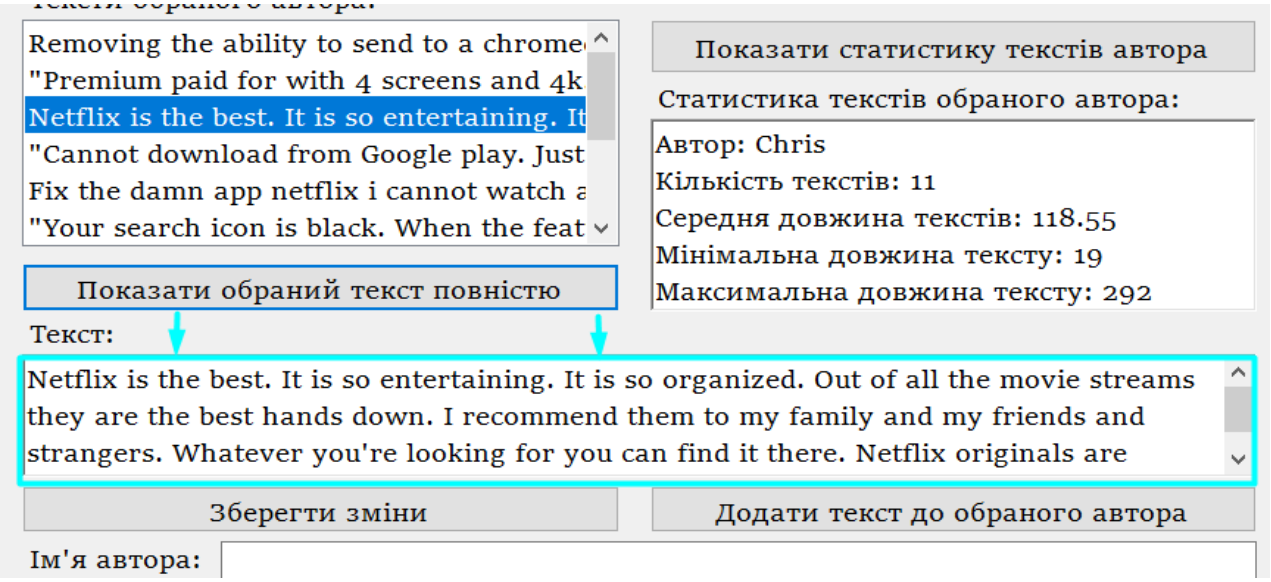


Рисунок 3.10 – Деталізація обраного тексту

Також з даної форми можна перейти до підсистеми інтелектуального визначення авторства за текстом, для чого необхідно натиснути однойменну кнопку на інтерфейсі (рисунок 3.11).

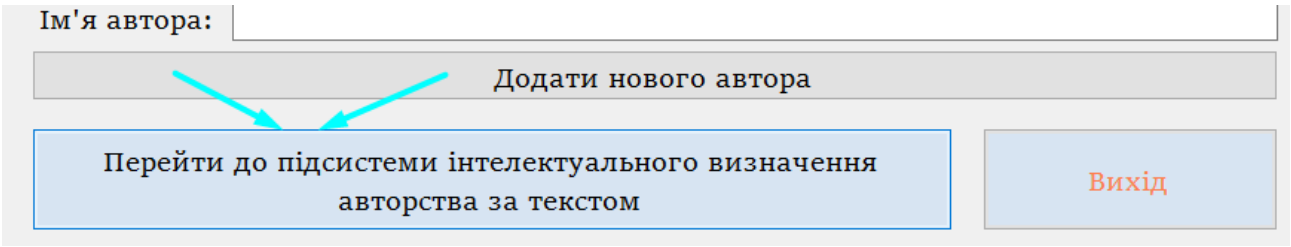


Рисунок 3.11 – Кнопка переходу до підсистеми інтелектуального визначення авторства

Вигляд підсистеми інтелектуального визначення авторства за текстом наведено на рисунку 3.12.

Можна обрати текст для аналізу з бази даних, тексти будуть виведені в текстове поле «Загальний перелік текстів для аналізу», обрати з переліку текст, який необхідно проаналізувати та натиснути кнопку «Показати весь обраний текст» (рисунок 3.13).

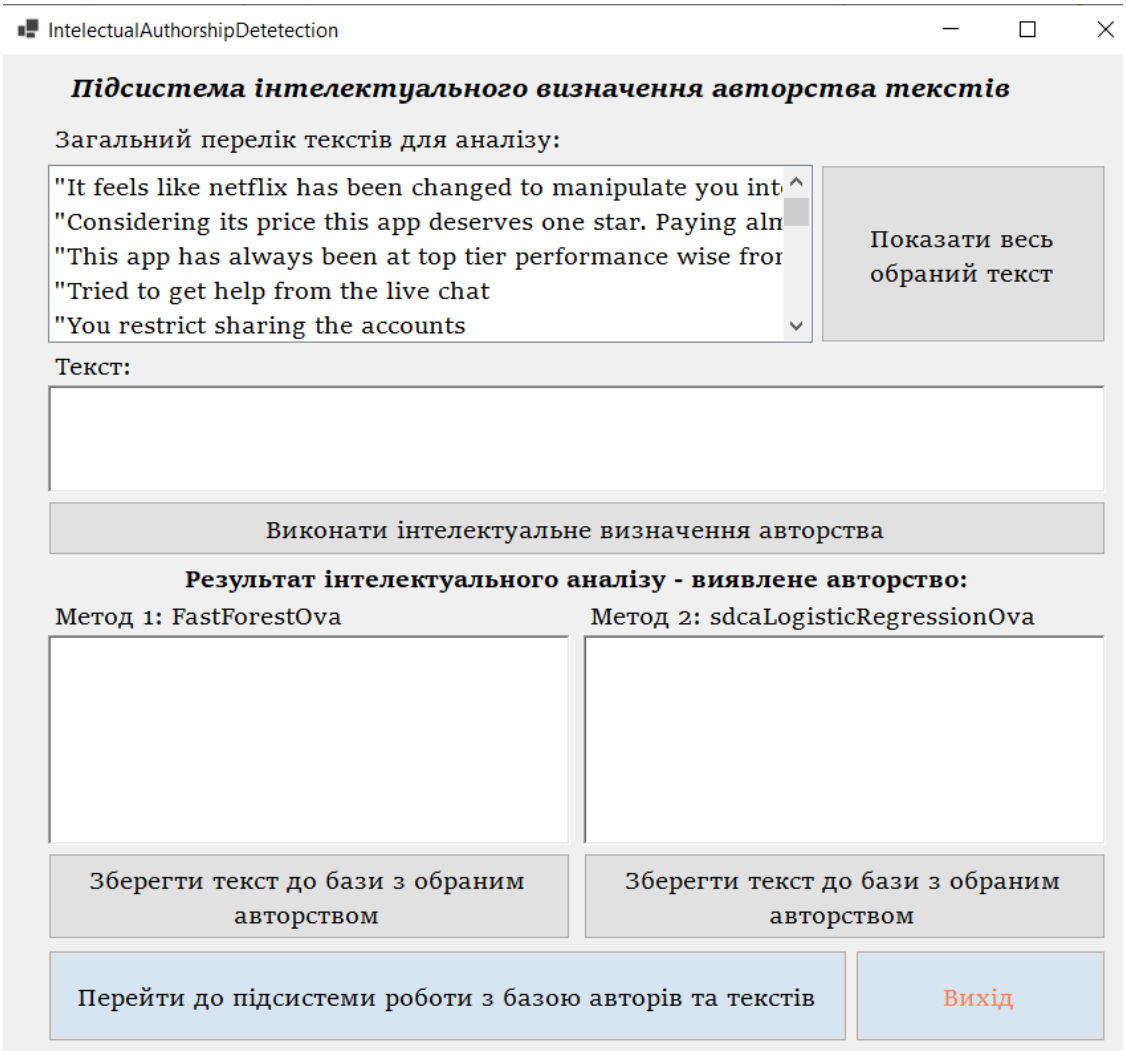


Рисунок 3.12 – Підсистема інтелектуального визначення авторства за текстом

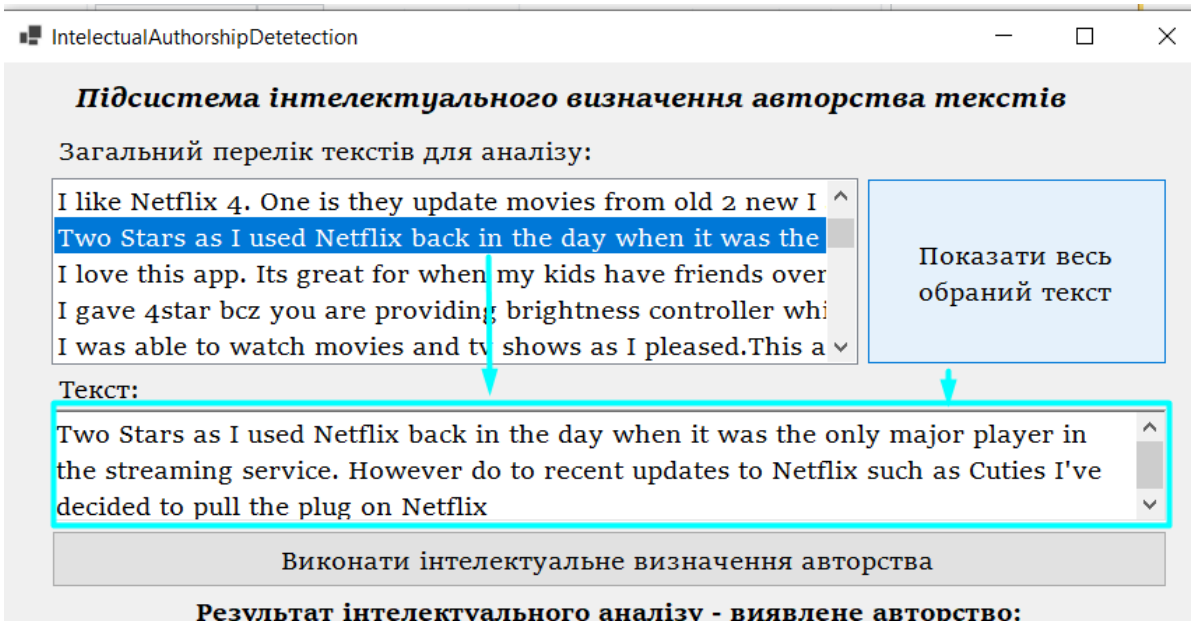


Рисунок 3.13 – Повний текст для аналізу

Також є можливість ввести текст «вручну», розмістивши уведений фрагмент для аналізу в полі «Текст:». Після уведення тексту, для автоматизованого визначення авторства необхідно натиснути кнопку «Виконати інтелектуальне визначення авторства». Після чого буде виконано автоматизоване визначення авторства за двома альтернативними підходами (рисунок 3.14).

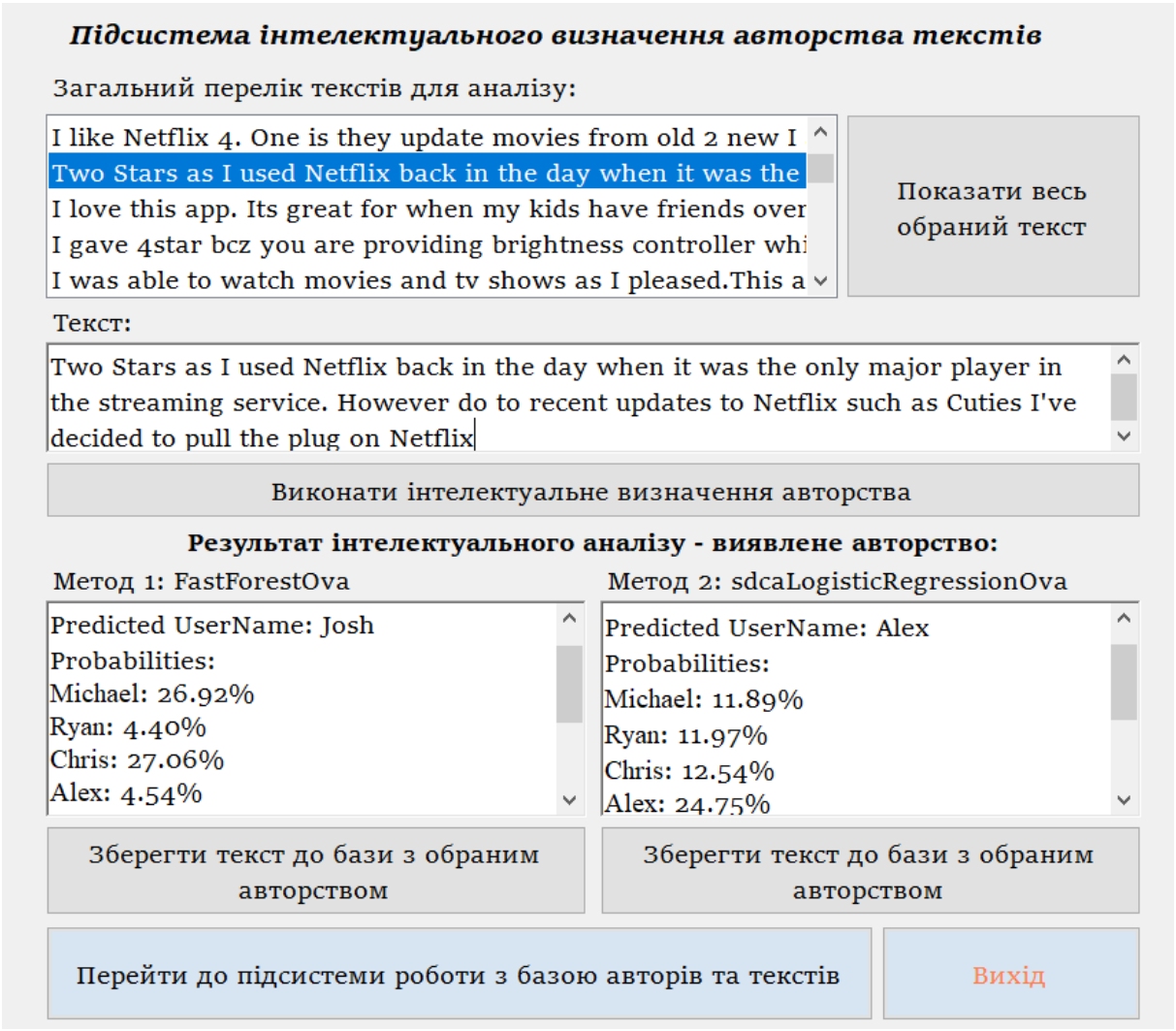


Рисунок 3.14 – Визначення авторства альтернативними моделями машинного навчання

Після завершення етапу інтелектуального аналізу у задачі виявлення авторства користувач побачить визначене авторство, а також відсоткові оцінки приналежності тексту до інших наявних в базі авторів.

Проаналізований текст можна зберегти до бази, як текст автоматизовано визначеного автора. Для цього необхідно натиснути на кнопку «Зберегти текст

до бази з обраним авторством» під моделлю, результат якої влаштовує користувача. Для виходу з програми необхідно натиснути кнопку «Вихід».

Отже, було проведено аналіз функціональності інтелектуальної системи визначення авторства за стилем написання, та описано основні аспекти для плідної роботи зі створеним програмним продуктом.

3.7 Результати досліджень методу інтелектуального визначення авторства текстів за стилем написання

Для дослідження ефективності методу інтелектуального визначення авторства текстів за стилем написання буде використано створене та протестоване програмне забезпечення.

Буде використано дві альтернативних моделі машинного навчання, FastForestOva та sdcaLogisticRegressionOva, між якими буде здійснено порівняння. Для порівняння авторських текстів буде використано розмічений попередньо набір даних, на якому буде обраховано основні метрики, такі як точність, влучність та повнота.

У ході експерименту було задіяно дані 8-ми авторів, тексти яких були попередньо розмічені, але не брали участь у навчанні та тестуванні. На тестових даних за макрометрикою точності було отримано значення 76,86 % для FastForestOva та 72,38 % для sdcaLogisticRegressionOva. Дані решти метрик наведено в таблиці 3.3.

Таблиця 3.3 – Значення макро-метрик оцінки ефективності моделей

	FastForestOva	sdcaLogisticRegressionOva
Точність	76,86 %	72,38 %
Влучність	73,49 %	68,97 %
Повнота	74,62 %	69,02 %

Дані з таблиці 3.3 наглядно проілюстровані на діаграмі (рисунок 3.15).

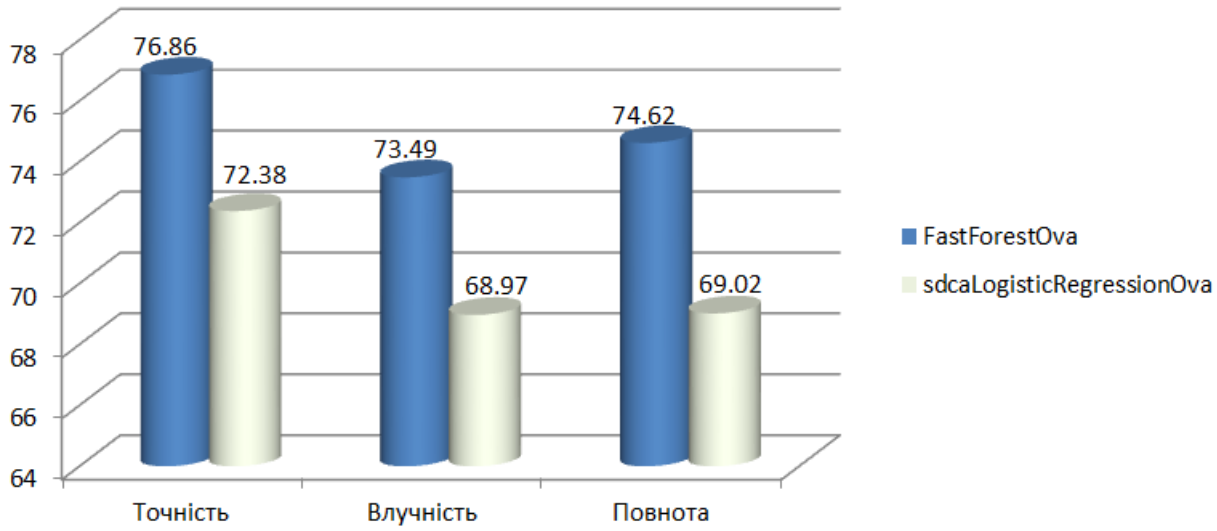


Рисунок 3.15 – Значення макро-метрик для альтернативних моделей машинного навчання

Оскільки модель машинного навчання FastForestOva показала вищі результати по всім 3-м метрикам, її також окремо було досліджено за метриками для кожного з 8-ми авторів. Дані дослідження для моделі машинного навчання FastForestOva наведено в таблиці 3.4.

Таблиця 3.4 – Значення мікро-метрик оцінки ефективності FastForestOva

Ідентифікатор автора	**Влучність**	**Повнота**
Michael	82.46 %	79.30 %
Ryan	68.65 %	71.70 %
Chris	75.38 %	76.58 %
Alex	64.85 %	69.49 %
J	78.68 %	78.90 %
Josh	72.93 %	74.05 %
A	69.98 %	72.27 %
M	74.92 %	74.55 %

Дані дослідження з таблиці 3.3 для наочності подано у вигляді діаграми на рисунку 3.16.

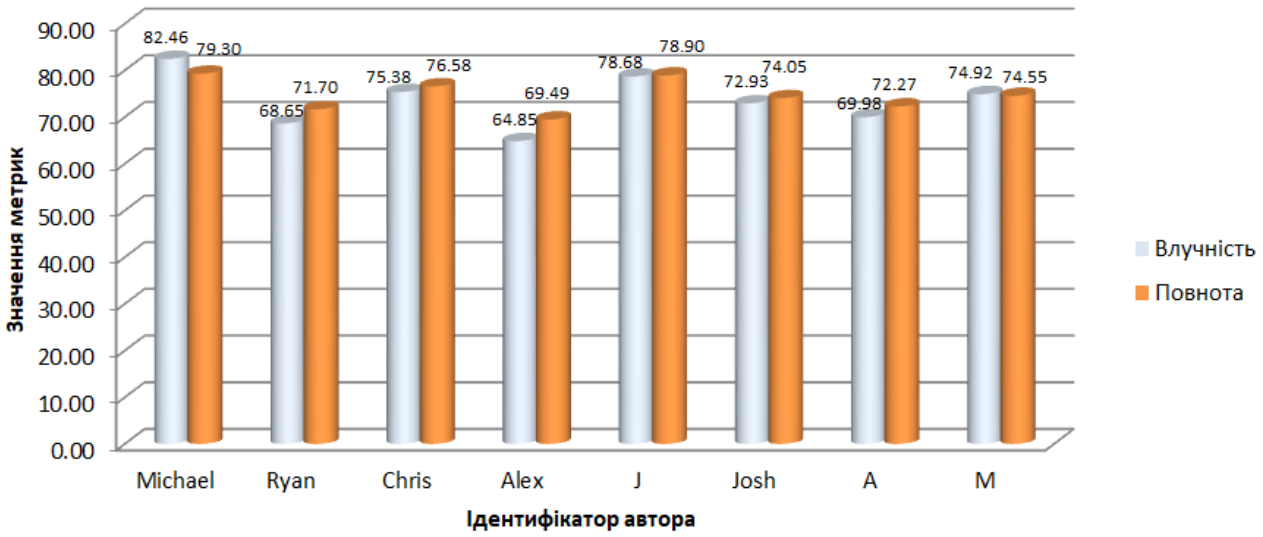


Рисунок 3.16 – Значення мікро-метрик для кожного автора

Автор з ідентифікатором Michael має найвищі показники влучності (82.46%) і повноти (79.30%), що свідчить про те, що модель FastForestOva найкраще розпізнає тексти цього автора. Високі значення обох метрик свідчать про те, що модель майже не допускає помилок у передбаченні текстів Michael.

Автор з ідентифікатором Alex має найнижчі показники влучності (64.85%) і повноти (69.49%), що вказує на те, що модель машинного навчання FastForestOva має найбільші труднощі з ідентифікацією текстів цього автора. Це свідчить про значну кількість хибно позитивних і хибно негативних передбачень для текстів Alex.

Автори з ідентифікаторами Ryan і A також мають нижчі, ніж середні, показники влучності та повноти. Це свідчить про те, що модель має деякі труднощі з правильною ідентифікацією текстів цих авторів, але не так сильно, як з Alex.

Автори з ідентифікаторами J, Chris, Josh і M мають показники влучності та повноти близькі до середніх. Це свідчить про те, що модель машинного навчання має помірну точність у розпізнаванні текстів цих авторів, допускаючи певну кількість помилок.

Зважаючи на предметну область, та на те, що тексти є доволі короткими, отримані результати можна оцінити як хороші. Вдалося досягти точності 76,86 % для експерименту з 8-ми авторами.

Для покращення отриманих результатів необхідно збільшувати кількість даних для навчання, а також можна спробувати інші нейромережеві архітектури, на кшталт BERT. Також у подальших дослідженнях планується також розширити кількість класів.

3.8 Висновки до розділу 3

Було визначено шляхи дослідження та засобів створення програмного забезпечення, для дослідження ефективності інтелектуального визначення авторства текстів за стилем написання було прийнято рішення використати дві альтернативних моделі машинного навчання, між якими буде здійснено порівняння. Обрано використати розмічений попередньо набір даних, на якому буде обраховано основні метрики, такі як точність, влучність та повнота.

Здійснено вибір засобів розробки інформаційної системи інтелектуального визначення авторства текстів за стилем написання. Обрано такий набір засобів: платформа .NET, середовище програмування Visual Studio 2022, мова програмування C#, СКБД SQLserver.

Наведено діаграму класів інтелектуальної системи визначення авторства текстів за стилем написання та описано функціональне призначення програмних складових. За наведеною структурою програмних складових виконано програмну реалізацію складових інтелектуальної системи визначення авторства текстів за стилем написання та описано особливості з програмування та використання підсистеми навчання моделей машинного навчання.

Проведено тестування основних функцій програмної реалізації інтелектуальної системи визначення авторства текстів за стилем написання, яке показало що весь заявлений функціонал працює коректно. Тестування проводилось з використанням юніт-тестів та тест-кейсів. Наведено

підтвердження успішного проходження тестів у вигляді скріншотів роботи програми. Проведено аналіз функціональності інформаційної системи інтелектуального визначення авторства текстів за стилем написання, та описано основні аспекти для плідної роботи зі створеним програмним продуктом.

Було проведено дослідження ефективності методу інтелектуального визначення авторства текстів за стилем написання, для якого обрано використати створене та протестоване програмне забезпечення.

У ході експерименту було задіяно дані 8-ми авторів, тексти яких були попередньо розмічені, але не брали участь у навчанні та тестуванні. На тестових даних за макрометрикою точності було отримано значення 76,86 % для FastForestOva та 72,38 % для sdcaLogisticRegressionOva. Модель машинного навчання FastForestOva показала кращі результати, і її було досліджено ще на мікрометрики.

Зважаючи на предметну область, та на те, що тексти є доволі короткими, отримані результати можна оцінити як хороші, а для покращення отриманих результатів необхідно збільшувати кількість даних для навчання, а також можна спробувати інші нейромережеві архітектури, на кшталт BERT. Також у подальших дослідженнях планується також розширити кількість класів

Загальні висновки

Мету кваліфікаційної роботи бакалавра, спрощення роботи систем експертизи за рахунок автоматизованого визначення авторства текстів за стилем написання, було виконано.

Для досягнення поставленої мети було поставлено та вирішено такі завдання:

– виконано аналіз інформаційних моделей в області інтелектуального визначення авторства текстів за стилем написання;

– проведено огляд теоретичних підходів, а також обрано підхід для інтелектуального визначення авторства текстів за стилем написання у вигляді машинного навчання;

– виконано аналіз існуючих публікацій за напрямком дослідження;

– проведено аналіз існуючого програмного забезпечення області інтелектуального визначення авторства текстів за стилем написання;

– створено метод інтелектуального визначення авторства текстів за стилем написання, що призначений для інтелектуального відстеження зміни поведінки взламаних акаунтів користувачів, а також може бути використаний для відслідковування фактів несанкціонованих текстових запозичень. Метод працює шляхом перетворення вхідних даних у форматі тексту для визначення авторства та попередньо натренованої моделі машинного навчання для інтелектуального визначення авторства у вихідні дані у форматі визначеного автора та оцінки приналежності тексту до визначеного автора, а також оцінки приналежності тексту до решти авторів з бази.;

– описано інформаційну структуру системи для інтелектуального визначення авторства текстів за стилем написання;

– обрано набір даних для інтелектуального визначення авторства текстів;

– створено відповідну програмну реалізацію на основі створеного методу;

– виконано тестування створеної інформаційної системи інтелектуального визначення авторства текстів за стилем написання;

– виконано дослідження ефективності створеного методу інтелектуального визначення авторства текстів за стилем написання з використанням розробленого ПЗ, що показало на тестових даних за макрометрикою точності значення 76,86 % для FastForestOva та 72,38 % для sdcaLogisticRegressionOva.

Зважаючи на специфіку предметної області та те, що тексти є доволі короткими, отримані результати можна вважати задовільними. Для подальшого покращення ефективності моделі необхідно збільшити обсяг навчальних даних. Крім того, варто розглянути застосування альтернативних нейромережевих архітектур, таких як BERT. У майбутніх дослідженнях також планується розширити кількість класів для класифікації.

Перелік посилань

1. nature.com. Authorship identification using ensemble learning. URL: https://www.nature.com/articles/s41598-022-13690-4

2. legalaid.gov.ua. Особливості правового регулювання авторського права в Україні. URL: https://legalaid.gov.ua/novyny/osoblyvosti-pravovogo-regulyuvannya-avtorskogo-prava-v-ukrayini/

3. ips.ligazakon.net. ЗАКОН УКРАЇНИ Про внесення змін до Кримінального та Кримінально-процесуального кодексів України щодо встановлення кримінальної відповідальності за наклеп. URL: https://ips.ligazakon.net/document/JF4XQ00A

4. detector.media. Авторське право під час війни: як медіа використовувати чужі фото та відео. URL: https://detector.media/production/article/204538/2022-11-04-avtorske-pravo-pid-chas-viyny-yak-media-vykorystovuvaty-chuzhi-foto-ta-video/

5. mediamaker.me. Що кажуть закони різних країн про авторські права та штучний інтелект. URL: https://mediamaker.me/hto-cze-zrobyv-shho-kazhut-zakony-riznyh-krayin-pro-avtorski-prava-ta-shtuchnyj-intelekt-3416/

6. zakon.rada.gov.ua. Бернська конвенція про охорону літературних і художніх творів. URL: https://zakon.rada.gov.ua/laws/show/995_051

7. geneva.mfa.gov.ua. Всесвітня організація інтелектуальної власності (ВОІВ). URL: https://geneva.mfa.gov.ua/posolstvo/2610-wipo

8. zakon.rada.gov.ua. Про авторське право і суміжні права. URL: https://zakon.rada.gov.ua/laws/show/3792-12#Text

9. HUANG, Baixiang; CHEN, Canyu; SHU, Kai. Can Large Language Models Identify Authorship?. Arxiv preprint arxiv:2403.08213, 2024

10. Uchendu, A., Le, T., Shu, K., & Lee, D. Authorship attribution for neural text generation. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 8384-8395, 2020

11. Fabien, M., Villatoro-Tello, E., Motlicek, P., & Parida, S. (2020, December). BertAA: BERT fine-tuning for Authorship Attribution. In Proceedings of the 17th International Conference on Natural Language Processing (ICON) (pp. 127-137).

12. Uchendu, Adaku, et al. "Authorship attribution for neural text generation." Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020

13. Alpaydin, Ethem. Machine learning, MIT press, 2021, URL: https://books.google.com.ua/books?hl=uk&lr=&id=2nQJEAAAQBAJ&oi=fnd&pg=PR7&dq=machine+learning&ots=fI_aOb-Dqo&sig=QhNn0h7_qbZo1UpoO7VvQEowiuU&redir_esc

14. JANIESCH, Christian; Zschech, Patrick; Heinrich, Kai. Machine learning and deep learning. Electronic Markets, 2021, 31.3: 685-695.

15. Morris, Christopher, et al. "Tudataset: A collection of benchmark datasets for learning with graphs." arXiv preprint arXiv:2007.08663, 2020

16. Стойко, І. І., & Поливода, А. В. Машинне навчання та способи його використання. Матеріали XII Міжнародної науково-практичної конференції молодих учених та студентів „Актуальні задачі сучасних технологій, 2023, 326-327.

17. Salloum, Said, et al. A systematic literature review on phishing email detection using natural language processing techniques. IEEE Access, 2022, 10: 65703-65727.

18. MA, Thae Ma; Yamamori, Kunihito; Thida, Aye. A comparative approach to Naïve Bayes classifier and support vector machine for email spam classification. In: 2020 IEEE 9th Global Conference on Consumer Electronics (GCCE). IEEE, 2020. p. 324-326.

19. Ramezani, Reza. A language-independent authorship attribution approach for author identification of text documents. Expert Systems with Applications, 2021, 180: 115139.

20. Lavanya, P. M., and E. Sasikala. "Deep learning techniques on text classification using Natural language processing (NLP) in social healthcare network: A comprehensive survey." 2021 3rd international conference on signal processing and communication (ICPSC). IEEE, 2021.

21. Ezen-Can, A. (2020). A Comparison of LSTM and BERT for Small Corpus. arXiv preprint arXiv:2009.05451.

22. GAO, Jing. Network intrusion detection method combining CNN and BiLSTM in cloud computing environment. Computational intelligence and neuroscience, 2022

23. Ulyanovska, Y., Firsov, O., Kostenko, V., Pryadka, O. (2024). Study of the process of identifying the authorship of texts written in natural language. Technology Audit and Production Reserves, 2 (2 (76)), 32–37. doi: https://doi.org/10.15587/2706-5448.2024.301706

24. Deep Learning based Authorship Identification Chen Qian Tianchang He Rao Zhang Department of Electrical Engineering Stanford University, Stanford

25. neoneuro.com. Authorship Attribution Tool. URL: https://neoneuro.com/products/authorship-attribution

26. NeoNeuro. Authorship Attribution. URL: https://www.google.com/url?sa=i&url=https%3A%2F%2Fneoneuro.com%2Fproducts%2Fauthorship-attribution&psig=AOvVaw3iEV2s7gnI0NoRAtmhyt_n&ust=1717068394968000&source=images&cd=vfe&opi=89978449&ved=0CBQQjhxqFwoTCODwrp_hsoYDFQAAAAAdAAAAABAE

27. docs.expert.ai. Writeprint detection. URL: https://docs.expert.ai/nlapi/v2/guide/detection/writeprint

28. Netflix Reviews [DAILY UPDATED]. URL: https://www.kaggle.com/datasets/ashishkumarak/netflix-reviews-playstore-daily-updated/data , Spotify Reviews. URL: https://www.kaggle.com/datasets/ashishkumarak/spotify-reviews-playstore-daily-update

29. ML.NET Machine learning library from Microsoft. URL: https://mareks-082.medium.com/ml-net-machine-learning-library-from-microsoft-39d265761b34

30. Windows Forms. URL: https://en.wikipedia.org/wiki/Windows_Forms

31. Accuracy vs. Precision vs. Recall in Machine Learning: What is the Difference. URL: https://encord.com/blog/classification-metrics-accuracy-precision-recall/].

32. Overview of .NET Framework. URL: https://learn.microsoft.com/en-gb/dotnet/framework/get-started/overview

33. Microsoft Visual Studio. URL: https://uk.wikipedia.org/wiki/Microsoft_Visual_Studio

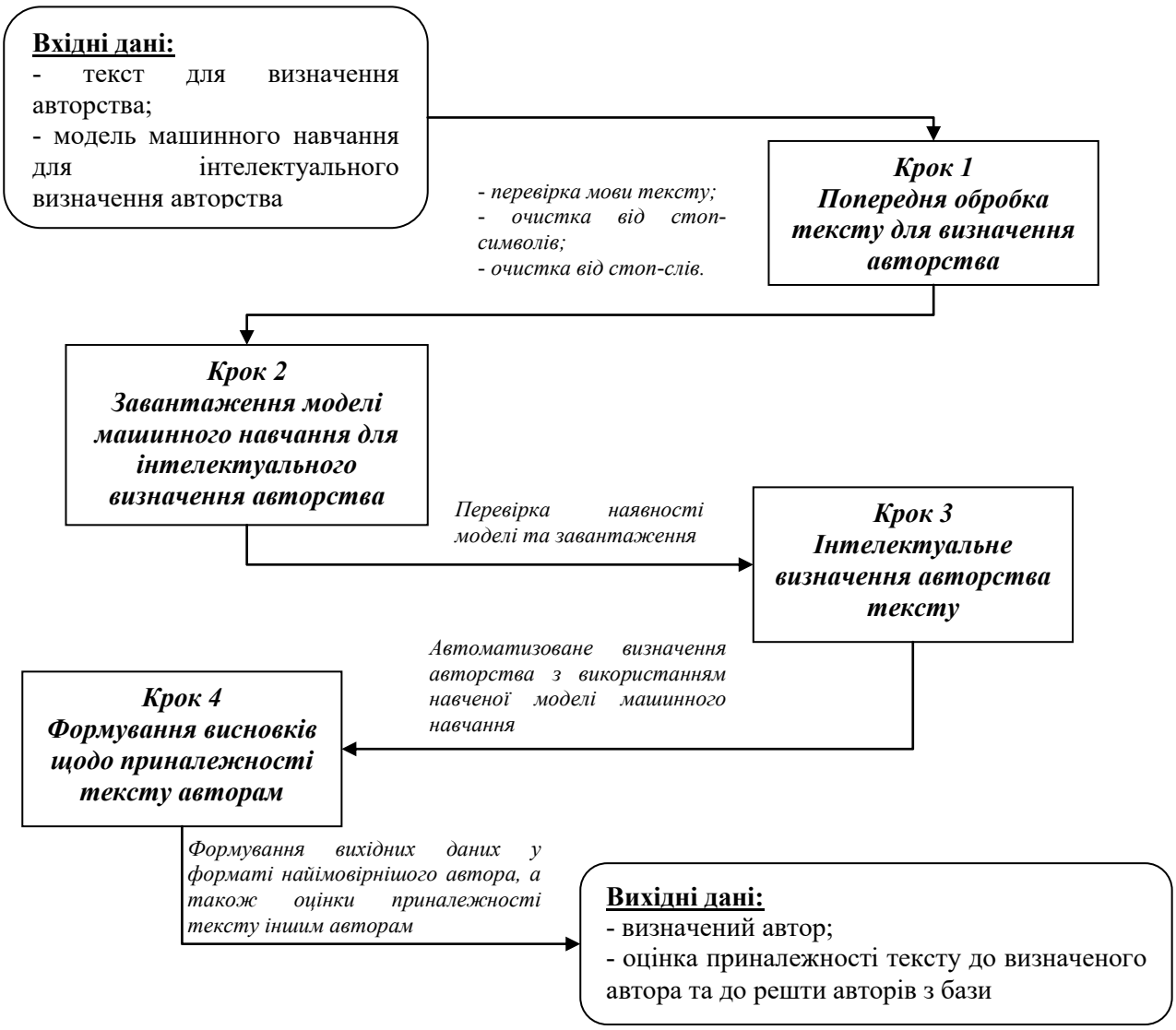
34. 7 причин, чому навчання C# – гарна інвестиція у кар'єру в IT. URL: https://www.volynpost.com/articles/2265-7-prychyn-chomu-navchannia-c--garna-investyciia-u-karieru-v-it

35. Microsoft SQL Server. URL: https://uk.wikipedia.org/wiki/Microsoft_SQL_Server

ДОДАТКИ

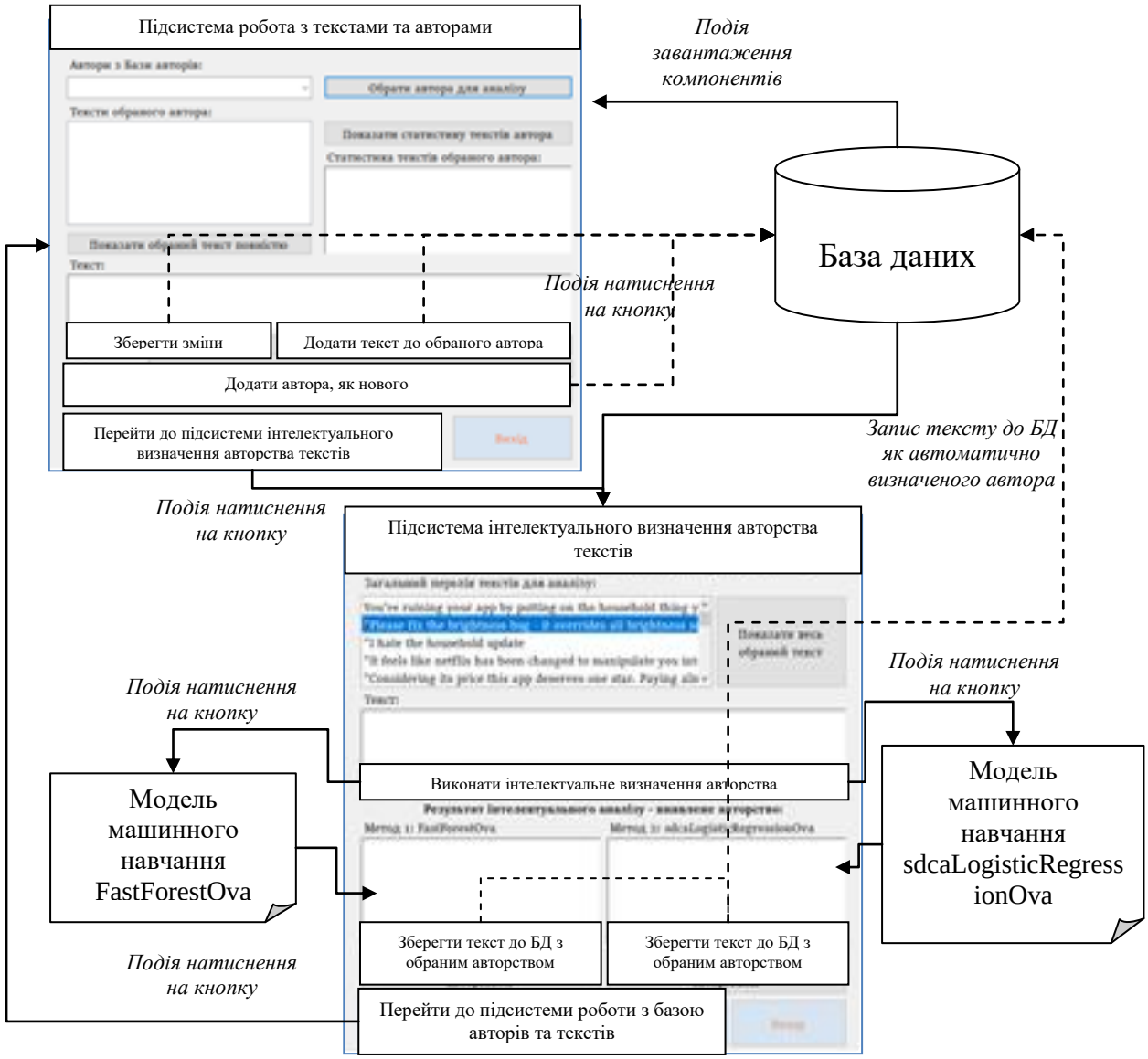
Додаток А

Кроки методу інтелектуального визначення авторства текстів



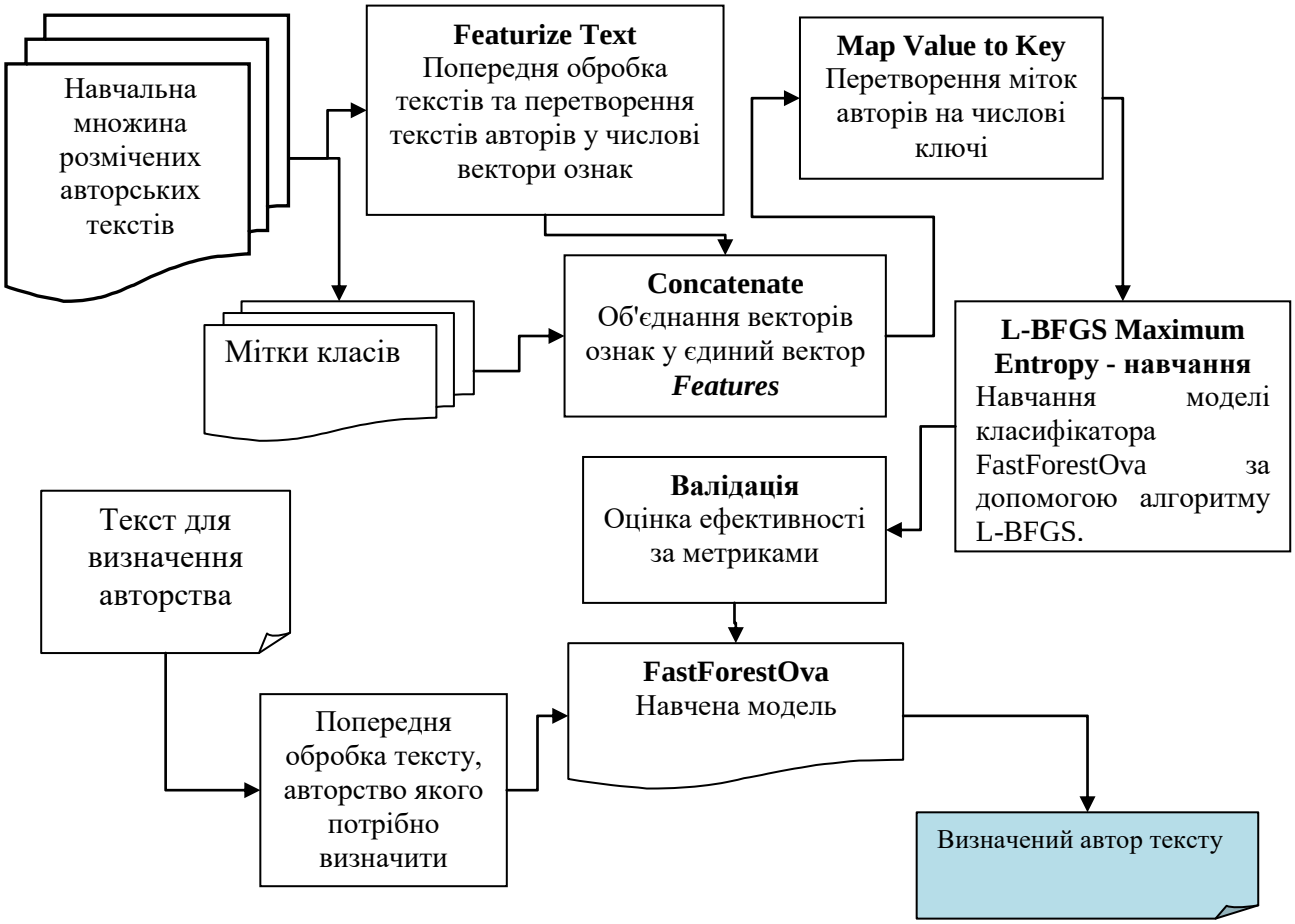
Додаток Б

Автоматизація обробки потоків даних інтелектуальної системи визначення авторства



Додаток В

Пайплайн типової моделі машинного навчання на прикладі FastForestOva



Додаток Г

Презентаційний матеріал

КВАЛІФІКАЦІЙНА РОБОТА БАКАЛАВРА

МЕТОД ІНТЕЛЕКТУАЛЬНОГО ВИЗНАЧЕННЯ АВТОРСТВА ТЕКСТІВ ЗА СТИЛЕМ НАПИСАННЯ

Виконав:
студент групи КН-20-2
Богдан ШПОРТ

Керівник:
ст. викладач каф. КН
Тетяна СКРИПНИК

Актуальність

Актуальність застосування методів інтелектуального визначення авторства текстів за стилем написання зростає в умовах сучасного інформаційного суспільства, де цифрова комунікація є важливою складовою багатьох сфер життя. Зокрема, одним із ключових завдань є інтелектуальне відстеження зміни поведінки взламаних акаунтів користувачів. Виявлення таких змін дозволяє своєчасно ідентифікувати випадки компрометації акаунтів та запобігти потенційним негативним наслідкам, зокрема поширенню дезінформації або шахрайству.

Пошук першоджерел та каналів розповсюдження пропаганди є ще однією важливою областю застосування методів інтелектуального визначення авторства текстів. У сучасному світі, де інформаційні війни та пропаганда стали звичними явищами, ідентифікація джерел пропагандистських матеріалів є критично важливою для забезпечення інформаційної безпеки та формування об'єктивної суспільної думки. Технології аналізу стилю написання текстів можуть суттєво сприяти виявленню та блокуванню пропагандистських ресурсів.

Відслідковування фактів несанкціонованих текстових запозичень також є важливою проблемою, особливо в академічному середовищі, де дотримання принципів академічної доброчесності є фундаментальним. Використання методів інтелектуального аналізу текстів для виявлення плагіату дозволяє не тільки захистити авторські права, але й забезпечити справедливість і об'єктивність у наукових дослідженнях. Таким чином, розробка та вдосконалення методів інтелектуального визначення авторства текстів є актуальною та важливою задачею в сучасному цифровому світі.

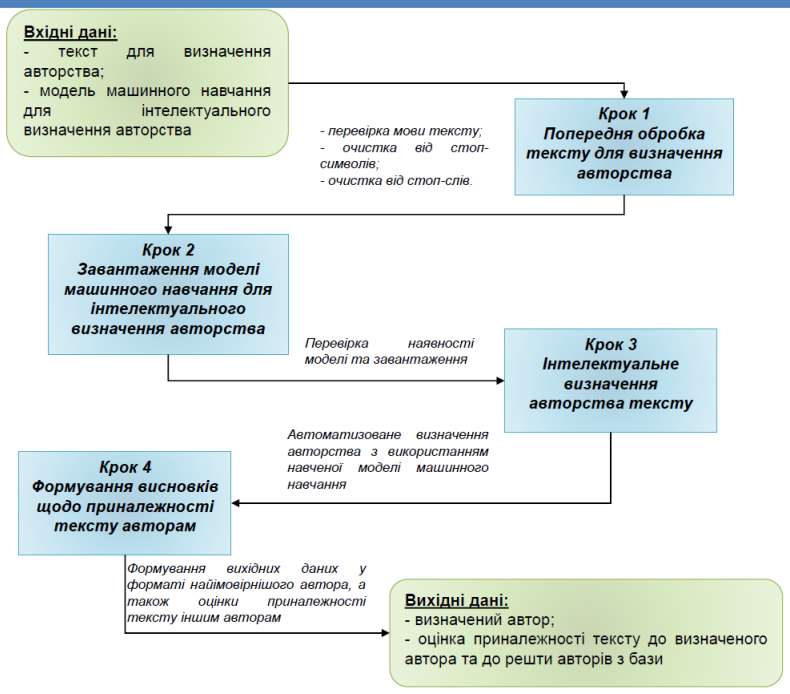
Мета і задачі роботи

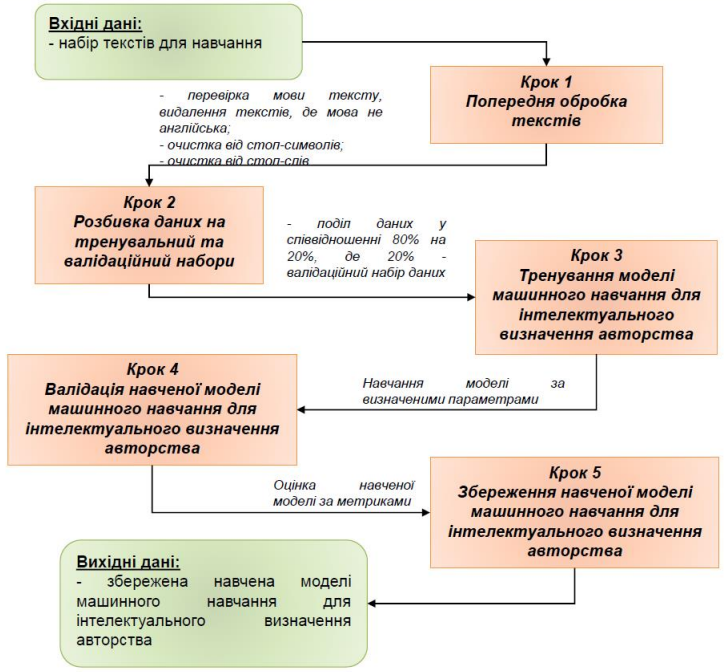
Метою кваліфікаційної роботи бакалавра є спрощення роботи систем експертизи за рахунок автоматизованого визначення авторства текстів за стилем написання.

Для досягнення поставленої мети слід вирішити такі **завдання**:

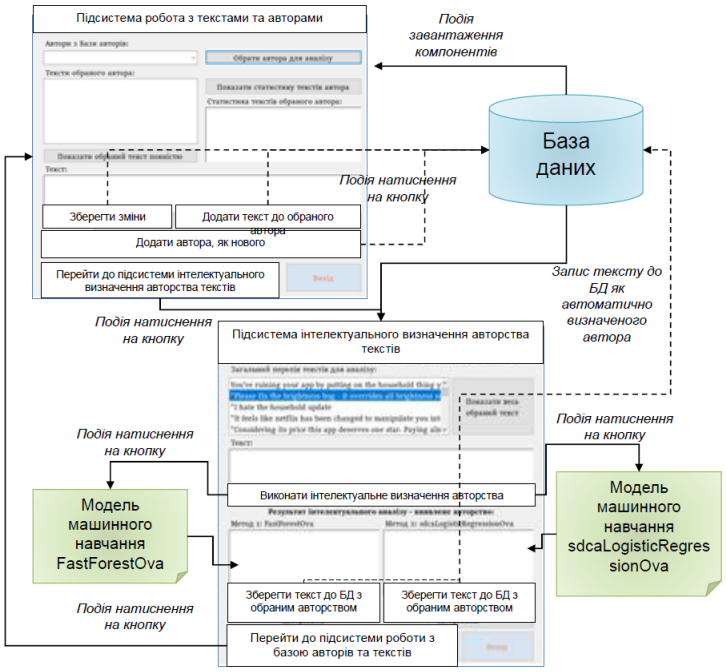
- виконати аналіз інформаційних моделей області інтелектуального визначення авторства текстів за стилем написання;
- виконати огляд теоретичних підходів, а також обрати підхід для інтелектуального визначення авторства текстів за стилем написання;
- виконати аналіз існуючих публікацій за напрямком дослідження;
- провести аналіз існуючого програмного забезпечення області інтелектуального визначення авторства текстів за стилем написання;
- створити метод інтелектуального визначення авторства текстів за стилем написання;
- описати інформаційну структуру системи для інтелектуального визначення авторства текстів за стилем написання;
- обрати набір даних для інтелектуального визначення авторства текстів;
- створити відповідну програмну реалізацію на основі створеного методу;
- виконати тестування створеного програмного забезпечення;
- виконати дослідження ефективності створеного методу інтелектуального визначення авторства текстів за стилем написання з використанням розробленого ПЗ.

Кроки методу інтелектуального визначення авторства текстів

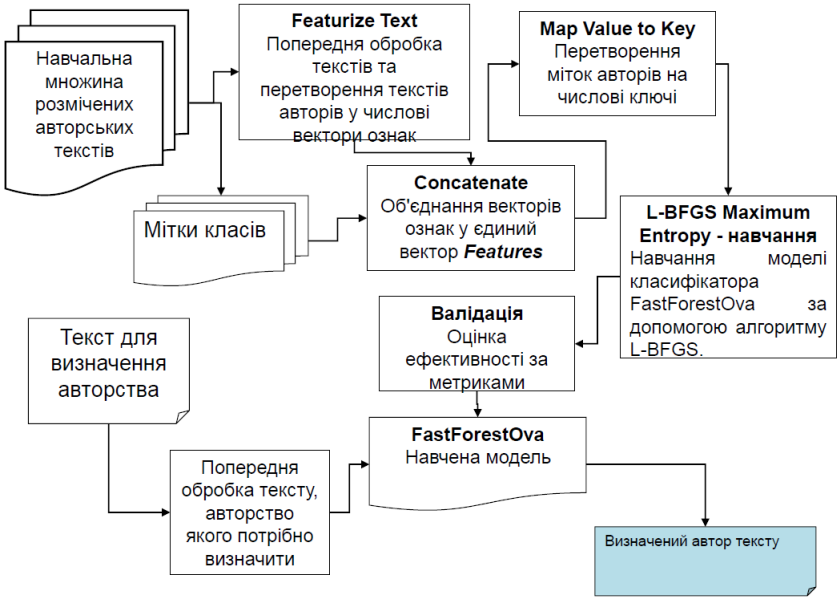




Кроки отримання навченої моделі машинного навчання для інтелектуального визначення авторства



Автоматизація обробки потоків даних інтелектуальної системи визначення авторства

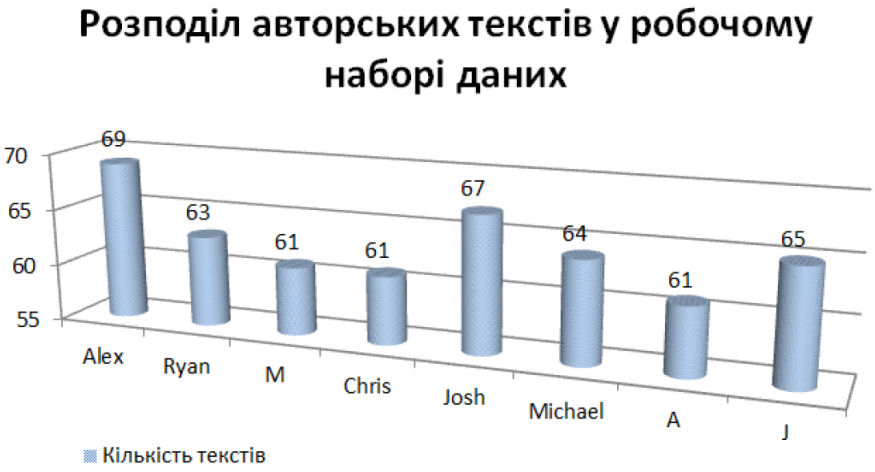


Пайплайн типової моделі машинного навчання на прикладі FastForestOva

Функціональна структура інтелектуальної системи визначення авторства текстів

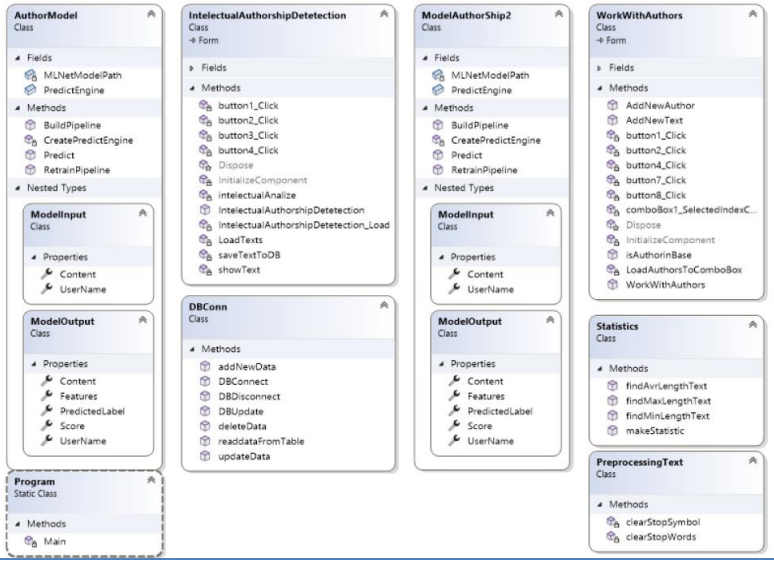
Даталогічна модель бази даних для методу інтелектуального визначення авторства текстів за стилем написання

Набір даних дослідження

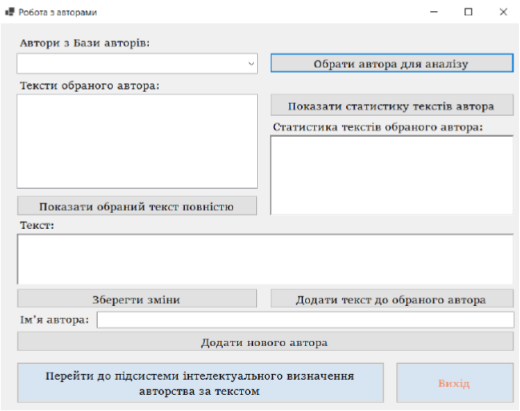


Для навчання моделей машинного навчання що спроможні виявляти авторство текстів за стилем написання було використано два набори даних: «Netflix Reviews» та «Spotify Reviews»

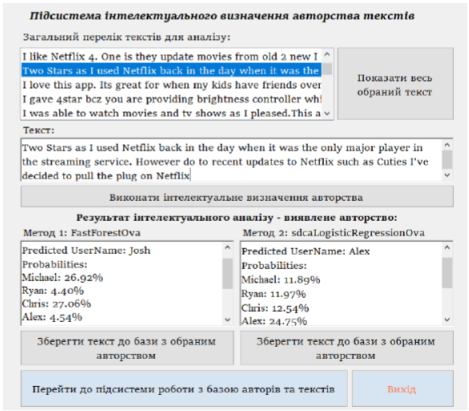
Діаграма класів інтелектуальної системи визначення авторства



Інтелектуальна системи визначення авторства



Підсистема роботи з текстами та авторами



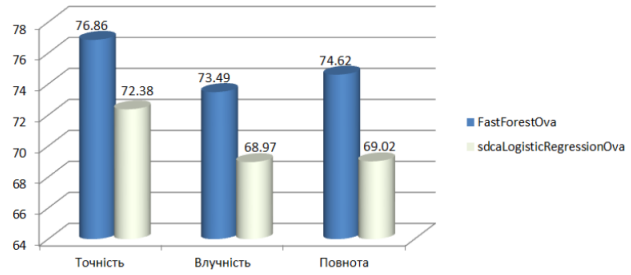
Визначення авторства альтернативними моделями машинного навчання

Результати досліджень

У ході експерименту було задіяно дані 8-ми авторів, тексти яких були попередньо розмічені, але не брали участь у навчанні та тестуванні. На тестових даних за макрометрикою точності було отримано значення 76,86 % для FastForestOva та 72,38 % для sdcaLogisticRegressionOva

	FastForestOva	sdcaLogisticRegressionOva
Точність	76,86 %	72,38 %
Влучність	73,49 %	68,97 %
Повнота	74,62 %	69,02 %

Значення макро-метрик оцінки ефективності моделей

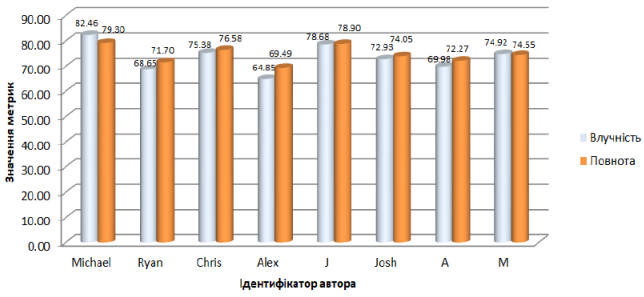


Значення макро-метрик для альтернативних моделей машинного навчання

Результати досліджень

Ідентифікатор автора	Влучність	Повнота
Michael	82.46 %	79.30 %
Ryan	68.65 %	71.70 %
Chris	75.38 %	76.58 %
Alex	64.85 %	69.49 %
J	78.68 %	78.90 %
Josh	72.93 %	74.05 %
A	69.98 %	72.27 %
M	74.92 %	74.55 %

Значення мікро-метрик оцінки ефективності FastForestOva



Значення мікро-метрик для кожного автора

Висновки

Метою кваліфікаційної роботи бакалавра було спрощення роботи систем експертизи за рахунок автоматизованого визначення авторства текстів за стилем написання. Мету було досягнуто.

Для досягнення поставленої мети було поставлено та вирішено такі завдання:

- виконано аналіз інформаційних моделей в області інтелектуального визначення авторства текстів за стилем написання;
- проведено огляд теоретичних підходів, а також обрано підхід для інтелектуального визначення авторства текстів за стилем написання у вигляді машинного навчання;
- виконано аналіз існуючих публікацій за напрямком дослідження;
- проведено аналіз існуючого програмного забезпечення області інтелектуального визначення авторства текстів за стилем написання;
- створено метод інтелектуального визначення авторства текстів за стилем написання, що призначений для інтелектуального відстеження зміни поведінки взламаних акаунтів користувачів, а також може бути використаний для відслідковування фактів несанкціонованих текстових запозичень;
- описано інформаційну структуру системи для інтелектуального визначення авторства текстів за стилем написання;
- обрано набір даних для інтелектуального визначення авторства текстів;
- створено відповідну програмну реалізацію на основі створеного методу;
- виконано тестування створеного програмного забезпечення;
- виконано дослідження ефективності створеного методу інтелектуального визначення авторства текстів за стилем написання з використанням розробленого ПЗ.

ДЯКУЮ ЗА УВАГУ!

Anti-Plagiarism v-15.257

Максимальне співпадіння з одним документом 3.0%

Словники перевірки: en_US, ru_RU, ua_UA. **Помилок в документах: 8%**

	Документ		Сумарний збіг по Базі Даних	
ID: 132036 Назва: КВАЛІФІКАЦІЙНА РОБОТА БАКАЛАВРА на тему Метод інтелектуального визначення авторства текстів за стилем написання Додано в БД: 2024-06-21 Автора: Богдан ШПОРТ Керівники: Тетяна СКРИПНИК Консультанти: Опоненти:	Символи	Лексеми	Символи	Лексеми
	80302	1168	4642 (6%)	73 (6%)

Джерело плагіату

ID	Опис	Наявність плагіату в документі	
		Символи	Лексеми

UNICHECK
by Turnitin

Ім'я користувача:
Кафедра КН

ID перевірки:
1016379890

Дата перевірки:
21.06.2024 07:43:23 EEST

Тип перевірки:
Doc vs Internet + Library

Дата звіту:
21.06.2024 13:53:56 EEST

ID користувача:
100005671

Назва документа: КН-20-2 Шпорт_ЗАПИСКА

Кількість сторінок: 69 **Кількість слів:** 12268 **Кількість символів:** 101344 **Розмір файлу:** 1.50 MB **ID файлу:** 1016189017

Виявлено модифікації тексту (можуть впливати на відсоток схожості)

12.1%
Схожість

Найбільша схожість: 4.27% з джерелом з Бібліотеки (ID файлу: 1016181930)

6.53% Джерела з Інтернету	736	Сторінка 71
8.29% Джерела з Бібліотеки	101	Сторінка 75

0% Цитат

Вилучення цитат вимкнене

Вилучення списку бібліографічних посилань вимкнене

0%
Вилучень

Немає вилучених джерел

Модифікації

Виявлено модифікації тексту. Детальна інформація доступна в онлайн-звіті.

Підозріле форматування	11 сторінок

РІШЕННЯ ЕКСПЕРТНОЇ КОМІСІЇ КАФЕДРИ КОМП'ЮТЕРНИХ НАУК
ПРО ДОПУСК КВАЛІФІКАЦІЙНОЇ РОБОТИ ДО ЗАХИСТУ

Підтверджуємо ознайомлення з результатом звіту подібності щодо роботи, генерованого системою виявлення текстових збігів/ідентичності/схожості:

Назва: Метод інтелектуального визначення авторства текстів за стилем написання

Автор: *студент групи КН-20-2 Богдан Шпорт*

Спеціальність: *122 – Комп'ютерні науки*

Освітня програма: *освітньо-професійна*

Науковий керівник: *ст.викладач каф. КН Тетяна Скрипник*

Після аналізу звіту подібності зроблено такий висновок:

№	Висновок	Позначка про відповідність
1	Запозичення, виявлені в роботі, є законними і не є плагіатом. Робота приймається до захисту.	***відповідає***
2	Виявлені запозичення не є плагіатом, розміщені в розділах, які не описують безпосередньо авторське дослідження, але кількість цитат перевищує обсяг, виправданий поставленою метою роботи Робота приймається до захисту, але має бути відкоригована. Відкоригований варіант має бути поданий на кафедру за 2 дні до захисту, разом із заявою щодо самостійності виконання письмової роботи та ідентичності друкованої та електронної версії роботи	
3	Виявлені запозичення не є плагіатом, але частково розміщені в розділах, які описують безпосередньо авторське дослідження, а кількість цитат перевищує обсяг, виправданий поставленою метою роботи. В зв'язку з цим мета роботи та поставлені завдання не були досягнені. Робота може бути допущена до захисту (наступного року) після того як буде відкоригована та допрацьована і успішно пройде повторну перевірку на академічний плагіат.	
4	Робота містить навмисні текстові спотворення, передбачувані спроби укриття запозичень або інші прояви академічного плагіату. Робота містить фабрикацію або фальсифікацію даних. Робота не допускається до захисту.	

Підтвердження:

Запозичення, виявлені в роботі Владислава Держака, не є плагіатом, оскільки: запозичення розміщені в розділі огляду існуючих підходів, не описують безпосередньо авторську роботу і не стосуються її результатів; усі запозичення фрагментарні; до запозичень входять фрагменти програмного коду, що не мають авторства і містять поширені конструкції; серед запозичень знаходяться загальновідомі терміни, скорочення.

Обсяг запозичень, визначений системами виявлення збігів/ідентичності/схожості, складає:

- за системою Anti-Plagiarism: 3%;

- за системою Unicheck: 12,1 %.

Керівник роботи _____ *Тетяна СКРИПНИК*

Гарант ОП _____ *Олександр МАЗУРЕЦЬ*

Завідувач кафедри КН _____ *Олександр БАРМАК*

ХМЕЛЬНИЦЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ
МОН УКРАЇНИ

Кафедра комп'ютерних наук

РЕЦЕНЗІЯ

на кваліфікаційну роботу бакалавра

студента *гр. КН-20-2 Шпорта Богдана Віталійовича*

за темою: *Метод інтелектуального визначення авторства текстів за стилем написання*

1. Актуальність обраної теми

У сучасному інформаційному суспільстві, де цифрова комунікація є ключовим елементом багатьох сфер життя, актуальність застосування методів інтелектуального визначення авторства текстів за стилем написання значно зростає. Дані методи стають незамінними інструментами для забезпечення інформаційної безпеки, зокрема для виявлення компрометації користувацьких акаунтів. Здатність системи аналізувати стиль письма дозволяє виявляти аномалії та своєчасно ідентифікувати несанкціоноване використання акаунтів., що допомагає запобігти потенційним загрозам, таким як поширення дезінформації або шахрайські дії, що мають негативні наслідки для користувачів та організацій.

2. Повнота розкриття мети та завдань роботи

У кваліфікаційній роботі бакалавра автор всебічно розкриває поставлену мету роботи, ретельно виконуючи всі завдання, визначені в межах обраної теми. Кожне завдання детально опрацьовано, що дозволяє досягти повного розуміння досліджуваного питання.

3. Зміст кожного розділу роботи

Робота містить три розділи. В першому розділі подано Характеристика предметної області визначення авторства текстів за стилем написання, виконано огляд теоретичних підходів щодо задачі визначення авторства текстів за стилем написання та розглянуто існуючі рішення. У другому розділі описано створений метод інтелектуального визначення авторства текстів за стилем написання та наведено функціональну структуру інтелектуальної системи визначення авторства текстів та взаємозв'язок компонентів. У третьому розділі визначено шляхи дослідження та засоби створення інформаційної системи інтелектуального визначення авторства текстів. За визначеним планом досліджено метод інтелектуального визначення авторства текстів за стилем написання.

4. Оцінка розробленої інформаційної системи, її практична цінність

Розроблена система призначена для адміністраторів соціальних мереж і може використовуватися також у наукових дослідженнях. Автоматизація процесу визначення авторства текстів за стилем написання дозволяє ефективно виявляти факти несанкціонованого запозичення текстів та відслідковувати зміни поведінки в зламаних акаунтах користувачів. Основними напрямками практичного використання цієї інформаційної системи є автоматизоване визначення авторства текстів за стилістичними ознаками.

5. Якість оформлення кваліфікаційної роботи бакалавра

Оформлення роботи відповідає необхідним нормам та вимогам, які ставляться до оформлення кваліфікаційних робіт. Структура роботи, правильність використання наукової термінології, цитування та посилання на використані джерела, а також стиль написання відповідає встановленим стандартам спеціальності.

6. Недоліки кваліфікаційної роботи бакалавра

При аналізі існуючих програмних засобів та наукових рішень до визначення авторства текстів за стилем написання, у деяких наведено лише логотипи програм, а не їх вигляд. На сторінці 54 подано два рисунки підряд, без основного тексту записки. Проте дані зауваження не впливають на якість отриманих результатів.

7. Загальний висновок (допускається чи не допускається до захисту), та оцінка на яку заслуговує кваліфікаційна робота.

Враховуючи рівень виконання та забезпечення усіх необхідних вимог, робота може бути допущена до захисту. Рекомендована оцінка ***«відмінно»****.*

Рецензент Валерій МАРТИНЮК

ХМЕЛЬНИЦЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ
МОН УКРАЇНИ

Кафедра комп'ютерних наук

ВІДГУК НАУКОВОГО КЕРІВНИКА

на кваліфікаційну роботу бакалавра

студента *гр. КН-20-2 Шпорта Богдана Віталійовича*

за темою *Метод інтелектуального визначення авторства текстів за стилем написання*

1. Актуальність теми

Актуальність застосування методів інтелектуального визначення авторства текстів за стилем написання значно зростає в умовах сучасного інформаційного суспільства, де цифрова комунікація відіграє важливу роль у багатьох сферах життя. Одним із ключових завдань є інтелектуальне відстеження змін поведінки користувачів у зламаних акаунтах. Виявлення таких змін дозволяє своєчасно ідентифікувати випадки компрометації акаунтів та запобігти потенційним негативним наслідкам, таким як поширення дезінформації або вчинення шахрайських дій. Використання цих методів забезпечує підвищення безпеки інформаційного простору та зменшує ризики, пов'язані з цифровими загрозами.

2. Відповідність роботи предметній області Стандарту спеціальності 122 Комп'ютерні науки

За стандартом, а саме описом предметної області, об'єктом дослідження є процес визначення авторства текстів за стилем написання NLP-засобами. Метою роботи є спрощення роботи систем експертизи за рахунок автоматизованого визначення авторства текстів за стилем написання. При вирішенні поставленої задачі використано методи та засоби машинного навчання для роботи з текстовою інформацією. Враховуючи це, результати виконання кваліфікаційної роботи бакалавра повністю відповідають стандарту бакалавра спеціальності 122 – Комп'ютерні науки.

3. Професійні та особистісні якості бакалавра

Шпорт Богдан Віталійович, працюючи над кваліфікаційною роботою бакалавра, проявив цілеспрямованість і дисциплінованість, а також продемонстрував належний рівень знань та навичок у сфері "Комп'ютерні науки".

4. Ступінь самостійності під час виконання кваліфікаційної роботи

Кваліфікаційна робота бакалавра студента Шпорта Богдана Віталійовича виконана здобувачем самостійно, із зазначенням посилань на використані джерела.

5. Ступінь оволодіння методами дослідження

Студент продемонстрував високий рівень володіння методами дослідження, які були застосовані у його роботі. Він не лише успішно застосував ці методи, але й показав глибоке розуміння їх теоретичних основ та практичних аспектів. Його вміння обирати і використовувати відповідні методики сприяли досягненню якісних та надійних результатів у дослідженні.

6. Повнота та якість розкриття теми роботи

Мета дослідження була повністю досягнута, а отримані результати ретельно розкриті та обґрунтовані. У роботі представлено детальний аналіз і пояснення кожного етапу дослідження, що підтверджує обґрунтованість висновків. Кожен результат підкріплено відповідними даними та теоретичними положеннями, що свідчить про глибокий і всебічний підхід до розв'язання поставлених завдань.

7. Логічність, послідовність, аргументованість, літературна грамотність викладення матеріалу

Матеріал роботи представлено в логічній та послідовній формі, з чітким обґрунтуванням кожного положення. Використана мова і стиль відповідають встановленим вимогам для наукових робіт у галузі 122 – Комп'ютерні науки, що сприяє зрозумілості та доступності викладеного матеріалу. Кожний розділ роботи логічно пов'язаний з попереднім і наступним, забезпечуючи цілісність та системність дослідження.

8. Можливість практичного застосування кваліфікаційної роботи бакалавра, окремих її частин

Основними напрямками практичного застосування розробленої інформаційної системи є автоматизоване визначення авторства текстів на основі аналізу стилю написання. Впровадження такої системи дозволяє значно підвищити точність і ефективність процесу визначення авторства, знижуючи вплив людського фактору та мінімізуючи можливість помилок.

9. Висновок про можливість допуску кваліфікаційної роботи бакалавра до захисту, на яку оцінку заслуговує робота

*Враховуючи рівень виконання та забезпечення усіх необхідних вимог, робота може бути допущена до захисту. Рекомендована оцінка «**відмінно**» .*

Керівник _____ *ст.викладач каф. КН Тетяна СКРИПНИК*