Хмельницький національний університет
Факультет інформаційних технологій
Кафедра комп'ютерних наук

КВАЛІФІКАЦІЙНА РОБОТА БАКАЛАВРА

на тему *Спосіб для визначення семантично невідповідних україномовних коментарів до описів товарів інтернет-магазину для задач аналізу купівельного попиту*

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

Виконав: *студент 3 курсу, група КНс-20-1* — Підпис — *А.Л. Семенишен*
Курс, група виконавця — Підпис — Ініціали, прізвище

Керівник: *викладач кафедри КН* — Підпис — *М.О. Молчанова*
Науковий ступінь, посада — Підпис — Ініціали, прізвище

Нормоконтроль: *к.т.н., доцент кафедри КН* — Підпис — *Р.О. Багрій*
Науковий ступінь, посада — Підпис — Ініціали, прізвище

До захисту допускаю:
Зав. кафедри КН, д.т.н., професор — Підпис — *О.В. Бармак*
Науковий ступінь, посада — Підпис — Ініціали, прізвище

05 06 2023 р.

ХМЕЛЬНИЦЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ

Факультет *інформаційних технологій*
Кафедра *комп'ютерних наук*
Освітній ступінь *бакалавр*
Галузь знань *12 – Інформаційні технології*
Спеціальність *122 – Комп'ютерні науки*
Освітня програма *освітньо-професійна програма підготовки бакалавра*

ЗАТВЕРДЖУЮ
Завідувач кафедри *комп'ютерних наук*

_____
(підпис)
д.т.н., професор О.В. Бармак
«*06*» *03* *2023* року

ЗАВДАННЯ
НА КВАЛІФІКАЦІЙНУ РОБОТУ БАКАЛАВРА

1. Тема кваліфікаційної роботи бакалавра: *«Спосіб для визначення семантично невідповідних україномовних коментарів до описів товарів інтернет-магазину для задач аналізу купівельного попиту»*

2. Завдання видано студенту *Семенишен Андрій Леонідович*
(прізвище, ім'я, по батькові)

3. Керівник роботи *викладач кафедри КН Молчанова Марина Олексіївна*
(посада, прізвище, ім'я, по батькові)

4. Затверджено наказом університету від «*01*» *03* *2023* р. № *5*

5. Дата видачі завдання студенту: «*03*» *03* *2023* р.

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

Провести аналіз предметної області, здійснити огляд методів пошуку ключових слів, визначити особливості аналізу україномовного контенту для задач електронної комерції. Розробити спосіб для визначення семантичної невідповідності україномовних коментарів до описів товарів інтернет-магазину. Спроєктувати структуру застосунку для визначення семантично невідповідних україномовних коментарів до описів товарів інтернет-магазину й структуру відповідної бази даних. Створити застосунок що реалізує розроблений спосіб та виконати його тестування.

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*	*грудень 2022*	виконано
2	*Ознайомлення з предметною областю, формулювання мети та задач дослідження, визначення об'єкта та предмета дослідження*	*січень 2023*	виконано
3	*Робота над розділом 1 – Характеристика предметної області та постановка задачі*	*січень 2023*	виконано
4	*Робота над розділом 2 – Спосіб для визначення семантичної невідповідності україномовних коментарів до описів товарів інтернет-магазину*	*березень 2023*	виконано
5	*Робота над розділом 3 – Програмна реалізація для визначення семантично невідповідних україномовних коментарів до описів товарів інтернет-магазину*	*квітень 2023*	виконано
6	*Оформлення пояснювальної записки згідно вимог*	*травень 2023*	виконано
7	*Попередній захист кваліфікаційної роботи бакалавра*	*травень 2023*	виконано
8	*Захист кваліфікаційної роботи бакалавра на засіданні Екзаменаційної комісії*	*червень 2023*	виконано

Виконавець: *студент 3 курсу, група КНс-20-1* (Курс, група виконавця) _____ (Підпис) *А.Л. Семенишен* (Ініціали, прізвище)

Керівник: *викладач кафедри КН* (Науковий ступінь, посада) _____ (Підпис) *М.О. Молчанова* (Ініціали, прізвище)

Анотація

Тема кваліфікаційної роботи бакалавра: *Спосіб для визначення семантично невідповідних україномовних коментарів до описів товарів інтернет-магазину для задач аналізу купівельного попиту*

Виконавець кваліфікаційної роботи бакалавра: *студент групи КНс-20-1 Семенишен Андрій Леонідович*

Керівник кваліфікаційної роботи бакалавра: *викладач кафедри КН Молчанова Марина Олексіївна*

Кваліфікаційна робота бакалавра містить:

Пояснювальна записка				Кількість додатків
Сторінок	Рисунків	Таблиць	Джерел інформації	
69	36	18	25	4

Метою кваліфікаційної роботи бакалавра є розробка способу для визначення семантичної невідповідності україномовних коментарів до описів товарів інтернет-магазину для задач аналізу купівельного попиту.

Результатом виконання кваліфікаційної роботи бакалавра є розроблений спосіб для визначення семантично невідповідних україномовних коментарів до описів товарів інтернет-магазину. Також в результаті роботи було створено відповідне програмне забезпечення для програмного пошуку невідповідностей серед існуючих користувацьких коментарів, що дасть можливість виконання аналізу купівельного попиту.

Ключові слова: семантичний аналіз, TF-IDF, нейромережевий класифікатор, користувацькі коментарі.

Виконавець: *студент 3 курсу, група КНс-20-1* (Курс, група виконавця) _____ (Підпис) *А.Л. Семенишен* (Ініціали, прізвище)

Зміст

Перелік скорочень

Скорочення, термін, позначення	Пояснення
КРБ	Кваліфікаційна робота бакалавра
КН	Комп’ютерні науки
NLP	Natural language processing
ШІ	Штучний інтелект
SA	Semantic analysis
API	Application Programming Interface
БД	База даних
МН	Машинне навчання

Вступ

Кваліфікаційна робота бакалавра присвячена розробці способу для визначення семантично невідповідних україномовних коментарів до описів товарів інтернет-магазину для задач аналізу купівельного попиту.

Актуальність. Інтелектуальний аналіз тексту, як один із напрямів ШІ є одним із значимих напрямів сучасних досліджень. Починаючи від категоризації текстів, пошуку інформації, обробки змін у колекціях текстів, до розробки засобів представлення інформації для поточних користувачів – все це є задачами Інтелектуального аналізу тексту.

Переважною більшістю інформації з веб-джерел є звичайний текст, який або погано структурований, або зовсім неструктурований. Тому, отримання значущої інформації є одним із провідних завдань інтелектуального аналізу тексту, що є процесом одержання якісної інформації з частково-структурованих та неструктурованих даних.

Окрім проблеми неструктурованих даних, є ще проблема семантичної не відповідності україномовних коментарів до описів товарів інтернет-магазину, виявлення яких може допомогти у проведенні аналізу купівельного попиту.

Об’єкт дослідження – процес семантичного аналізу україномовних коментарів до текстових описів товарів інтернет-магазину.

Предмет дослідження – моделі, методи, алгоритми та засоби для виявлення семантичної невідповідності україномовних коментарів до описів товарів інтернет-магазину.

Мета кваліфікаційної роботи бакалавра – розробка способу для визначення семантичної невідповідності україномовних коментарів до описів товарів інтернет-магазину для задач аналізу купівельного попиту.

Завдання кваліфікаційної роботи бакалавра – виконати аналіз сучасних методів NLP в області аналізу коротких текстових даних; розробити спосіб для визначення семантичної невідповідності україномовних коментарів до описів товарів інтернет-магазину для задач аналізу купівельного попиту;

створити до розробленого способу для визначення семантичної невідповідності україномовних коментарів до описів товарів відповідне програмне забезпечення з відповідною функціональністю для визначення семантичної невідповідності україномовних коментарів до описів товарів; провести тестування розробленої програмної реалізації.

Розділ 1 Характеристика предметної області та постановка задачі

1.1 Семантичний аналіз текстів та методи пошуку ключових слів

Хоча обробка природної мови сьогодні не є новою наукою, технологія стрімко розвивається завдяки зростанню інтересів до зв'язку між людьми, а також наявності великих даних, потужних обчислень та великого обсягу розширених алгоритмів [1].

Людина може говорити та писати англійською, іспанською чи китайською, у той час як рідна мова комп'ютера більш відома як машинний код або машинна мова, яка здебільшого незрозуміла для більшості людей. На найнижчому рівні технічного пристрою комунікація відбувається не за допомогою слів, а через мільйони нулів та одиниць, які викликають логічні дії.

Обробка природної мови належить галузі інформатики, а точніше до галузі ШІ, що займається наданням комп'ютерам здатності розуміти текст чи мовлення майже так само, як це можуть зробити люди [2].

NLP поєднує у собі комп'ютерну лінгвістику (моделювання людської мови на основі правил) зі статистичними моделями, моделями машинного та глибокого навчання. Разом такі технології дозволяють комп'ютерам обробляти людську мову у формі тексту чи відповідно, голосових даних та «розуміти» її повне значення разом із наміром і почуттям мовця або ж письменника.

NLP має керування комп'ютерними програмами, що перекладають текст з однієї мови на іншу, а також реагують на голосові команди і швидко підсумовують доволі великі обсяги тексту навіть у реальному часі. Є доволі великі шанси, що кожна людина, так чи інакше взаємодіяла з NLP у формі голосових систем GPS, або ж цифрових помічників чи програмного забезпечення для диктування мови в текст, чат-ботів для обслуговування клієнтів чи інших зручностей для споживачів. Проте NLP також відіграє велику роль у корпоративних рішеннях, що допомагають оптимізувати бізнес-операції, підвищити продуктивність працівників і спростити критично важливі бізнес-процеси.

Семантикою є розділ лінгвістики, який вивчає смислове значення одиниць мови. Семантика задається певною математичною моделлю, що описує деякі обчислення, які є можливими для мови. Семантичний аналіз дає точне або словникове значення із структур що створені синтаксичним аналізом. Основною метою семантичного аналізу є мінімізація структури синтаксису та знаходження їх значення завдяки пошуку синонімів, пошуку сенсу слова, перекладу на інші мови, а також заповнення баз знань. Завдання семантичного аналізу на сьогоднішній день в повній мірі не вирішено. Його можна зустріти в системах перекладу, у системах перевірки граматики, чат ботах тощо. Незважаючи на доволі багату теорію в області SA, застосування у сучасних сферах робочих середовищ знаходять тільки методи аналізу, що були засновані на статистичних характеристиках слів та словосполучень тексту, що аналізувався [3].

Семантичний аналіз є підполем для обробки природної мови, яке намагається зрозуміти значення природної мови. Розуміння природної мови може здатися людям доволі простим процесом. Проте через величезну складність і суб’єктивність людської мови її інтерпретація і розуміння є доволі складним завданням для машин. SA природної мови фіксує значення даного тексту, беручи до уваги контекст, логічну структуру речень, а також граматичні ролі [4].

Семантичний аналіз природної мови умовно можна розділити на дві великі частини:

1. Лексико-семантичний аналіз, що передбачає розуміння значення кожного слова тексту окремо. В основному ця частина стосується вилучення зі словника значення, що має нести конкретне слово в тексті.

2. Аналіз композиційної семантики, що передбачає розуміння сенсу. Хоча знати значення кожного слова у тексті є важливим, цього зазвичай недостатньо, щоб повністю зрозуміти значення тексту.

Машинам бракує довідкової системи, щоб зрозуміти значення слів, речень і документів. Усунення неоднозначності та ідентифікація значення слів

може забезпечити краще розуміння мовних даних для машинного навчання. Ось як працює кожна частина семантичного аналізу [5]:

– Лексичним аналізом є процес читання потоку символів, ідентифікації лексем та перетворення їх на лексеми, що можуть бути прочитаними машини.

– Граматичним аналізом є співвідношення послідовності лексем (слів) і застосовування до них формальної граматики, щоб стало можливе позначення частин мови.

– Синтаксичний аналіз виконує аналіз або розбирає синтаксис та застосовує правила граматики, щоб забезпечити контекст значенням на рівні слів та речень.

– Семантичним аналізом використовується все вищезазначене, щоб зрозуміти значення слів та виконати інтерпретацію структури речень, щоб машини могли розуміти мову, аналогічно людям.

Для подальшої обробки текстової інформації зазвичай використовують ключові слова.

Вилучення ключових слів зазвичай використовується для вилучення ключової інформації з серії абзаців або документів. Вилучення ключових слів є автоматизованим методом вилучення найбільш релевантних слів та фраз із уведеного тексту. Це метод аналізу тексту, що передбачає автоматичне виділення найважливіших слів та виразів. Це допомагає для підсумовування змісту тексту та визначенні ключових питань, що обговорюються – наприклад, протокол наради, тощо [6].

Однією з популярних задач є задача пошуку оцінок продукту. Щоб переглянути всі дані та знайти терміни, які найкраще визначають кожен відгук, можна використати вилучення ключових слів. Можна помітити, які теми викликають найбільше обговорень серед споживачів, а автоматизація процесу заощадить багато часу персоналу.

Є досить багато інструментів для пошуку ключових слів, найпоширеніші з яких буде розглянуто нижче.

Yake. Для автоматичного вилучення ключових слів у Yake текстові функції використовуються без нагляду. YAKE є базовим неконтрольований автоматичним методом виділення ключових слів, що визначає найбільш релевантні ключові слова в тексті за допомогою текстових статистичних даних з окремих текстів. Ця техніка не покладається на словники, зовнішні корпуси, розмір тексту, мову чи домен і не потребує навчання певному набору документів. Основні характеристики алгоритму Yake такі:

- неконтрольований підхід;
- корпусно-незалежний;
- незалежно від домену та мови;
- єдиний документ.

Хмара слів. Величина кожного слова представляє його частоту або релевантність у хмарі слів, що є інструментом візуалізації даних для візуалізації текстових наборів даних. Хмару слів можна використовувати для підкреслення важливих текстових даних. Дані з вебсайтів соціальних мереж часто аналізуються за допомогою хмар слів.

Що більший і жирніший термін з’являється в хмарі слів, то більше разів він з’являється в джерелі текстових даних (наприклад, у промові, публікації в блозі чи базі даних). Хмара слів є набором слів різних розмірів. Чим частіше термін з’являється в документі і чим він важливіший, тим він більший і жирніший. Це чудові способи вилучення найважливіших частин текстових даних, таких як публікації в блогах і бази даних.

KeyBert. KeyBERT є базовою та простою у використанні технікою вилучення ключових слів, що генерує ключові слова та ключові фрази, найбільш схожі на певний документ, використовуючи вбудовування BERT. Він використовує BERT-вбудовування та базову косинусну подібність, щоб знайти піддокументи у документі, що є найбільш схожими на батьківський документ.

BERT застосовують для вилучення вбудованих документів для отримання представлення на рівні документа. Далі вилучаються вбудовування слів для N-граммових слів/фраз. Нарешті, він використовує косинусну подібність, щоб

знайти слова/фрази, що будуть найбільше схожі на документ. Тоді можна визначити найбільш порівнювані терміни як ті, що найкраще описують увесь документ.

Оскільки алгоритм побудований на BERT, KeyBert генерує вбудовування за допомогою попередньо підготовлених моделей на основі трансформатора huggingface. Модель повністю MiniLM-L6-v2 використовується за замовчуванням для вбудовування [6].

IDF. Для зниження значущості слів, що зустрічаються майже у всіх документах колекції, вводять інверсну частоту терміну IDF (inverse document frequency) що є логарифмом відношення числа всіх документів до документів, які містять певне слово *t*. Значення цього параметра буде менше, що частіше слово зустрічається у документах бази документів. Тому, для слів, що зустрічаються у великій кількості документів значення IDF буде близьким до нуля (якщо слово зустрічається у всіх документах колекції, IDF буде рівним нулю), що допомагає знайти важливі слова. Визначається за формулою [7]:

$$IDF = \log \frac{|D|}{|d_i \in t|} \tag{1}$$

Параметр TF (з англ. term frequency) є відношенням числа разів, яке певне слово *t* зустрілося в документі *d* до довжини документа. Нормалізація довжиною документа потрібна насамперед для того, щоб зрівняти в правах короткі та довгі документи. Визначається за формулою:

$$TF = \frac{|D|}{n_t} \tag{2}$$

Коефіцієнт *TF-IDF* буде рівний добутку *TF* та *IDF*. *TF* відіграє роль підвищуючого множника, *IDF* відповідно, понижуючого. Тоді ваговими параметрами відповідної векторної моделі певного документа можна прийняти коефіцієнти *TF* * *IDF* слів, що входять до нього.

Для того, щоб ваги перебували в інтервалі від 0 до 1, а вектори документів мали рівну довжину, значення *TF*IDF* зазвичай проходить нормалізацію по косинусу.

Варто зазначити, що дана формула оцінює значимість терміна лише з погляду частоти входження до документа, не враховуючи порядок дотримання термінів у документі та їх синтаксичну роль; інакше кажучи, семантика документа зводиться до лексичної семантики термінів, що входять до нього, а відповідна композиційна семантика не розглядається.

Ключовими словами у такому випадку будуть слова, що набрали найбільшу вагу. Тоді слова з малою вагою взагалі можна не враховувати при класифікації.

Отож, для пошуку ключових слів для визначення семантично невідповідних україномовних коментарів до описів товарів інтернет-магазину для задач аналізу купівельного попиту доцільно буде використати *TF*IDF*.

1.2 Існуючі підходи до класифікації текстових даних

Сучасний світ тоне у великих обсягах інформації, кількість якої невпинно зростає. Що породжує для людини складність аналізувати, обробляти та класифікувати дані за категоріями. Однак, окрім зростання інформації є і одночасне зростання доступної обчислювальної потужності комп’ютерів, що дозволяють у свою чергу використовувати сучасні методи для вирішення задачі класифікації.

Процес класифікації текстів може займати доволі багато часу, адже для вирішення такої задачі необхідно не тільки прочитати текст, а й проаналізувати його зміст і обрати категорію, що підійде із множини доступних категорій. Коли ж мова йде про глобальні задачі та великі масиви текстових даних, то часто виникають проблеми із браком людських ресурсів для їх обробки [8].

Для класифікації текстів необхідно пройти такі етапи:

– ознайомитись із переліком категорій, на які можна розділити пропоновані тексти;

– ознайомитись із переліком текстів та проаналізувати кожен текст;

– на основі виконаного аналізу віднести текст до конкретної категорії.

Щоб автоматизувати даний процес, необхідно вирішити задачу машинного навчання. Задачі машинного навчання умовно поділяються на 2 типи: навчання з учителем (supervised learning) та навчання без учителя (unsupervised learning). При виборі навчання з учителем, машина заздалегідь знає результати роботи алгоритму. При навчанні ж без учителя програма сама намагається зрозуміти суть алгоритму. Для вирішення будь-якої задачі машинного навчання з учителем потрібно мати початковий набір даних або ж як ще його називають, датасет. В процесі автоматизації система працює з документами і виконує аналіз їх змісту. У кінці автоматизації буде отримано систему, яка дозволятиме користувачеві самостійно створювати моделі машинного навчання на основі створених дата сетів, навчати та аналізувати їх ефективність.

Існує велика кількість програм та інструментів для аналізу числових даних, проте їх доволі мало для текстів. Мультиноміальна наївна класифікація Байєса є однією з найпопулярніших класифікацій навчання з учителем, яка використовується для аналізу категоризованих текстових даних [9].

Класифікація текстових сьогодні даних набуває значної популярності, оскільки є величезна кількість інформації, що доступна в електронній пошті, документах, на вебсайтах тощо, яку потрібно аналізувати та класифікувати. Знання контексту навколо певного типу тексту допомагає визначити сприйняття програмного забезпечення або ж програмного забезпечення користувачами, що збираються ним користуватися [9].

Мультиноміальний наївний алгоритм Байєса є імовірнісним методом навчання, що переважно використовується для обробки природної мови. Алгоритм заснований на теоремі Байєса що передбачає тег тексту, наприклад тему електронного листа або ж газетної статті. Алгоритм обчислює ймовірність кожного тега для даної вибірки, а потім видає тег категорії із найвищою ймовірністю як вихідні дані.

Наївний класифікатор Байєса є набором багатьох алгоритмів, де всі алгоритми мають деякий спільний принцип, а саме: кожна ознака, що класифікується, не пов’язана з жодною іншою ознакою. Таким чином, наявність або відсутність функції не впливає на наявність або відсутність іншої функції.

Naive Bayes є потужним алгоритмом, що використовується для аналізу текстових даних та задач із кількома класами. Щоб зрозуміти роботу алгоритму, важливо спершу зрозуміти концепцію теореми Байєса.

Теорема Байєса була сформульована Томасом Байєсом і обчислює ймовірність події на основі попередніх знань про умови, щ обули пов’язані з подією. Формульно це виражається так:

$$P(A \mid B) = P(A) * P(B \mid A)/P(B) \quad (3)$$

де обчислюється ймовірність класу A, коли предиктор B вже надано.

P(B) є попередньою ймовірність B, P(A) є пріоритетною ймовірність класу A, P(B|A) є появою предиктора B із заданою ймовірністю класу A. Формула (3) допомагає підрахувати ймовірність появи тегів у тексті.

SVM є ще одним алгоритмом навчання з учителем. Варто зазначити, що SVM може застосовуватися і для завдань регресії, проте головна мета SVM як класифікатора це знаходження рівняння роздільної гіперплощини (рисунок 1.1).

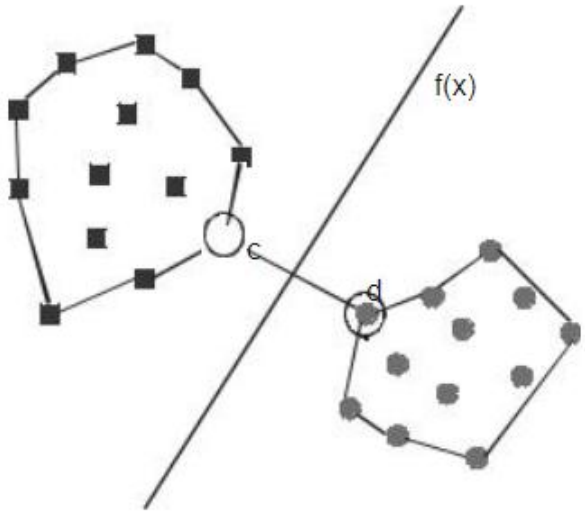


Рисунок 1.1 – Графічна інтерпритація роботи SVM

Ідея методу ґрунтується на припущенні про те, що найкращим способом поділу точок у m-мірному просторі є m−1 площина (задана параметризація f(x) = 0), що рівновіддалена від точок, які належать різним класам.

Як можна помітити, для вирішення такого завдання достатньо провести площину, рівновіддалену від найближчих один до одного точок, що належать до різного класу. На рисунку 1.1 такими точками є точки c і d. Даний метод інтерпретує об'єкти та відповідні їм точки у просторі, як вектори розміру m. Іншими словами, незалежні змінні, які характеризують об'єкти, є координатами векторів. Найближчі один до одного вектори, що стосуються різних класів, називаються векторами підтримки (support vectors) [10].

Для класифікації коментарів для векторизації україномовних коментарів до описів товарів інтернет-магазину для задач аналізу купівельного попиту буде використано SVM.

1.3 Аналіз джерел україномовного контенту для задачі класифікації коментарів описів товарів

Сфера електронної комерції в 2021 році показала значне зростання, що на думку експертів є справжнім проривом у свідомості клієнтів. Якщо до 2021 року онлайн-покупки за допомогою мобільних програм були у пріоритеті у зумерів, то наразі до них долучились представники старшого покоління, а також долучилася аудиторія у віком 25-40 років [11].

Однак, 2022 вніс значні корективи у роботі електронної комерції, український онлайн-бізнес майже сягнув дна на початку березня, проте досить швидко відштовхнувся та почав підлаштовуватися під нові реалії [12].

З середини березня минулого року, маркетологи спостерігали помітне збільшення сесій: українці відійшли від попереднього шоку, почала відновлюватись логістика, добігала кінця активна фаза переселення.

Вже з травня 2022 деякі категорії товарів навіть повернулися до довоєнних показників.

Одним із видів електронної комерції є інтернет-магазин, який є електронною комерційною платформою, яка дозволяє покупцям здійснювати покупки товарів та послуг через мережу інтернет. Інтернет-магазин може мати різні форми та розміри – від невеликого магазину з декількома товарами до доволі великої онлайн-платформи з великою кількістю не лише товарів, а і послуг.

Значною перевагою інтернет-магазину є зручність та швидкість здійснення покупок без виходу з дому чи офісу, а також можливості порівняти ціни та характеристики товарів з різних інтернет-магазинів. Переваги інтернет-магазину є і у власників. До прикладу, можливість знизити витрати на оренду приміщень та оплату праці персоналу, а також працювати з більшою аудиторією клієнтів за рахунок доступності в мережі інтернет.

Через електронну комерцію можна реалізовувати товари та послуги. Під товарами інтернет-магазину розуміють товари, які продаються в мережі інтернет через електронний магазин. Товарами можуть бути як фізичні товари, такі як одяг, харчові продукти, електроніка, книги, тощо або ж цифрові товари, накшталт музики, фільмів, програмного забезпечення, електронних книг та інші. Такі товари зазвичай замовляються за допомогою комп'ютера або мобільного пристрою, з допомогою інтернет-браузера чи спеціальної програми. Замовлення оплачується онлайн або при отриманні товару, наприклад, через «Нова пошта» накладним платежем. Це на сьогодні популярний засіб покупок, оскільки дозволяє зручно та швидко знайти та придбати потрібний товар без виходу з дому.

Проте, купівля товарів в інтернеті все ще асоціюється з можливими ризиками: продавець може виявитись шахраєм, або ж надійде неякісний товар [13]. Відгуки допомагають подолати такі страхи, адже якщо люди позитивно оцінюють інтернет-магазин, то і рівень довіри до нього зростає.

Відгуки вирішують різні завдання:

– можуть допомогти визначитися із кольором, розміром, фасоном виробу. До прикладу, якщо у відгуках користувачі пишуть, що взуття

малорозмірне, клієнт знаючи цю інформацію замовить кросівки більшого розміру, та уникне проблем з обміном;

– дозволяють працювати з запереченнями клієнта. Часто клієнтами у полі для відгуків ставляться запитання (і магазин у коментарях відповідає на них), що у свою чергу значить, що у інших відвідувачів сайту, які читатимуть дані відгуки, уже залишиться менше сумнівів та заперечень;

– відгуки підвищують конверсію. Головне завдання відгуків є продавати зазначені товари.

Нижче наведено приклад відгуку до мобільного телефону Xiaomi магазину «Rozetka» (рисунок 1.1).

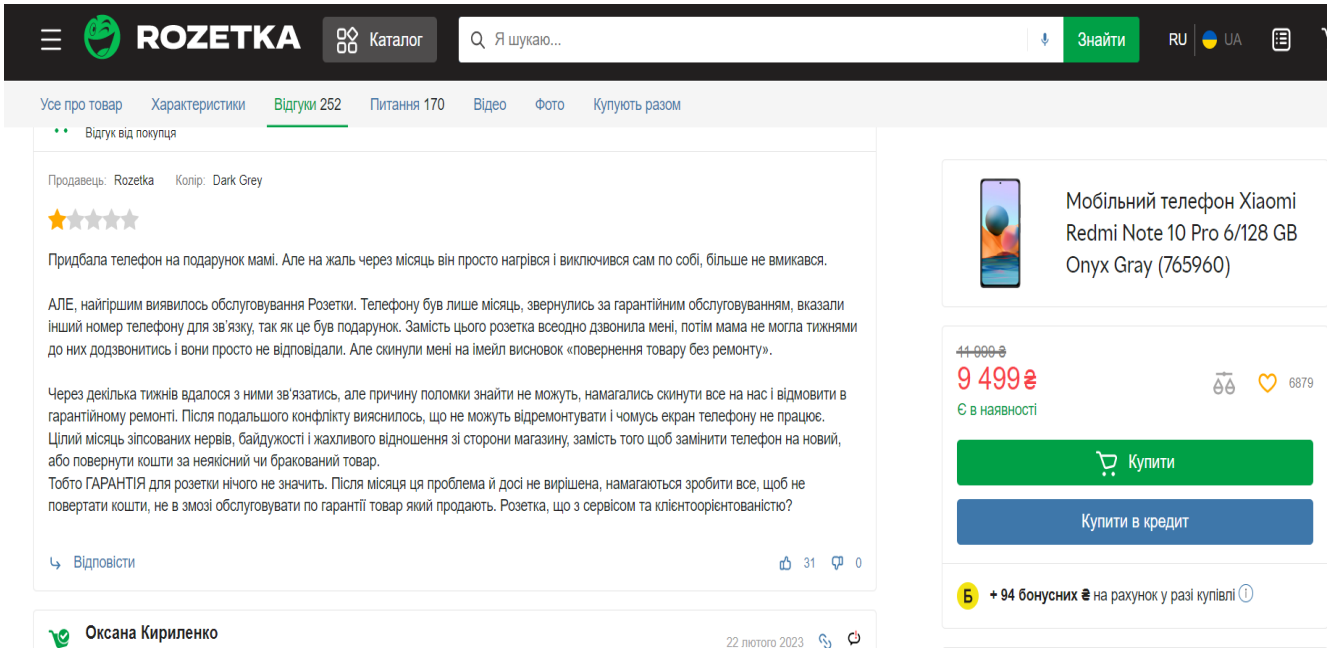


Рисунок 1.1 – Ілюстрація відгука магазину «Розетка»

Проте, часом буває так, що користувач хибно оцінює товар, або коментує не зовсім не той товар, який хотів. На рисунку 1.2 проілюстровано негативний відгук клієнта, проте виставлена максимальна оцінка.

Рисунок 1.2 – Невідповідність відгуку і оцінки

Отож, використовуючи бінарну класифікацію україномовних відгуків до описів товарів у роботі з інтернет-магазином можна створити більш об’єктиву оцінку товарів та відкинути для користувачів невалідний контент.

1.4 Аналіз існуючих рішень для подібних задач

Галузь класифікації текстової інформації на сьогоднішній день є доволі популярною, тому на ринку програмних продуктів наявні деякі рішення.

Програма KNIME [14] має модуль для класифікації текстів, який визначає повністю автоматизовану вебпрограму, що позначатиме дані користувача за допомогою активного навчання. Модуль розроблено для того, щоб бізнес-аналітики легко переглядали документи, які потрібно позначати будь-якою кількістю класів. У кожній ітерації користувач буде позначати нові документи, а модель буде навчатися за допомогою вже позначених екземплярів. З кожною новою ітерацією модель відбирає найбільш невизначені документи за допомогою вузла оцінювача ентропії. Коли користувач буде задоволений продуктивністю, що була досягнута з доступними мітками, він може вийти з циклу та виконати експорт моделі, щоб позначити решту екземплярів. Зовнішній вигляд програмного продукту проілюстровано на рисунку 1.3.

Також класифікація текстів є одним із компонентів Azure Cognitive Service для мовної служби. Це хмарна служба API, яка застосовує логіку

машинного навчання для створення власних моделей для вирішення завдань класифікації тексту.

Дозволяє користувачам створювати власні моделі штучного інтелекту для класифікації текстів на основі класів, заздалегідь визначених користувачем. Якість даних із мітками значною мірою впливає на продуктивність моделі. Щоб спростити створення та налаштування моделі, служба пропонує налаштований вебпортал, доступ до якого можна отримати за допомогою Language Studio [15].

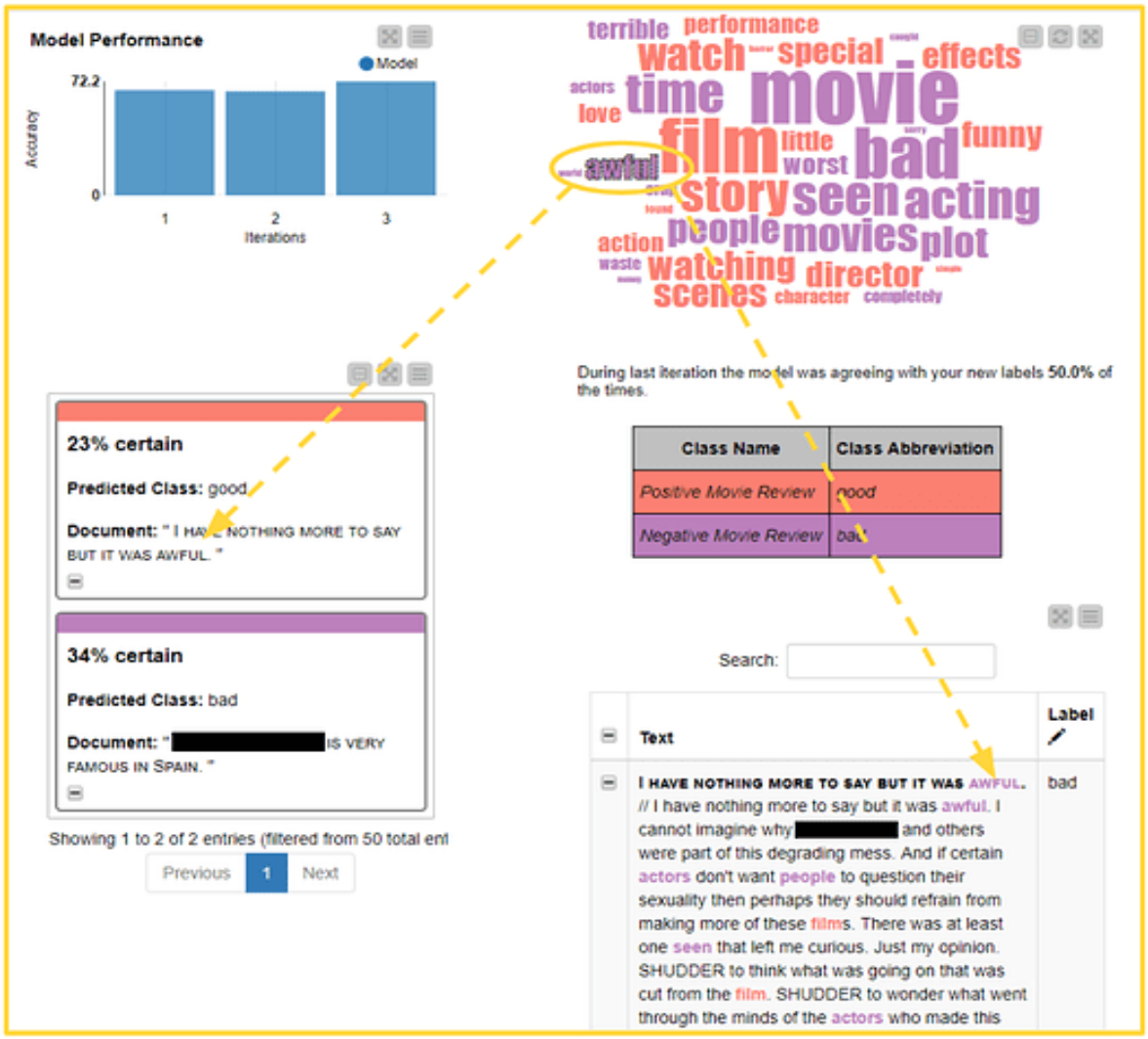


Рисунок 1.3 – Модуль для класифікації текстів «KNIME»

Користувацька класифікація текстів підтримує проєкти двох типів:

– Класифікація за однією міткою, коли кожному документу з набору даних призначається лише один клас. Наприклад, сценарій фільму можна класифікувати лише як "Романтичний фільм" чи "Комедія".

– Класифікація за декількома мітками, коли кожному документу з набору даних можна призначити декілька класів. Наприклад, сценарій фільму можна класифікувати як "Комедія" або "Романтичний фільм" та "Комедія" [16].

Отже, зважаючи на популярність предметної області, існує певна кількість програмних рішень, чи надбудов для швидкої реалізації класифікації текстів, проте більшість з них є платними і не мають достатнього функціоналу для вирішення задачі класифікації семантично невідповідних україномовних коментарів до описів товарів інтернет-магазину.

1.5 Постановка задачі

Метою роботи є розробка способу для визначення семантичної невідповідності україномовних коментарів до описів товарів інтернет-магазину для задач аналізу купівельного попиту.

У рамках досягнення поставленої мети ставляться наступні задачі:

– розробити спосіб для визначення семантичної невідповідності україномовних коментарів до описів товарів інтернет-магазину;

– створити набір даних, що використовується для описів товарів у інтернет-магазинах українською мовою;

– розробити архітектуру нейронної мережі для визначення семантичної невідповідності україномовних коментарів до описів товарів інтернет-магазину;

– спроєктувати структуру застосунку для визначення семантично невідповідних україномовних коментарів до описів товарів інтернет-магазину й структуру відповідної БД;

– розробити інформаційну систему визначення семантичної невідповідності україномовних коментарів;

– провести тестування створеного програмного забезпечення;

– провести дослідження ефективності запропонованого способу.

Розділ 2 Спосіб для визначення семантичної невідповідності україномовних коментарів до описів товарів інтернет-магазину

2.1 Особливості аналізу україномовного контенту

Аналіз україномовного контенту є однією із важливих складових частин ефективної роботи електронної комерції, оскільки дозволяє зрозуміти поведінку та потреби українських користувачів.

Однією з основних особливостей аналізу україномовного контенту є те, що українській мові притаманна досить складна граматика та велика кількість варіантів написання одного слова, що значно ускладнює задачу обробки тексту. Крім того, аналізуючи україномовний контент необхідно враховувати культурні та соціальні особливості української аудиторії [17]. Наприклад, використання специфічних слів або фраз, що можуть бути зрозумілі тільки українцям, може допомогти залучити більше клієнтів, а також покращити їхнє сприйняття бренду.

Аналіз коротких текстових даних відрізняється від аналізу довгих текстів, які складаються із десятків речень, абзаців та тематичних розділів. Короткі тексти значно складніше піддаються або взагалі не підлягають категоризації. Однак, нейромережі глибокого навчання, а також векторні представлення документів значно краще працюють з даними невеликого розміру, тому їх застосування до коротких текстів спроможне дати кращі результати в задачах семантичного аналізу. Аналіз україномовного контенту для задач електронної комерції є значно складнішим, ніж аналіз англомовних аналогів, так як тут присутній непрямий порядок слів, що ускладнює знаходження зв’язків між словами, а також велика кількість форм слова, що часто потребує попередньої обробки для нормалізованого представлення слів. Також розвиток алгоритмів обробки україномовного контенту уповільнює нестача україномовних корпусів та актуальних датасетів [18]. Тож у рамках роботи необхідно створити датасет, що буде у подальшому використано для навчання та тестування нейромережі.

2.2 Визначення семантичної невідповідності україномовних коментарів до описів товарів за допомогою нейронної мережі

Для визначення семантичної невідповідності україномовних коментарів до описів товарів за допомогою набору моделей на базі машинного навчання пропонується підхід, проілюстрований на рисунку 2.1.

Рисунок 2.1 – Схема способу визначення семантичної невідповідності україномовних коментарів до описів товарів

Пропонований спосіб призначений для перетворення вхідних даних у формі текстів-коментарів для побудови україномовного датасету, що

використовується для опису товарів у інтернет-магазинах для навчання нейромережевого класифікатора та тестових текстів-коментарів, невідповідність до опису товару яких потрібно визначити у кінцеві дані у вигляді числової оцінки відповідності коментаря до обраного товару.

Виконання способу визначення семантичної невідповідності україномовних коментарів до описів товарів у ході роботи проходить ключові етапи, першим з яких є етап побудови датасету україномовних відгуків та описів товарів у інтернет-магазинах. Коментарі до товарів пропонується отримувати технологією веб-скрапінгу, що дозволяє автоматично отримати дані з веб-сторінок. Дана технологія дає можливість програмно отримувати текстовий контент, зображення, посилання, таблиці та іншу інформацію з веб-сайтів. Файли для навчання нейромережі формуються за схемою, наведеною на рисунку 2.2.

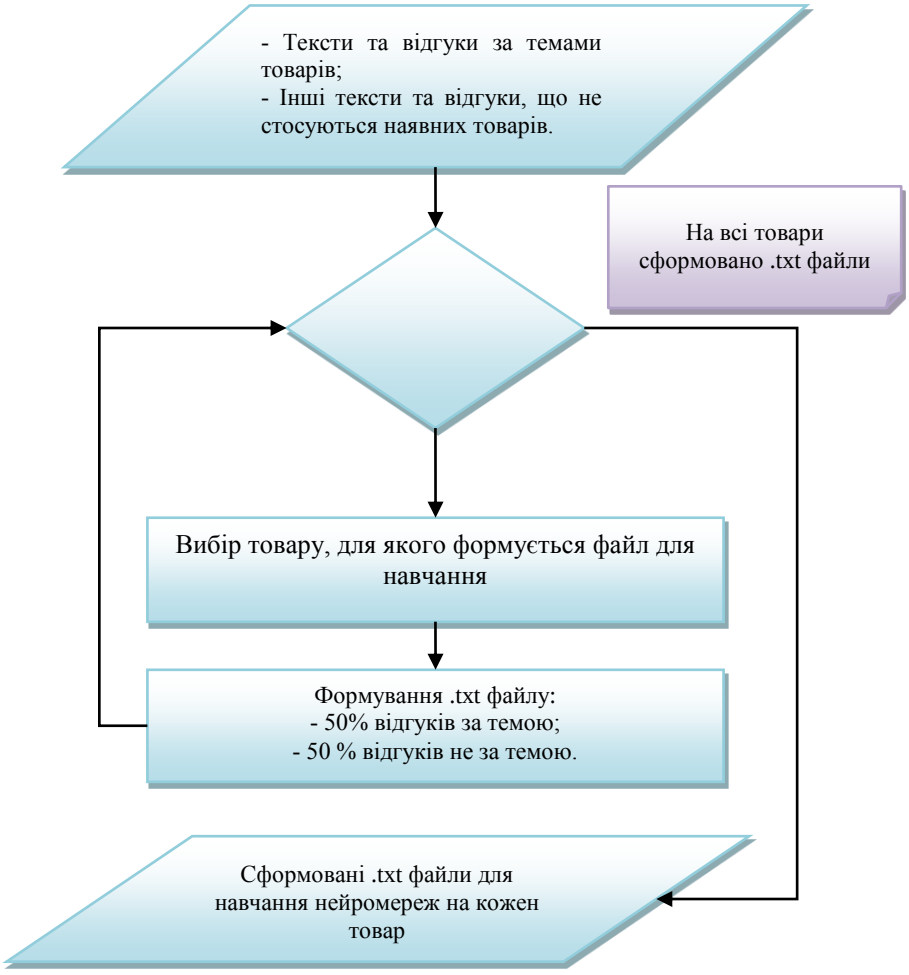


Рисунок 2.2 – Схема формування файлів навчання нейромережі

На цьому ж етапі здійснюється попередня обробка тексту та виконується формування файлу для навчання нейромережі. Файли для навчання формуються для кожного з товарів інтернет-магазину за принципом 50% релевантних коментарів та 50% не релевантних. Якщо коментар стосується розпізнаваного класу, ставиться оцінка 1, якщо не стосується – 0. Приклад сформованого розміченого файлу наведено на рисунку 2.3.

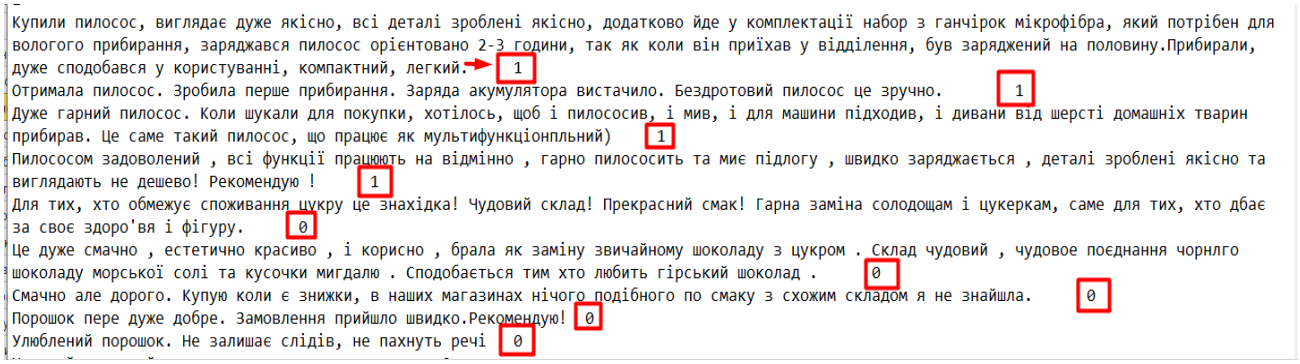


Рисунок 2.3 – Приклад розміченого файлу для навчання

Наступним етапом є побудова навчальної та тестової множин для кожного з товарів. Сформований файл для навчання у подальшому ділиться на 2 за принципом 20 % на 80 %. Де 80 % – дані файлу для навчання і 20 % – для тестування.

Далі йде етап побудови нейромережевих моделей для кожного з товарів. На кожен товар для подальшої ідентифікації навчається відповідна нейромережева модель. Для навчання кожної нейромережевої моделі використовуються набори даних, що описані на етапах 1 та 2. Навчені моделі зберігаються.

Відповідно, збережені навчені нейромережеві моделі проходять процес оцінювання ефективності за метриками Acuracy, оцінка F1, та метрикою ConfusionMatrix.

Наступним етапом є оцінка семантичної невідповідності для обраного товару, що у свою чергу є вихідними даними пропонованого методу.

Отже, запропоновано спосіб для визначення семантичної невідповідності україномовних коментарів до описів товарів, вхідними даними якого є тексти-коментарі для побудови моделі української мови, що використовується для опису товарів у інтернет-магазинах та тестові тексти-коментарі, невідповідність до опису товару яких потрібно визначити, а вихідними оцінка семантичної невідповідності для обраного товару, що призначений для визначення семантичної невідповідності україномовних коментарів.

2.3 Вибір архітектури бінарного класифікатора

Для визначення семантичної невідповідності україномовних коментарів до описів товарів запропоновано використовувати машинне навчання за визначеним переможцем серед пропонованих алгоритмів навчання. Підхід проілюстровано на рисунку 2.4.

Вхідними даними є текстовий набір даних за обраним товаром. Після завантаження навчальних даних по черзі відбувається навчання за різними алгоритмами (FastTree, AveragedPerceptron, FastForest, LinearSvm).

Для кожного алгоритму навчання визначається його ефективність, що вимірюється метриками. Якщо зважена оцінка за метриками для поточного класифікатора вища, ніж для попереднього, такий класифікатор запам’ятовується як найкращий і так поки не переберуться усі чотири варіанти.

На виході буде збережено класифікатор з найкращими показниками за метриками (рисунок 2.4).

Рисунок 2.4 – Ілюстрація підходу до підбору класифікатора

Отже, наведеним чином здійснюється вибір оптимального класифікатора для задачі визначення семантичної невідповідності україномовних коментарів до описів товарів.

2.4 Проєктування структури застосунку для визначення семантично невідповідних україномовних коментарів до описів товарів інтернет-магазину

За описаним підходом до визначення семантичної невідповідності україномовних коментарів до описів товарів за допомогою набору моделей на базі машинного навчання, пропонується така структура застосунку (рисунок 2.5).

Модуль бази даних призначений для збереження даних про товари, відгуків до них та датасету і навчених моделей нейромережі.

Нейромережевий модуль призначений для здійснення навчання моделей для обраних товарів, оцінювання якості навчених моделей та збереження навчених моделей, які у подальшому будуть використані для оцінки семантичної невідповідності україномовних коментарів до описів товарів.

Модуль оцінки семантичної відповідності до описів товарів інтернет-магазинів призначений для вибору товару, відгуки до якого будуть оцінюватись та безпосереднього здійснення оцінка рівня відповідності відгуку.

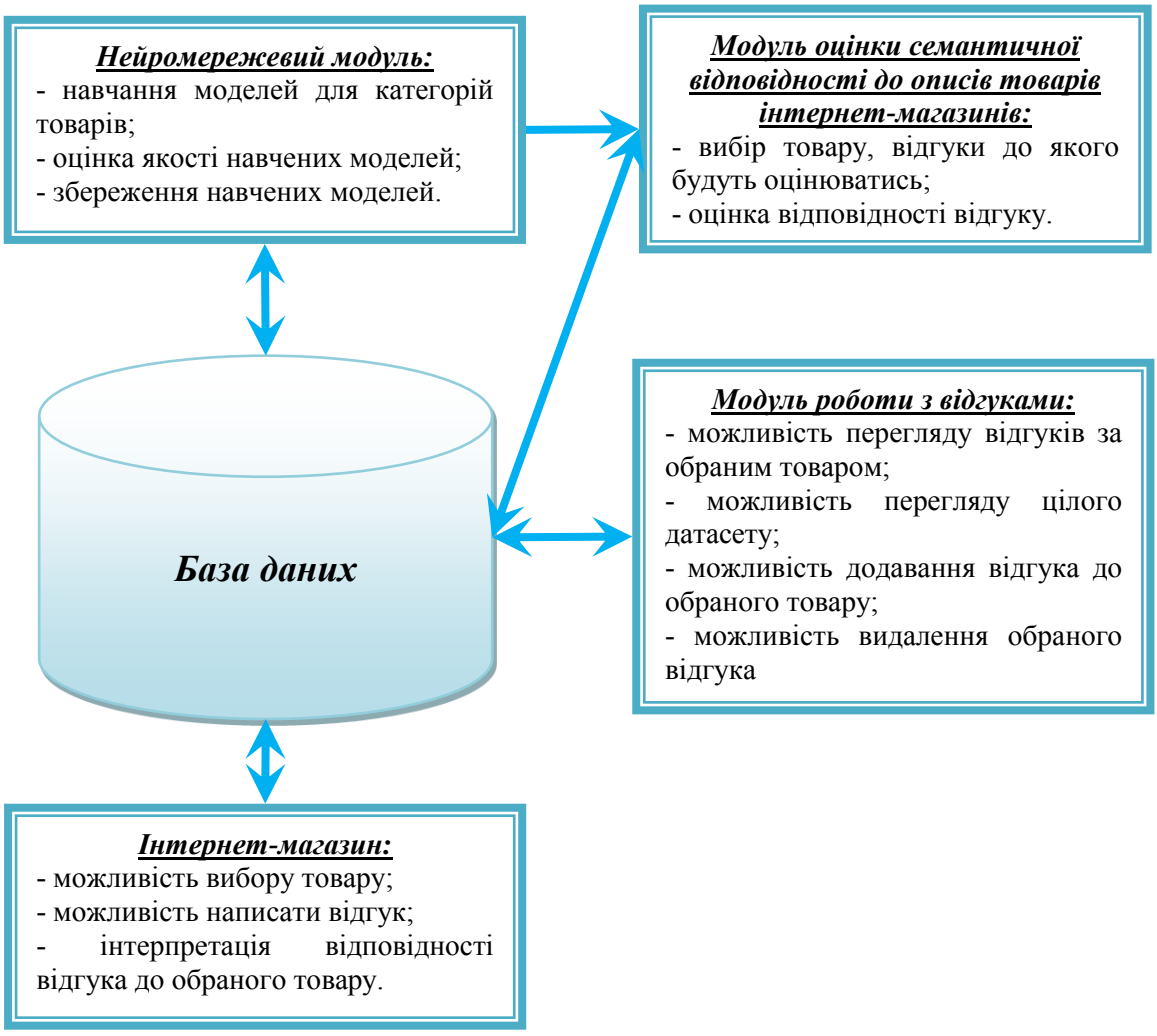


Рисунок 2.5 – Схема структури застосування для визначення семантично невідповідних україномовних коментарів до описів товарів інтернет-магазину

Модуль роботи з відгуками призначений для реалізації можливості перегляду відгуків за обраним товаром, реалізації можливості перегляду цілого

датасету, реалізації можливості додавання відгука до обраного товару, реалізації можливості видалення обраного відгука.

Модуль інтернет-магазину є інтернет-магазином класичної архітектури з можливістю вибору товарів, можливістю написати відгук про товар та можливістю оцінки рівня відповідності відгуків до залишених раніше відгуків про товар.

Отже, наведеним чином було спроєктовано структуру застосунку для визначення семантично невідповідних україномовних коментарів до описів товарів інтернет-магазину.

2.5 Даталогічна модель бази даних

Для програмної реалізації способу для визначення семантично невідповідних україномовних коментарів до описів товарів інтернет-магазину для задач аналізу купівельного попиту було реалізовано відповідну базу даних, що є це засобом збирання та впорядкування інформації [19].

Таблиця «Products» (таблиця 2.1) містить інформацію про наявні продукти, їх ідентифікатор, опис, категорію товару та ціни на них.

Таблиця 2.1 – Атрибути таблиці «Products»

№ п/п	Назва атрибуту	Тип даних	Опис
1.	ID	int	Первинний ключ, числовий ідентифікатор для однозначного визначення запису таблиці.
2.	Name	Varchar(50)	Назва товару
3.	Description	text	Опис товару
4.	Category	Varchar(50)	Категорія товару
5.	Price	double	Ціна товару

Таблиця «CommentAnalysis» (таблиця 2.2) містить інформацію про результати аналізу текстових коментарів. Таблиця зберігатиме: зміст коментаря, його семантичну оцінку, нейромережу та заключення про коментар.

Таблиця 2.2 – Атрибути таблиці «Comment_Analysis»

№ п/п	Назва атрибуту	Тип даних	Опис
1.	ID	int	Первинний ключ, числовий ідентифікатор для однозначного визначення запису таблиці.
2.	FK_Comment	int	Вторинний ключ, посилання на таблицю Comments для співставлення із відповідним записом про коментарі
3.	Sentiment_score	double	Числова оцінка семантичної відповідності коментаря до опису товару
4.	FK_Network	int	Вторинний ключ, посилання на таблицю Neural_Networks для співставлення із відповідним записом про нейромережу, що використовувалась при аналізі коментарів
5.	FK_Conclusion	int	Вторинний ключ, посилання на таблицю Conclusions для співставлення із відповідним записом про розгорнуте заключення про коментар.

Таблиця «Comments» (таблиця 2.3) містить інформацію про коментарі, які користувачі залишають щодо продуктів. Таблиця зберігатиме: посилання на таблиці з продуктами та користувачами, а також текст та дату коментаря.

Таблиця 2.3 – Атрибути таблиці «Comments»

№ п/п	Назва атрибуту	Тип даних	Опис
1.	ID	int	Первинний ключ, числовий ідентифікатор для однозначного визначення запису таблиці.
2.	FK_Product	int	Вторинний ключ, посилання на таблицю Products для співставлення із відповідним записом про наявні товари
3.	FK_User	int	Вторинний ключ, посилання на таблицю Users для співставлення із відповідним записом про користувачів, які залишають коментар
4.	Text	text	Текст коментаря
5.	Date	datetime	Дата написання коментаря

Таблиця «Users» (таблиця 2.4) містить інформацію про користувачів, які залишали коментарі під товарами. Таблиця зберігатиме ім’я та прізвище користувача, а також електронну пошту та пароль від особистого кабінету.

Таблиця 2.4 – Атрибути таблиці «Users»

№ п/п	Назва атрибуту	Тип даних	Опис
1.	ID	int	Первинний ключ, числовий ідентифікатор для однозначного визначення запису таблиці.
2.	First_name	Varchar(50)	Ім’я користувача
3.	Last_name	Varchar(50)	Прізвище користувача
4.	Email	string	Електронна пошта користувача
5.	Password	string	Пароль користувача

Таблиця «Conclusions» (таблиця 2.5) містить інформацію про заключну оцінку коментарів, їх відповідність, а також посилання на вторинні таблиці з очікуваним та актуальним значенням того чи іншого коментаря.

Таблиця 2.5 – Атрибути таблиці «Conclusions»

№ п/п	Назва атрибуту	Тип даних	Опис
1.	ID	int	Первинний ключ, числовий ідентифікатор для однозначного визначення запису таблиці.
2.	Text	text	Розгорнутий висновок щодо аналізу коментаря.
3.	FK_Expected_meaning	Int	Вторинний ключ, посилання на таблицю Expected_meanings для співставлення із відповідним записом про очікуване значення коментаря
4.	FK_Actual_meaning	Int	Вторинний ключ, посилання на таблицю Actual_meanings для співставлення із відповідним записом про актуальне значення коментаря
5.	Relevance	double	Відповідність коментаря

Таблиця «Expected_meanings» (таблиця 2.6) містить визначену нейромережевою моделлю очікувану тему відгуку.

Таблиця 2.6 – Атрибути таблиці «Expected_meanings»

№ п/п	Назва атрибуту	Тип даних	Опис
1.	ID	int	Первинний ключ, числовий ідентифікатор для однозначного визначення запису таблиці.
2.	Theme	Varchar(50)	Назва теми коментаря

Таблиця «Neural_Networks» (таблиця 2.7) містить загальну інформацію про нейромережі, які використовуються при оцінці коментарів. Таблиця включає в себе наступні атрибути: назва нейромережі; посилання на вторинні таблиці (тип нейромережі, який використовувався та датасет, на якому тренувалась певна нейромережа); час тренування нейромережі; функції точності, втрат, F1-норми та матрицю сплутувань; шлях до збереженої моделі нейромережі.

Таблиця 2.7 – Атрибути таблиці «Neural_Networks»

№ п/п	Назва атрибуту	Тип даних	Опис
1.	ID	int	Первинний ключ, числовий ідентифікатор для однозначного визначення запису таблиці.
2.	Name	Varchar(50)	Назва
3.	FK_Type	Int	Вторинний ключ, посилання на таблицю Neural_Network_Type для співставлення із відповідним записом про тип нейромережі, який використовувався
4.	FK_Training_dataset	Int	Вторинний ключ, посилання на таблицю Training_dataset для співставлення із відповідним записом про датасети, які використовувались для тренування нейромережі
5.	Training_time	datatime	Час, який було витрачено на навчання нейромережі
6.	Accuracy	double	Параметр для визначення точності роботи нейромережі
7.	Loss_function	double	Параметр для визначення функції втрат нейромережі
8.	Confusion_matrix	double	Параметр матриці сплутувань нейромережі
9.	F1_score	double	Значення параметру F1-норма.
10.	Path	varchar	Шлях до збереженої моделі нейромереж

Таблиця «Actual_meanings» (таблиця 2.8) містить визначену нейромережевою моделлю актуальну тему відгуку.

Таблиця 2.8 – Атрибути таблиці «Actual_meanings»

№ п/п	Назва атрибуту	Тип даних	Опис
1.	ID	int	Первинний ключ, числовий ідентифікатор для однозначного визначення запису таблиці.
2.	Theme	Varchar(50)	Назва теми коментаря

Таблиця «Datasets» (таблиця 2.9) містить інформацію про датасети, які використовувались для тренування нейромереж, а саме: назву, вагу, тренувальні та файли для валідації.

Таблиця 2.9 – Атрибути таблиці «Datasets»

№ п/п	Назва атрибуту	Тип даних	Опис
1.	ID	int	Первинний ключ, числовий ідентифікатор для однозначного визначення запису таблиці.
2.	Name	Varchar(50)	Назва датасету
3.	Weight	double	Вага датасету
4.	Training_files	Varchar	Файли, які використовуються у тренуванні нейромережі
5.	Validation_files	Varchar	Файли, які використовуються у валідації нейромережі

Таблиця «Neural_Network_Type» (таблиця 2.10) містить інформацію про тип нейромережі та її назву.

Таблиця 2.10 – Атрибути таблиці «Neural_Network_Type»

№ п/п	Назва атрибуту	Тип даних	Опис
1.	ID	int	Первинний ключ, числовий ідентифікатор для однозначного визначення запису таблиці.
2.	Name	Varchar(50)	Назва типу нейромережі

Таблиця «Trained_models» (таблиця 2.11) містить інформацію про тренувальні моделі нейромереж, дату їх проведення та посилання на вторинну таблицю з видами мовних моделей.

Таблиця 2.11 – Атрибути таблиці «Trained_models»

№ п/п	Назва атрибуту	Тип даних	Опис
1.	ID	int	Первинний ключ, числовий ідентифікатор для однозначного визначення запису таблиці.
2.	FK_Model	int	Вторинний ключ, посилання на таблицю Training_Models для співставлення із відповідним записом про тренувальні моделі для нейромережі
3.	Date	datetime	Дата проведення тренування

Таблиця «Language_models» (таблиця 2.12) містить інформацію про мовні моделі, які використовувались у тренуванні нейромереж. Має наступні атрибути: назва, опис моделі та дата проведення тренування.

Таблиця 2.12 – Атрибути таблиці «Language_models»

№ п/п	Назва атрибуту	Тип даних	Опис
1.	ID	int	Первинний ключ, числовий ідентифікатор для однозначного визначення запису таблиці.
2.	Name	Varchar(50)	Назва мовної моделі
3.	Description	text	Опис мовної моделі

На рисунку 2.6 наведено схему бази даних для програмного застосунку на базі способу для визначення семантично невідповідних україномовних коментарів до описів товарів інтернет-магазину для задач аналізу купівельного попиту.

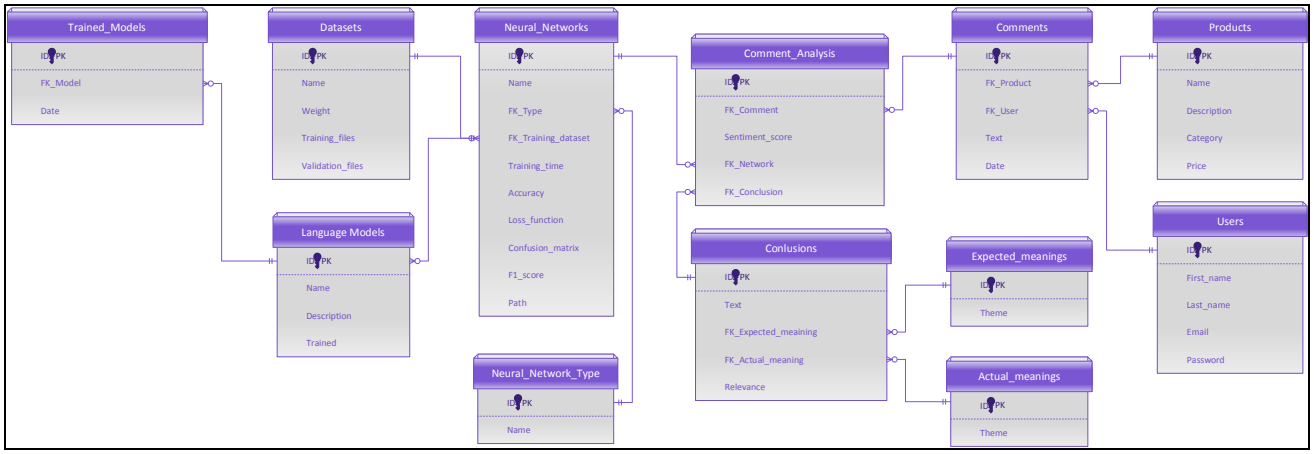


Рисунок 2.6 – Схема бази даних для програмного застосунку

Отже, було створено базу системи визначення семантично невідповідних україномовних коментарів до описів товарів інтернет-магазину для задач аналізу купівельного попиту. Було представлено загальну структуру бази даних із зв'язками між ними. Кожна таблиця була детально описана, вказано поля та їх призначення. Також було надано візуальне представлення бази даних у вигляді діаграми.

Отримана база даних може бути використана для ефективного аналізу купівельного попиту та виявлення невідповідностей україномовних коментарів

до товарів, що дозволяє покращити якість обслуговування та задоволення потреб клієнтів україномовного інтернет-магазину.

2.6 Метрики оцінювання ефективності визначення семантичної невідповідності україномовних коментарів до описів товарів

Для оцінювання машинного розв'язку задач двійкової класифікації використовуються метрики. Однією із ключових метрик є матриця плутанини.

Є кілька ключових термінів, які повинні бути враховані, коли мова йде про продуктивність моделей класифікації. Ці терміни найкраще описуються та визначаються за допомогою матриці плутанини. Матриця плутанини або матриця помилок є одним із ключових понять, коли йде мова про проблеми класифікації. Ця матриця є матрицею NxN і є табличним представленням прогнозів моделі проти фактичних значень. Вигляд матриці сплутувань показано на рисунку 2.7 [20].

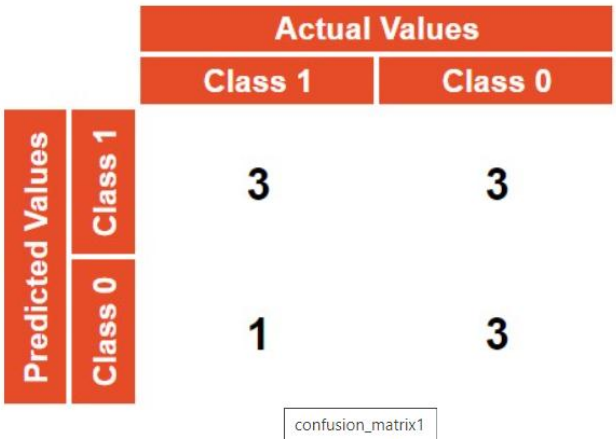


Рисунок 2.7 – Матриця сплутувань [21]

Кожен стовпець і рядок присвячені одному класу. На одному кінці показано фактичні значення, а на іншому – прогнозовані значення. У наведеному вище прикладі з 4 значень, позначених як клас 0, модель правильно класифікувала 3 значення та неправильно класифікувала 1 значення. Ця модель

із 6 значень, позначених як клас 1, правильно позначена як 3 і неправильно класифікована як 3.

Якщо клас 1 названо позитивним, а клас 0 – негативним, тоді 3 зразки, прогнозовані як клас 0, називаються істинно негативними, а 1 зразок, прогнозований як клас 1, називають хибнонегативними. 3 зразки, правильно класифіковані як клас 1, називаються справді позитивними, а ці 3 неправильно класифіковані екземпляри називаються хибнопозитивними (рисунок 2.8).

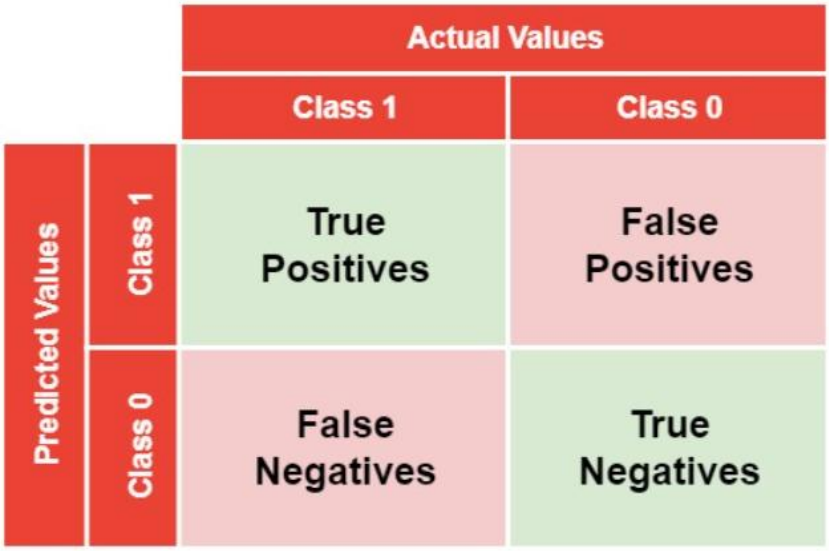


Рисунок 2.8 – Пояснення до матриці сплутувань [22]

Матриця сплутувань не лише дає деталі про те, як працює розроблена модель прогнозування, але й на основі концепцій, викладених у цій матриці, можна будувати деякі інші показники. Деякі з них є Precision та Recall.

Precision є дуже корисним показником, і містить більше інформації, ніж Acuracy. По суті можна точно відповісти на питання: «Яка частка позитивних ідентифікацій була правильною?». Розраховується для кожного класу окремо за формулою (1):

$$\Pr ecision = \frac{TruePositives}{(TruePositives + FalsePositives)} \quad (1)$$

Recall можна описати як здатність класифікатора знаходити всі позитивні зразки [23]. За допомогою цього показника можна отримати відповідь на

питання: «Яку частку фактичних позитивних результатів було визначено правильно?» Він визначається як частка зразків із певного класу, які правильно передбачені моделлю та математично обраховується за формулою (2):

$$\mathrm{Re}\,call = \frac{TruePositives}{(TruePositives + FalsePositives)} \tag{2}$$

Precision і Recall різні, але водночас доволі схожі. Точність є мірою релевантності результату, тоді як Recall є мірою того, скільки справді релевантних результатів повертається. На початку може бути важко розшифрувати різницю між цими двома.

Ще однією базовою метрикою, яку легко зрозуміти є точність (Acuracy). Розраховується як кількість правильних прогнозів, поділена на загальну кількість прогнозів [24]. Якщо помножити це на 100, то буде отримано точність у відсотках. Розраховується за формулою (3):

$$Acuracy = \frac{NumberOfCorrect\ \mathrm{Pr}ediction}{TotalNumberOfSimples} \tag{3}$$

Ще одним показником для оцінювання ефективності є оцінка F1, який поєднує в собі точність і запам’ятовування [25]. Він представляє їх середнє гармонійне. Для двійкової класифікації можена визначити даний показник за допомогою формули (4):

$$F1Score = 2 * \frac{\mathrm{Pr}ecision * \mathrm{Re}\,call}{(\mathrm{Pr}ecision + \mathrm{Re}\,call)} \tag{4}$$

Отже, для оцінювання ефективності нейромережевого визначення семантичної невідповідності україномовних коментарів до описів товарів будуть використані метрики Acuracy, оцінка F1, та метрика ConfusionMatrix.

Розділ 3 Програмна реалізація для визначення семантично невідповідних україномовних коментарів до описів товарів інтернет-магазину

3.1 Структура модулів інформаційної системи, їх взаємозв'язок

Інформаційна система визначення семантично невідповідних україномовних коментарів до описів товарів інтернет-магазину складається із прикладного застосування, діаграма класів якого проілюстрована на рисунку 3.1, та інтернет-магазину класичної архітектури з можливістю вибору товарів, можливістю написати відгук про товар та можливістю оцінки рівня відповідності відгуків до залишених раніше відгуків про товар.

Прикладне застосування має в собі три основні модулі, що призначені для навчання класифікатора для обраного товару та роботи з відгуками.

Головним модулем є модуль оцінки відповідності відгука до обраного товару («*EvaluateReviewNN*»), що має методи для визначення семантично невідповідних відгуків до товарів. Для завантаження навченої моделі на обраний товар використовується метод *loadModel()*. Клас використовує як вхідні дані навчені у нейромережевому модулі моделі. Метод *PreprocessText()* призначений для попередньої обробки текстових даних користувацького відгуку, його векторизації методом *TF-IDF*. Метод *GetReviesForCategory()* призначений для для виводу переліку вже збережених у базі відгуків, що присвячені обраному товару. Метод *DisplayRevew()* призначений для виводу тексту обраного відгуку по обраному товару.

Нейромережевий модуль реалізовано класом «*NeuralNetworkModul*». Даний клас містить методи для навчання та збереження нейромереж. Метод *AutoTrainModel()* призначений для автоматичного вибору класифікатора, який найкраще виконує класифікацію. Метод *TestModel()* призначений для тестування ефективності нейромережі за метриками. Метод *UseModel()* використовується для використання навчених моделей.

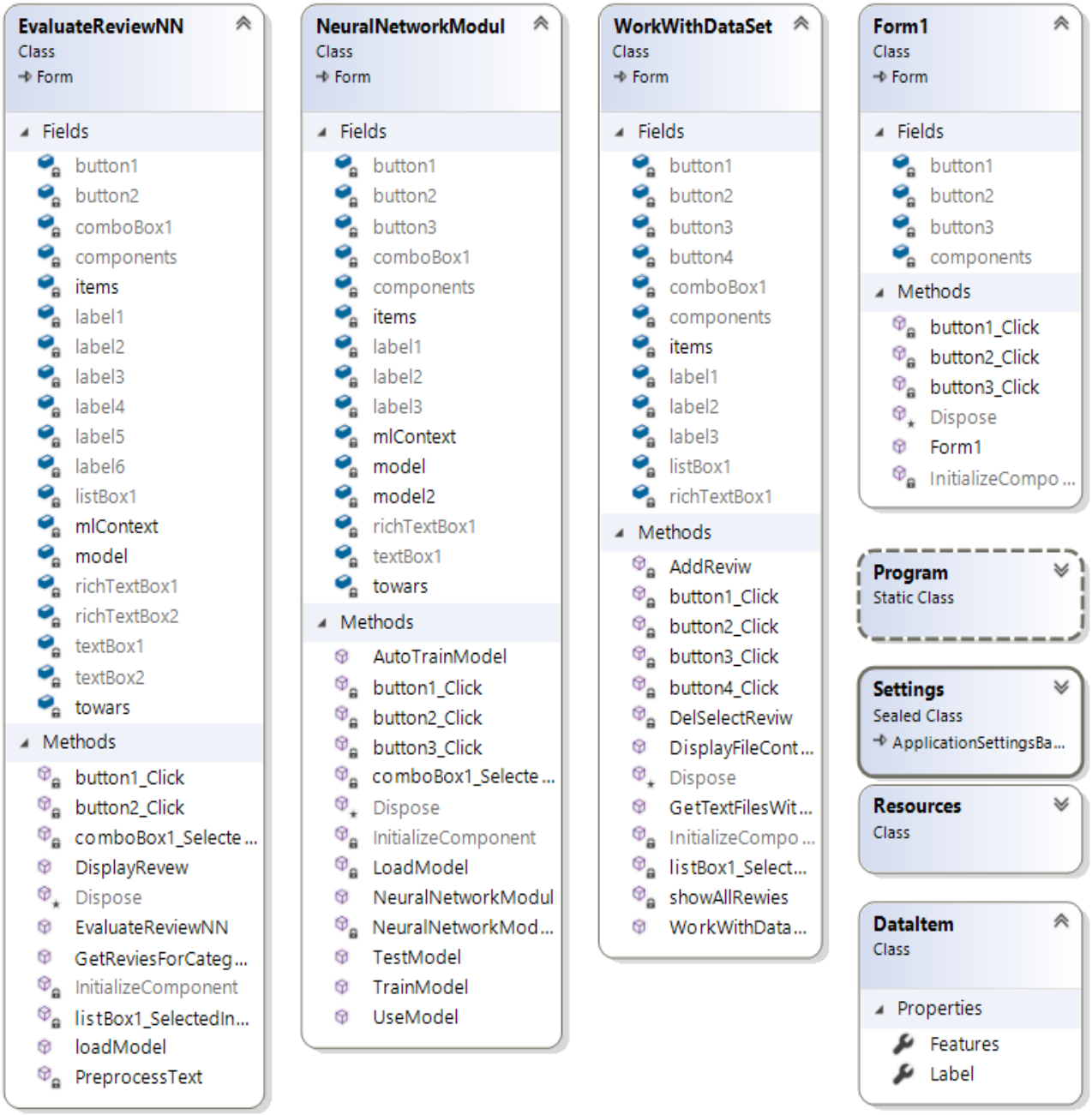


Рисунок 3.1 – Діаграма класів застосунку

Клас «*DataItem*» містить ознаки для структури файлу навчання, яка складається із ознак та мітки.

Клас «*WorkWithDataSet*» призначений для взаємодії користувача з набором даних відгуків та описами товарів. Він містить основні методи для додавання відгука до обраного товару (метод «*AddReviw()*»), для видалення обраного відгука з бази (метод «*DelSelectReviw()*»). Метод *«GetTextFilesWithKeyword()»* призначений для відображення відгуків, виключно

тих що стосуються обраного товару. Метод *showAllRewies()* призначений для відображення усіх відгуків датасету.

Отже, таким чином розроблено структуру інформаційної системи визначення семантично невідповідних україномовних коментарів до описів товарів інтернет-магазину, що складається з ряду модулів для її функціонування.

3.2 Особливості розробки складових інформаційної системи визначення семантичної невідповідності україномовних коментарів

Для визначення невідповідності україномовних коментарів до товарів було створено відповідну програмну реалізацію у формі окремого додатку до інтернет-магазину класичної архітектури з можливістю вибору товарів, можливістю написати відгук про товар та можливістю оцінки рівня відповідності відгуків до залишених раніше відгуків про товар, а саме Windows Form застосування дослідницького характеру.

Застосування дослідницького характеру складається із 3х основних модулів: «Нейромережевого модуля», що призначений для здійснення навчання моделей для обраних товарів, оцінювання якості навчених моделей та збереження навчених моделей, які у подальшому будуть використані для оцінки семантичної невідповідності україномовних коментарів до описів товарів, «Модуля оцінки семантичної відповідності до описів товарів інтернет-магазинів», що призначений для оцінювань відгуків обраного товару та безпосереднього здійснення оцінка рівня відповідності відгуку, а також «Модуль роботи з відгуками» для роботи з відгуками.

Для реалізації нейромережевого модуля було використано бібліотеку «*Microsoft.ML*», що створена для машинного навчання компанією Microsoft. Вона надає розширений функціонал для побудови моделей машинного навчання для визначення семантично невідповідних україномовних коментарів і включає в себе різноманітні алгоритми, метрики та інструменти для підготовки даних і оцінки моделей. *Microsoft.ML* підтримує різні види задач машинного навчання,

включаючи бінарну класифікацію, яка використовується для роботи моделі для визначення семантично невідповідних україномовних коментарів.

Для завантаження даних для навчання моделі (файлу розмітки), використовується метод *Data.LoadFromTextFile()* класу *MLContext*, що представляє собою контекст для виконання операцій машинного навчання, таких як завантаження даних, побудова моделей, навчання, оцінка та застосування моделей.

Оскільки у процесі навчання користувач не знає наперед, який з алгоритмів покаже найкращий результат, створюється масив *trainers* типу *IEstimator<ITransformer>[]*, який містить різні алгоритми навчання для бінарної класифікації для визначення семантично невідповідних україномовних коментарів до описів товарів.

Далі для кожної моделі в циклі створюється пайплайн, що в Microsoft.ML використовується для послідовного виконання операцій обробки даних та навчання моделі машинного навчання. Він дозволяє послідовно поєднувати різні компоненти обробки даних та моделі, щоб створити повну послідовність дій для розв'язання задачі виявлення семантично невідповідних україномовних коментарів.

Після чого відбувається оцінка метриками, де зберігається найкращий навчений зразок. Вищеописаний програмний код призначений для виконання натиснення на кнопку «Навчити модель» (рисунок 3.2). Перед натисненням на кнопку «Навчити модель» необхідно обрати потрібний товар з випадаючого списку, а також увести шлях до файлу навчання нейромережі із розміткою. Або обрати потрібний файл через провідник, натиснувши кнопку «Обрати файл через провідник».

Серед алгоритмів, які будуть виконуватись для бінарної класифікації є наступні: *FastTree(), AveragedPerceptron(), FastForest(), LinearSvm().*

FastTree є алгоритмом, який доступний для використання в бібліотеці ML.NET, яка надає інструменти для розробки моделей машинного навчання.

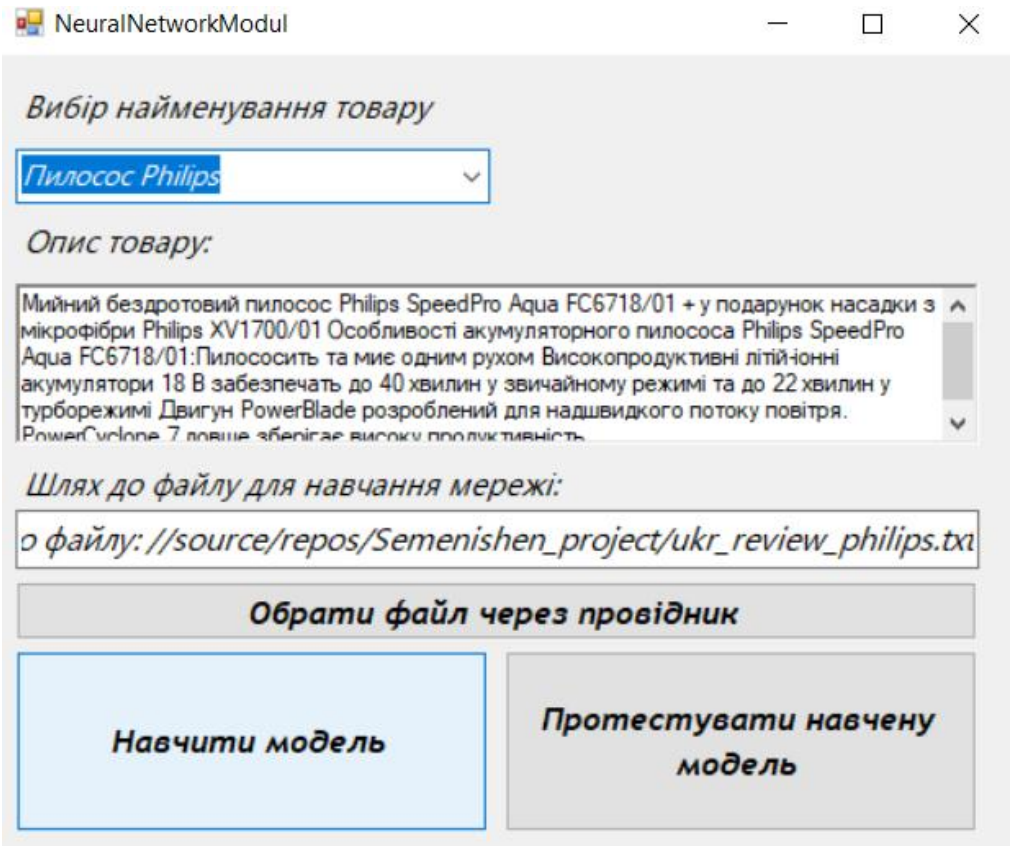


Рисунок 3.2 – Інтерфейс нейромережевого модуля

У *ML.NET* реалізація *FastTree* надає можливість використовувати цей алгоритм для завдань бінарної класифікації та регресії. *FastTree* використовується у складі пайплайну навчання моделі, де його можна поєднувати з іншими компонентами обробки даних та моделями. Цей алгоритм базується на ідеї рішучих дерев, але відрізняється від класичного дерева рішень, так як використовує метод швидкої апроксимації для ефективного навчання та передбачення.

Алгоритм *Averaged Perceptron* є модифікованою версією класичного алгоритму *Perceptron*. Він є лінійним бінарним класифікатором, який навчається в процесі ітеративного оновлення ваг та порогів. Алгоритм *Averaged Perceptron* використовується для багатокласової класифікації, де для кожного класу будується окремий бінарний класифікатор, який розпізнає цей клас проти всіх інших класів.

FastForest є алгоритмом машинного навчання, який доступний в бібліотеці ML.NET. Цей алгоритм використовується для завдань класифікації та

регресії. Він базується на ідеї випадкових лісів (*Random Forest*) і є розширенням алгоритму *FastTree*.

FastForest використовує техніку випадкових лісів (*Random Forest*), яка полягає в побудові ансамблю дерев рішень. Кожне дерево навчається на підмножині вибірки даних та використовує підмножину ознак. Після навчання, модель комбінує прогнози кожного дерева, щоб отримати кінцевий результат.

LinearSvm (лінійний метод опорних векторів) є алгоритмом машинного навчання, доступним в бібліотеці ML.NET. Він використовується для задач бінарної класифікації, де модель навчається розділяти дані на два класи за допомогою гіперплощини. Алгоритм побудований на основі лінійної моделі, яка намагається знайти оптимальну гіперплощину, що розділяє дані двох класів. Ця модель досить проста, але ефективна для багатьох задач класифікації.

Після збереження найкращої версії моделі, можна оцінити її результати. Для цього було запрограмовано кнопку «Протестувати навчену модель» (рисунок 3.3).

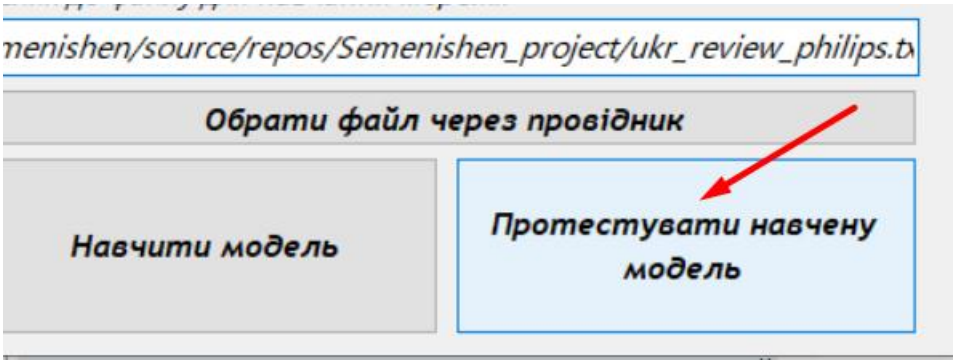


Рисунок 3.3 – Кнопка для тестування навченої моделі

В обробнику натиснення на кнопку викликаються методи завантаження збереженої моделі, а також методи виклику метрик для ілюстрацій результатів навчання. Результат виведення матриці плутанини проілюстровано на рисунку 3.4.

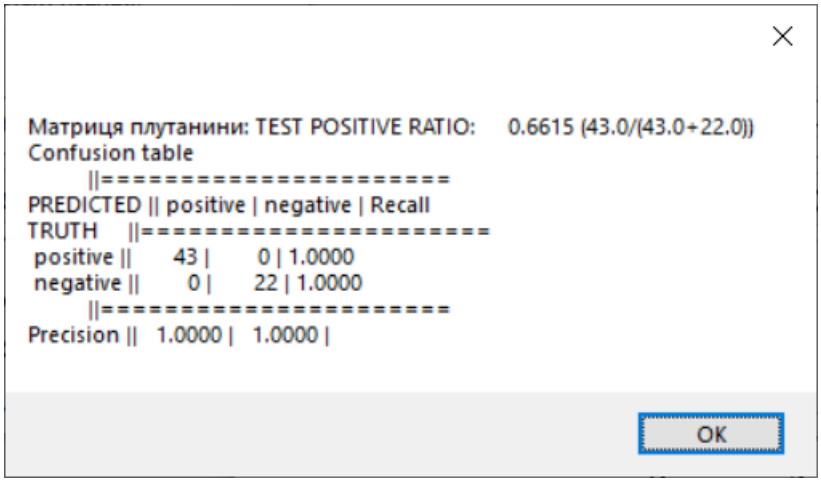


Рисунок 3.4 – Виведення матриці плутанини

Головним модулем дослідницького програмного забезпечення є «Модуль оцінки семантичної відповідності до описів товарів інтернет-магазинів». Він використовує моделі, навчені попереднім розглянутим модулем і призначений для визначення невідповідних відгуків до обраного товару. Результат виводу оцінки відповідності відгука продемонстровано на рисунку 3.5.

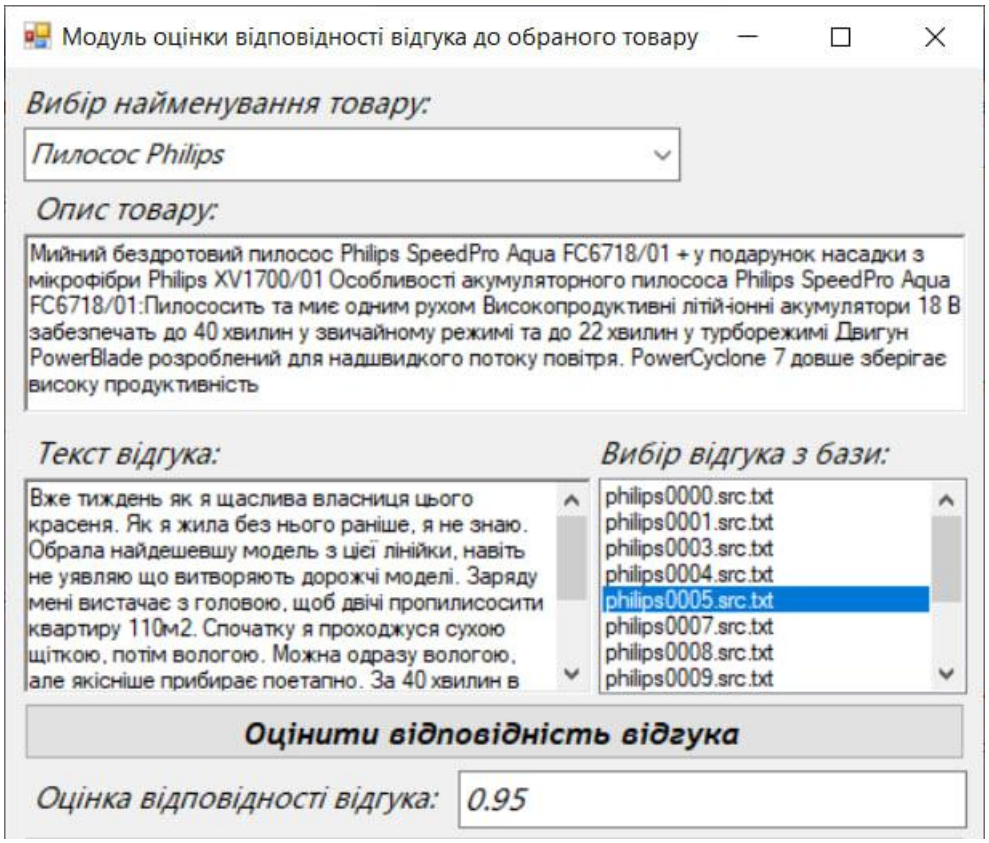


Рисунок 3.5 – Оцінка відгукадо пилососа «Philips»

Також можна подивитись параметри класифікатора, який здійснює оцінку відповідності відгука до обраного товару. Параметри класифікатора будуть виведені по натисненні на кнопку «Показати параметри класифікатора» (рисунок 3.6).

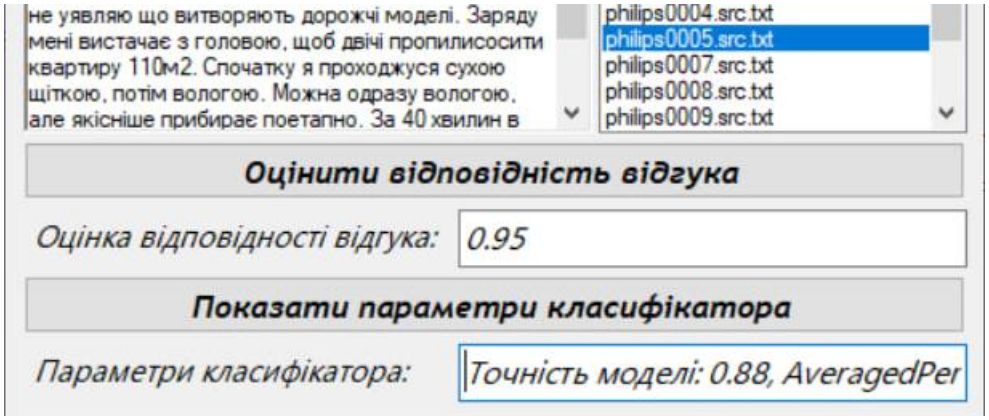


Рисунок 3.6 – Відображення параметрів класифікатора

Розроблений спосіб дозволяє визначати відповідність відгуків користувачів до побутових товарів інтернет-магазину (рисунок 3.7).

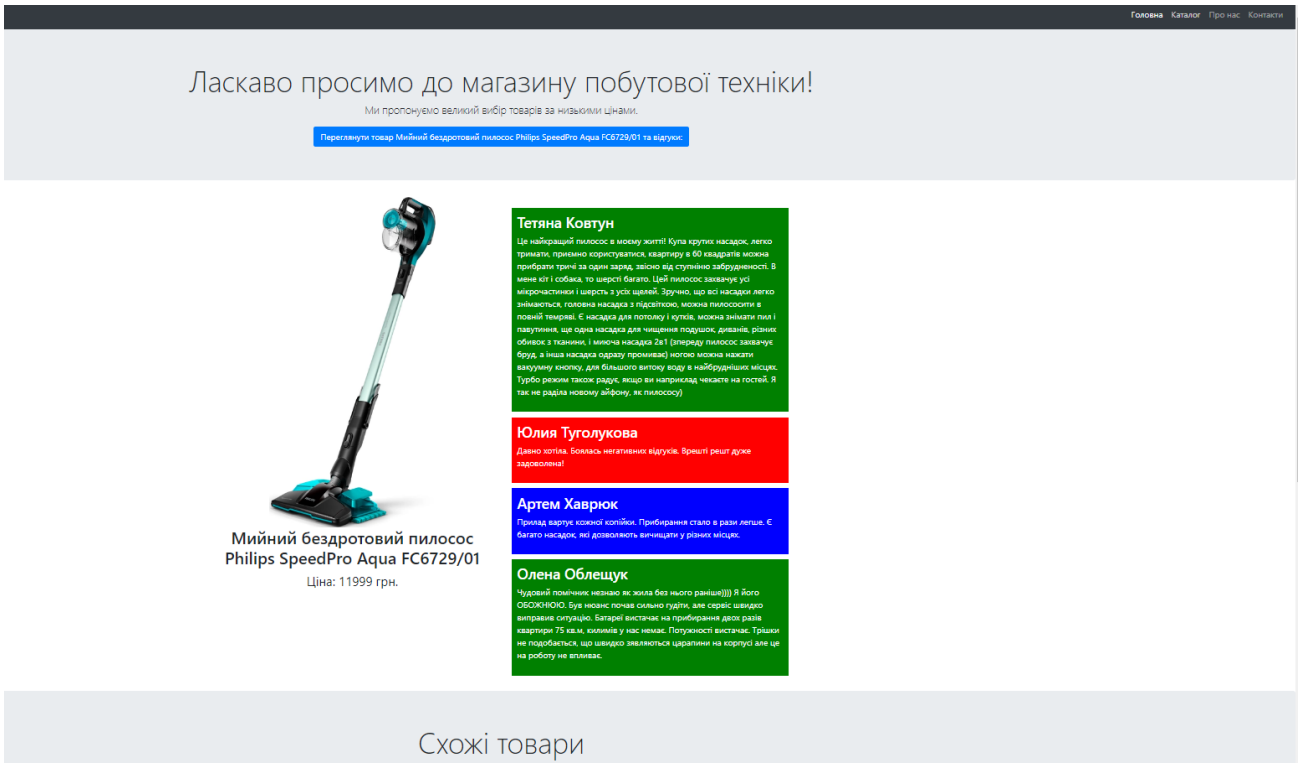


Рисунок 3.7 – Вигляд обраного товару інтернет-магазину

Зеленим позначено відгуки, що стосуються опису товару. Відгуки що відповідають опису товару вважаються такими, відсоток приналежності яких вищий за 0,7 за нейромережевою оцінкою. Такі що стосуються товару, проте неоднозначно – від 0,4 до 0,69 (позначено синім). І такі що не стосуються товару – до 0,39 (позначено червоним).

Отже, таким чином було описано особливості розробки складових інформаційної системи визначення семантично невідповідних україномовних коментарів.

3.3 Тестування інформаційної системи визначення семантичної невідповідності україномовних коментарів

Створена інформаційна система визначення семантично невідповідних україномовних коментарів до описів товарів інтернет-магазину, що складається із прикладного застосування, та опціонально інтернет-магазину класичної архітектури (не входить до складу роботи) повинна бути протестована засобами тест-кейсів з метою виявлення можливих дефектів в роботі.

Першим тестовим випадком є перевірка відображення файлів датасету, що складається із коментарів та описів товарів. Кроки тестового випадку наведені у таблиці 3.1.

Таблиця 3.1 – Тест-кейс 00001

Тест-кейс ID: 00001	**Приорітет:** 1	**Створено:** 1.05.2023, Семенишен А.Л.
Назва: Перевірка відображення файлів датасету		
Кроки		**Очікуваний результат**
1. Запустити дослідницький модуль інформаційної системи 2. Натиснути кнопку «Модуль роботи з датасетом» 3. Натиснути на кнопку «Показати весь датасет»		Відкривається головний екран дослідницького модуля Відкривається модуль роботи з датасетом У вікні «Оберіть файл для перегляду» будуть відображені всі файли що містять відгуки та описи товарів
Результат виконання тест-кейсу: перевірку пройдено успішно.		

Після запуску застосування та виконання кроків, описаних у таблиці 3.1, можна переконатись, що заявлений функціонал працює коректно. Результат зображено на рисунку 3.8.

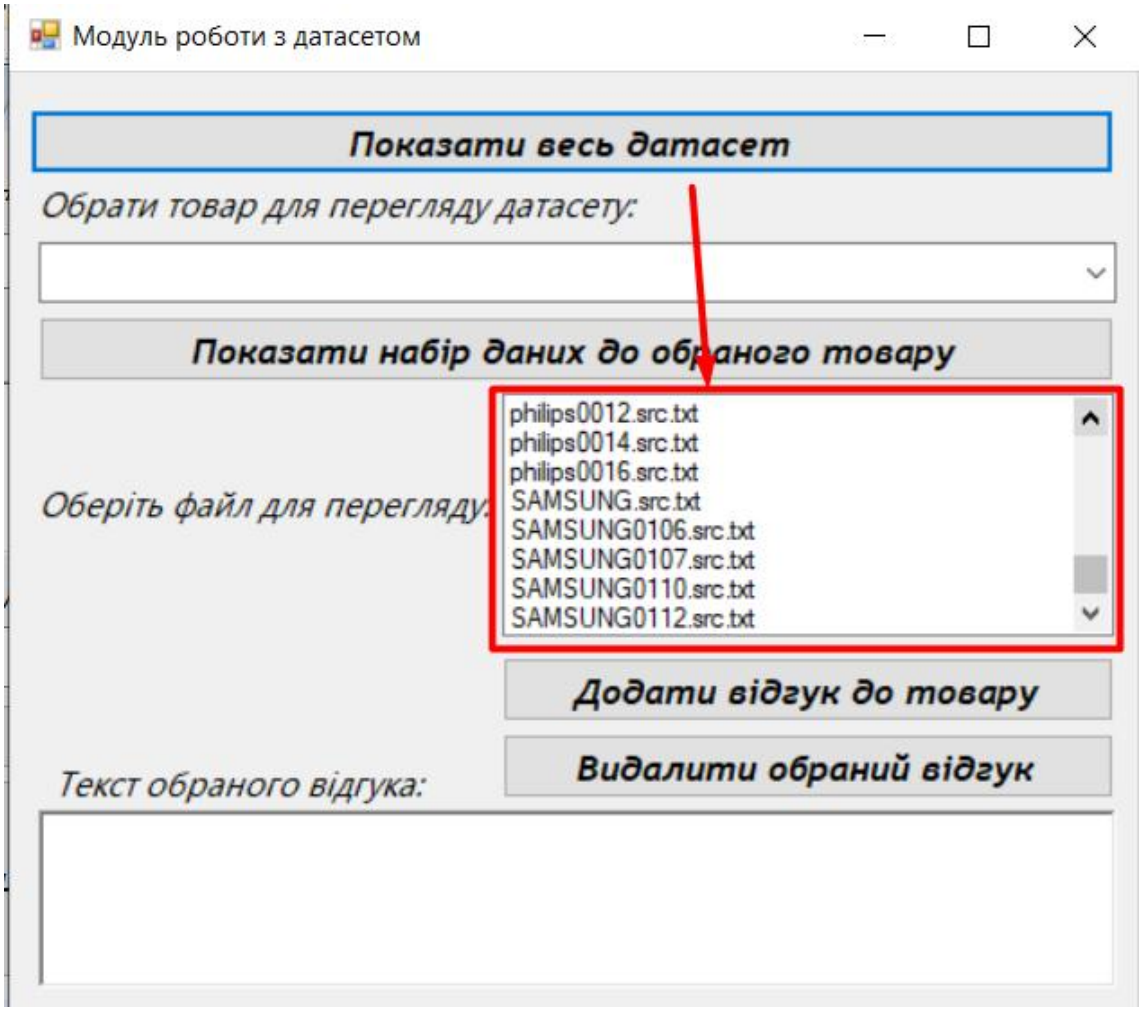


Рисунок 3.8 – Відображення файлів відгуків датасета

Наступним тестовим випадком буде перевірка відображення файлів, що стосуються лише обраного товару. Кроки тестового випадку наведені у таблиці 3.2.

Таблиця 3.2 – Тест-кейс 00002

Тест-кейс ID: 00002	**Приорітет:** 1	**Створено:** 3.05.2023, Семенишен А.Л.
Назва: Перевірка відображення файлів-відгуків до обраного товару		
Кроки		**Очікуваний результат**
1. Запустити дослідницький модуль інформаційної системи 2. Натиснути кнопку «Модуль роботи з датасетом» 3. З випадаючого переліку «Обрати товар для перегляду датасета» обрати «Пральна машина вузька Sumsung». 4. Натиснути на кнопку «Показати набір даних до обраного товару»		Відкривається головний екран дослідницького модуля Відкривається модуль роботи з датасетом З випадаючого списку «Обрати товар для перегляду датасета» відображено вибір «Пральна машина вузька Sumsung». Виведено набір даних до товару «Пральна машина вузька Sumsung».
Результат виконання тест-кейсу: перевірку пройдено успішно.		

Після запуску застосування та виконання кроків, що описані у таблиці 3.2, можна переконатись, що відображення файлів-відгуків до обраного товару працює коректно. Підтвердження результату тестування зображено на рисунку 3.9.

Наступним тестовим випадком буде перевірка відображення вмісту файлів-відгуків, що стосуються лише обраного товару. Кроки тестового випадку наведені у таблиці 3.3.

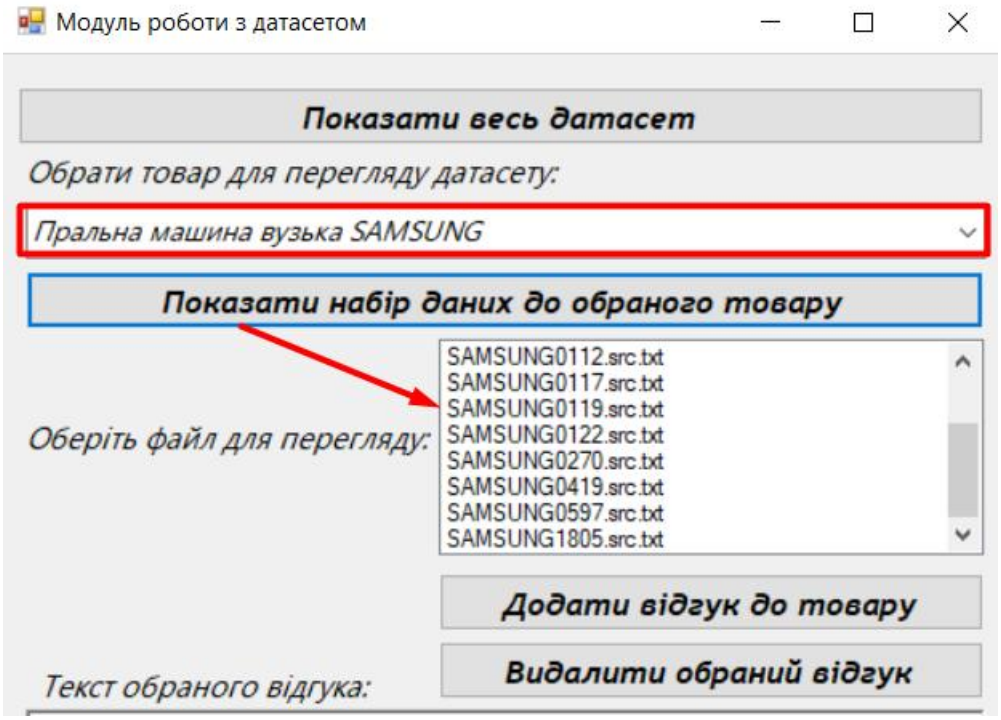


Рисунок 3.9 – Відображення файлів відгуків до обраного товару

Таблиця 3.3 – Тест-кейс 00003

Тест-кейс ID: 00003	**Приорітет:** 1	**Створено:** 3.05.2023, Семенишен А.Л.
Назва: Перевірка відображення вмісту файлів-відгуків до обраного товару		
Кроки		**Очікуваний результат**
1. Запустити дослідницький модуль інформаційної системи 2. Натиснути кнопку «Модуль роботи з датасетом» 3. З випадаючого переліку «Обрати товар для перегляду датасета» обрати «Пральна машина вузька Samsung». 4. Натиснути на кнопку «Показати набір даних до обраного товару» 5. Обрати файл для перегляду «SUMSUNG0110».		Відкривається головний екран дослідницького модуля Відкривається модуль роботи з датасетом З випадаючого списку «Обрати товар для перегляду датасета» відображено вибір «Пральна машина вузька Samsung». Виведено набір даних до товару «Пральна машина вузька Sumsung». Вміст відгуку відображено у полі «Текст обраного відгука».
Результат виконання тест-кейсу: перевірку пройдено успішно.		

Після запуску застосування та виконання кроків, що описані у таблиці 3.3, можна переконатись, що відображення вмісту файлів-відгуків до обраного товару працює коректно. Підтвердженням результату проведеного тестування є відображення вмісту файлу, що проілюстровано на рисунку 3.10.

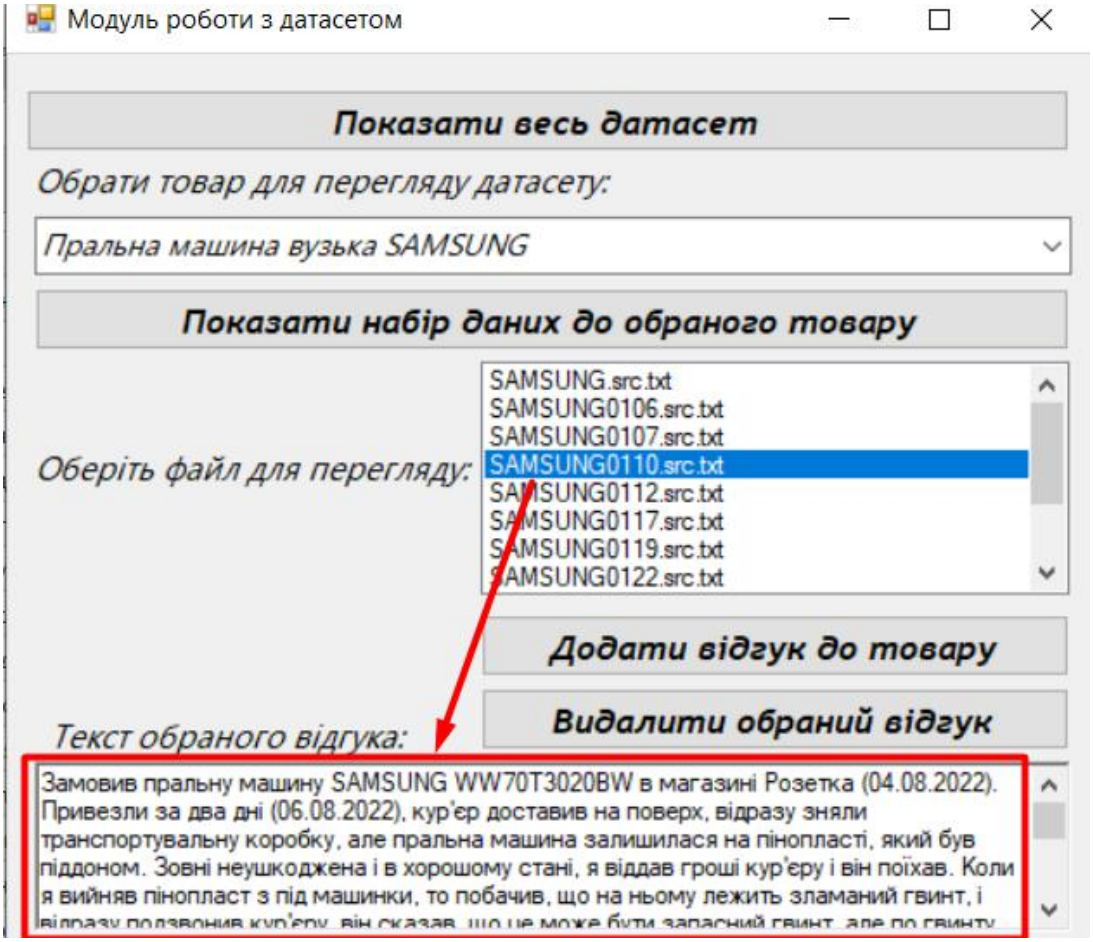


Рисунок 3.10 – Відображення вмісту файлів-відгуків до обраного товару

Наступним тестовим випадком буде перевірка успішного збереження моделі після навчання для обраного товару. Для виконання тесту необхідно виконати кроки з таблиці 3.4.

Таблиця 3.4 – Тест-кейс 00004

Тест-кейс ID: 00004	**Приорітет:** 1	**Створено:** 3.05.2023, Семенишен А.Л.
Назва: Перевірка успішного збереження моделі після завершення навчання		
Кроки		**Очікуваний результат**
1. Запустити дослідницький модуль інформаційної системи 2. Натиснути кнопку «Нейромережевий модуль» 3. З випадаючого переліку «Найменування товару» обрати «Пилосос Philips». 4. Натиснути на кнопку «Длях до файлу» і обрати шлях до розміченого файлу «ukr_review_philips». 5. Натиснути на кнопку «Навчити модель»		Відкривається головний екран дослідницького модуля Відкривається нейромережевий модуль З випадаючого переліку «Найменування товару» обрано «Пилосос Philips». Відкриється файловий провідник, у якому при обиранні файлу «ukr_review_philips» буде відображено шлях до нього. Розпочнеться навчання моделі з подальшим виводом повідомлення «Модель навчена та збережена», в папці з застосунком є модель «save_model_philips».
Результат виконання тест-кейсу: перевірку пройдено успішно.		

Після запуску застосування та виконання кроків, що описані у таблиці 3.4, можна переконатись, що навчена на наборы даних для товару «Пилосос Philips». Підтвердженням результату збереження моделі є зображення 3.11.

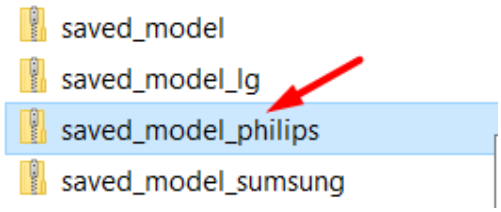


Рисунок 3.11 – Збережена модель у файловій системі

Також, як було вище зазначено, користувач буде бачити результат у вигляді повідомлення «Модель навчена та збережена», що проілюстровано на рисунку 3.12.

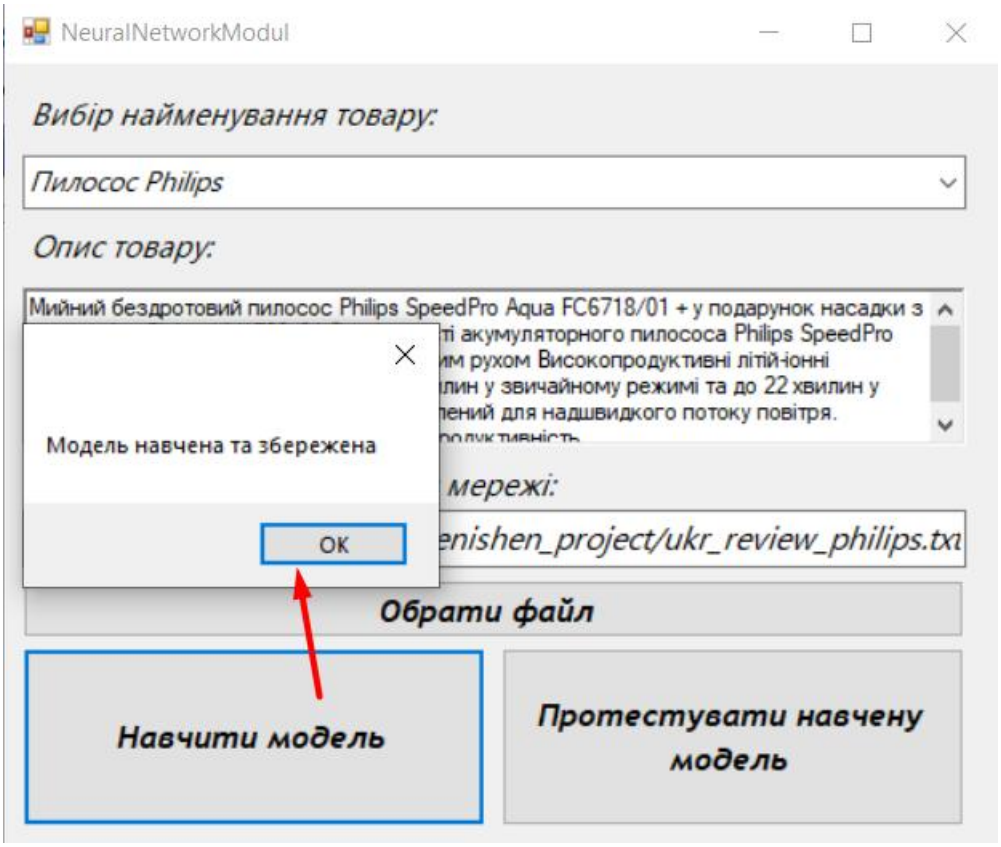


Рисунок 3.12 – Завершення навчання моделі

Наступним тестовим випадком буде перевірка оцінки відповідності відгуку до обраного товару. Для виконання тесту необхідно виконати кроки, описані в таблиці 3.5.

Таблиця 3.5 – Тест-кейс 00005

Тест-кейс ID: 00005	**Приорітет:** 1	**Створено:** 5.05.2023, Семенишен А.Л.
Назва: Перевірка оцінки відповідності відгуку до обраного товару		
Кроки		**Очікуваний результат**
1. Запустити дослідницький модуль інформаційної системи		Відкривається головний екран дослідницького модуля
2. Натиснути кнопку «Модуль визначення рівня невідповідності відгуку до опису товару»		Відкривається модуль визначення рівня невідповідності відгуку до опису товару
3. З випадаючого переліку «Вибір найменування товару» обрати «Пилосос Philips».		З випадаючого переліку «Найменування товару» обрано «Пилосос Philips».
4. Вибрати відгук з бази з назвою «philips0012».		Відкриється текст відгуку з бази.
5. Натиснути на кнопку «Оцінити відповідність відгука».		Прогнозована оцінка: відгук відповідний з оцінкою більше 0.6
Результат виконання тест-кейсу: перевірку пройдено успішно.		

Після запуску застосування та виконання кроків, що описані у таблиці 3.5, можна переконатись, що навчена модель на наборі даних для товару «Пилосос Philips» успішно, і справляється із завданням оцінки відповідності відгуку. Підтвердженням успішного проходження тесту є рисунок 3.12.

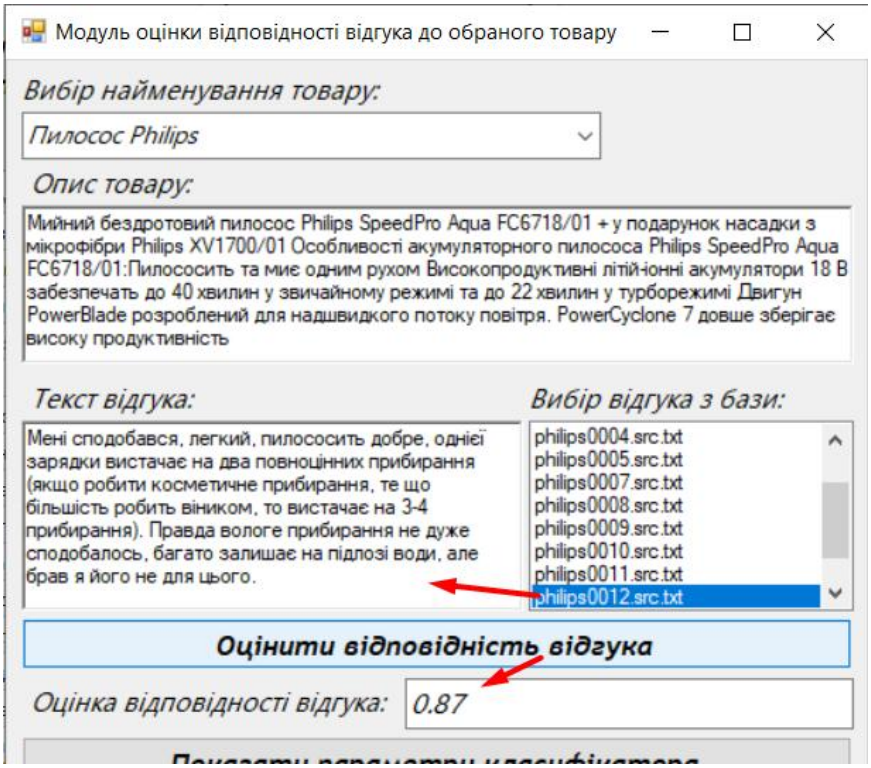


Рисунок 3.12 – Виявлення невідповідних відгуків

Якщо ж ввести у текстове поле «Текст відгуку» інший текст, що не стосується опису товару, оцінка відгуку буде меншою за 0.5. Результат виконання зображено на рисунку 3.13.

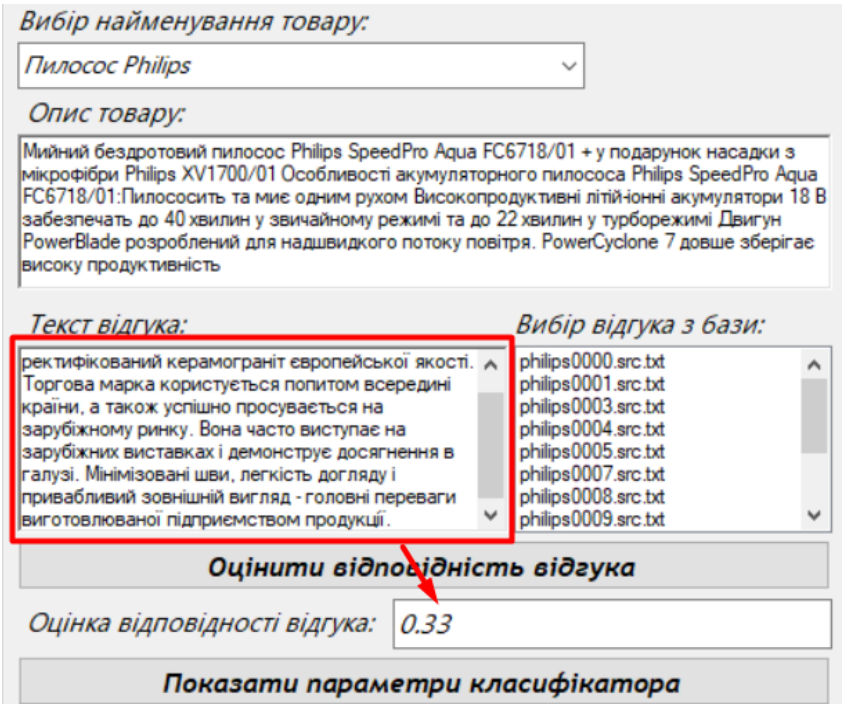


Рисунок 3.13 – Невідповідний відгук до опису товару

Показник 0.33 обумовлений тим, що все ж таки мова йде про товар, просто про інший. Проте, зважаючи на поріг в 0.5, класифікатор працює коректно.

Отже, було здійснено тестування інформаційної системи визначення семантично невідповідних україномовних коментарів до описів товарів інтернет-магазину. Відповідно до проведеного тестування некоректно працюючих функцій не виявлено.

3.4 Інструкція користувача до реалізованої інформаційної системи

Для зручності користування інформаційною системою визначення семантичної невідповідності україномовних коментарів до опису товарів інтернет-магазинів з метою аналізу купівельного попиту було створено інструкцію користувача. Після запуску програми користувач буде бачити стартовий екран для взаємодії із користувацькими модулями (рисунок 3.14).

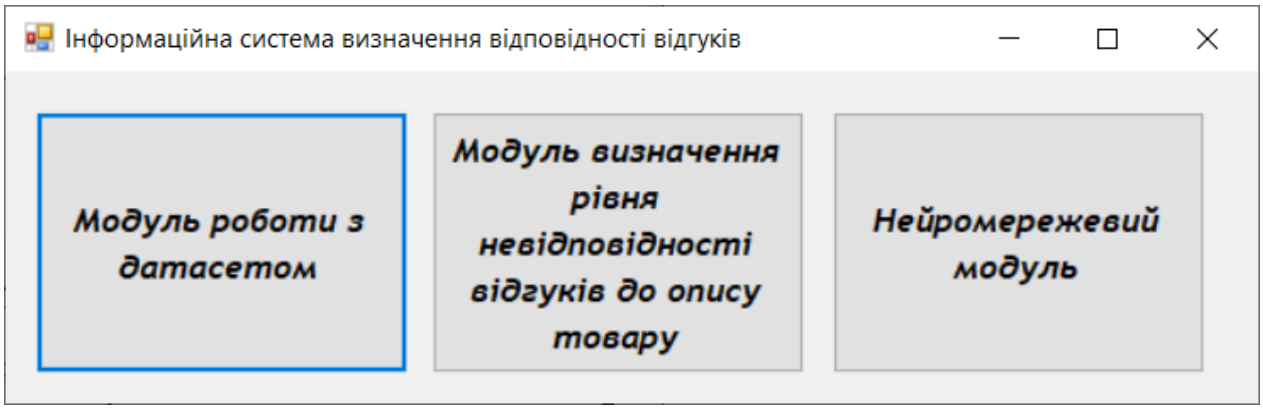


Рисунок 3.14 – Стартовий екран

При натисканні на кнопку «Модуль роботи з датасетом», користувач перейде на форму роботи з датасетом, інтерфейс якого зображено на рисунку 3.15.

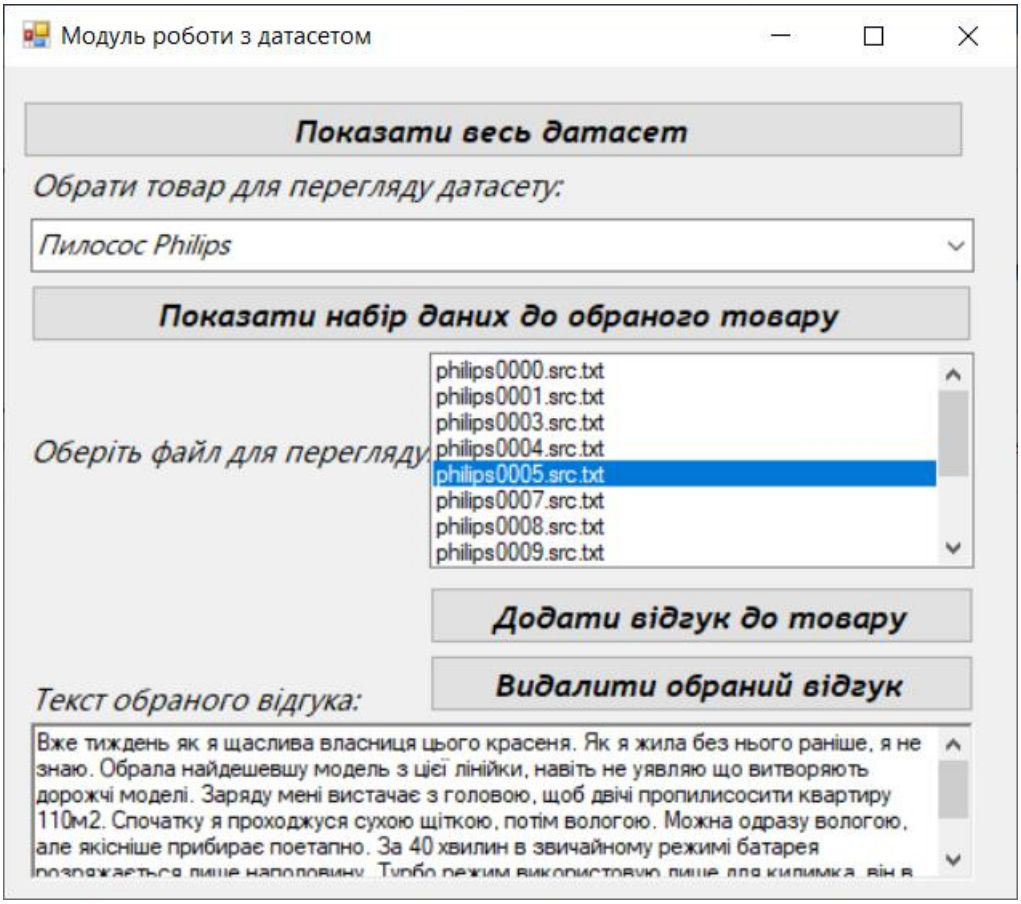


Рисунок 3.15 – Інтерфейс модуля роботи з датасетом

У рамках модуля є можливість перегляду як окремих відгуків до обраних товарів, так і переглянути усі наявні записи про товари. Щоб побачити весь датасет необхідно натиснути «Показати весь датасет» (рисунок 3.16).

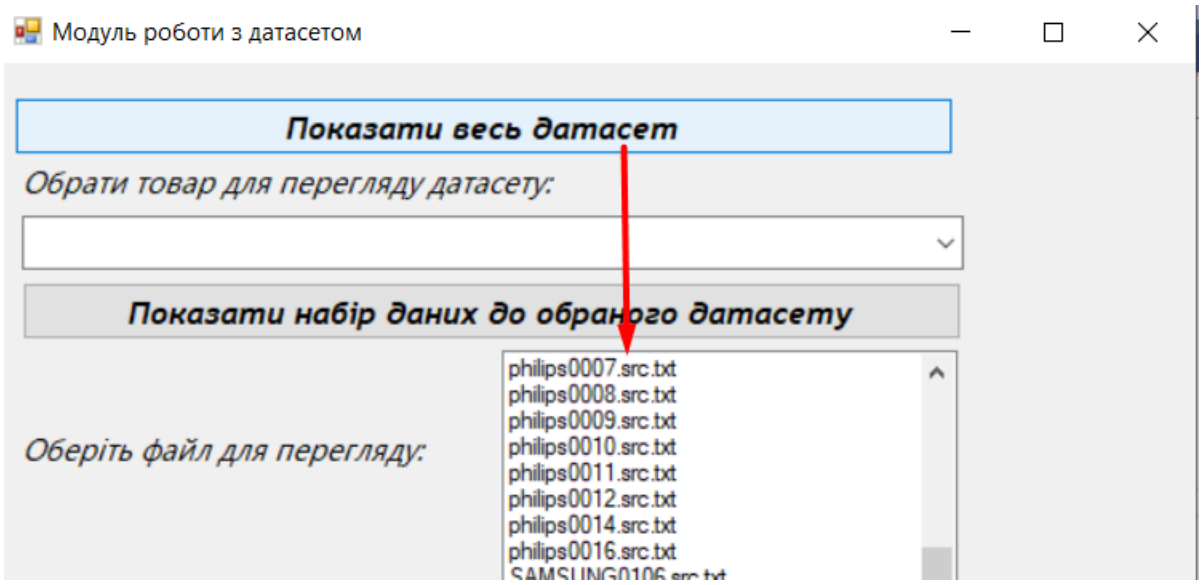


Рисунок 3.16 – Результат натиснення на кнопку «Показати весь датасет»

Для відображення відгуків до обраного товару необхідно у переліку товарів обрати товар, відгуки до якого потрібно переглянути (рисунок 3.17).

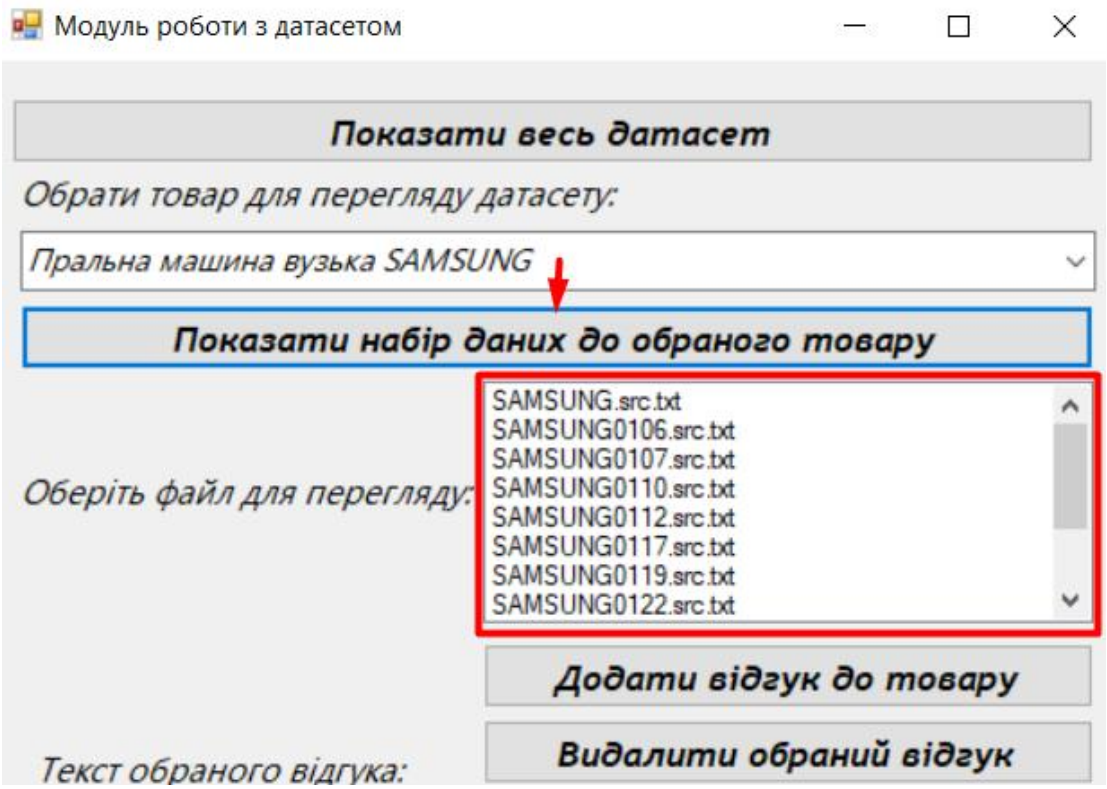


Рисунок 3.17 – Відгуки до товару «Пральна машина Samsung»

Також у рамках цього модуля є можливість додавати відгук до товару, для цього необхідно написати текст у поле «Текст обраного відгука» і натиснути кнопку «Додати відгук до товару». Написаний відгук буде збережено до бази.

Окрім додавання нового відгуку можна видаляти наявні. Для видалення наявного відгуку необхідно обрати відгук, який потрібно видалити з переліку наявних відгуків та натиснути кнопку «Видалити обраний відгук».

Для переходу на нейромережевий модуль необхідно в головному вікні (рисунок 3.13) натиснути кнопку «Нейромережевий модуль». Відбудеться перехід до нейромережевого модуля, інтерфейс якого зображено на рисунку 3.18.

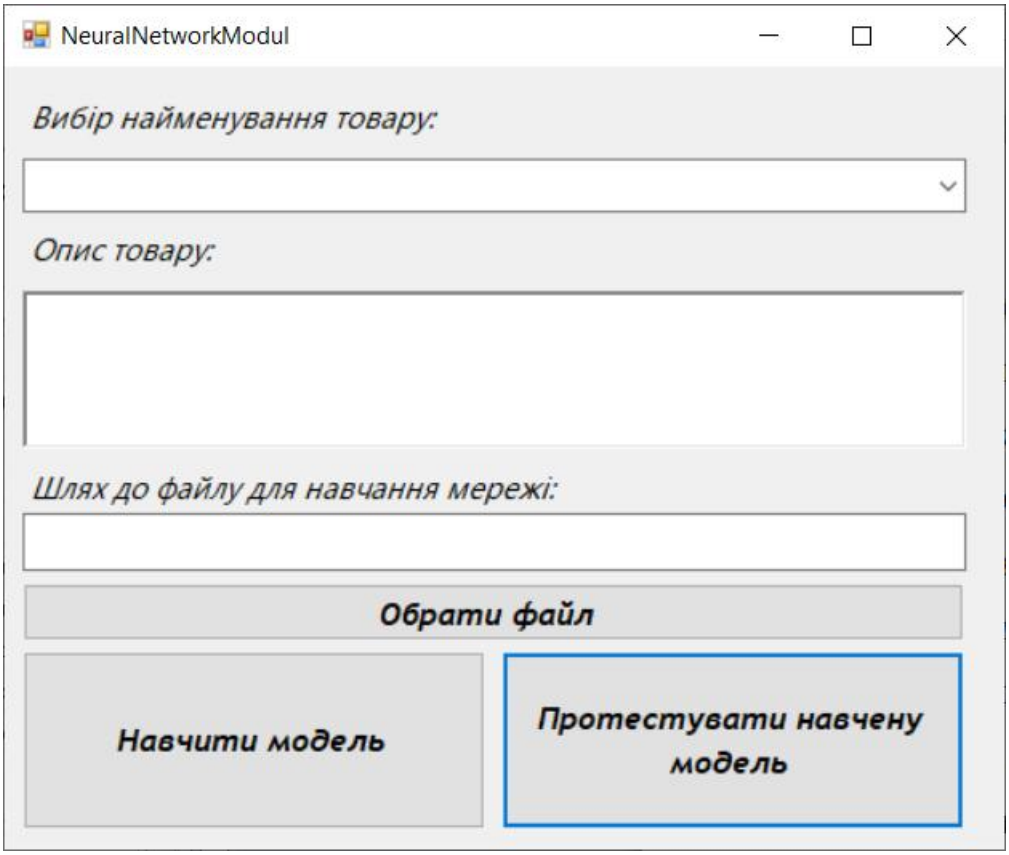


Рисунок 3.18 – Інтерфейс нейромережевого модуля

Для реалізації навчання нейромережі спершу потрібно обрати товар, для якого потрібно навчити модель. Після вибору товару одразу автоматично користувач побачить його опис (рисунок 3.19).

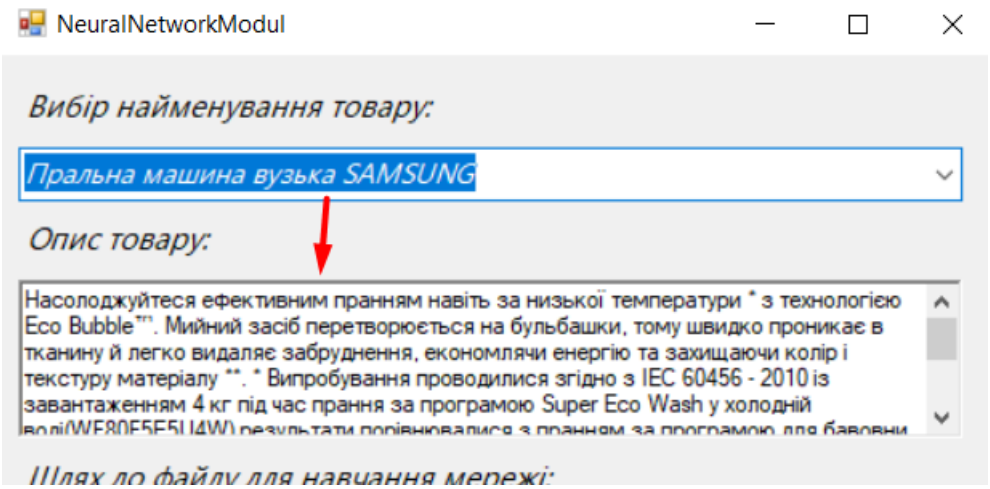


Рисунок 3.19 – Виведення опису за обраним товаром

Далі для навчання необхідно обрати розмічений файл. Це можна зробити, увівши шлях до нього, або ж натиснувши кнопку «Обрати файл». Після цього можна починати процес навчання класифікатора. Після завершення навчання користувач побачить повідомлення про успішне завершення навчання (рисунок 3.20)

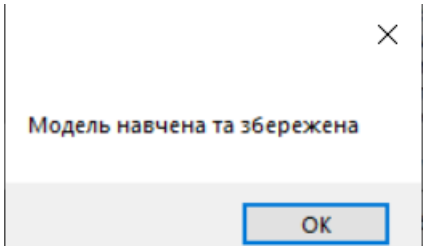


Рисунок 3.20 – Повідомлення про завершення навчання

Наступним повідомленням буде виведення алгоритму навчання, що виявився найкращим для даної задачі. Приклад повідомлення про обраний автоматизовано алгоритм навчання зображено на рисунку 3.21.

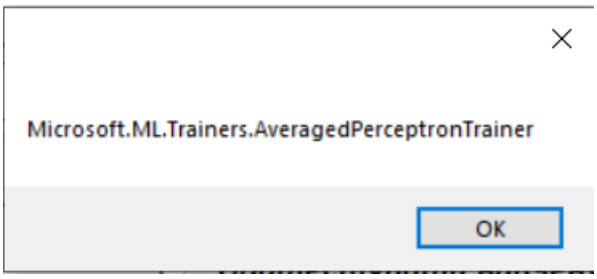


Рисунок 3.21 – Виведення використаного алгоритму-переможця

Після завершення процесу навчання користувач може переглянути статистику за метриками по навченій збереженій моделі. Для перегляду даних На рисунку 3.22 зображено дані по метрикам навченої моделі.

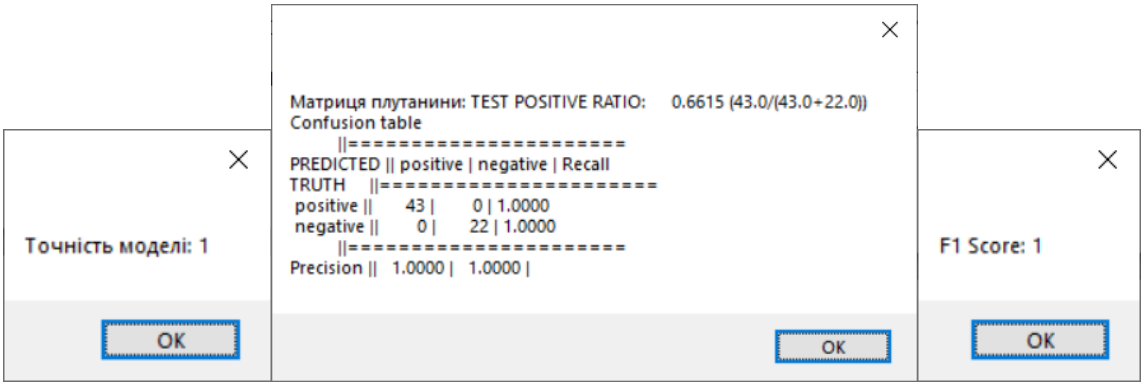


Рисунок 3.22 – Дані метрик по навченій моделі

Для переходу на модуль визначення семантично невідповідних україномовних коментарів потрібно з стратового екрану натиснути кнопку «Модуль визначення рівня відповідності відгуків до описів товару». Стартовий екран модуля проілюстровано на рисунку 3.23.

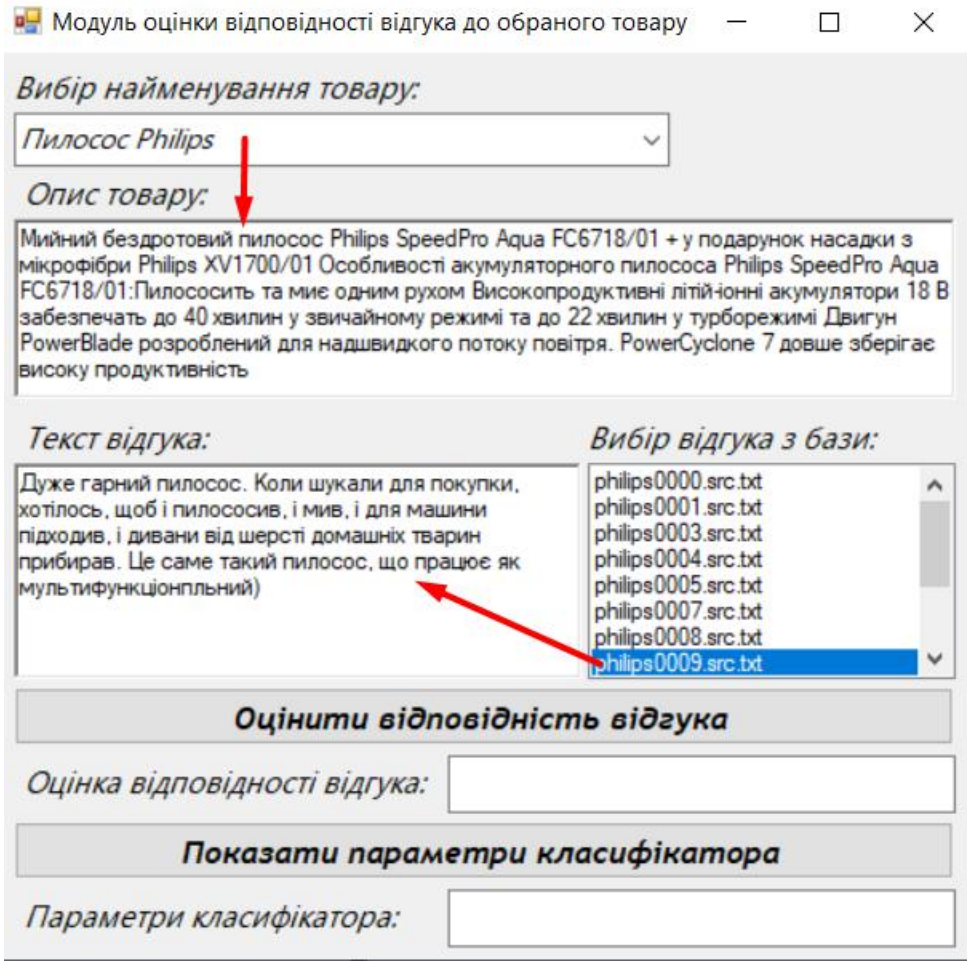


Рисунок 3.23 – Стартовий екран модуля оцінки відповідності відгука

Обравши з випадаючого списку найменування товару товар, до якого буде виведено опис, можна далі увести в поле для відгука текст, який потрібно оцінити, або ж обрати існуючі тексти з бази. На рисунку 3.24 показано здійснену оцінку відгука, якого немає у базі, та який не належить обраному товару.

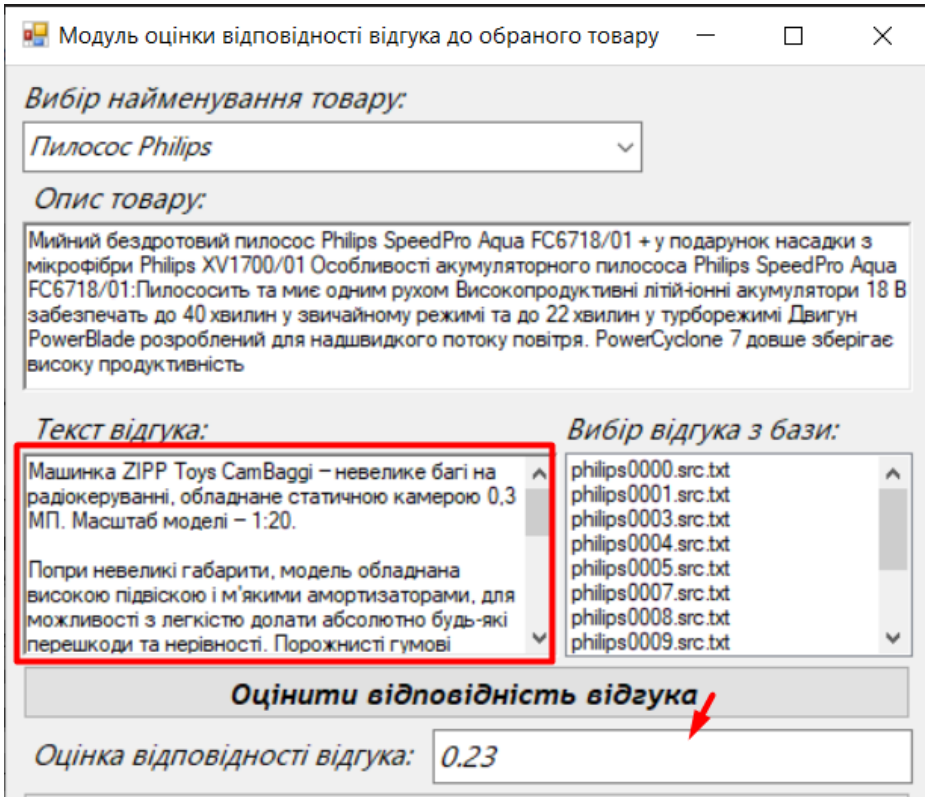


Рисунок 3.24 – Оцінка невідповідного відгука до обраного товару

Отже, таким чином було створено інструкцію користувача для інформаційної системи визначення семантичної невідповідності україномовних коментарів до опису товарів інтернет-магазинів з метою аналізу купівельного попиту.

3.5 Дослідження ефективності способу для визначення семантичної невідповідності україномовних коментарів до описів товарів інтернет-магазину

Для дослідження ефективності способу для визначення семантичної невідповідності україномовних коментарів до описів товарів інтернет-магазину

було проведено дослід на коментарях до товару, яких не було у навчальній та тестових вибірках.

Дослідження проводилось на 4-х товарах, до кожного з яких було по 25 відгуків. Дані по ідентифікації зразків, що були відібрані експертами наведено в таблиці 3.6.

Таблиця 3.6 – Дані по ідентифікації відгуків

Товар	**Коректно ідентифіковані відгуки**	**Некоректно ідентифіковані відгуки**	**% коректно-ідентифікованих відгуків**
Мийний бездротовий пилосос Philips SpeedPro Aqua	24	1	96
Вузька пральна машина Samsung	24	1	96
Кавомашина KRUPS Evidence Eco-Design	23	2	92
Мікрохвильова піч BOSCH FFL020MW0	25	0	100

Рузультат експерименту також наведено у вигляді діаграми (рисунок 3.25).

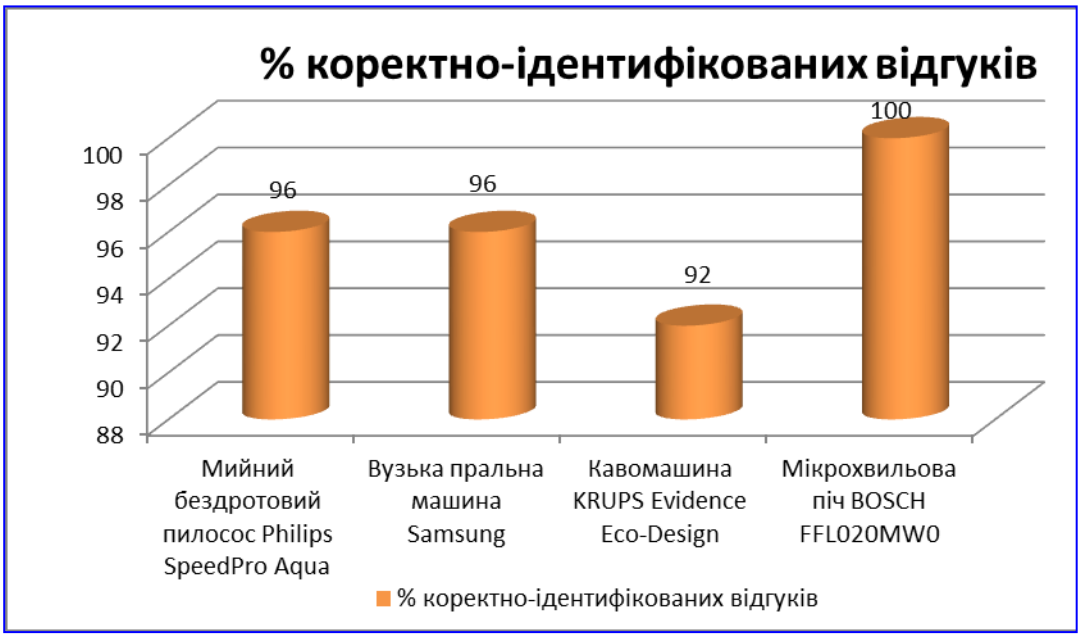


Рисунок 3.25 – Діаграма відсотків коректно-ідентифікованих відгуків товарів інтернет-магазину

Як видно з таблиці 3.6 та рисунка 3.25 – найгірше ідентифікувались відгуки до кавомашини «KRUPS Evidence Eco-Design», а найкраще – до мікрохвильової печі «BOSCH FFL020MW0».

Отже, було проведено дослідження ефективності способу для визначення семантичної невідповідності україномовних коментарів до описів товарів інтернет-магазину.

Висновки

Метою кваліфікаційної роботи бакалавра було розробити спосіб для визначення семантичної невідповідності україномовних коментарів до описів товарів інтернет-магазину для задач аналізу купівельного попиту.

У рамках досягнення поставленої мети були поставлені та виконані такі задачі:

– розроблено спосіб для визначення семантичної невідповідності україномовних коментарів до описів товарів інтернет-магазину;

– створено набір даних, що використовується для описів товарів у інтернет-магазинах українською мовою за допомогою використання технології вебскрапінгу;

– обрано архітектуру бінарного класифікатора для визначення семантичної невідповідності україномовних коментарів до описів товарів інтернет-магазину;

– спроєктовано структуру застосунку для визначення семантично невідповідних україномовних коментарів до описів товарів інтернет-магазину й структуру відповідної БД;

– розроблено інформаційну систему визначення семантичної невідповідності україномовних коментарів;

– проведено тестування створеного програмного забезпечення, яке показало, що всі функції працюють згідно заявленого функціоналу;

– проведено дослідження ефективності створеного способу для визначення семантичної невідповідності україномовних коментарів до описів товарів інтернет-магазину.

Основним результатом виконання кваліфікаційної роботи бакалавра є розроблений спосіб для визначення семантично невідповідних україномовних коментарів до описів товарів інтернет-магазину для задач аналізу купівельного попиту, що призначений для перетворення вхідних даних у формі текстів-коментарів для побудови україномовного датасету, що використовується для

опису товарів у інтернет-магазинах для навчання нейромережевого класифікатора та тестових текстів-коментарів, невідповідність до опису товару яких потрібно визначити, у кінцеві дані у вигляді числової оцінки відповідності коментаря до обраного товару. Також в результаті роботи було створено відповідне програмне забезпечення для програмного пошуку невідповідностей серед існуючих користувацьких коментарів, що дасть можливість виконання аналізу купівельного попиту. Відповідно, всі поставлені завдання кваліфікаційної роботи бакалавра було виконано.

Перелік посилань

1. Natural Language Processing (NLP). URL: https://www.sas.com/en_us/insights/analytics/what-is-natural-language-processing-nlp.html

2. What is natural language processing? URL: https://www.ibm.com/topics/natural-language-processing

3. Методи та засоби семантичного аналізу текстів. URL: https://ela.kpi.ua/bitstream/123456789/39855/1/Myhal_magistr.pdf

4. Sentiment Analysis Main Tasks and Applications: A Survey. URL: http://xml.jips-k.org/full-text/view?doi=10.3745/JIPS.04.0120

5. What Semantic Analysis Means to Natural Language Processing. URL: https://www.expert.ai/blog/natural-language-process-semantic-analysis-definition/

6. Keyword Extraction Methods from Documents in NLP. URL: https://www.analyticsvidhya.com/blog/2022/03/keyword-extraction-methods-from-documents-in-nlp

7. Глобальна статистика, модель TF*IDF. URL: https://studbooks.net/2056135/informatika/globalnaya_statistika_model_tfidf

8. Розв’язання задачі класифікації текстів методами обробки природньої мови та машинного навчання. URL: https://naukajournal.org/index.php/naukajournal/article/download/1817/1867

9. Multinomial Naive Bayes Explained: Function, Advantages & Disadvantages, Applications in 2023. URL: https://www.upgrad.com/blog/multinomial-naive-bayes-explained/

10. Support Vector Machines (SVM). URL: https://studfile.net/preview/16440083/page:32/

11. Електронна комерція-2022: на що варто звернути увагу українським компаніям. URL: https://interfax.com.ua/news/blog/793857.html

12. Український e-commerce під час війни – дослідження. URL: https://ain.ua/2022/07/01/ukrayinskyj-e-commerce-pid-chas-vijny/

13. Як інтернет-магазину працювати з відгуками: покроковий алгоритм. URL: https://fulfillmentmtp.com.ua/ua/blog/tpost/1z05mldbb1-yak-nternet-magazinu-pratsyuvati-z-vdguk

14. Guided Labeling for Document Classification. URL: https://forum.knime.com/t/guided-labeling-for-document-classification/17277

15. What is custom text classification? URL: https://learn.microsoft.com/en-us/azure/cognitive-services/language-service/custom-text-classification/overview

16. Binary Classification Metrics. URL: https://rubikscode.net/2021/04/12/machine-learning-with-ml-net-evaluation-metrics/

17. Information system for converting audio in Ukrainian language into its textual representation using nlp methods and machine learning. URL: https://www.researchgate.net/publication/368557538_Information_system_for_converting_audio_in_Ukrainian_language_into_its_textual_representation_using_nlp_methods_and_machine_learning

18. Контент-аналіз україномовних текстів, написаних природньою мовою. URL: http://ekmair.ukma.edu.ua/bitstream/handle/123456789/22442/Brus_Bakalavrska_robota.pdf?sequence=1&isAllowed=y

19. Основні відомості про бази даних. URL: https://support.microsoft.com/uk-ua/office/основні-відомості-про-бази-даних-a849ac16-07c7-4a31-9948-3c8c94a7c204

20. 20 Popular Machine Learning Metrics. Part 1: Classification & Regression Evaluation Metrics. URL: https://towardsdatascience.com/20-popular-machine-learning-metrics-part-1-classification-regression-evaluation-metrics-1ca3e282a2ce

21. What is a Confusion Matrix in Machine Learning? URL: https://www.simplilearn.com/tutorials/machine-learning-tutorial/confusion-matrix-machine-learning

22. Confusion Matrix. URL: https://devopedia.org/confusion-matrix

23. Data Labeling: The Authoritative Guide. URL: https://scale.com/guides/data-labeling-annotation-guide#high-quality-data-annotations

24. Training and Building Machine Learning Models. URL: https://scale.com/guides/model-training-building

25. Перегляд точності та продуктивності прогнозних скорингових моделей. URL: https://learn.microsoft.com/uk-ua/dynamics365/sales/scoring-model-accuracy

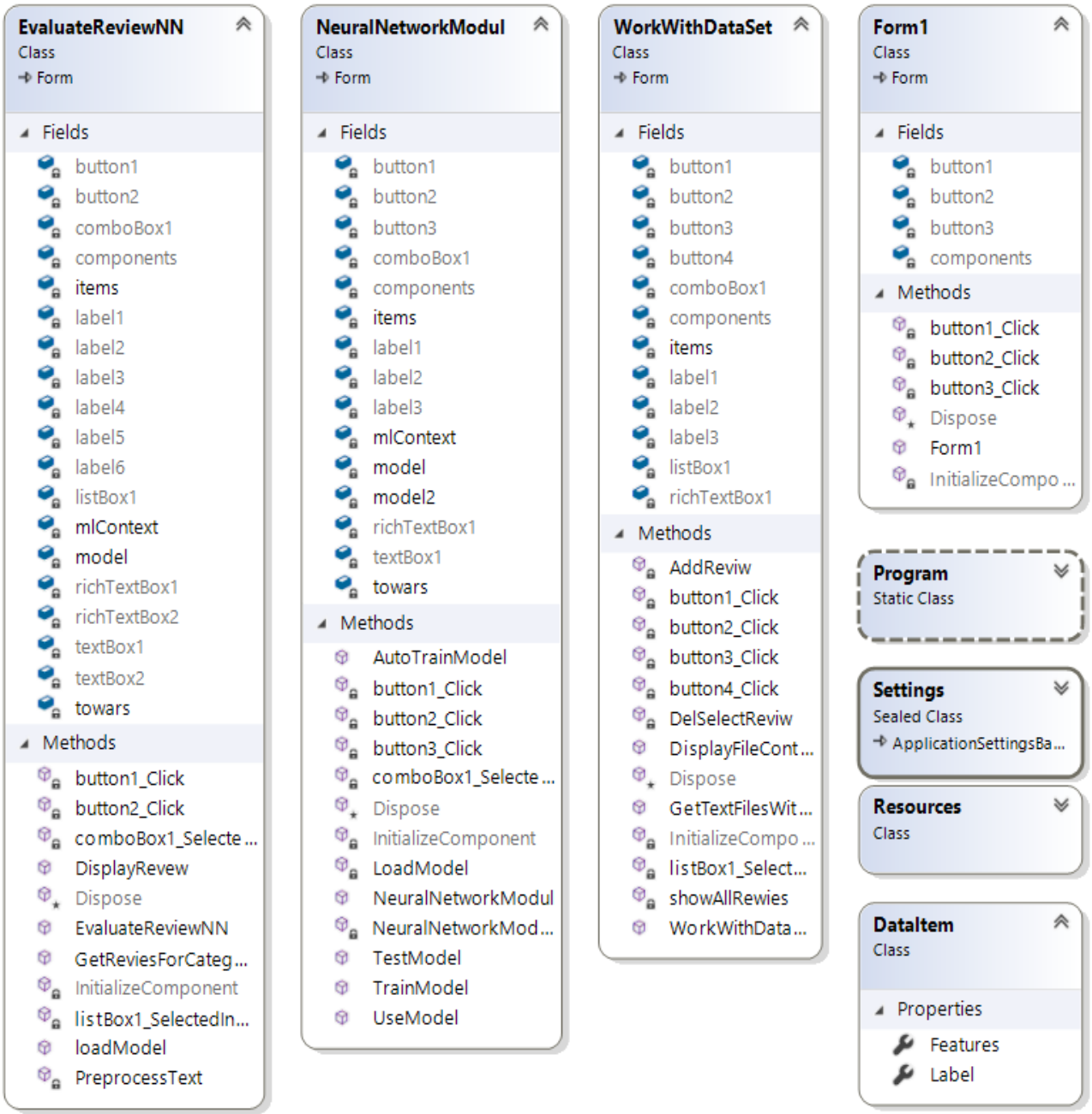
ДОДАТКИ

Додаток А

Схема способу визначення семантичної невідповідності україномовних коментарів до описів товарів за допомогою нейронної мережі

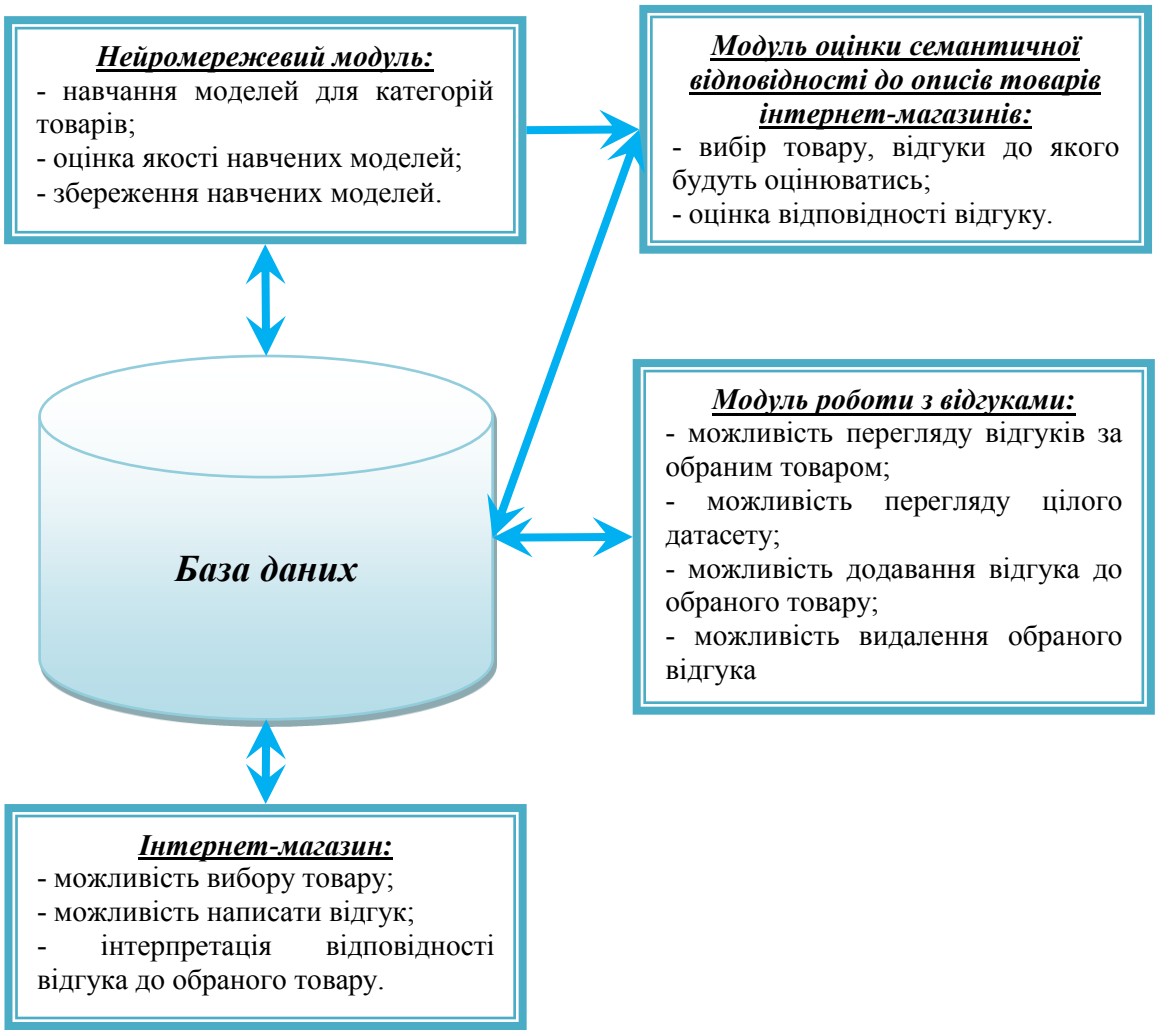
Додаток Б

Діаграма класів інформаційної системи визначення семантичної невідповідності україномовних коментарів до описів товарів інтернет-магазину



Додаток В

Схема структури застосування для визначення семантично невідповідних україномовних коментарів до описів товарів інтернет-магазину



Додаток Г

Презентаційний матеріал

КВАЛІФІКАЦІЙНА РОБОТА БАКАЛАВРА

СПОСІБ ДЛЯ ВИЗНАЧЕННЯ СЕМАНТИЧНО НЕВІДПОВІДНИХ УКРАЇНОМОВНИХ КОМЕНТАРІВ ДО ОПИСІВ ТОВАРІВ ІНТЕРНЕТ-МАГАЗИНУ ДЛЯ ЗАДАЧ АНАЛІЗУ КУПІВЕЛЬНОГО ПОПИТУ

Виконав:
студент 3 курсу, групи КНс-20-1
Семенишен Андрій Леонідович
Керівник:
викладач кафедри КН
Молчанова Марина Олексіївна

Актуальність

Інтелектуальний аналіз тексту, як один із напрямів ШІ є одним із значимих напрямів сучасних досліджень. Починаючи від категоризації текстів, пошуку інформації, обробки змін у колекціях текстів, до розробки засобів представлення інформації для поточних користувачів – все це є задачами Інтелектуального аналізу тексту.

Переважною більшістю інформації з веб-джерел є звичайний текст, який або погано структурований, або зовсім неструктурований. Тому, отримання значущої інформації є одним із провідних завдань інтелектуального аналізу тексту, що є процесом одержання якісної інформації з частково-структурованих та неструктурованих даних.

Окрім проблеми неструктурованих даних, є ще проблема семантичної не відповідності україномовних коментарів до описів товарів інтернет-магазину, виявлення яких може допомогти у проведенні аналізу купівельного попиту.

Об’єкт та предмет дослідження

Об’єкт дослідження – процес семантичного аналізу україномовних коментарів до текстових описів товарів інтернет-магазину.

Предмет дослідження – моделі, методи, алгоритми та засоби для виявлення семантичної невідповідності україномовних коментарів до описів товарів інтернет-магазину.

Мета і задачі роботи

Метою кваліфікаційної роботи бакалавра є розробка способу для визначення семантичної невідповідності україномовних коментарів до описів товарів інтернет-магазину для задач аналізу купівельного попиту.

У рамках досягнення поставленої мети ставляться наступні задачі:

- розробити спосіб для визначення семантичної невідповідності україномовних коментарів до описів товарів інтернет-магазину;
- створити набір даних, що використовується для описів товарів у інтернет-магазинах українською мовою;
- підібрати архітектуру нейронної мережі для визначення семантичної невідповідності україномовних коментарів до описів товарів інтернет-магазину;
- спроєктувати структуру застосунку для визначення семантично невідповідних україномовних коментарів до описів товарів інтернет-магазину й структуру відповідної БД;
- розробити інформаційну систему визначення семантичної невідповідності україномовних коментарів;
- провести тестування створеного програмного забезпечення;
- провести дослідження ефективності запропонованого способу.

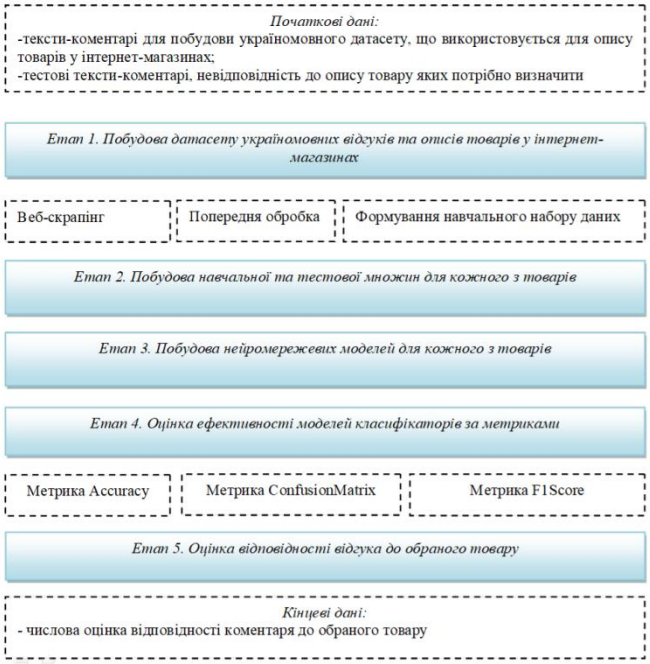


Схема способу визначення семантичної невідповідності україномовних коментарів до описів товарів за допомогою машинного навчання

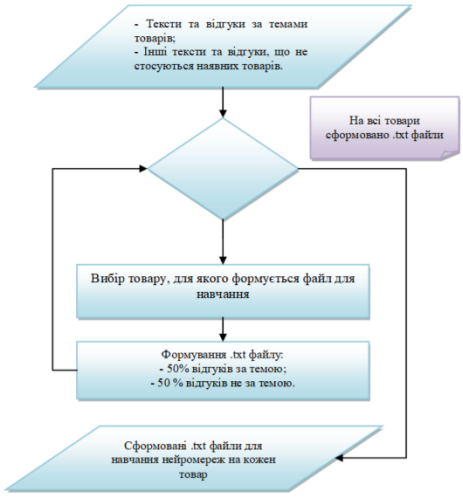


Схема формування файлів навчання нейромережі

Приклад сформованого розміченого файлу

Купили пилосос, виглядає дуже якісно, всі деталі зроблені якісно, додатково йде у комплектації набор з ганчірок мікрофібра, який потрібен для вологого прибирання, заряджався пилосос орієнтовано 2-3 години, так як коли він приїхав у відділення, був заряджений на половину.Прибирали, дуже сподобався у користуванні, компактний, легкий. 1
Отримала пилосос. Зробила перше прибирання. Заряда акумулятора вистачило. Бездротовий пилосос це зручно. 1
Дуже гарний пилосос. Коли шукали для покупки, хотілось, щоб і пилососив, і мив, і для машини підходив, і дивани від шерсті домашніх тварин прибирав. Це саме такий пилосос, що працює як мультифункціонпльний) 1
Пилососом задоволений , всі функції працюють на відмінно , гарно пилососить та миє підлогу , швидко заряджається , деталі зроблені якісно та виглядають не дешево! Рекомендую ! 1
Для тих, хто обмежує споживання цукру це знахідка! Чудовий склад! Прекрасний смак! Гарна заміна солодощам і цукеркам, саме для тих, хто дбає за своє здоро'вя і фігуру. 0
Це дуже смачно , естетично красиво , і корисно , брала як заміну звичайному шоколаду з цукром . Склад чудовий , чудовое поєднання чорнлго шоколаду морської солі та кусочки мигдалю . Сподобається тим хто любить гірський шоколад . 0
Смачно але дорого. Купую коли є знижки, в наших магазинах нічого подібного по смаку з схожим складом я не знайшла. 0
Порошок пере дуже добре. Замовлення прийшло швидко.Рекомендую! 0
Улюблений порошок. Не залишає слідів, не пахнуть речі 0

Схема підходу до підбору класифікатора

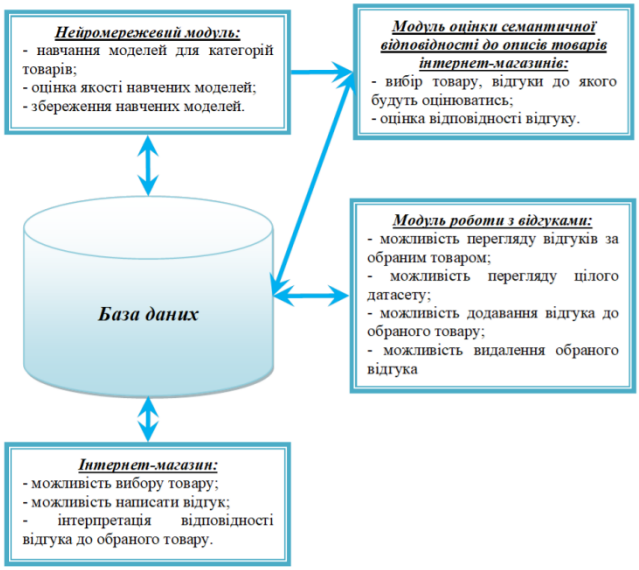


Схема структури застосування для визначення семантично невідповідних україномовних коментарів до описів товарів інтернет-магазину

Засоби розробки

Для дослідницької частини:

Фреймворк: ***.NET***

Мова програмування: ***C#***

*Інтерфейс користувача: «**WindowsForms**»*

Для побудови нейромережевої моделі, навчання та оцінки її продуктивності:

- *«**Microsoft.ML**»*, метрики для оцінювання: «***Acuracy***», оцінка «***F1***», та Метрика «***ConfusionMatrix***»

*Для реалізації бази даних: «**MSSQLServer**»*

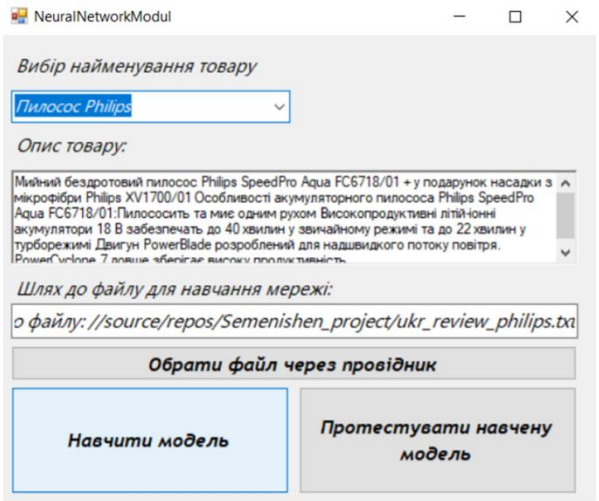
*Для реалізації інтернет-магазину: платформа «**ASP.NET**», архітектурний шаблон «**MVC 5**», фронтенд частина «**HTML, CSS, Bootstrap**»*

Програмна реалізація способу визначення семантичної невідповідності україномовних коментарів до описів товарів

Діаграма класів

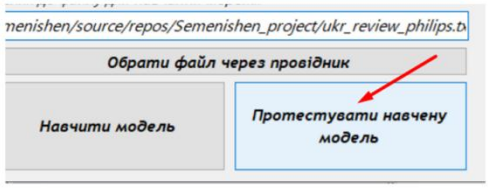
Програмна реалізація способу визначення семантичної невідповідності україномовних коментарів до описів товарів

Нейромережевий модуль

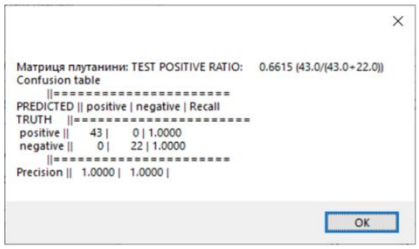


Програмна реалізація способу визначення семантичної невідповідності україномовних коментарів до описів товарів

Нейромережевий модуль

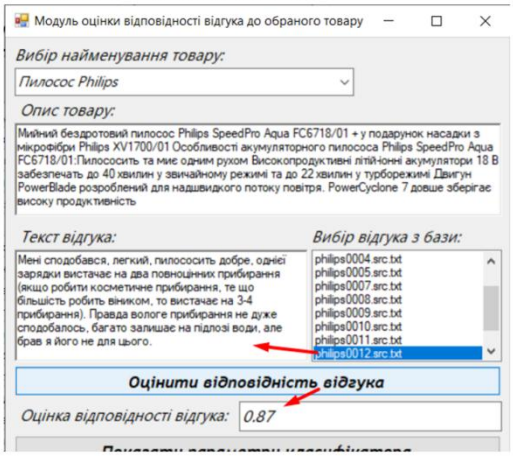


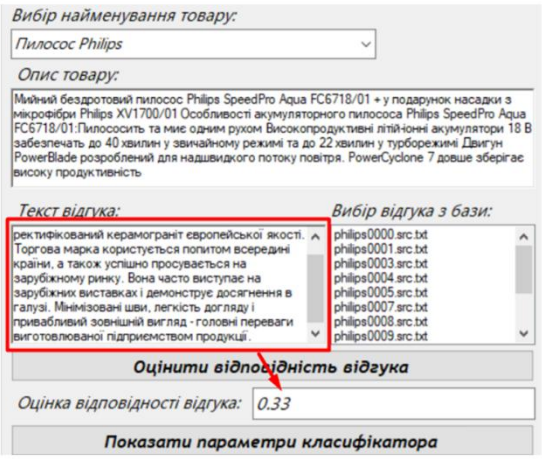
Тестування навченої моделі



Програмна реалізація способу визначення семантичної невідповідності україномовних коментарів до описів товарів

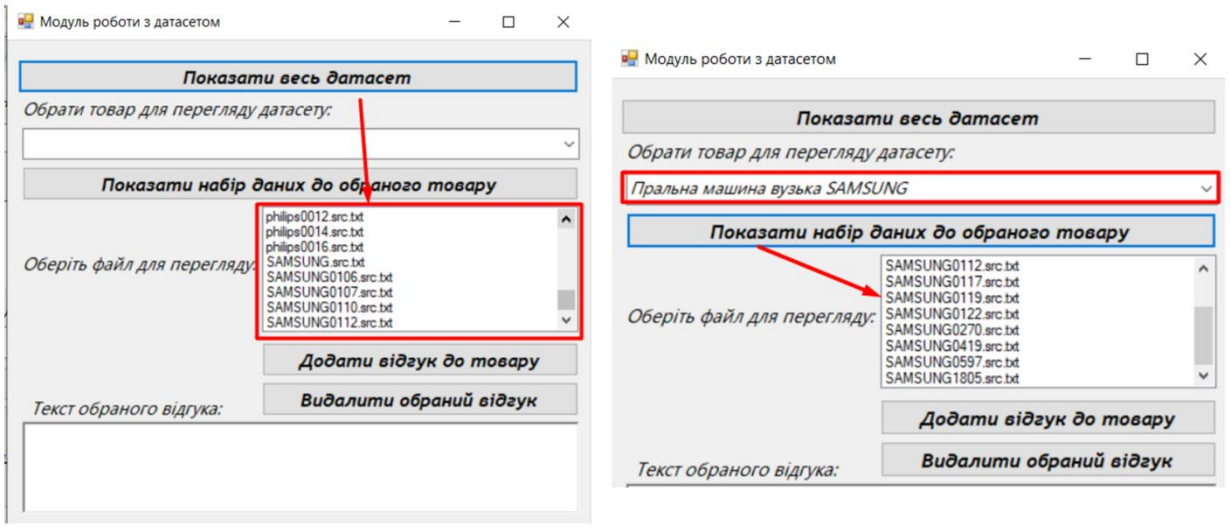
Модуль оцінки відповідності відгука до обраного товару





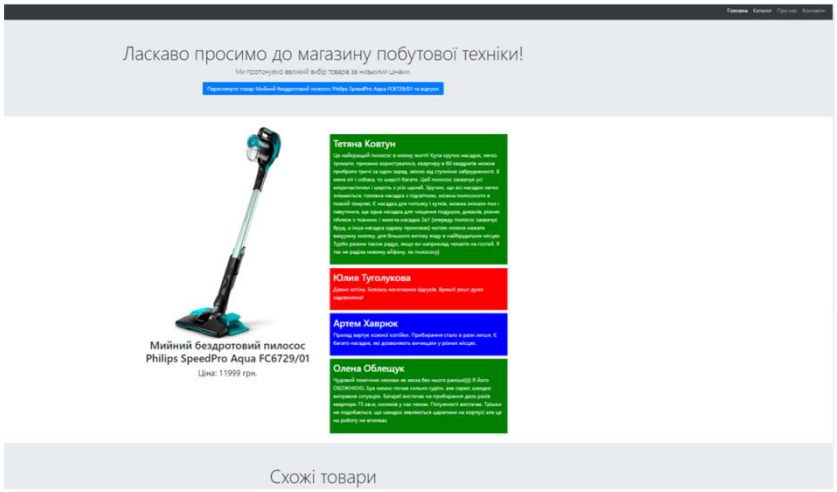
Програмна реалізація способу визначення семантичної невідповідності україномовних коментарів до описів товарів

Модуль роботи з датасетом



Програмна реалізація способу визначення семантичної невідповідності україномовних коментарів до описів товарів

Модуль інтернет-магазину



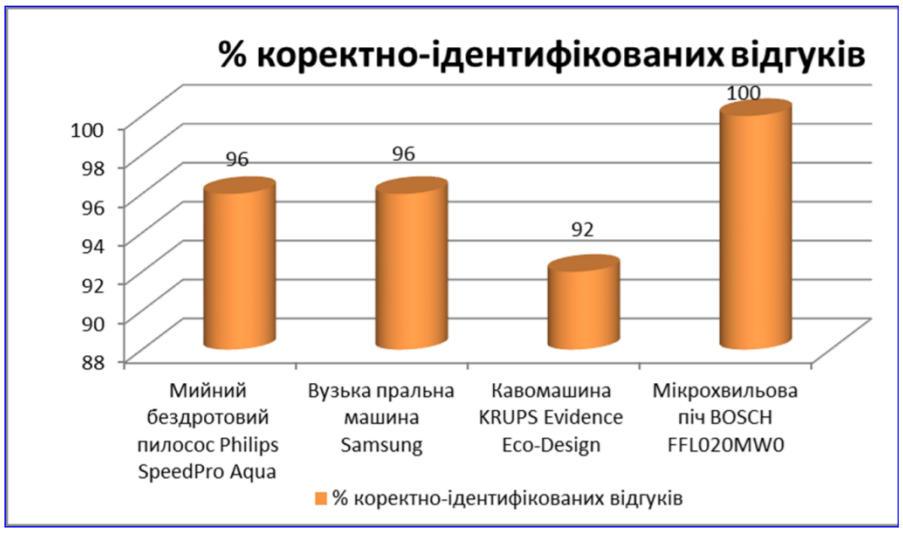
Дослідження ефективності розробленого способу

Таблиця коректно-ідентифікованих відгуків товарів інтернет-магазину

Товар	Коректно ідентифіковані відгуки	Некоректно ідентифіковані відгуки	% коректно-ідентифікованих відгуків
Мийний бездротовий пилосос Philips SpeedPro Aqua	24	1	96
Вузька пральна машина Samsung	24	1	96
Кавомашина KRUPS Evidence Eco-Design	23	2	92
Мікрохвильова піч BOSCH FFL020MW0	25	0	100

Дослідження ефективності розробленого способу

Діаграма відсотків коректно-ідентифікованих відгуків товарів інтернет-магазину



Висновки

Результатом виконання кваліфікаційної роботи бакалавра є розроблений спосіб для визначення семантичної невідповідності україномовних коментарів до описів товарів інтернет-магазину для задач аналізу купівельного попиту.

У рамках досягнення поставленої мети були поставлені та виконані такі задачі:

- розроблено спосіб для визначення семантичної невідповідності україномовних коментарів до описів товарів інтернет-магазину;
- створено набір даних, що використовується для описів товарів у інтернет-магазинах українською мовою за допомогою використання технології вебскрапінгу;
- обрано архітектуру бінарного класифікатора для визначення семантичної невідповідності україномовних коментарів до описів товарів інтернет-магазину;
- спроєктовано структуру застосунку для визначення семантично невідповідних україномовних коментарів до описів товарів інтернет-магазину й структуру відповідної БД;
- розроблено інформаційну систему визначення семантичної невідповідності україномовних коментарів;
- проведено тестування створеного програмного забезпечення, яке показало, що всі функції працюють згідно заявленого функціоналу;
- проведено дослідження ефективності створеного способу для визначення семантичної невідповідності україномовних коментарів до описів товарів інтернет-магазину.

Отже, завдання кваліфікаційної роботи бакалавра виконано у повному обсязі.

ДЯКУЮ ЗА УВАГУ!

UNICHECK

Ім'я користувача:
Кафедра КН

ID перевірки:
1015421728

Дата перевірки:
05.06.2023 09:31:00 EEST

Тип перевірки:
Doc vs Internet + Library

Дата звіту:
05.06.2023 09:47:01 EEST

ID користувача:
100005671

Назва документа: **КНс-20-1 Семенишен**

Кількість сторінок: **66** Кількість слів: **10060** Кількість символів: **77127** Розмір файлу: **2.68 MB** ID файлу: **1015084188**

Виявлено модифікації тексту (можуть впливати на відсоток схожості)

9.46%
Схожість

Найбільша схожість: **3.84%** з джерелом з Бібліотеки (ID файлу: **1015079726**)

8.34% Джерела з Інтернету	381	Сторінка 68
5.89% Джерела з Бібліотеки	79	Сторінка 70

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	21 сторінка

Mon Jun 05 08:44:12 EEST 2023, Петровський Сергій Степанович, Хмельницький національний університет, ХНУ

Anti-Plagiarism v-15.257

Максимальне співпадіння з одним документом 37.0%

Словники перевірки: en_US, ru_RU, ua_UA. **Помилок в документах: 10%**

	Документ		Сумарний збіг по Базі Даних	
ID: 114684 Назва: КВАЛІФІКАЦІЙНА РОБОТА БАКАЛАВРА Додано в БД: 2023-06-05 Автора: А.Л. Семенишен Керівники: М.О. Молчанова Консультанти: Опоненти:	Символи	Лексеми	Символи	Лексеми
	59776	918	23935 (40%)	377 (41%)

Джерело плагіату

ID	Опис	Наявність плагіату в документі	
		Символи	Лексеми
114082	Назва: ЗВІТ з професійної практики Додано в БД: 2023-05-26 Автора: Семенишен А.Л Керівники: Скрипник Т.К. Консультанти: Опоненти:	22147 (37.0%)	343 (37.0%)

РІШЕННЯ ЕКСПЕРНОЇ 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: *Спосіб для визначення семантично невідповідних україномовних коментарів до описів товарів інтернет-магазину для задач аналізу купівельного попиту*

Автор: *студент групи КНс-20-1 Семенишен Андрій Леонідович*

Спеціальність: *122 – Комп'ютерні науки*

Освітня програма: *освітньо-професійна*

Науковий керівник: *викладач Молчанова М.О.*

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 Семенишена А.Л., не є плагіатом, оскільки: запозичення розміщені в розділі огляду існуючих підходів, не описують безпосередньо авторську роботу і не стосуються її результатів; усі запозичення фрагментарні; серед запозичень знаходяться загальновідомі терміни, скорочення та матеріали статей.

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

- за системою Anti-Plagiarism: 37%, з яких 37% є посиланням на власний звіт з професійної практики, що є допустимими запозиченнями які відносяться до описаних вище;

- за системою Unicheck: 9.46 %,

Керівник роботи _____ *Марина МОЛЧАНОВА*

Гарант ОП _____ *Олександр МАЗУРЕЦЬ*

Завідувач кафедри КН _____ *Олександр БАРМАК*

ХМЕЛЬНИЦЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ
МОН УКРАЇНИ

Кафедра комп'ютерних наук

ВІДГУК НАУКОВОГО КЕРІВНИКА

на кваліфікаційну роботу бакалавра

студента *гр. КНс-20-1 Семенишен Андрій Леонідович*

за темою: *Спосіб для визначення семантично невідповідних україномовних коментарів до описів товарів інтернет-магазину для задач аналізу купівельного попиту*

1. Актуальність теми

Розробка способу для визначення семантично невідповідних україномовних коментарів до описів товарів інтернет-магазину для задач аналізу купівельного попиту є актуальною, оскільки спостерігається збільшення кількості інтернет-магазинів та зростання онлайн-купівель, а аналіз покупок та відгуків користувачів є важливим інструментом для сучасного ведення бізнесу.

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

Кваліфікаційна робота бакалавра повністю відповідає предметній області Стандарту спеціальності 122 Комп'ютерні науки. Вона включає розробку програмного забезпечення, застосування алгоритмів навчання класифікаторів та засобів і методів комп'ютерних наук для аналізу невідповідних україномовних коментарів до описів товарів інтернет-магазину

3. Професійні та особистісні якості бакалавра

У процесі виконання кваліфікаційної роботи бакалавра проявилися високі професійні та особистісні якості. Бакалавр виявив достатні знання в галузі комп'ютерних наук, а також продемонстрував вміння працювати з складними даними та алгоритмами.

4. Ступінь самостійності під час виконання кваліфікаційної роботи

Бакалавр продемонстрував достатній рівень самостійності під час виконання кваліфікаційної роботи. В тому числі, він самостійно досліджував та розробляв алгоритми для визначення семантично невідповідних україномовних коментарів та аналізував результати дослідження.

5. Ступінь оволодіння методами дослідження

Бакалавр успішно оволодів методами дослідження в рамках своєї кваліфікаційної роботи. Він провів аналіз наявних даних, розробив експериментальну методологію та використовував статистичні методи для оцінки результатів та виведення висновків.

6. Повнота та якість розкриття теми роботи

Тема роботи була повноцінно розкрита з використанням наукових джерел, методів дослідження та практичних експериментів. Бакалавр детально описав методи для визначення семантично невідповідних україномовних коментарів, розробив та реалізував програмний засіб, провів експерименти та аналіз результатів.

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

У роботі викладено матеріал логічно та послідовно. Аргументи були чіткими та обґрунтованими. Текст має науковий стиль та літературну грамотність, що забезпечує зрозумілість та доступність матеріалу.

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

Розроблений спосіб для визначення семантично невідповідних україномовних коментарів до описів товарів інтернет-магазину для задач аналізу купівельного попиту може мати практичне застосування у сфері онлайн-торгівлі. Його можна використовувати для швидкого та ефективного виявлення невідповідних україномовних коментарів до описів товарів інтернет-магазину.

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

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

Керівник _____ *викладач кафедри КН Марина МОЛЧАНОВА*

ХМЕЛЬНИЦЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ
МОН УКРАЇНИ

Кафедра комп'ютерних наук

РЕЦЕНЗІЯ

на кваліфікаційну роботу бакалавра

студента *гр. КНс-20-1 Семенишена Андрія Леонідовича*

за темою: *Спосіб для визначення семантично невідповідних україномовних коментарів до описів товарів інтернет-магазину для задач аналізу купівельного попиту*

1. Актуальність обраної теми

Аналіз семантично невідповідних коментарів дозволить інтернет-магазинам виявити недоліки в описах товарів і вжити заходи для поліпшення якості обслуговування, що може позитивно вплинути на задоволеність покупців. Тому даний напрямок досліджень є беззаперечно актуальним.

2. Повнота розкриття мети та завдань роботи

У даній кваліфікаційній роботі бакалавра мета та завдання були повністю розкриті. Автор детально описав процес розробки та програмної реалізації способу для визначення семантично невідповідних україномовних коментарів до описів товарів інтернет-магазину для задач аналізу купівельного попиту. Достатній рівень деталізації та науковий підхід дозволяють чітко розуміти суть та значення даної роботи.

3. Зміст кожного розділу роботи

Кожен розділ роботи бакалавра вміщує в собі інформацію, що безпосередньо стосується теми, від теоретичних аспектів аналізу купівельного попиту до практичної реалізації способу та опису інформаційної системи. Чітка структура та зв'язок між розділами допомагають зрозуміти послідовність дослідження та реалізації проєкту.

4. Оцінка розробленої інформаційної системи, її практична цінність

Розроблена інформаційна система для визначення семантично невідповідних україномовних коментарів до описів товарів інтернет-магазину має достатню оцінку потенціалу застосування та практичну цінність. Вона демонструє достатньо високу точність та швидкість виявлення невідповідних коментарів до описів товарів інтернет-магазину. Це може суттєво полегшити роботу інтернет-маркетологів та вживання заходів для поліпшення якості обслуговування, що може позитивно вплинути на задоволеність сервісом покупців.

5. Якість оформлення кваліфікаційної роботи бакалавра

Оформлення кваліфікаційної роботи бакалавра є на високому рівні. Вона включає всі необхідні розділи, таблиці, графіки та посилання на використану літературу. Чітка структура, належне використання наукового стилю та коректна організація тексту сприяють зрозумілому сприйняттю та оцінці роботи.

6. Недоліки кваліфікаційної роботи бакалавра

Хоча кваліфікаційна робота бакалавра демонструє достатній потенціал та практичне досягнення, можна відзначити деякі недоліки. Наприклад, не наведено обмежень на використання запропонованого способу. Розширення цих аспектів може збагатити практичну цінність роботи та підвищити її наукову цінність. Також недоліком є відсутність програмної інтеграції модуля оцінки відповідності україномовних коментарів до описів товарів з програмною системою інтернет-магазину.

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 «добре».

Рецензент Ямпенко О.М., доц. каф. ІПЗ