Хмельницький національний університет
Факультет інформаційних технологій
Кафедра комп'ютерних наук

КВАЛІФІКАЦІЙНА РОБОТА БАКАЛАВРА

на тему *Метод автоматизованої тематичної класифікації коротких текстових повідомлень*

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 4 курсу, група КН-18-1* Курс, група виконавця	Підпис	*О.В. Здоровик* Ініціали, прізвище
Керівник:	*старший викладач кафедри КН* Науковий ступінь, посада	Підпис	*Т.К. Скрипник* Ініціали, прізвище
Нормоконтроль:	*к.т.н., доцент кафедри КН* Науковий ступінь, посада	Підпис	*Р.О. Багрій* Ініціали, прізвище

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

16 червня 2022 р.

Хмельницький 2022

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

ЗАТВЕРДЖУЮ
Завідувач кафедри *комп'ютерних наук*

(підпис)
д.т.н., професор О.В. Бармак
«*25*» *березня* *2022* року

ЗАВДАННЯ
НА КВАЛІФІКАЦІЙНУ РОБОТУ БАКАЛАВРА

1. Тема кваліфікаційної роботи бакалавра: *«Метод автоматизованої тематичної класифікації коротких текстових повідомлень»*

2. Завдання видано студенту *Здоровику Олександру Васильовичу*
(прізвище, ім'я, по батькові)

3. Керівник роботи *старший викладач каф. КН Скрипник Тетяна Казимирівна*
(посада, прізвище, ім'я, по батькові)

4. Затверджено наказом університету від «*01*» *березня* *2022* р. № *18*

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

Мета роботи – розробка методу автоматизованої тематичної класифікації коротких текстових повідомлень й відповідної інформаційної системи. При тематичній класифікації слід враховувати параметри тексту. Потрібно забезпечити функції обробки інформації, внесення та перегляду даних системи, й одержання результатів роботи системи, таких як: визначення унікальних слів текстів класифікацій, визначення ключових слів текстів, визначення семантично важливих значень для кожного ключового слова класифікації, визначення класифікації для тестового тексту коротких текстових повідомлень.

Виконавець: *студент 4 курсу, група КН-18-1* (Курс, група виконавця) _____ (Підпис) *О.В. Здоровик* (Ініціали, прізвище)

Керівник: *старший викладач кафедри КН* (Науковий ступінь, посада) _____ (Підпис) *Т.К. Скрипник* (Ініціали, прізвище)

Анотація

Тема кваліфікаційної роботи бакалавра: *«Метод автоматизованої тематичної класифікації коротких текстових повідомлень»*

Виконавець кваліфікаційної роботи бакалавра: *студент групи КН-18-1 Здоровик Олександр Васильович*

Керівник кваліфікаційної роботи бакалавра: *старший викладач каф. КН Скрипник Тетяна Казимирівна*

Кваліфікаційна робота бакалавра містить:

Пояснювальна записка				Кількість додатків
Сторінок	Рисунків	Таблиць	Джерел інформації	
61	34	10	52	2

Метою кваліфікаційної роботи бакалавра є розробка інформаційної системи автоматизованої тематичної класифікації коротких текстових повідомлень. Для розробки інформаційної системи було використано мову програмування C#, а також систему керування базами даних MS SQL Server.

Розроблена система призначена для власників соціальних мереж, чатів, форумів та інших платформ на яких розташовані короткі текстові повідомлення. Реалізована автоматизація тематичної класифікації коротких текстових повідомлень дозволяє підвищити ефективність обробки інформації, визначення ключових слів, класифікацій та пришвидшити цей процес.

Напрямками практичного використання розробленої інформаційної системи визначено автоматизовану обробку інформації, визначення семантично важливих показників та класифікації коротких текстових повідомлень.

Ключові слова: короткі текстові повідомлення, класифікація, інформаційна система, ключові слова.

Виконавець: *студент 4 курсу, група КН-18-1* (Курс, група виконавця) _____ (Підпис) *О.В. Здоровик* (Ініціали, прізвище)

Зміст

Перелік скорочень 3
Вступ 4
Розділ 1 Характеристика предметної області: аналіз моделей, методів та реалізацій 6
1.1 Аналіз інформаційних моделей 6
1.2 Огляд теоретичних підходів до розв’язку подібних задач 12
1.3 Аналіз існуючих програмних рішень 17
1.4 Мета, задачі та вимоги до реалізації інформаційної системи 21
Розділ 2 Проєктування інформаційної системи 23
2.1 Аналіз та автоматизація обробки потоків даних 23
2.1.1 Метод автоматизованої тематичної класифікації коротких текстових повідомлень 23
2.1.2 Функціональна структура інформаційної системи 26
2.2 Інформаційна структура системи 28
2.2.1 Проектна архітектура системи та взаємозв’язок компонентів 28
2.2.2 Інформаційна модель 29
2.3 Вибір засобів розробки інформаційної системи 32
2.3.1 Вибір середовища програмування 33
2.3.2 Аналіз засобів створення програмного забезпечення 34
2.3.3 Вибір мови програмування 35
2.3.4 Вибір фреймворку 36
2.3.5 Вибір СКБД 37
Розділ 3 Програмна реалізація інформаційної системи 38
3.1 Структура та функціональне призначення програмних складових системи 38
3.2 Особливості реалізації програмних складових системи 40
3.3 Тестування інформаційної системи 42
3.4 Інструкція користувача 49
3.5 Вимоги до розгортання інформаційної системи 55
Висновки 56
Перелік посилань 58
Додатки

Перелік скорочень

Скорочення, термін, позначення	Пояснення
БД	База даних
ІС	Інформаційна система
ІТ	Інформаційні технології
КРБ	Кваліфікаційна робота бакалавра
КН	Комп’ютерні науки
ПЗ	Пояснювальна записка
ПП	Програмний продукт
СКБД	Система керування базами даних
ХНУ	Хмельницький національний університет.
DE	Disperce Evaluation
TF	Term Frequency
TF-IDF	Term Frequency – Inverse Document Frequency

Вступ

Сучасний світ базується на використанні інтернету та інформаційних ресурсів, які допомагають людству полегшити своє життя. Масштаби інформаційних потоків збільшуються, саме тому тексти почали класифікувати, відносити до категорій та певних класів.

Проблема обробки текстів, пошуку їх семантичних величин та значень залишається актуальною і досі. Вона ускладнюється ще й тим, що для якісної автоматизованої обробки інформації необхідно створити комплекс методів та алгоритмів, які працюючи разом будуть виконувати поставлену задачу.

Пошук інформації відбувається за ключовими словами, які є у кожній статті, на кожному сайті або ж на сторінках соціальних мереж, де ключові слова в основному називаються – хештегами. Саме хештеги (теги) дозволяють класифікувати невелику або ж навпаки більш масштабну інформацію.

Короткі текстові повідомлення – інформація, яку найчастіше передають у соціальних мережах, месенджерах або чатах. Класифікація таких повідомлень є більш важкою, в порівнянні з науковими статтями або ж сторінками сайтів чи науковими публікаціями. Ускладнення полягають у тому, що короткі текстові повідомлення мають меншу кількість символів, такий параметр тексту впливає на точність обробки тексту, коректність аналізу та визначення семантично важливих значень.

Для розв’язання завдання класифікації коротких текстових повідомлень застосовують різні методи автоматизованої класифікації. Серед таких методів використовують DE, TF, TF-IDF, BM25 та інші. Найбільш популярним залишається частотний метод TF-IDF.

Класифікація інформації, а зокрема, класифікації коротких текстових повідомлень допомагає пришвидшити пошук потрібних даних в інтернеті, фільтрувати повідомлення за класифікаціями, хештегами, визначати найбільш популярні повідомлення, що актуально у соціальних мережах таких як Twitter та

Instagram. Також, відкривається можливість сортування інформації за тегами, а не лише за датою додавання, та найбільшою кількість вподобань.

Мета кваліфікаційної роботи бакалавра – створення та прикладна програмна реалізація методу автоматизованої тематичної класифікації коротких текстових повідомлень.

Об'єкт дослідження – процес класифікації текстових повідомлень.

Предмет дослідження – інформаційні технології, моделі, методи й засоби для автоматизованої класифікації коротких текстових повідомлень.

Розділ 1 Характеристика предметної області: аналіз моделей, методів та реалізацій

1.1 Аналіз інформаційних моделей

На сьогоднішній день, інтернет є невід'ємною частиною життя більшості людей. Інтернет використовується у багатьох сферах, починаючи від пошуку інформації та купівлі різних товарів, закінчуючи звичайним спілкуванням у месенджерах та інших чатах, форумах [1].

Спілкування в інтернеті має свої особливості, зокрема:

– є можливість спілкуватись анонімно, при тому будь-хто може прочитати саме повідомлення та відповісти на нього свої, але автор буде невідомим;

– спілкування онлайн призводить до знецінення невербального спілкування;

– відсутність реальної картини про співрозмовника, спілкуючись онлайн співрозмовник може презентувати себе у яскравому образі, а зустрівшись в реальності розчарувати своїм виглядом або манерами;

– захоплення віртуальність, а саме створення віртуального світу, така особливість дозволяє користувачу забути про реальність, що негативно впливає на розвиток та плинність часу [2].

Інтернет надає людям повну свободу дій, тому спілкування в інтернеті стало однією з головних аспектів сучасного життя. Для зручності надсилання повідомлень було створено соціальні мережі. Саме у соціальних мережах користувачі можуть не лише обмінюватись повідомленнями, а і переглядати фото, відео та іншу інформацію про співрозмовника.

Соціальна мережа – це певна соціальна структура, за допомогою якої відбуваються різні соціальні взаємозв'язки, зокрема, спілкування, розміщення інформації чи зображень, створення коментарів та надсилання повідомлень [3]. Користувачами соціальних мереж можуть бути як окремі самостійні користувачі, так і організації чи фірми, що просувають свій бізнес.

Соціальні мережі можна класифікувати за різними напрямками, наприклад, за доступністю, за аудиторією, за призначенням та за платформою [4]. За доступністю соціальні мережі поділяються на відкриті, закриті та змішані. Більшість мереж є відкритими, що приваблює користувачів досить швидко та у великій кількості. Проте, існують і закриті соціальні мережі, вони базуються на спеціальній бізнес-моделі, яка не передбачає публічності. Також, існують і змішані мережі, які майже не користуються популярністю серед користувачів, через те, що вони мають певні перепони, що ускладнюють роботу з ними.

За аудиторією соціальні мережі можна поділити на широкі та нішеві (тематичні). Широкі соцмережі об’єднують усіх користувачів незалежно від їх походження чи інтересів. Всередині таких мереж можуть існувати і свої підтеми та групи. Так, до прикладу, соціальна мережа Facebook, яка використовується для спілкування між різними користувачами, включає в себе групи, спільноти та сторінки, які об'єднують людей по інтересах та вподобаннях [5].

Нішеві або тематичні соціальні мережі – це мережі, які направлені на об’єднання людей, що мають спільні інтереси, мету або відносяться до однієї ніші. Наприклад, мережа «The-Dots», яка групує людей за їх навичками у мистецтві та творчості.

За призначенням соціальні мережі поділяють на:

– інформаційні, які інформують про новини, акції та інші повсякденні речі;

– освітні, які дозволяють спілкуватись студентам (ResearchGate);

– мережі для знайомств (Badoo, How About We і Tinder);

– мультимедійні, для обміну фото та відео контентом (Instagram, Tik-Tok);

– мережі обміну повідомленнями (Twitter і Facebook);

– торгівельні мережі.

За платформою соціальні мережі поділяють на комп’ютерні, мобільні та змішані. Комп’ютерні соціальні мережі дозволяють отримувати доступ до мережі лише з комп’ютеру. Мобільні надають доступ лише з мобільного

пристрою, а змішані дозволяють користуватись соціальною мережею як з комп'ютера, так і з мобільного пристрою.

Месенджер один з видів обміну повідомленнями у інтернет середовищі. Месенджер – це служба зв'язку, для обміну текстовими повідомленнями між комп'ютерами, мобільними пристроями та іншими пристроями користувачів через мережу інтернет [6].

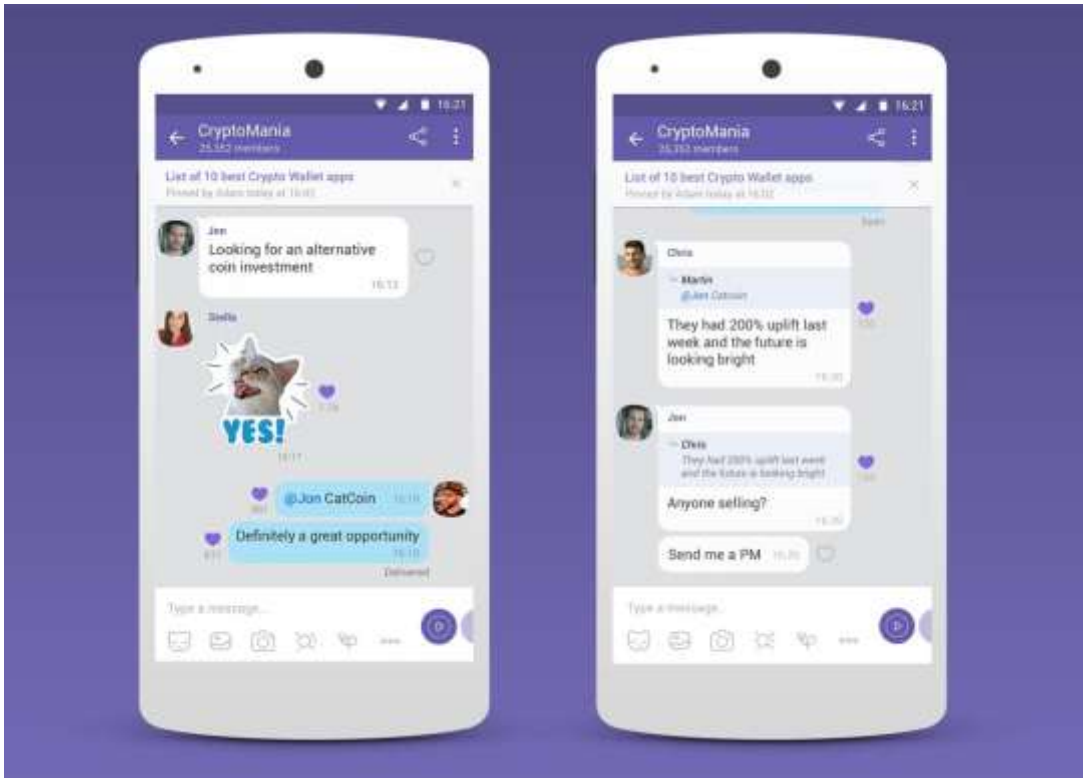


Рисунок 1.1 – Інтерфейс месенджера «Viber» [8]

Серед популярних розглядають такі месенджери, як «Viber», «Facebook Messenger» та «Telegram» [7]. Viber є стандартним по функціональності месенджерем, який встановлений у понад 900 млн активних користувачів. Мережа обміну повідомленнями має досить високий рейтинг у сервісах завантаження додатків, таких як: Google Play (4.3) та App Store (4.5). У застосунку можна обмінюватись текстовими, голосовими, фото та відео повідомленнями, надсилати стікери та gif-файли, а також, відправляти графіті [8]. Viber надає можливість виконувати аудіо або і відео дзвінки з досить хорошою передачею зображення та звуку. Переваги такої мережі у тому, що в

ній можна об’єднувати людей у групи, спільноти, і секретні чати. Секретні чати це особливий вид чатів, завдяки якому можна встановити час через який повідомлення самознищаться, і такі повідомлення користувач не зможе переслати іншим користувачам. Інтерфейс програми представлено на рисунку 1.1.

Ще одним популярним месенджером серед користувачів виділяють «Facebook Messenger». Ця мережа має понад 1 млрд активних користувачів, така популярність викликана тим, що даний сервіс для обміну повідомленнями використовується майже в усьому світі [9]. По функціональності додаток схожий на попередній, проте «Facebook Messenger» має більш мінімалістичний інтерфейс. Незважаючи на популярність додатку, рейтинг у магазинах у нього досить низький відносно інших (Google Play (4.0), App Store (3.5)). Інтерфейс програми представлено на рисунку 1.2.

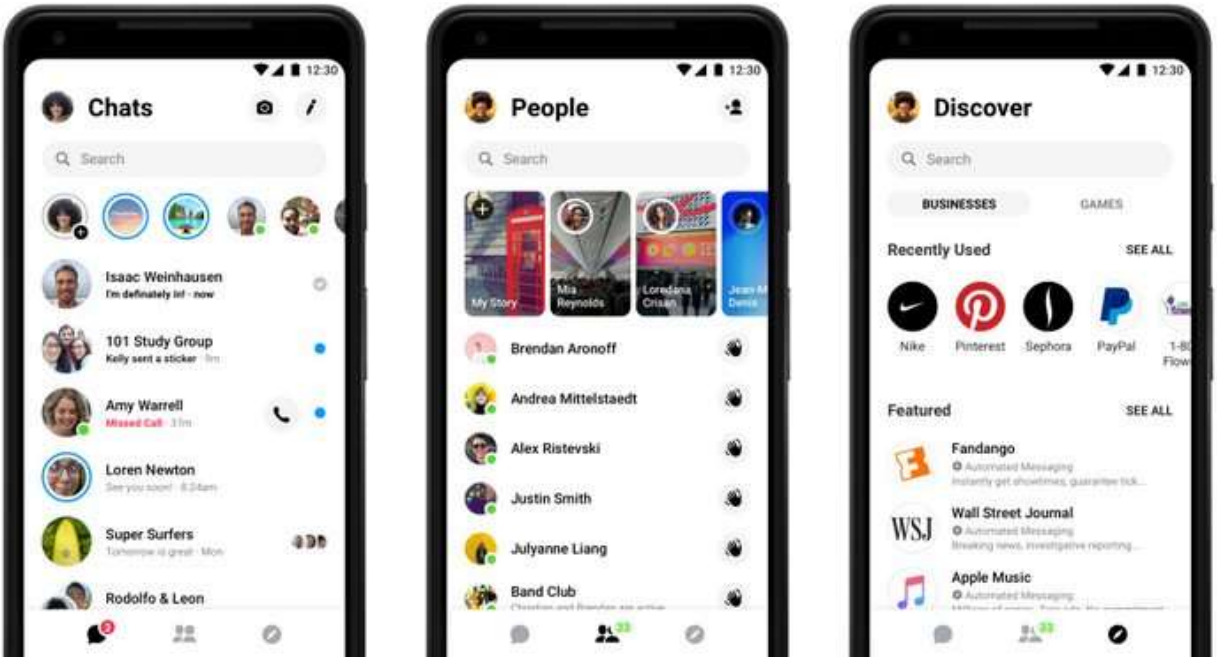


Рисунок 1.2 – Інтерфейс програми «Facebook Messenger» [9]

Telegram – месенджер, який за останній рік набрав досить велику аудиторію користувачів та підвищив свою популярність. Гарний інтерфейс, зручність у користуванні, простота та функціональність усе це поєднується у

даному додатку [10]. Розробники програми запевняють, що додаток має одну з найкращих систем безпеки та захисту повідомлень і даних. Через додаток можна виконувати усі функції, як і в попередніх, проте, також є можливість створювати та додавати свої набори стікерів, створювати канали та секретні чати, завантажувати різних форматів файли та слухати у додатку музику. Рейтинг додатку такий ж як і у Viber. Інтерфейс програми представлено на рисунку 1.3.

Рисунок 1.3 – Інтерфейс програми «Telegram» [10]

Чати – ще однин з сервісів обміну повідомленнями, який використовується у режимі реального часу. Чати за способом реалізації можна поділити на [11]:

– веб-чати (розміщені на веб-сторінці, оновлюється з певною заданою періодичністю);

– чати на IRC (спеціалізований протокол для чатів);

– чати на сторонніх протоколах;

– чат-програми для обміну даними в локальній мережі.

За сферою застосуванням чати поділяються на: all2all (групові), p2p (персональні), b2b (ділові) та b2c (споживацькі для клієнтів на сайтах).

Блог, ще один вид обміну інформацією. Сайт блогу дозволяє його автору розповідати про своє життя та новини, натомість отримувати коментарі від читачів та їх прихильність. Таким чином, блог стає середовищем обміну повідомленнями. Записи, що додаються у блог можуть містити як звичайний текст, так і зображення або відео контент [12].

Різновидом блогу є мікроблог, він дозволяє користувачу публікувати короткі текстові повідомлення або фото чи відео контент. Видимість дописів мікроблогу може бути відкрита як для усіх користувачів, так і для обмеженого кола, яке створює сам автор мікроблогу [13]. Різниця мікроблогу від звичайного блогу у тому, що у мікроблозі дописи користувача є значно меншими за обсягом, так текстові повідомлення можуть складатись з кількох слів або речень, і містити одне фото або відео.

Найпопулярнішою мережею, що містить в собі мікроблоги є Twitter [14]. Twitter дозволяє користувачам залишати короткі текстові повідомлення використовуючи для цього приємний та простий інтерфейс, смс та інші програми-клієнти [15]. Кожне повідомлення має обмеження в кількості символів – до 280 символів. Дописи які створюють користувачі називають «твітами» (рисунок 1.4). Мережу використовують не тільки для звичайного спілкування, а ще й для комерційних цілей, до прикладу фірми використовують соціальну мережу Twitter для розміщення реклами та оповіщення клієнтів про новинки та акції.

Короткі текстові повідомлення, такі як твіти у соціальній мережі Twitter, необхідно класифікувати, щоб було зручно їх шукати у пошуку або ж сортувати та впорядковувати. Класифікація – це певна система розподілу явищ або предметів, за однаковими ознаками та властивостями, на класи або групи [16]. Для кращої класифікації можна використовувати теги. Теги – це короткі пошукові запити, за допомогою яких користувачі можуть знайти текст публікації [17]. Вони мають містити якомога ширший опис публікації, щоб можна було

охопити велику кількість текстів. Процедура введення тегів, які описують текст називають тегуванням.

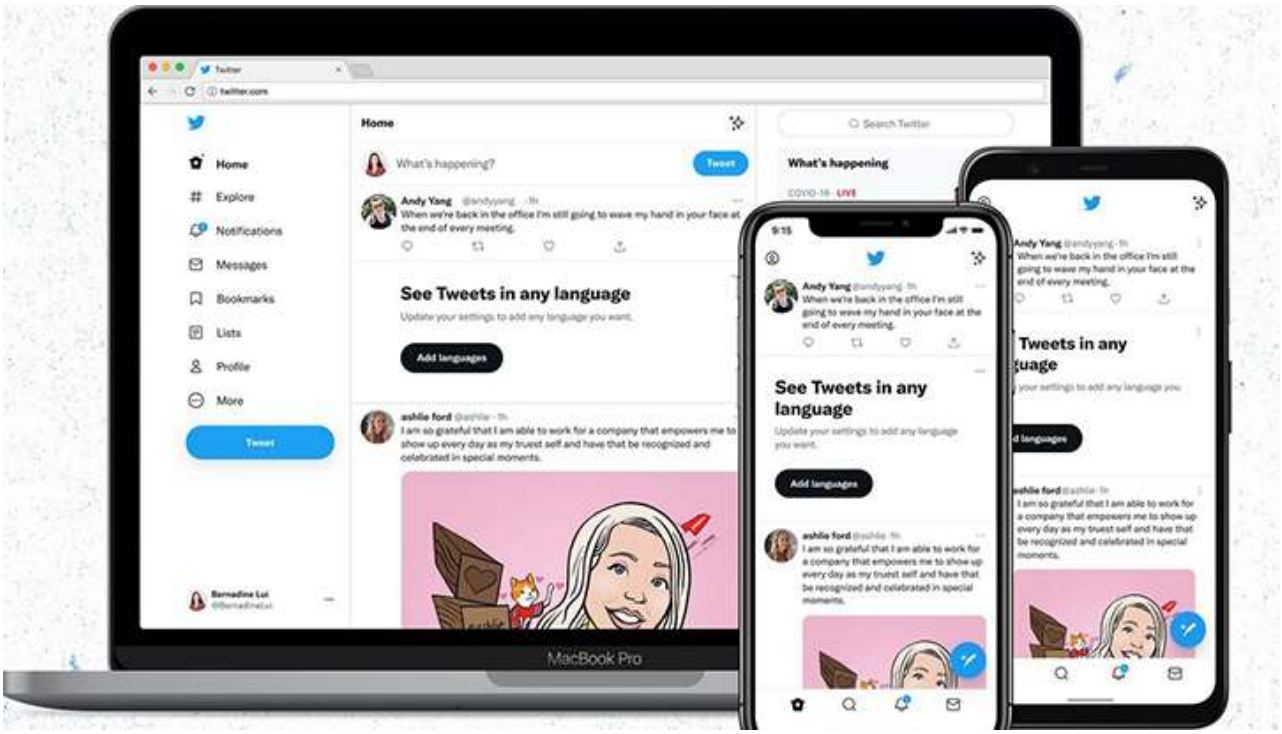


Рисунок 1.4 – Твіти та інтерфейс соціальної мережі Twitter [15]

Сутностями предметної області можна назвати короткі текстові повідомлення (назва, його текст, дата та час його створення), ключові слова, теги текстів та оцінка їх приналежності, кількість ключових слів у множині послідовних слів тексту, входження ключових слів у теги та оцінка їх семантичної важливості.

Текстових повідомлень є досить велика кількість, і ця цифра збільшується з кожним днем, з кожною хвилиною, тому для зручності пошуку та спілкування буде актуально створити програму для автоматизованої тематичної класифікації коротких текстових повідомлень.

1.2 Огляд теоретичних підходів до розв’язку подібних задач

Для класифікації коротких текстових повідомлень, таких як твіти у соціальній мережі Twitter, доцільно буде використовувати комп’ютерну

лінгвістику. Комп’ютерна лінгвістика – це вид мовознавства, за допомогою якого мова вивчається за використання комп’ютера [18]. Даний вид лінгвістики досліджує яким способом та засобами мова людини може опрацьовуватись та перетворюватись. У цій галузі науки враховуються та розглядаються характеристики мови та створюються алгоритми для автоматичної обробки мови. В основному аналіз проводиться за використання нейронних мереж та процесів, які взяті з логіки [19]. Дослідження, що проводяться в комп’ютерній лінгвістиці призводять до таких можливостей:

- накопичення, обробка та пошук інформації;
- переклад тексту з різних мов;
- аналіз тексту за змістом та тематикою;
- узагальнення тексту;
- створення ботів, які виконують завдання пов’язані з текстом.

Також, автоматизувались наступні процеси:

- складання та лінгвістична обробка машинних словників;
- виявлення та виправлення помилок при введенні текстів;
- індексування документів;
- класифікація певних документів.

До комп’ютерної лінгвістики входить такий вид аналізування текстів, як семантичний аналіз тексту. Семантичний аналіз тексту є одним із кроків алгоритму автоматизованого розуміння тексту, внаслідок якого визначаються семантичні відношення та формування семантичного представлення [20].

Одним з основних етапів для класифікації коротких текстових повідомлень є визначення масиву ключових слів. Ключові слова – це слова або словосполучення, що визначають зміст та тематику статті, допису або сторінки на сайті [21]. За допомогою ключових слів користувач може з певною точністю знайти необхідну інформацію. Ключові слова використовуються у пошуковій системі, такій як Google. Кожне ключове слово можна віднести до певної групи частотності: високочастотні (запити перевищують 3 тисячі разів у місяць), середньочастотні (запити від 1 до 3 тисяч разів у місяць) та низькочастотні

(менше 1000 разів у місяць). Низькочастотні ключові слова дозволяють швидко отримати цільову аудиторію по тематиці створеного сайту або статті.

Окрім поділу за частотою, ключові слова можна поділити за видами [22]:

– геозапити (у пошуковому запиті використовується конкретна локація, місто або країна);

– інформаційні (у пошуковому запиті вказується конкретна інформація);

– комерційні;

– навігаційні (у пошуковому запиті вказується сайт або його назва);

– мультимедійні (у пошуковому запиті вказується відповідний тип контенту);

– загальні.

Ключовими можуть бути як слова, так і словосполучення. Словосполучення – це поєднання двох або більше повнозначних слів, одне з яких - головне, а інше – залежне [23].

Слово або словосполучення, яке означає чітко визначене конкретне поняття певної галузі науки називається терміном [24]. За допомогою термінів у ключових словах, пошук потрібної інформації стає більш точним.

Для пошуку ключових слів використовуються алгоритми та методи. Існує велика кількість методів пошуку ключових слів, серед них:

– Bag-of-Words;

– TextRank;

– BM25;

– TF-IDF.

Bag-of-Words модель для визначення ознак з тексту для використання в моделюванні, до прикладу може використовуватись у алгоритмах машинного навчання. Модель було названо так, через те, що інформація про порядок слів та їх структуру у тексті документу ігнорується, а враховується лише те чи зустрічаються слова в документі вже вдруге чи ні. Зазвичай така модель використовується у класифікації документів [25]. Така модель має свої недоліки:

– вузький обсяг словника;

– важче моделювати розрідженість представлення;

– ігнорується семантика, тобто значення слів та контекст.

Метод TextRank визначає ключові слова за допомогою графової моделі ранжування [26]. Графова модель ранжування – це визначення важливості вершини графа, що знаходиться всередині, на основі усієї інформації яка обертається в графі. Головною ідеєю є голосування за вершину графа, тобто, коли вершина з'єднується з іншою вершиною вона віддає їй свій голос, таким чином, збільшується вага вершини. Чим більша вага вершини, тим важливіший її голос при голосуванні.

Токенізація – розділення тексту на послідовність символів, таких як букви, пробіли, знаки пунктуації, цифри, слова та фрази [27]. Мета та завдання токенізації у тому, щоб відокремити слова від синтаксичних знаків, цифр, букв та цифр разом, інтернет-адрес та ін.

Стемінг є процесом, що скорочує слова до їх основи, при цьому відбувається відкидання допоміжних частин, тобто, закінчення чи суфікси слова [28]. Стемінг використовується в морфології та в інформаційному пошуку.

Для побудови графу, необхідно спочатку обробити та підготувати текст, токенізувати його, провести стемінг та видалити стоп-слова. Кожне слово стає вершиною графа, два слова з'єднуються ребром, якщо вони зустрічаються поряд у тексті. Вигляд графа зображено на рисунку 1.5. Вага ребер на початку рівна нулю, при зустрічі слова у тексті вага збільшується на одиницю поділену на відстань між двома словами у тексті.

BM25 – функція ранжування, яка застосовується для первинної обробки тексту відповідно до релевантності пошукового запиту. Зв'язок між словами при такому методі не враховується [29].

TF-IDF – статистична міра, яка застосовується для оцінки важливості слова в тексті. TF є відношенням кількості входжень певного слова до загальної кількості слів у тексті. Тобто, після такого аналізу визначається оцінка важливості слова у конкретних межах окремого документу. IDF – обернена

частота, яка зменшує вагу слів що вживаються частіше інших, тобто, є широковживаними [30].

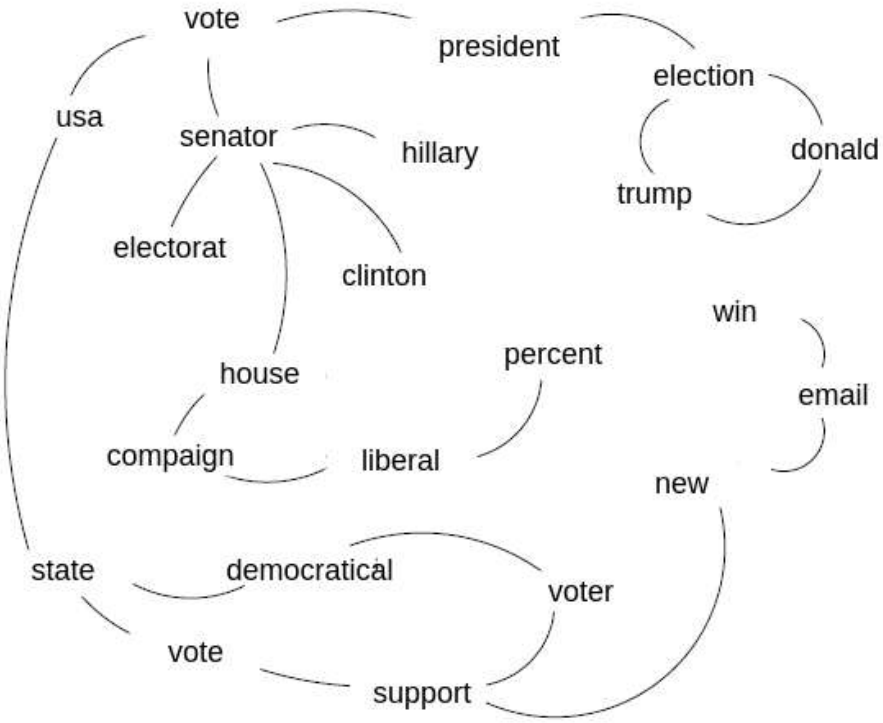


Рисунок 1.5 – Приклад графу створеного алгоритмом TextRank [26]

Для зберігання тексту повідомлень, та масивів оброблених слів необхідно використовувати базу даних. База даних – це певна структура, яка призначена для зміни, зберігання та обробки інформації яка залежить одна від одної [31].

Для нормальної та коректної роботи з базою даних потрібно використовувати СКБД (системи керування базами даних). СКБД виступає посередником між користувачем системи та базою даних. Користувачів які використовують одну і ту ж базу даних може бути багато, якщо кожен з них буде користуватись СКБД.

Таким чином, для розробки методу автоматизованої тематичної класифікації коротких текстових повідомлень буде використано метод TF-IDF та база даних для зберігання інформації та тексту.

1.3 Аналіз існуючих програмних рішень

Кількість інформації в інтернет ресурсах щодня збільшується, через це створюють програми для їх класифікації та сортування. Прикладом програмного забезпечення яке проводить семантичний аналіз текстів є сайт «SiteAnalyzer» [32]. Інтерфейс програми зображено на рисунку 1.6.

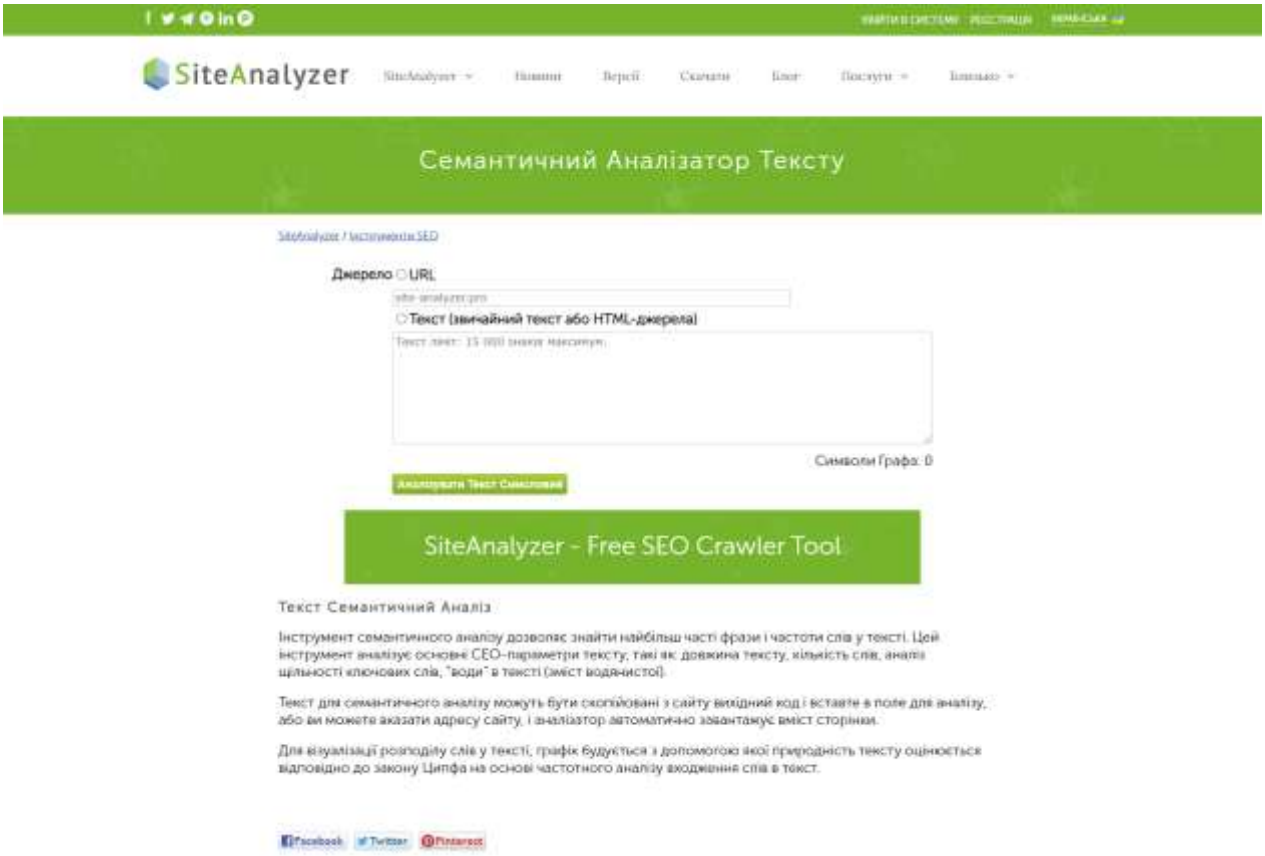


Рисунок 1.6 - Інтерфейс сайту «SiteAnalyzer» [32]

Даний сайт дозволяє провести семантичний аналіз тексту, тобто, знайти фрази та слова, які вживаються найчастіше, визначити довжину тексту, кількість слів, проаналізувати щільність ключових слів та зміст водянистості тексту.

Також, сайт дозволяє додавати текст вручну у відповідне вікно або ж подати посилання на джерело де розміщено відповідний текст. Система має як безкоштовні так і платні послуги, що дозволяють провести аналіз та замовити додаткові послуги пов’язані з просуванням та аналітикою.

Ще одним прикладом системи, яка використовує токенізацію, стемінг та інші методи для пошуку ключових слів, є система розроблена під час дослідження та виконання кваліфікаційної роботи на тему «Дослідження методів вилучення інформації з неструктурованих текстів на природній мові» [33]. Приклад збереження даних системи та оцінка точності використаної моделі зображено на рисунках 1.7-1.8.

	url	published	topic	title	text
0	https://tsn.ua/politika/pri-vsiy-lyubovi-do-ro...	2021-02-19	Політика	"Зі всієї любові до росіян": Лукашенко збрався...	Олександр Лукашенко заявив, що хоче розвивати ...
1	https://tsn.ua/politika/sankciyi-proti-medvedc...	2021-02-19	Політика	Санкції проти Медведчука та його оточення: чим...	У п'ятницю, 19 лютого, Рада національної безпе...
2	https://tsn.ua/politika/rishennya-rnbo-pro-nov...	2021-02-19	Політика	Рішення РНБО про нові санкції ґрунтується на м...	Рішення Ради національної безпеки і оборони (Р...
3	https://tsn.ua/politika/pochatok-novoyi-eri-kiz...	2021-02-19	Політика	"Початок нової ери, коли ворога нарешті назива...	Міністр культури та інформаційної політики Укр...
4	https://tsn.ua/politika/represiyi-ta-znischenn...	2021-02-19	Політика	"Репресії" та "знищення демократії": в ОПЗЖ ві...	У політичній партії "Опозиційна платформа – За...
5	https://tsn.ua/politika/za-sankciyi-proti-medv...	2021-02-19	Політика	За санкції проти Медведчука та його оточення о...	Під час голосування за санкції проти нардепа в...
6	https://tsn.ua/politika/u-nas-zgidno-konstituc...	2021-02-19	Політика	"У нас згідно з Конституцією всі рівні": Даніл...	Секретар РНБО Олексій Данілов подякував українсь...
7	https://tsn.ua/politika/sankciyi-proti-medvedc...	2021-02-19	Політика	Санкції проти Медведчука: Кравчук підтримав рі...	19 лютого на засіданні Ради національної безпе...
8	https://tsn.ua/politika/rnbo-doruchiv-povernut...	2021-02-19	Політика	РНБО доручила повернути у державасність частин...	Рада національної безпеки і оборони доручила К...
9	https://tsn.ua/politika/putin-zve-lukashenka-n...	2021-02-18	Політика	Путін кличе Лукашенка на кипіш у Сочі: чим Укр...	Хоч Лукашенко й пообіцяв не балотуватися на на...

Рисунок 1.7 – Читання даних та вивід перших десяти [33]

```
accuracy 0.922
                precision    recall  f1-score   support

     Політика       0.92      0.81      0.86       610
          АТО       0.90      0.91      0.91       584
    Економіка       0.92      0.98      0.95       626
      Туризм        0.96      0.96      0.96       591
  Наука та IT       0.91      0.96      0.93       589

     accuracy                           0.92      3000
    macro avg       0.92      0.92      0.92      3000
 weighted avg       0.92      0.92      0.92      3000
```

Рисунок 1.8 – Оцінка точності моделі [33]

Таким чином, видно що використана модель в системі досить точна (оцінка = 0.922). Система класифікує новини за тематикою, зокрема, по темах: «Політика», «АТО», «Економіка», «Туризм», та «Наука та ІТ». Такий програмний продукт є досить актуальним для сортування та пошуку відповідних новин, проте інтерфейс системи не є досить привабливим.

Одним з прикладів класифікації коротких текстових повідомлень є соціальна мережа Twitter, а саме її функції. У даній мережі є кілька класифікацій, наприклад, класифікація твітів за популярністю, за хештегами та за темами. Класифікація твітів за темою та хештегами зображено на рисунку 1.9 [34].

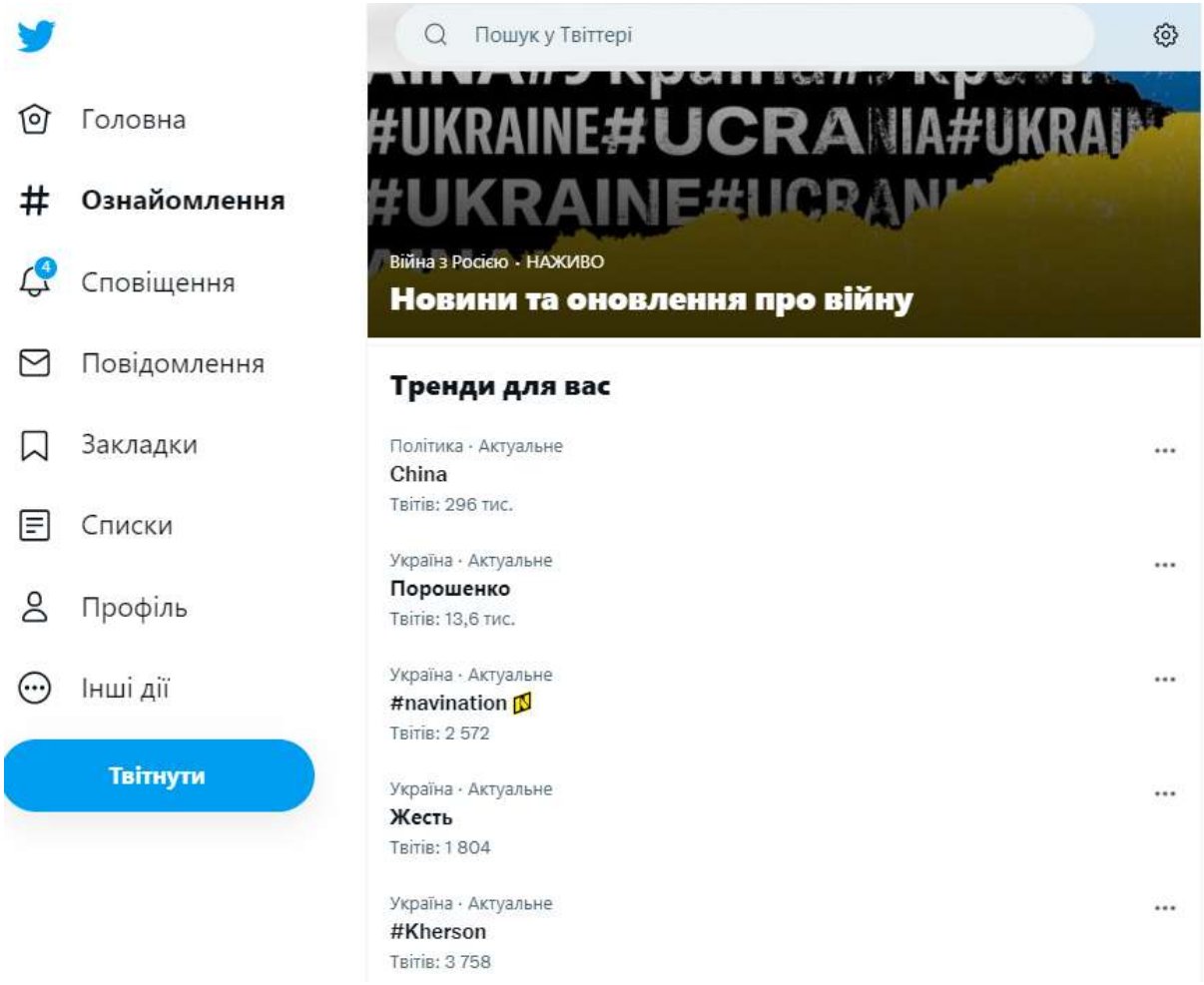


Рисунок 1.9 – Класифікація твітів за темою та хештегами [34]

Кожен допис у соціальній мережі Twitter пізніше буде відсортовано, тобто, буде проведена класифікація його за темою, що дозволить у подальшому знайти створений допис у відповідному розділі тематики.

Також, є ще кілька сервісів, які дозволяють виконувати пошук ключових слів. Серед них є такі: Keyword Planner та Serpstat. Keyword Planner – безкоштовний інструмент від Google. Планер дозволяє шукати ключові слова подібні до тих, що вводяться, також, визначати за популярністю ключові слова та дивитись частотність запитів [35]. Інтерфейс сторінки системи зображено на рисунку 1.10.

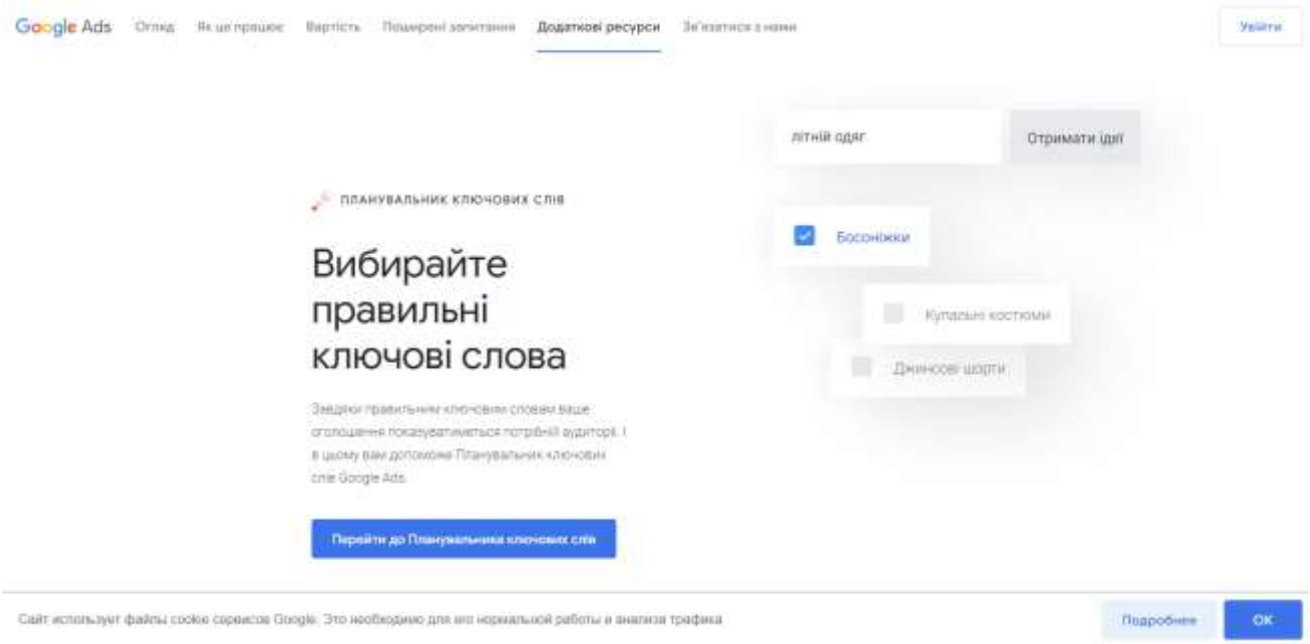


Рисунок 1.10 – Інтерфейс системи Keyword Planner [35]

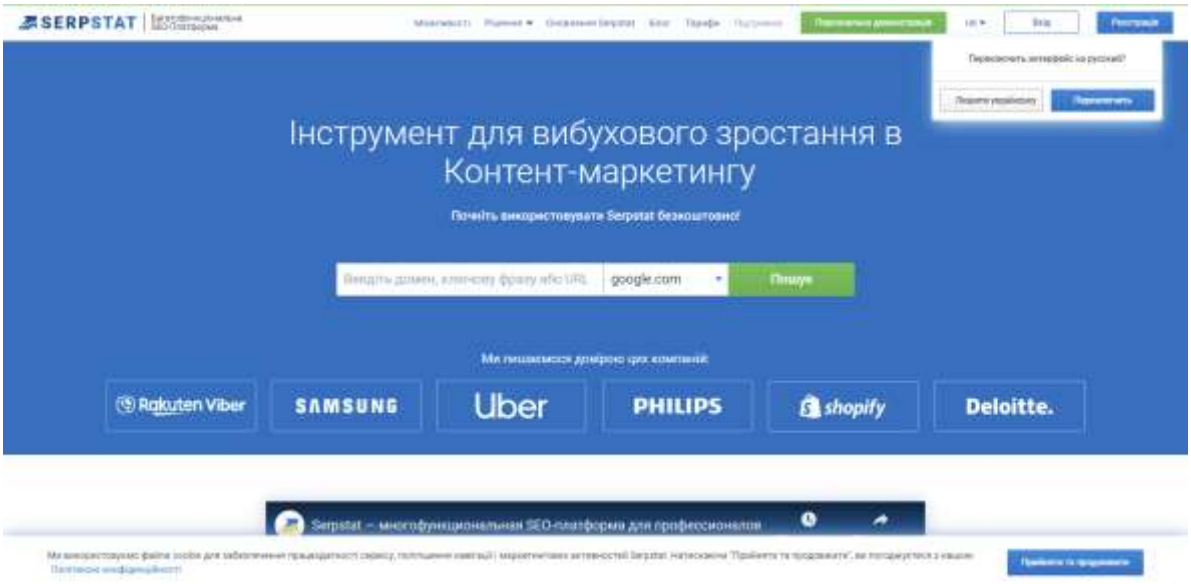


Рисунок 1.11 – Інтерфейс сервісу Serpstat [36]

Serpstat - платний сервіс, проте має обмежену безкоштовну версію. Така система має свої переваги, наприклад, є точна кількість запитів, можна побачити де ще відбувається пошук, тобто, на яких сайтах слово яке вводиться в пошук частіше відображається, простий інтерфейс дає сервісу більшу прихильність від користувачів [36]. Інтерфейс системи зображено на рисунку 1.11.

Що один, що другий сервіс мають платну і обмежену безкоштовну версію та свої переваги і недоліки, які кожен користувач визначає для себе.

Таким чином, для автоматизованої класифікації коротких текстових повідомлень необхідно створити систему яка буде об’єднувати певні функції з вище розглянутих програмних продуктів.

1.4 Мета, задачі та вимоги до реалізації інформаційної системи

Метою кваліфікаційної роботи бакалавра є розробка та прикладна програмна реалізація методу автоматизованої тематичної класифікації коротких текстових повідомлень. Для того, щоб досягнути поставлену мету необхідно вирішити наступні задачі:

1. Провести аналіз предметної області семантичного аналізу цифрових текстів.
2. Провести аналіз засобів спілкування в Інтернеті.
3. Огляд теоретичних підходів до розв’язання задач подібних до автоматизації формування тематичної класифікації коротких текстових повідомлень.
4. Обрати алгоритм початкової обробки коротких текстових повідолень.
5. Вдосконалити метод автоматизованої тематичної класифікації коротких текстових повідомлень.
6. Розробити інформаційну технологію автоматизованого пошуку та формування масиву ключових слів коротких текстових повідомлень.
7. Провести прикладне дослідження методу автоматизованої тематичної класифікації коротких текстових повідомлень і виконати аналіз результатів.

Розроблена інформаційна система автоматизованої тематичної класифікації коротких текстових повідомлень на платформі .NET, має виконувати наступні основні функції:

– розбивання тексту на слова та додаткові знаки, цифри;

– визначення загальних параметрів тексту: кількість слів, знаків;

– виділення з тексту лише слова, та зменшення регістру тексту;

– формування масиву унікальних слів;

– формування обробленого тексту;

– визначення позиції та частоти зустрічання кожного унікального слова;

– формування альтернативних текстів для кожного тегу;

– обрахунок значень TF-IDF для кожного унікального слова;

– сортування та обмеження значень TF-IDF;

– визначення ключових слів тексту;

– обробка тестового тексту короткого текстового повідомлення;

– визначення ключових слів тестового тексту, обрахунок їх частоти – CW;

– визначення приналежності ключових слів тестового тексту до кожного тегу.

– сортування слів тестового тексту по приналежності до тегів.

Розділ 2 Проєктування інформаційної системи

2.1 Аналіз та автоматизація обробки потоків даних

2.1.1 Метод автоматизованої тематичної класифікації коротких текстових повідомлень

Метод автоматизованої тематичної класифікації коротких текстових повідомлень у якості вхідних даних має базу текстів коротких текстових повідомлень та тестовий текст короткого текстового повідомлення, такий як твіт у соціальній мережі Twitter. У якості вихідних даних інформаційна система має ключові слова тегів (класифікацій) та визначений тег до якого відноситься тестовий текст. Тобто, інформаційна технологія методу автоматизованої тематичної класифікації коротких текстових повідомлень має дозволяти використовуючи відповідні алгоритми з вхідних даних у вигляді базу текстів отримувати вихідні дані у вигляді ключових слів та тегу тестового тексту повідомлень. На рисунку 2.1 зображена схема методу автоматизованої тематичної класифікації коротких текстових повідомлень.

На Кроці 1 методу автоматизованої тематичної класифікації коротких текстових повідомлень відбувається поелементна обробка тексту та визначення унікальних слів у тексті. Зокрема, відбувається розбивання тексту на слова та додаткові знаки, такі як цифри, знаки пунктуації, та інші, далі відбувається визначення загальних параметрів тексту (кількість слів, знаків), очищення тексту від додаткових знаків, символів, цифр. Далі переводиться текст у один регістр, виконується формування масиву унікальних слів тексту та формування обробленого тексту.

Крок 2 відповідає за підготовку текстів та пошук частоти зустрічей кожного унікального слова. На даному етапі відбувається пошук та запис позицій кожного унікального слова у тексті, визначення частоти зустрічання кожного унікального слова у тексті та формування альтернативних текстів для кожного тегу. Після цього кроку доступними для подальшого аналізу стають альтернативні тексти та частота унікальних слів.

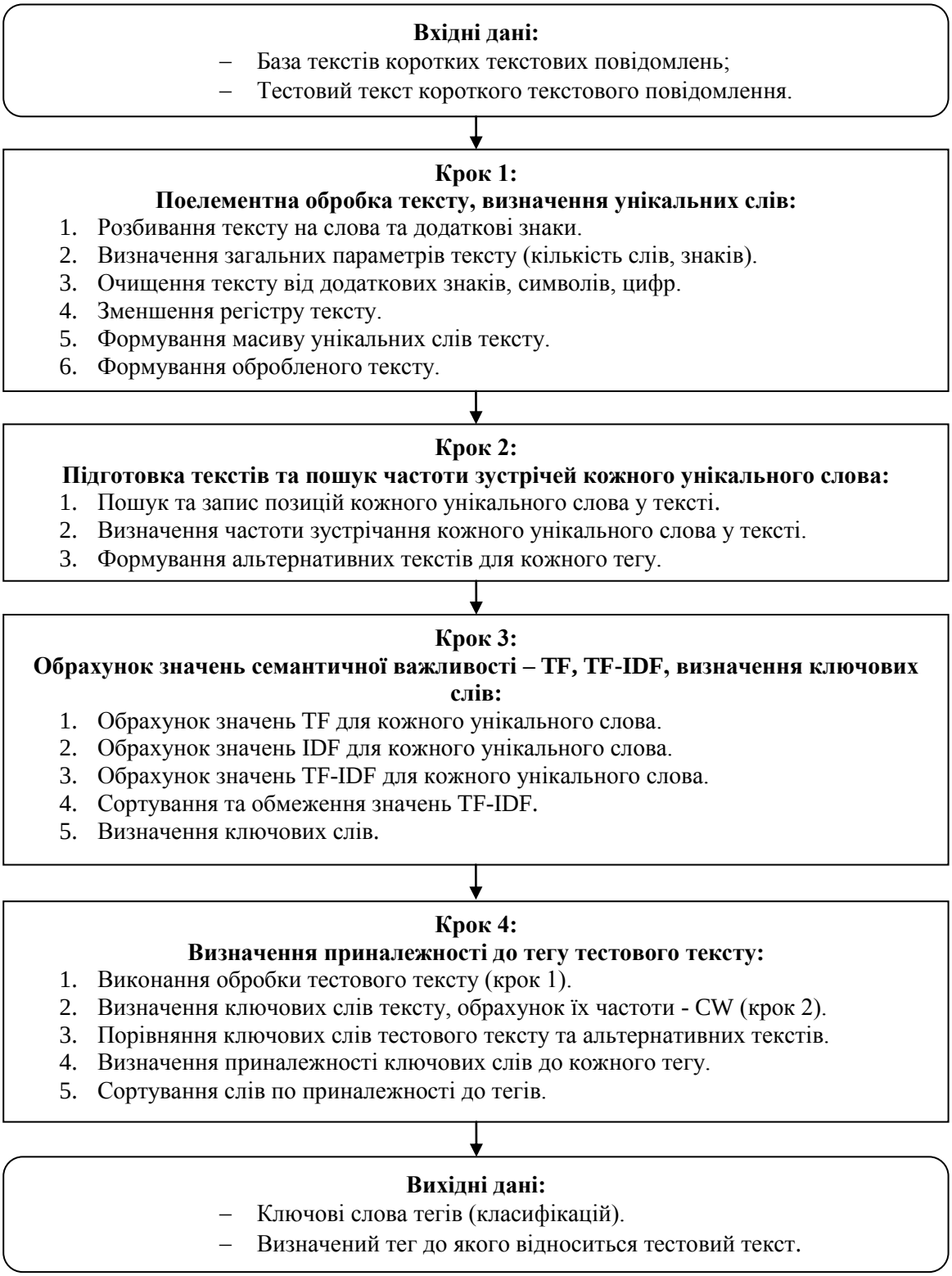


Рисунок 2.1 – Схема методу автоматизованої тематичної класифікації коротких текстових повідомлень

На Кроці 3 відбувається обрахунок значень семантичної важливості – TF, TF-IDF та визначення ключових слів. Зокрема, відбувається обрахунок значень

TF для кожного унікального слова, обрахунок значень IDF для кожного унікального слова, та обрахунок значень TF-IDF для кожного унікального слова.

Метод TF-IDF розуміє під собою у використанні статистичного показника, він призначений для оцінки значимості слів у контексті документа. TF (частота слова) відношення кількості входжень слова до загальної кількості слів у документі:

$$TF_i = \frac{n(i)}{\sum_k n_{ik}}$$

де n_i це кількість входжень слова в документі, а дільник це кількість всіх слів.

IDF обернена частота документа з якою слово зустрічається в документах колекції, застосування IDF зменшує вагу слів, які є найбільш вживаними:

$$IDF_i = \log \frac{D}{d_i}$$

де D – кількість документів колекції; d_i – кількість документів, в яких зустрічається дане слово.

Показник TF-IDF це добуток двох частот - TF та IDF:

$$TF - IDF = TF * IDF$$

Далі проводиться сортування та обмеження значень TF-IDF та визначення ключових слів тексту. Цей крок є завершальним для визначення ключових слів, таким чином на виході з цього кроку отримуємо ключові слова тегів (класифікацій).

Ще одним етапом, завершальним у даній інформаційній системі, є Крок 4, на якому відбувається визначення приналежності до тегу тестового тексту, виконання обробки тестового тексту, така ж як на кроці 1, далі відбувається

визначення ключових слів тексту та обрахунок їх частоти, CW (так як на кроці 2). Наступним є порівняння ключових слів текстового тексту та слів альтернативних текстів. Останнім відбувається визначення приналежності ключових слів до кожного тегу та сортування слів по приналежності до тегів. Таким чином, визначається тег до якого належить тестовий текст повідомлення.

Отже, метод автоматизованої тематичної класифікації коротких текстових повідомлень дозволяє перетворювати вхідні дані у вигляді базу текстів та показниками їх семантичної важливості у вихідні дані у вигляді ключових слів та тегу приналежності тестового тексту повідомлень.

2.1.2 Функціональна структура інформаційної системи

Діаграма активності, дозволяє відображати потік від однієї до іншої діяльностей. Сама діяльність це процеси які виконує інформаційна система, з’єднуючись між собою потоками, які виходять від виходу одного вузла до входу іншого.

Діаграма активності процесів інформаційної системи за методом автоматизованої тематичної класифікації коротких текстових повідомлень зображена на рисунку 2.2.

Діаграма послідовності дозволяє відображати взаємодії об’єктів, які впорядковуються за часом виконання. Метою діаграми послідовності є :

- зобразити поведінку системи у динаміці;
- описати потік повідомлень в системі;
- описати організацію структури об'єктів;
- описати взаємозв’язки між об'єктами.

Діаграма послідовності інформаційної системи автоматизованої тематичної класифікації коротких текстових повідомлень зображена на рисунку 2.3.

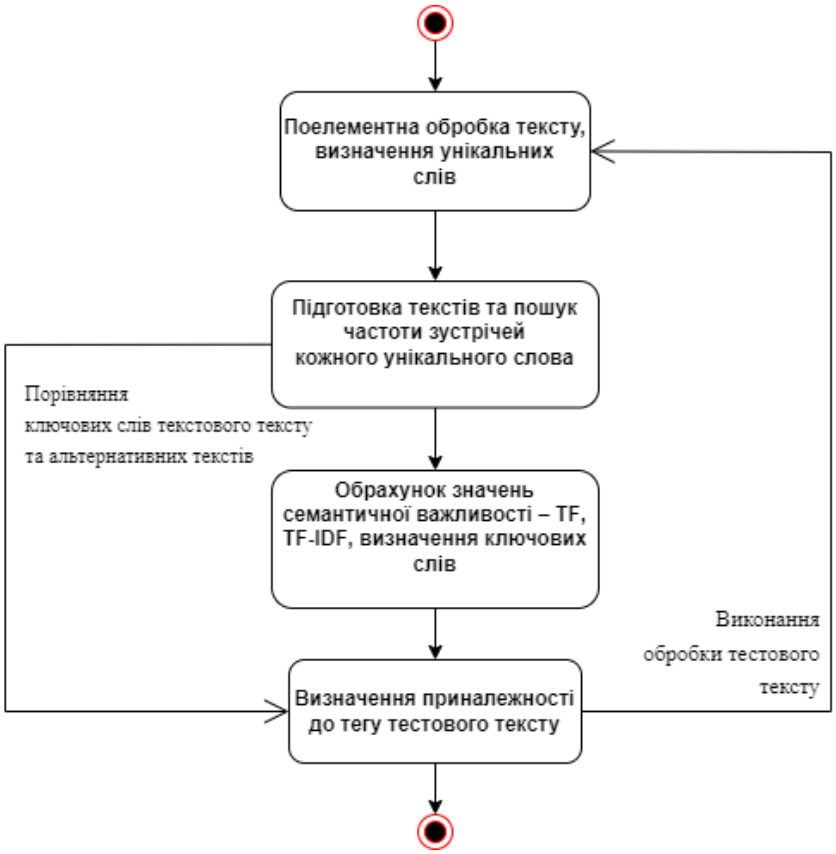


Рисунок 2.2 - Діаграма активності процесів інформаційної системи автоматизованої тематичної класифікації коротких текстових повідомлень

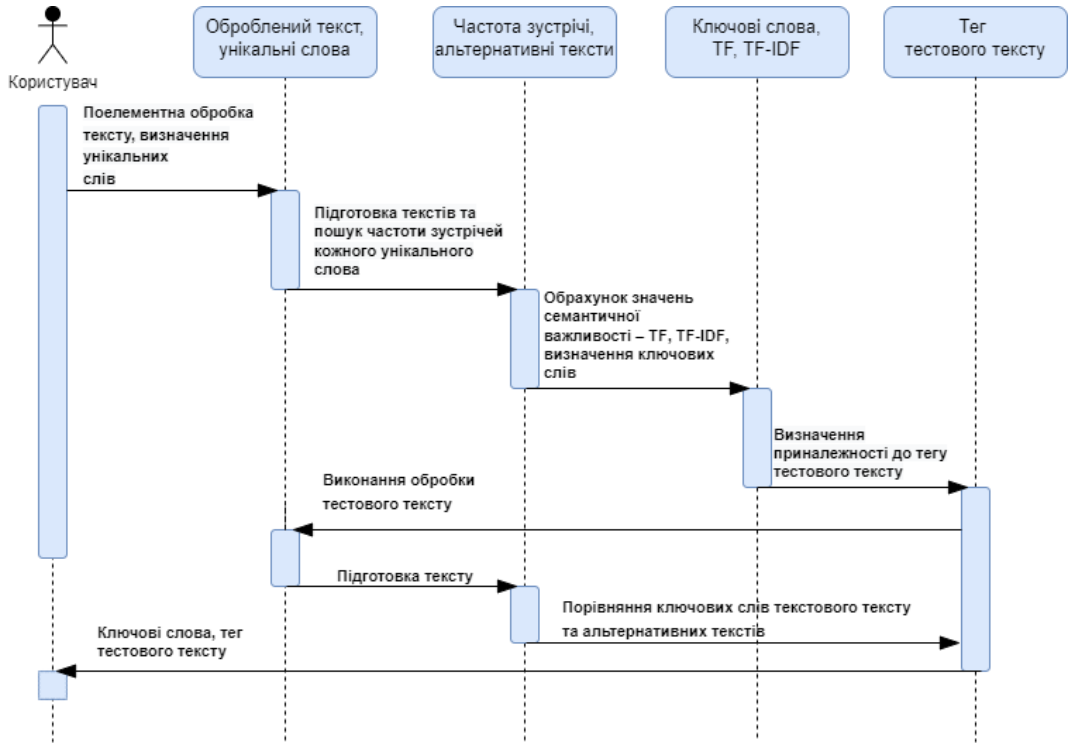


Рисунок 2.3 - Діаграма послідовності інформаційної системи автоматизованої тематичної класифікації коротких текстових повідомлень

Таким чином, завдяки діаграмам активності та послідовності можна переглянути структуру інформаційної системи автоматизованої тематичної класифікації коротких текстових повідомлень та побачити взаємодію між об’єктами цієї системи.

2.2 Інформаційна структура системи

2.2.1 Проектна архітектура системи та взаємозв’язок компонентів

Згідно з методом автоматизованої тематичної класифікації коротких текстових повідомлень було спроектовано відповідну структуру інформаційної системи, зображену на рисунку 2.4. Інформаційна система включає в себе чотири підсистеми та базу даних.

Підсистема роботи з текстом та унікальними словами виконує функції: роботи з текстом, роботи з параметрами тексту (кількість слів, знаків), очищення тексту від додаткових знаків та символів, зменшення регістру, роботи з масивом унікальних слів тексту та формування обробленого тексту.

Підсистема підготовки текстів та пошуку частоти зустрічей кожного унікального слова виконує функції: пошуку та запис позицій кожного унікального слова у тексті, визначення частоти зустрічання кожного унікального слова у тексті та формування альтернативних текстів для кожного тегу.

Підсистема обрахунку значень семантичної важливості – TF, TF-IDF та визначення ключових слів, виконує функції: обрахунку значень TF, IDF, TF-IDF для кожного унікального слова, сортування та обмеження значень TF-IDF та визначення ключових слів.

Підсистема визначення приналежності до тегу тестового тексту виконує функції: обробки тестового тексту, визначення ключових слів тексту, обрахунку їх частоти – CW, порівняння ключових слів тестового тексту та альтернативних текстів, визначення приналежності ключових слів до кожного тегу та сортування слів по приналежності до тегів.

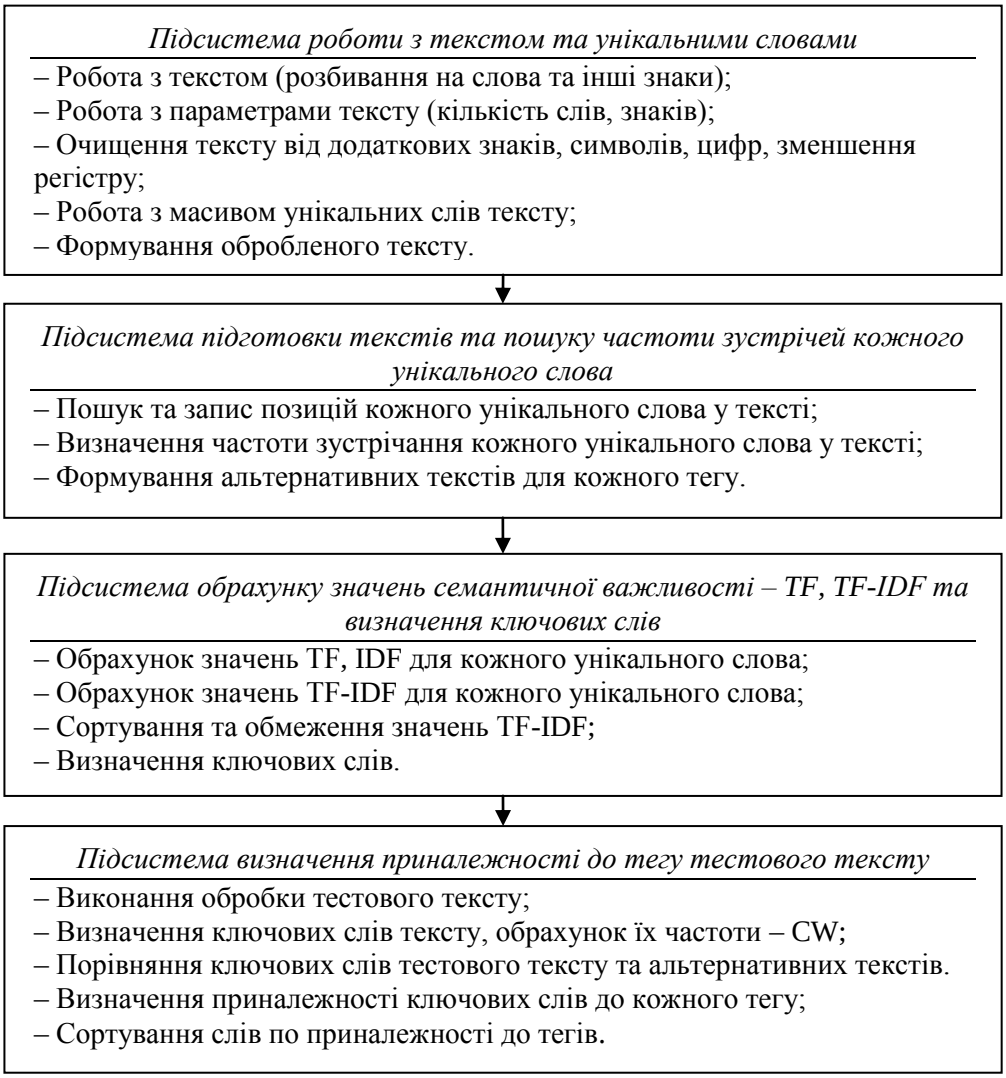


Рисунок 2.4 – Структура інформаційної системи автоматизованої тематичної класифікації коротких текстових повідомлень

Таким чином, підсистеми інформаційної системи виконують необхідні функції поставлені у меті завдання та взаємодіють з базою даних програмного продукту.

2.2.2 Інформаційна модель

Кожен програмний продукт сьогодення має містити в собі базу даних, тому її розробка є одним з головних етапів при створенні інформаційної системи. Для розуміння структури бази даних та наглядного прикладу було

створено діаграму бази даних, яка відображає перелік таблиць, їх атрибутів та зв'язки між ними. Структура бази даних представлена на рисунку 2.5.

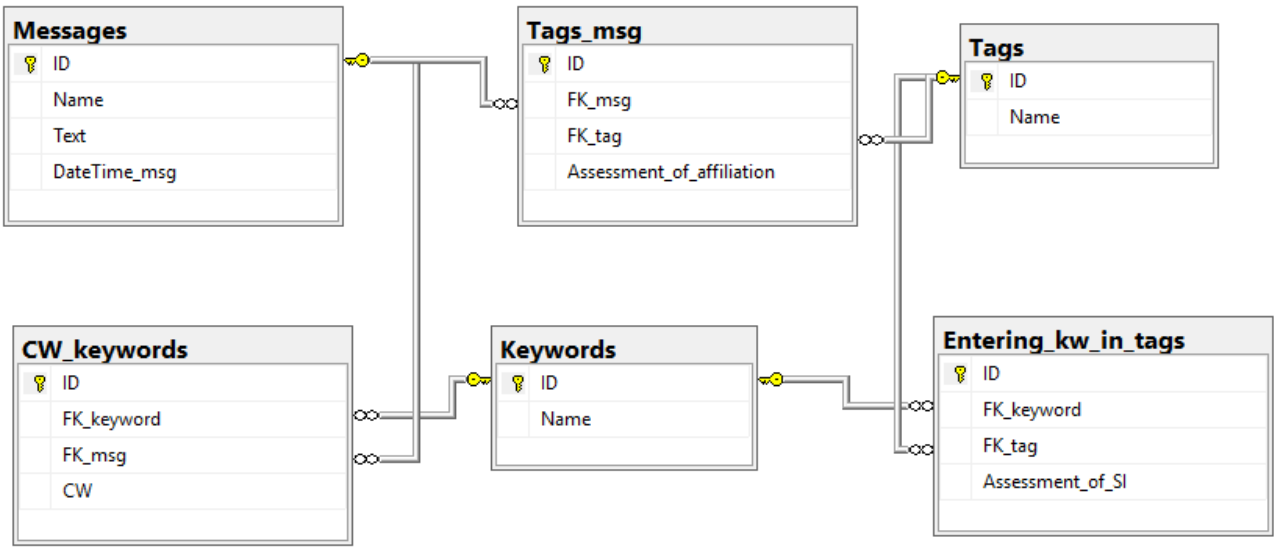


Рисунок 2.5 – Схема структури бази даних інформаційної системи автоматизованої тематичної класифікації коротких текстових повідомлень

Згідно з поставленим завданням, тобто, розробка інформаційної системи автоматизованої тематичної класифікації коротких текстових повідомлень, було створено базу даних, яка дозволяє додавати, зберігати, обробляти та видаляти короткі текстові повідомлення, теги за якими вони класифікуються та ключові слова, які виділяються під час обробки повідомлень.

База даних включає в себе такі таблиці: Messages, Tags, Tags_msg, CW_keywords, Keywords, Entering_kw_in_tags.

Таблиця 2.1 – Атрибути таблиці «Messages»

№ п/п	Назва атрибуту	Тип даних	Опис
1.	ID	int	Ключовий ідентифікатор
2.	Name	nvarchar(100)	Назва повідомлення
3.	Text	text	Текст повідомлення
4.	DateTime_msg	datetime	Дата та час повідомлення

Таблиця «Messages» зберігає дані про короткі текстові повідомлення, включає в себе такі атрибути: Id, Name, Text, DateTime_msg (таблиця 2.1).

Таблиця «Tags» зберігає дані про теги до яких відносяться повідомлення, включає в себе такі атрибути: Id, Name (таблиця 2.2).

Таблиця 2.2 – Атрибути таблиці «Tags»

№ п/п	Назва атрибуту	Тип даних	Опис
1.	ID	int	Ключовий ідентифікатор
2.	Name	nvarchar(100)	Назва тегу

Таблиця «Keywords» зберігає дані про ключові слова повідомлень, включає в себе такі атрибути: Id, Name (таблиця 2.3).

Таблиця 2.3 – Атрибути таблиці «Keywords»

№ п/п	Назва атрибуту	Тип даних	Опис
1.	ID	int	Ключовий ідентифікатор
2.	Name	nvarchar(100)	Ключове слово

Таблиця «Tags_msg» зберігає дані про повідомлення та тег до якого воно відноситься, включає в себе такі атрибути: Id, FK_msg, FK_tag, Assessment_of_affiliation (таблиця 2.4).

Таблиця 2.4 – Атрибути таблиці «Tags_msg»

№ п/п	Назва атрибуту	Тип даних	Опис
1.	ID	int	Ключовий ідентифікатор
2.	FK_msg	int	Повідомлення
3.	FK_tag	int	Тег
4.	Assessment_of_affiliation	float	Оцінка приналежності

Таблиця «CW_keywords» зберігає дані про кількість ключових слів у множині послідовних слів тексту, включає в себе такі атрибути: Id, FK_keyword, FK_msg, CW (Count of Words) (таблиця 2.5).

Таблиця 2.5 – Атрибути таблиці «CW_keywords»

№ п/п	Назва атрибуту	Тип даних	Опис
1.	ID	int	Ключовий ідентифікатор
2.	FK_keyword	int	Ключове слово
3.	FK_msg	int	Повідомлення
4.	CW	int	Кількість

Таблиця «Entering_kw_in_tags» зберігає дані про входження ключових слів у теги, включає в себе такі атрибути: Id, FK_keyword, FK_tag, Assessment_of_SI (Semantic Importance) (таблиця 2.6).

Таблиця 2.6 – Атрибути таблиці «Entering_kw_in_tags»

№ п/п	Назва атрибуту	Тип даних	Опис
1.	ID	int	Ключовий ідентифікатор
2.	FK_keyword	int	Ключове слово
3.	FK_tag	int	Тег
4.	Assessment_of_SI	float	Оцінка семантичної важливості

Отже, було реалізовано базу даних для інформаційної системи автоматизованої тематичної класифікації коротких текстових повідомлень.

2.3 Вибір засобів розробки інформаційної системи

Створення інформаційної системи автоматизованої тематичної класифікації коротких текстових повідомлень не може обійтись без середовища програмування, мови програмування, фреймворку на основі якого буде створено систему та СКБД за допомогою якої можна керувати даними.

2.3.1 Вибір середовища програмування

Середовище програмування розуміє під собою комплекс програм, які містять засоби що автоматизують процеси виконання додатків та дій користувача [37]. Засобами які включає в себе середовище розробки є: редактор текстів програм; довідково-інформаційна система; бібліотеки додатків; компілятор або інтерпретатор; покроковий виконавець програми. Для кожної коми програмування може існувати кілька середовищ розробки, які її підтримують.

Середовищем програмування було обрано Visual Studio 2019. Visual Studio є інтергованим середовищем розробки, яке було створене компанією Microsoft [38]. У ньому можна створювати додатки різного виду: консольні програми, з графічним інтерфейсом, такі як Windows Forms, Web-додатки.

Середовище програмування Visual Studio 2019 підтримує велику кількість мов програмування: Visual C#, Visual Basic, Visual F#, Visual C++, Python та інші. Також, можна розробляти додатки що підходять під мобільні телефони, на базі Android та IOS.

Перевагами такого середовища є, що воно є досить доступним, зручним та простим у використанні, також, існує велика кількість інформації про роботу з даним середовищем, тобто, інформаційна спільнота Visual Studio, дозволить з легкістю освоїти роботу з середовищем.

Кожен окремий додаток у середовищі називається рішенням (solution), будь-яке рішення може складатись з декількох проектів, які мають відкриватись одночасно та бути пов’язані у одному рішенні [39].

Отже, для реалізації програмного додатку методу автоматизованої тематичної класифікації коротких текстових повідомлень було обрано середовище розробки Visual Studio 2019.

2.3.2 Аналіз засобів створення програмного забезпечення

Кожен програмний додаток має свій застосунок, це може бути: веб-застосунок, мобільний додаток чи віконний додаток (Windows Forms).

Веб-застосунок являє собою програмний продукт, доступ до якого реалізований через веб-інтерфейс. Створення такого застосунку в основному орієнтована на комерційну організацію або компанію, тому функціональність таких застосунків включає досить потужні інструменти, які орієнтовані на бізнес [40]. Веб-застосунки мають доступ до синхронізації з будь-якою внутрішньою системою організації або ж компанії. Так, доступ до неї може мати не лише той хто працює з нею напряму, а ще й власник фірми чи продавець у компанії.

На базі веб-застосунку можна створити систему, яка буде мати більший функціонал чим внутрішня, до прикладу вона зможе виконувати функції: ведення обліку часу роботи всіх співробітників, виконання обліку вантажних або пасажирських перевезень, моніторинг діяльності компанії, керування роботою персоналу та нараховування заробітної плати.

Основними перевагами такого застосунку над іншими є:

– зміна розмірів та масштабів веб-системи;

– інтеграція з іншими системами;

– необмежений доступ до функціоналу кількох особам;

– зручне обслуговування системи.

Мобільний додаток - це система яка використовується на таких пристроях як мобільні телефони, планшети, які створені на платформі Android або IOS [41]. Так, як час який користувач проводить у мобільному телефоні щодня збільшується, такий тип застосунку є актуальним. Такий вид застосунку має свої переваги для користувачів:

– швидкий доступ до інформації;

– персоналізація, яка дозволяє в додатку зберегти свої налаштування, дані, систему сповіщення та інші;

– швидкі продажі, в основному актуально для людей які проводять продаж, або купівлю товарів онлайн.

Віконний додаток або форма являє собою додаток створений в середовищі розробки Visual Studio. Такий додаток ще називають десктоп додатком, він вимагає оператора, який буде працювати з програмою. Додатки створені на такому застосунку не вимагають підключення до інтернету, мають більш високу швидкодію, проте для користування таким додатком кільком користувачам, кожному потрібно завантажити програму на свій комп’ютер, але дані які кожен буде змінювати у себе на комп’ютері не будуть змінюватись у іншого користувача [42].

Отже, проаналізувавши вищезгадані типи застосунків для розробки інформаційної системи автоматизованої тематичної класифікації коротких текстових повідомлень було обрано віконний застосунок (Windows Form).

2.3.3 Вибір мови програмування

Для розробки інформаційної системи було обрано мову програмування C#. Ця мова використовує об'єктно-орієнтований підхід до програмування. Такий підхід дозволяє працювати з даними за принципом чорного ящика, тобто, системі не потрібно запам'ятовувати усю інформацію [43]. У мові присутня велика кількість конструкцій, які спрощують написання коду.

Одним з найбільших плюсів мови те, що у ній є велика кількість бібліотек та шаблонів, що пришвидшує розробку додатків. Мова має безпечну типізацію під платформу .NET. Синтаксис C# близький до C++ і Java.

Мова має строгу статичну типізацію, також, підтримує поліморфізм, перевантаження операторів, вказівники на функції класів, атрибути, події, властивості, винятки, коментарі у форматі XML.

Завдяки базуванні на мовах попередниках C# виключає певні моделі, що показали себе як проблематичні при розробці програмних систем: тому, C# не

підтримує множинне спадкування класів, як це робить мова програмування C++ або виведення типів як це робить Haskell [44].

Таким чином, для реалізації програмного додатку для автоматизованої тематичної класифікації коротких текстових повідомлень було обрано мову програмування C#.

2.3.4 Вибір фреймворку

Для сторення та програмної реалізації програмного продукту було обрано фреймворк - .NET. Основа платформи її багатомовне середовище використання - Common Language Runtime (CLR). Функціонал фреймворку доступний на будь-якій мові яку він підтримує [45]. Також, фреймворк дозволяє об'єднувати служби, що були написані різними мовами.

Кожна збірка в фреймворку має свій код – версію, для того щоб у програмного продукту не викликало конфлікту, версія фреймворку має бути однакова.

Використання фреймворку значно спрощує розробку інформаційної системи, завдяки перевагам [46]:

– автоматичний збирач сміття;

– асинхронна модель виконання коду;

– великий функціонал для створення інтерфейсу програми;

– простота використання;

– зручність підключення до бази даних.

– велика кількість підходів та технологій для розробки десктопних додатків;

– масштабна інформаційна складова, додаткова інформація про фреймворк.

Отже, для реалізації програмного додатку для автоматизованої тематичної класифікації коротких текстових повідомлень було обрано .NET фреймворк.

2.3.5 Вибір СКБД

Для роботи з даними інформаційної системи автоматизованої тематичної класифікації коротких текстових повідомлень необхідно використовувати систему керування базами даних, для даної розробки було обрано СКБД Microsoft SQL Server.

Дана СКБД спрощує передачу, додавання, обробку, та видалення даних, таких як тексти повідомлень, статей та іншої інформації. Система дозволяє додавати як структуровані так і неструктуровані дані [47]. СКБД Microsoft SQL Server має такі переваги:

– висока продуктивність;

– підтримка швидкої постійної пам’яті;

– висока безпека даних;

– гнучкість у виборі платформи, засобів доставки та мови програмування;

– інтелектуальний аналіз даних.

СКБД SQL Server ідеально підходить для операційної системи Windows. Крім досить широкого набору програмних засобів для розробки й адміністрування, актуальність серверу надає наявність високоінтелектуального процесора запитів і розвиненого діалекту мови SQL (Transact-SQL) [48].

Отже, для реалізації програмного додатку методу автоматизованої тематичної класифікації коротких текстових повідомлень було обрано СКБД Microsoft SQL Server.

Таким, чином, для створення інформаційної системи автоматизованої тематичної класифікації коротких текстових повідомлень було обрано таку комбінацію засобів розробки: мову програмування C#, платформу .NET, СКБД Microsoft SQL Server та середовище розробки Visual Studio 2019.

Розділ 3 Програмна реалізація інформаційної системи

3.1 Структура та функціональне призначення програмних складових системи

Структура системи складається з бази даних та відповідних сторінок, на яких виконуються функції системи. Структура інформаційної системи для автоматизованої тематичної класифікації коротких текстових повідомлень зображена на рисунку 3.1.

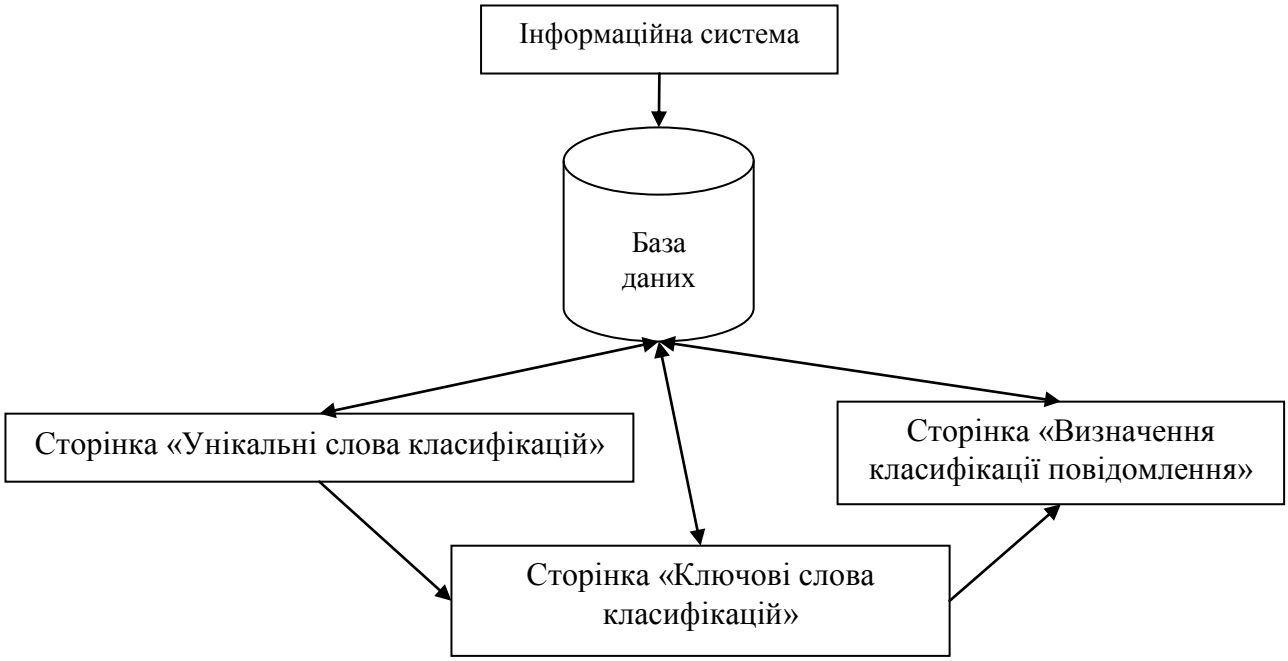


Рисунок 3.1 – Структура інформаційної системи автоматизованої тематичної класифікації коротких текстових повідомлень

Сторінка «Унікальні слова класифікацій» виконує функції обробки тексту, очищення його від усіх знаків окрім слів та визначає для кожного тегу (класифікації) унікальні слова. Текст який відправляється на обробку береться з бази даних інформаційної системи.

Сторінка «Ключові слова класифікацій» виконує функції визначення ключових слів, за допомогою визначення значень семантичної важливості

кожного унікального слова, та за допомогою альтернативних текстів, які беруться з бази даних.

Сторінка «Визначення класифікації повідомлення» виконує функції обробки тестового тексту повідомлення, визначення приналежності кожного ключового слова тестового тексту до слів альтернативних текстів усіх тегів та функцію визначення тегу тестового тексту повідомлення.

Інформаційна система автоматизованої тематичної класифікації коротких текстових повідомлень реалізована за допомогою класів, структура яких представлена на рисунку 3.2.

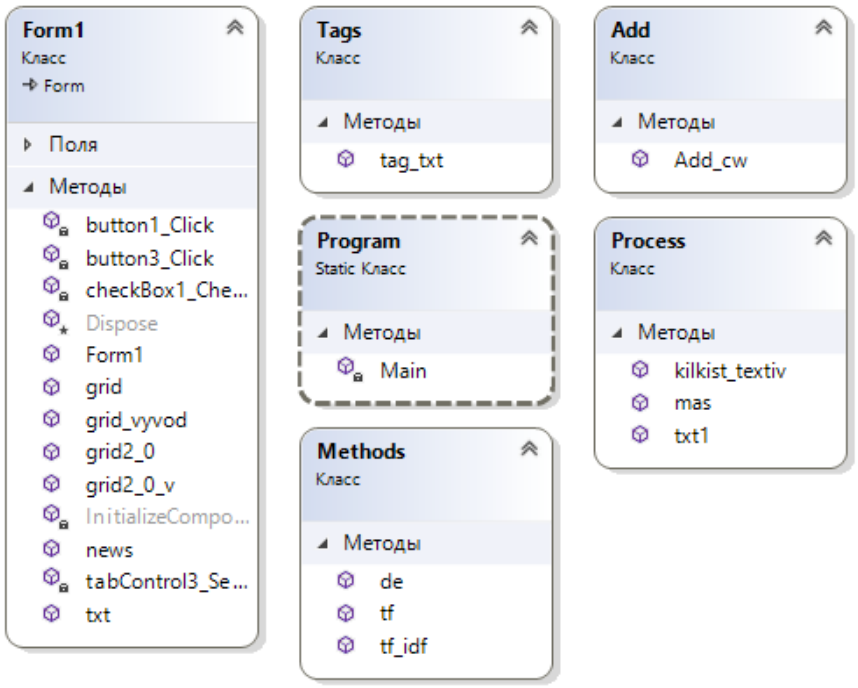


Рисунок 3.2 – Діаграма класів інформаційної системи

Робота інформаційної системи розпочинається з класу «Program», який вміщує в собі метод «Main». Клас «Form1» - форма на якій відбуваються усі маніпуляції в інформаційній системі, вона вміщує в собі методи що виводять дані на екран та дістають їх з бази даних.

Клас «Process» включає в себе методи, які обробляють вхідні тексти коротких текстових повідомлень системи, альтернативні тексти та тестовий

текст, забирає розділові знаки, знаки пунктуації, цифри та ін. Також, даний клас визначає унікальні слова в текстах.

Клас «Methods» вміщує методи, які обраховують значення DE, TF, IDF та TF-IDF. Класи «Add» та «Tags» складається з методів, які визначають тег тестового тексту, порівнюють унікальні слова тестового тексту з унікальними словами альтернативних текстів усіх тегів.

Таким чином, було реалізовано структуру інформаційної системи автоматизованої тематичної класифікації коротких текстових повідомлень, яка представлена за допомогою схеми та діаграми класів.

3.2 Особливості реалізації програмних складових системи

Інформаційна система автоматизованої тематичної класифікації коротких текстових повідомлень має забезпечувати виконання функцій очищення тексту від знаків, цифр та інших символів, окрім слів.

Виконання даної функції відбувається за допомогою наступного програмного коду:

```
public string[] txt1(string text, string[] mas, string text_new)
        {
            string[] mas_chys = { "0", "1", "2", "3", "4", "5", "6", "7", "8", "9", "\n", "\r", "+", "°" };
            int i, j, k = 0;

            for (i = 0; i < text.Length; i++)
            {
                if (Char.IsPunctuation(text[i]))
                {
                    k = 1;
                }
                else
                {  for (j = 0; j < mas_chys.Length; j++)
                    {
                        if (Convert.ToString(text[i]) == mas_chys[j])
                        {
                            k = 1;
                            break;
                        } }

                        for (j = 0; j < mas.Length; j++)
                        {

                            if (text[i] == Convert.ToChar(mas[j])) { k = 0; }
                        }
                }

                if (k == 0)
                {
                    text_new += Convert.ToString(text[i]);

                }
                k = 0;
            }

            return text_new.Split(' ');
        }
```

Обрахунок методів TF та TF-IDF для унікальних слів класифікацій відбувається за допомогою наступного програмного коду:

```
///TF
        public double[] tf(string[] nowy_mas, int[] mas_kilk, int zagal_kilk, double[] chast)
        {

            for (int i = 0; i < nowy_mas.Length; i++)
            {
                chast[i] = Convert.ToDouble(mas_kilk[i]) / zagal_kilk;
            }
            return chast;
        }
        ///TF-IDF
        public double[] tf_idf(string[] nowy_mas, int[] kilk_text, double[] idf)
        {

            for (int i = 0; i < nowy_mas.Length; i++)
            {
                idf[i] = Math.Log((7 / (kilk_text[i] + 1)));

            }
            return idf;
        }
```

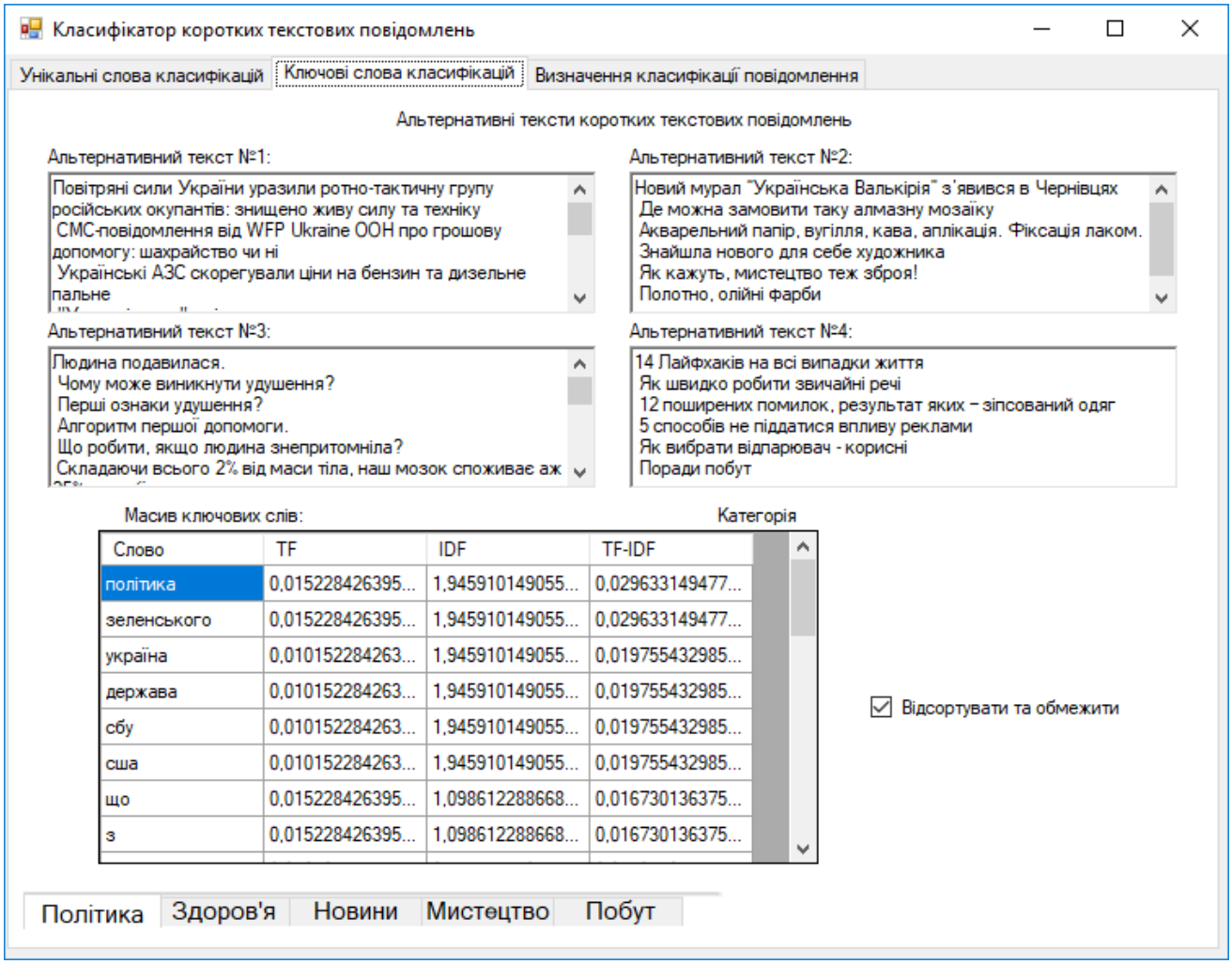


Рисунок 3.3 - Масив ключових слів класифікації «Політика»

Результат обрахунку методів виведених у таблицю зображено на рисунку 3.3.

Таким чином, за допомогою програмних кодів будується інформаційна система методу автоматизованої тематичної класифікації коротких текстових повідомлень, яка виконує усі необхідні функції.

3.3 Тестування інформаційної системи

Для коректності виконання усіх функцій інформаційної системи автоматизованої тематичної класифікації коротких текстових повідомлень, необхідно провести тестування. Тому, було розроблено чотири тест-кейси.

Перший випадок перевіряє коректність відображення масиву коротких текстових повідомлень відповідної класифікації (таблиця 3.1). Після запуску програми на вкладці «Унікальні слова класифікації» у полі «Масив коротких текстових повідомлень» має відображатись масив що відповідає відкритій класифікації, тобто «Політика». Після вибору іншої класифікації текст у полі має змінитись на текст відповідного тегу (рисунок 3.4 – 3.5).

Таблиця 3.1 – Тест-кейс AT0001

<table>
<tr><td>Тест-кейс ID: AT0001</td><td>Пріоритет: 1</td><td>Створено: 28.05.2022, О.В. Здоровик</td></tr>
<tr><td colspan="3">Назва: Перевіряється коректність відображення масиву коротких текстових повідомлень
Вхідні дані: Класифікація = «Політика», «Здоров’я»</td></tr>
<tr><td colspan="2">Кроки</td><td>Очікуваний результат</td></tr>
<tr><td colspan="2">1. Запустити додаток
2. Обрати класифікацію «Політика»
3. Порівняти фактичний результат з очікуваним
4. Обрати класифікацію «Здоров’я»
5. Порівняти фактичний результат з очікуваним</td><td>Відображення коректних даних.</td></tr>
<tr><td colspan="3">Результат виконання тест-кейсу: пройдено успішно</td></tr>
</table>

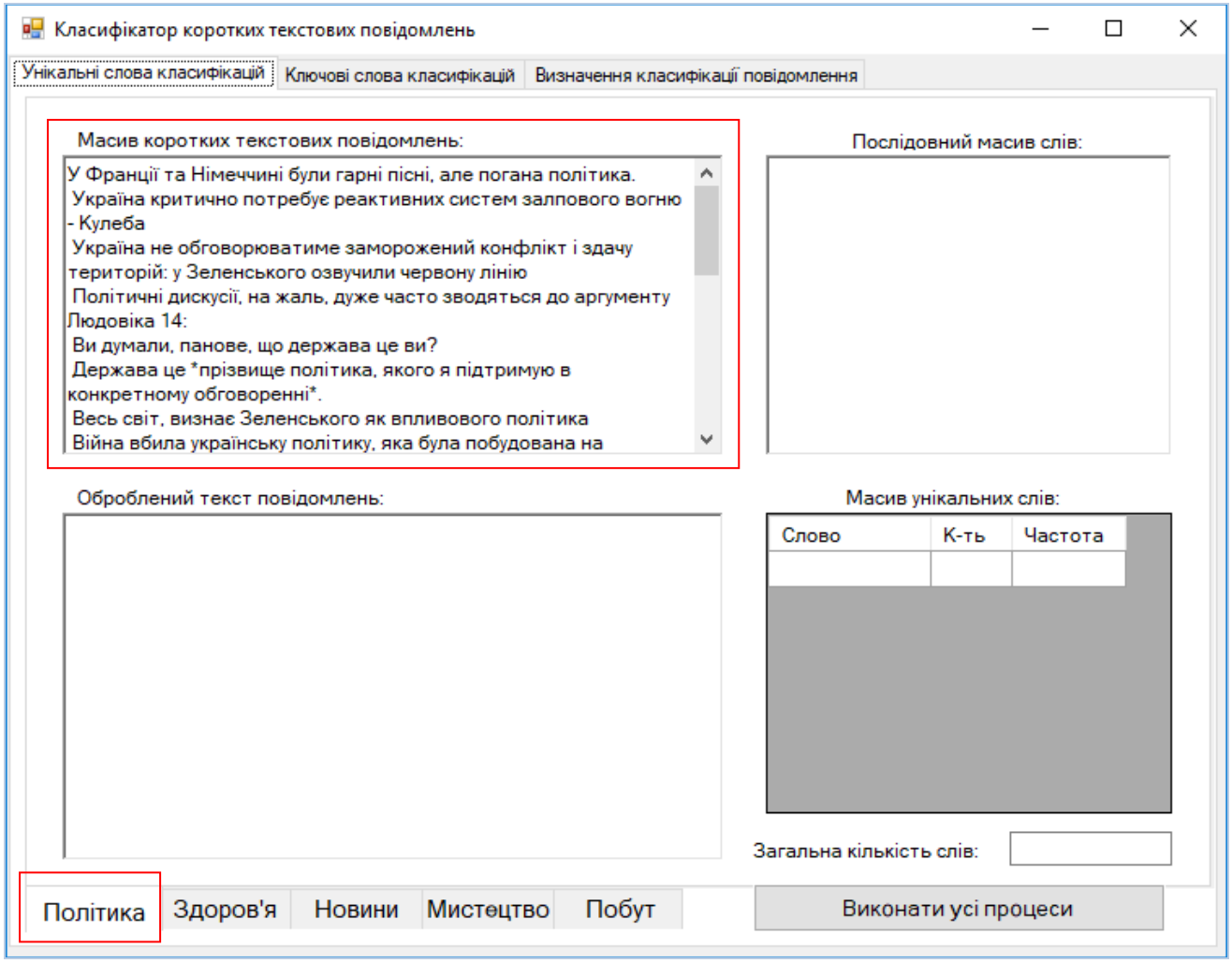


Рисунок 3.4 – Масив коротких текстових повідомлень тегу «Політика»

Другий випадок перевіряє коректність обробки масиву коротких текстових повідомлень (таблиця 3.2). Після натискання на кнопку «Виконати усі процеси» відбувається обробка текстів та виведення результатів у відповідні поля. При зміні класифікації мають змінюватись дані у послідовному та унікальному масивах слів, обробленому тексті повідомлень та загальна кількість слів (рисунок 3.6-3.7).

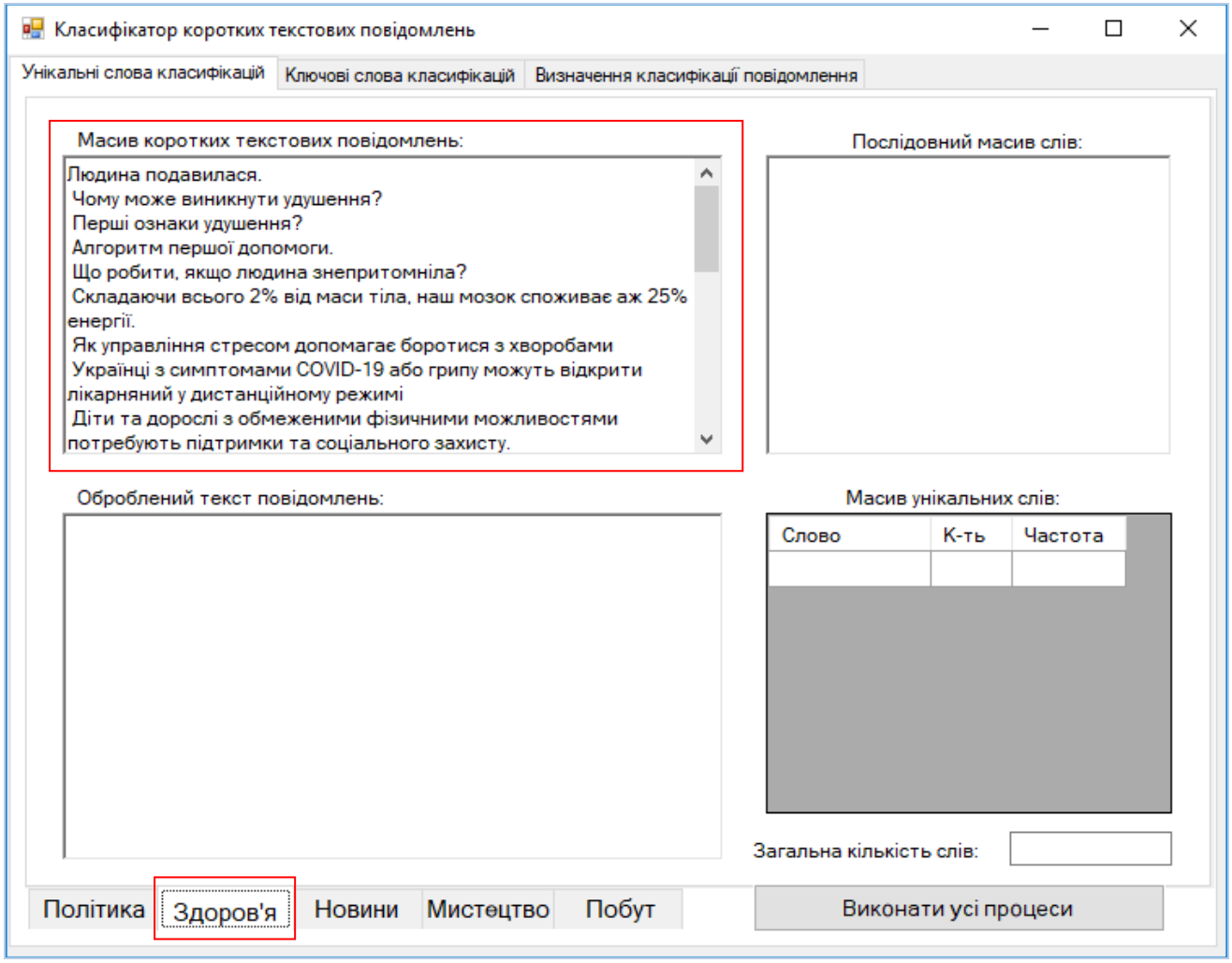


Рисунок 3.5 – Масив коротких текстових повідомлень тегу «Здоров’я»

Таблиця 3.2 – Тест-кейс AT0002

Тест-кейс ID: AT0002	**Пріоритет:** 1	**Створено:** 28.05.2022, О.В. Здоровик
Назва: Перевіряється коректність відображення оброблених данних **Вхідні дані:** Класифікація = «Політика», «Здоров’я»		
Кроки		**Очікуваний результат**
1. Запустити додаток 2. Натиснути кнопку «Виконати усі процеси» 3. Обрати класифікацію «Політика» 4. Порівняти фактичний результат з очікуваним 5. Обрати класифікацію «Здоров’я» 6. Порівняти фактичний результат з очікуваним		Відображення коректних даних.
Результат виконання тест-кейсу: пройдено успішно		

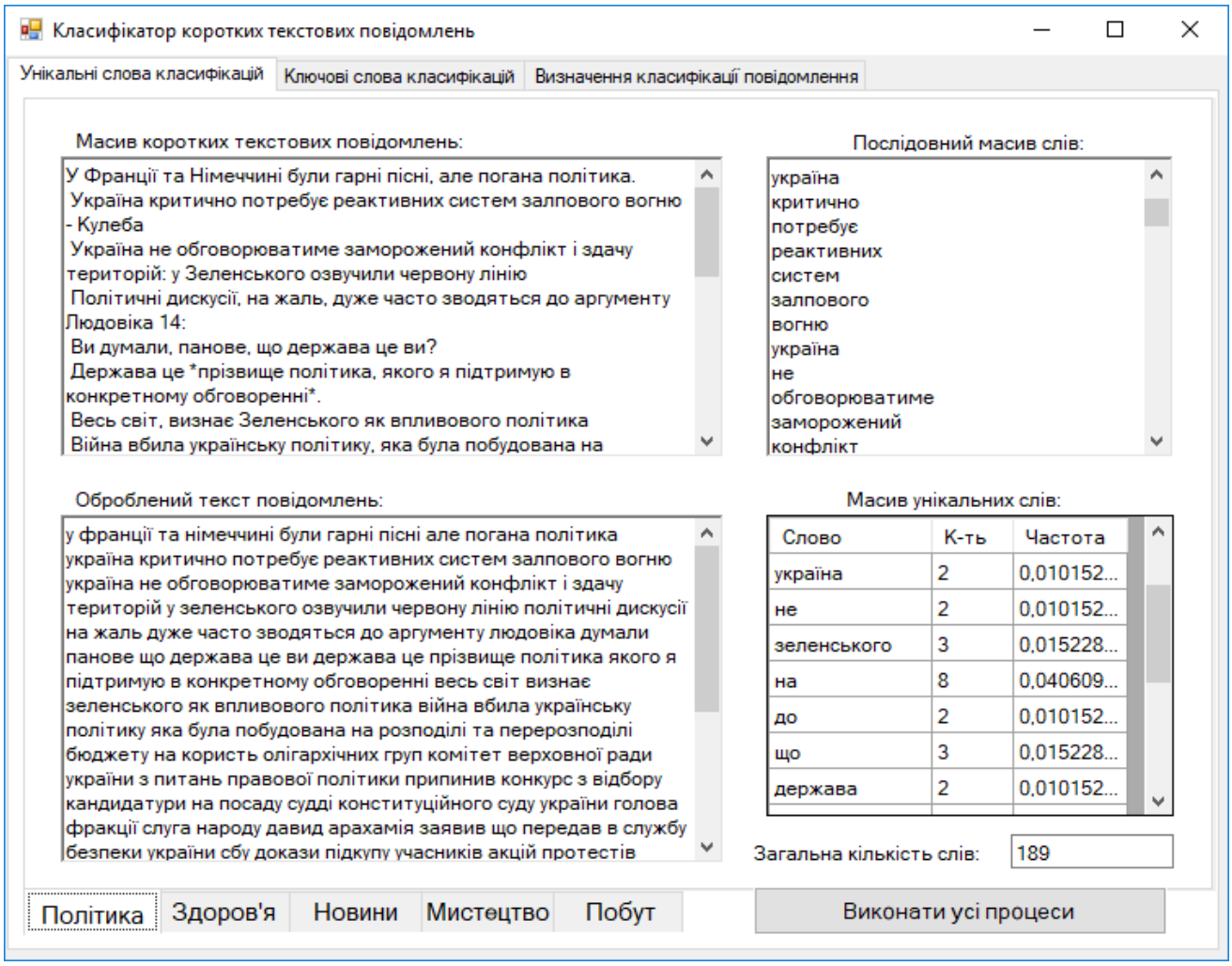


Рисунок 3.6 – Відображення оброблених даних класифікації «Політика»

Третій тестовий випадок перевіряє коректність відображення альтернативних текстів та масиву ключових слів класифікацій (таблиця 3.3). Після натискання на вкладку «Ключові слова класифікацій» та натискання на необхідний тег, мають змінюватись альтернативні тексти, що відповідають обраній класифікації, та масив ключових слів із значеннями TF, IDF та TF-IDF (рисунок 3.8-3.9).

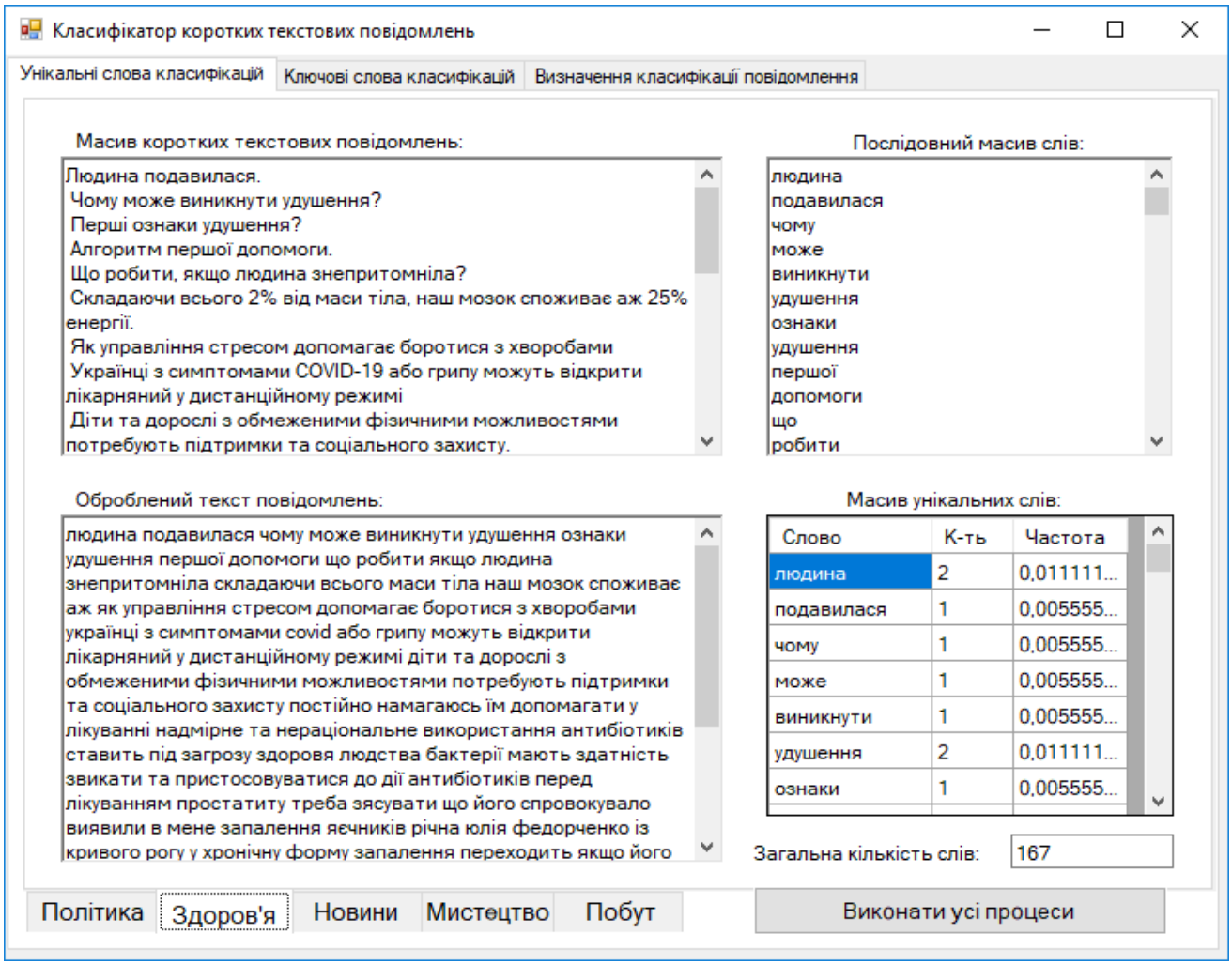


Рисунок 3.7 – Відображення оброблених даних класифікації «Здоров'я»

Таблиця 3.3 – Тест-кейс AT0003

Тест-кейс ID: AT0003	**Пріоритет:** 1	**Створено:** 28.05.2022, О.В. Здоровик
Назва: Перевіряється коректність відображення альтернативних текстів та масиву ключових слів **Вхідні дані:** Класифікація = «Політика», «Новини»		
Кроки		**Очікуваний результат**
1. Запустити додаток 2. Натиснути кнопку «Виконати усі процеси» 3. Перейти на вкладку «Ключові слова класифікацій» 4. Обрати класифікацію «Політика» 5. Порівняти фактичний результат з очікуваним 6. Обрати класифікацію «Новини» 7. Порівняти фактичний результат з очікуваним		Відображення коректних даних.
Результат виконання тест-кейсу: пройдено успішно		

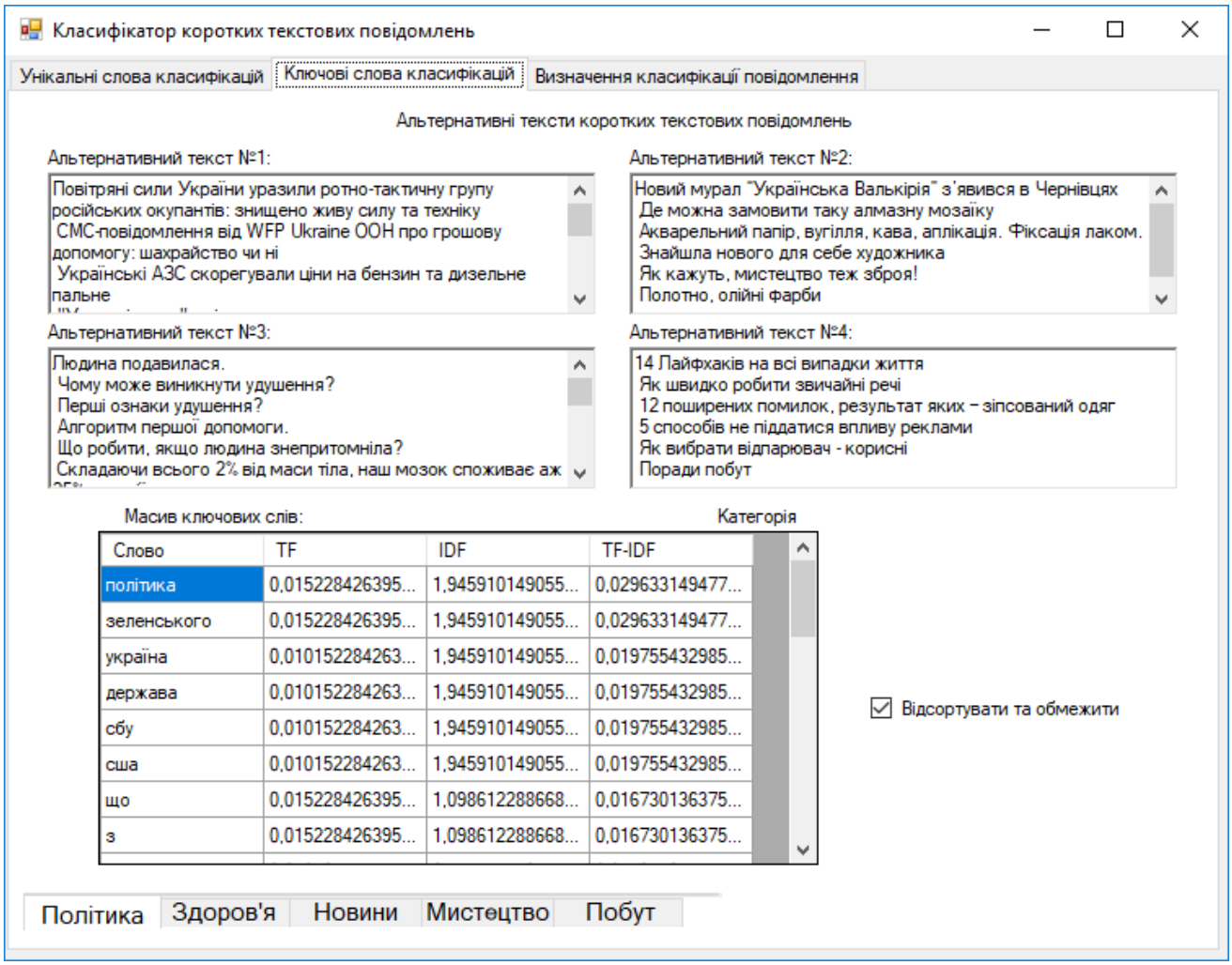


Рисунок 3.8 - Відображення альтернативних текстів та масиву ключових слів класифікації «Політика»

Останній тест-кейс перевіряє коректність аналізу тестового тексту коротких текстових повідомлень (таблиця 3.4). Після переходу на вкладку «Визначення класифікації повідомлення» та натискання на кнопку «Проаналізувати тестовий текст» має проводитись аналіз тестового тексту, порівняння унікальних слів зі словами усіх класифікацій та виведення оцінки та відсотку відповідності до тегів (рисунок 3.10). За приклад було взято текст який відповідає класифікації – «Новини».

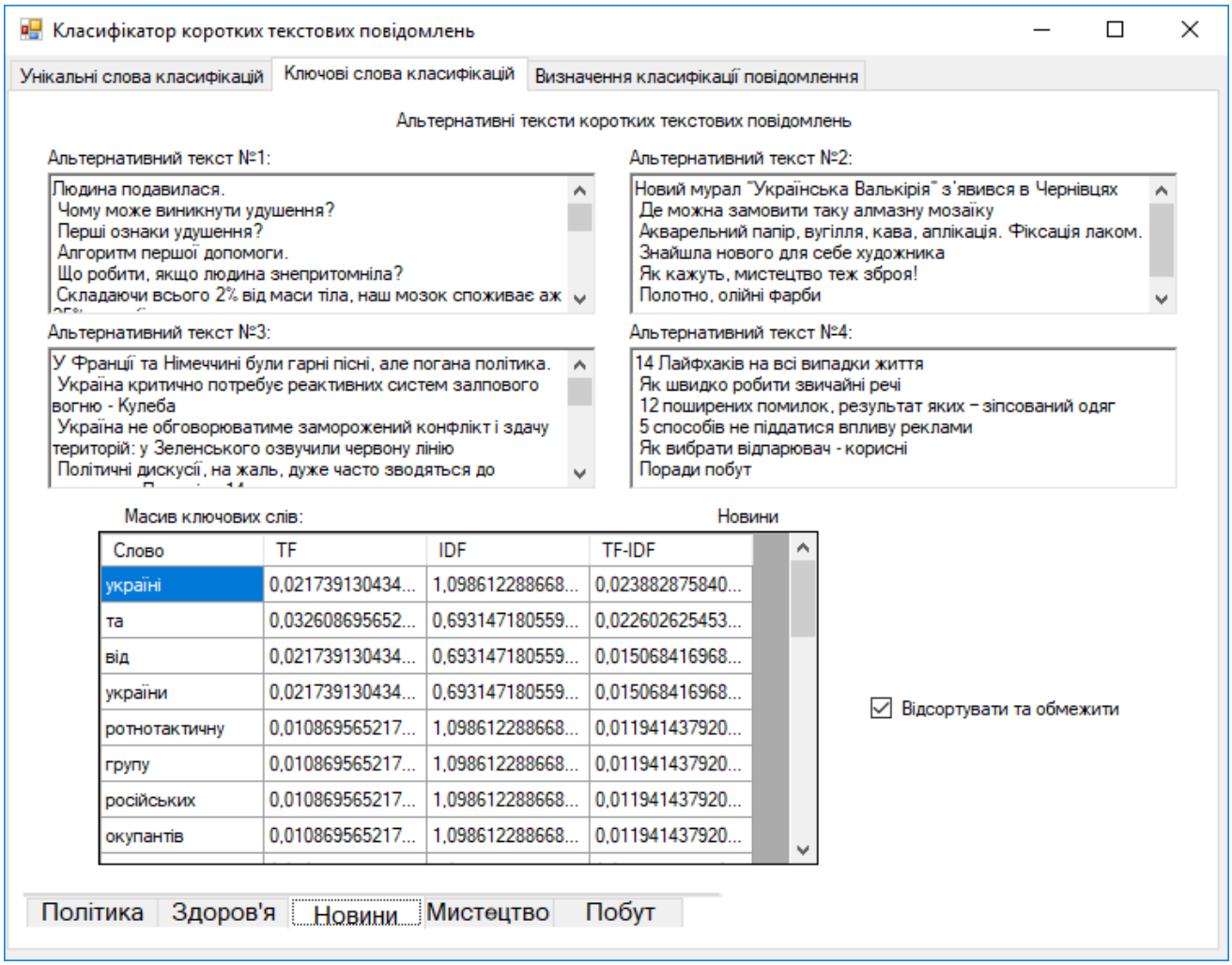


Рисунок 3.9 - Відображення альтернативних текстів та масиву ключових слів класифікації «Новини»

Таблиця 3.4 – Тест-кейс AT0004

Тест-кейс ID: AT0004	**Пріоритет:** 1	**Створено:** 28.05.2022, О.В. Здоровик
Назва: Перевіряється коректність аналізу тестового тексту **Вхідні дані:** Класифікація = «Новини», тестовий текст		
Кроки		**Очікуваний результат**
1. Запустити додаток 2. Натиснути кнопку «Виконати усі процеси» 3. Перейти на вкладку «Визначення класифікації повідомлення» 4. Натиснути на кнопку «Проаналізувати тестовий текст» 5. Порівняти фактичний результат з очікуваним		Відображення коректних даних.
Результат виконання тест-кейсу: пройдено успішно		

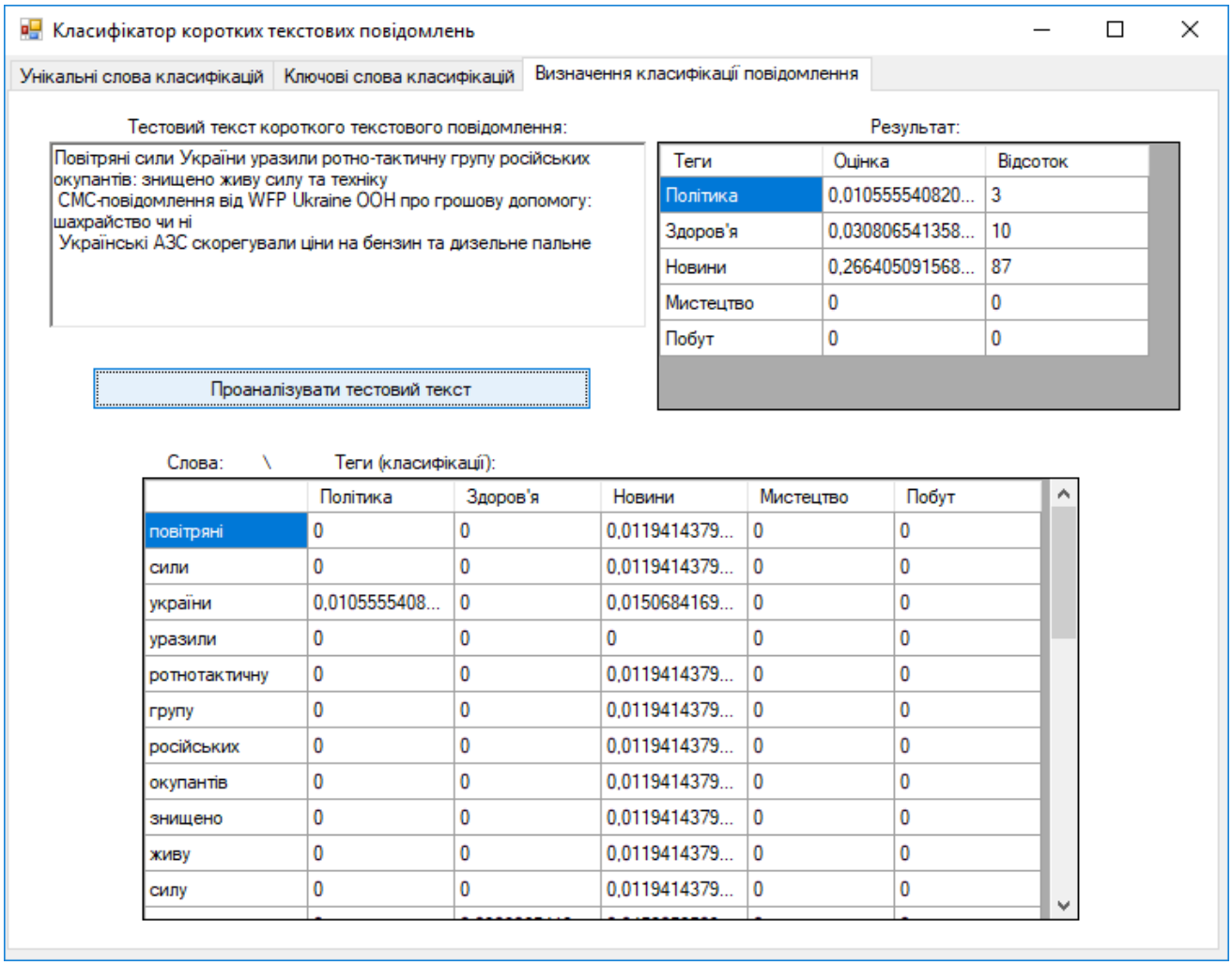


Рисунок 3.10 – Відображення аналізу тестового тексту категорії «Новини»

Отже, було проведено дослідження коректності виконання функцій системи. Тестування завершились успішно. Наочно було показано, що створений функціонал реалізовано у відповідності до поставлених задач.

3.4 Інструкція користувача

Робота користувача з інформаційною системою автоматизованої тематичної класифікації коротких текстових повідомлень розпочинається з початкової сторінки «Унікальні слова класифікацій». На ній користувач має можливість, переглянути масив коротких текстових повідомлень, який додається з бази даних та змінюється відповідно до обраного тегу (класифікації) (рисунки 3.11–3.12).

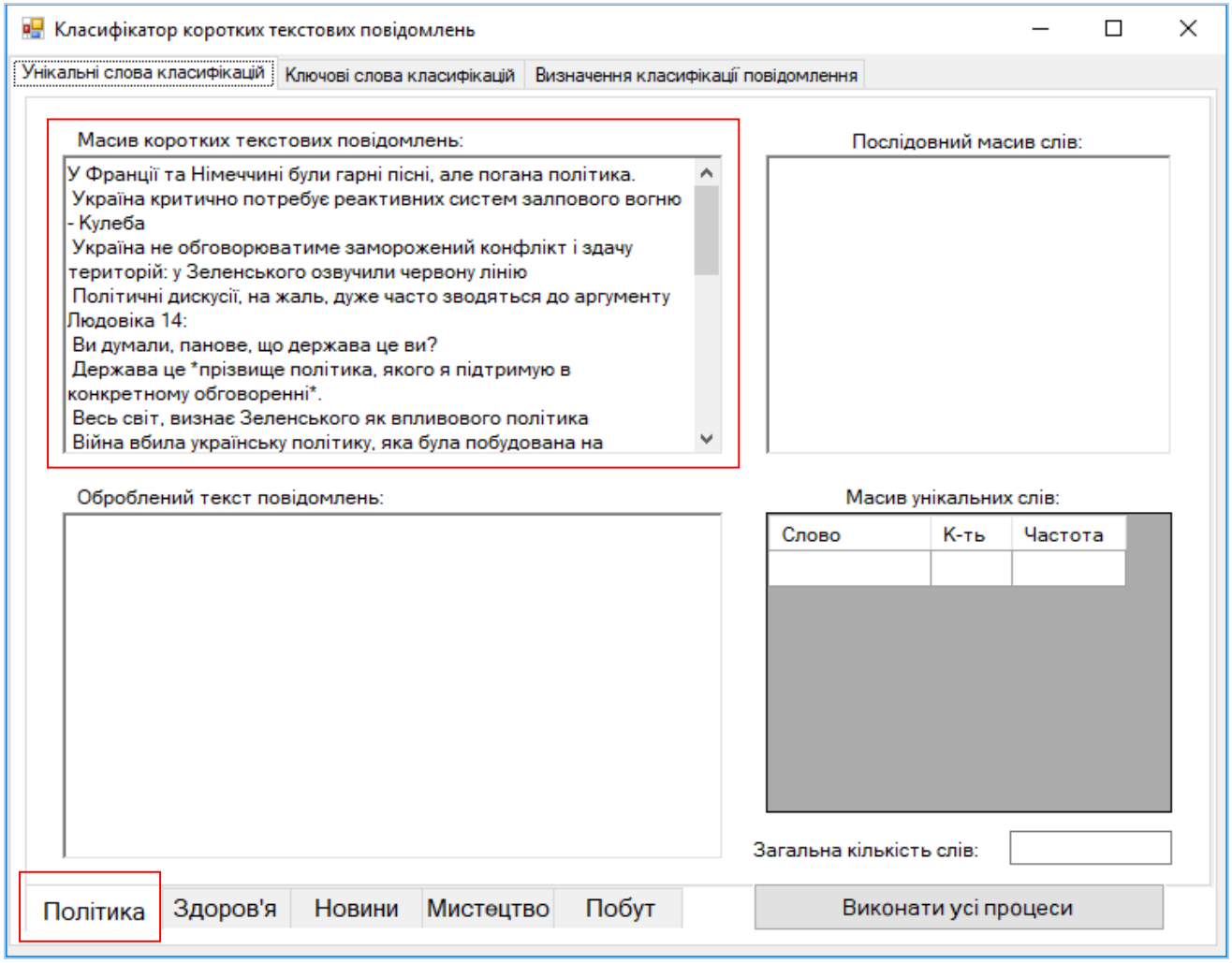


Рисунок 3.11 – Масив коротких текстових повідомлень тегу «Політика»

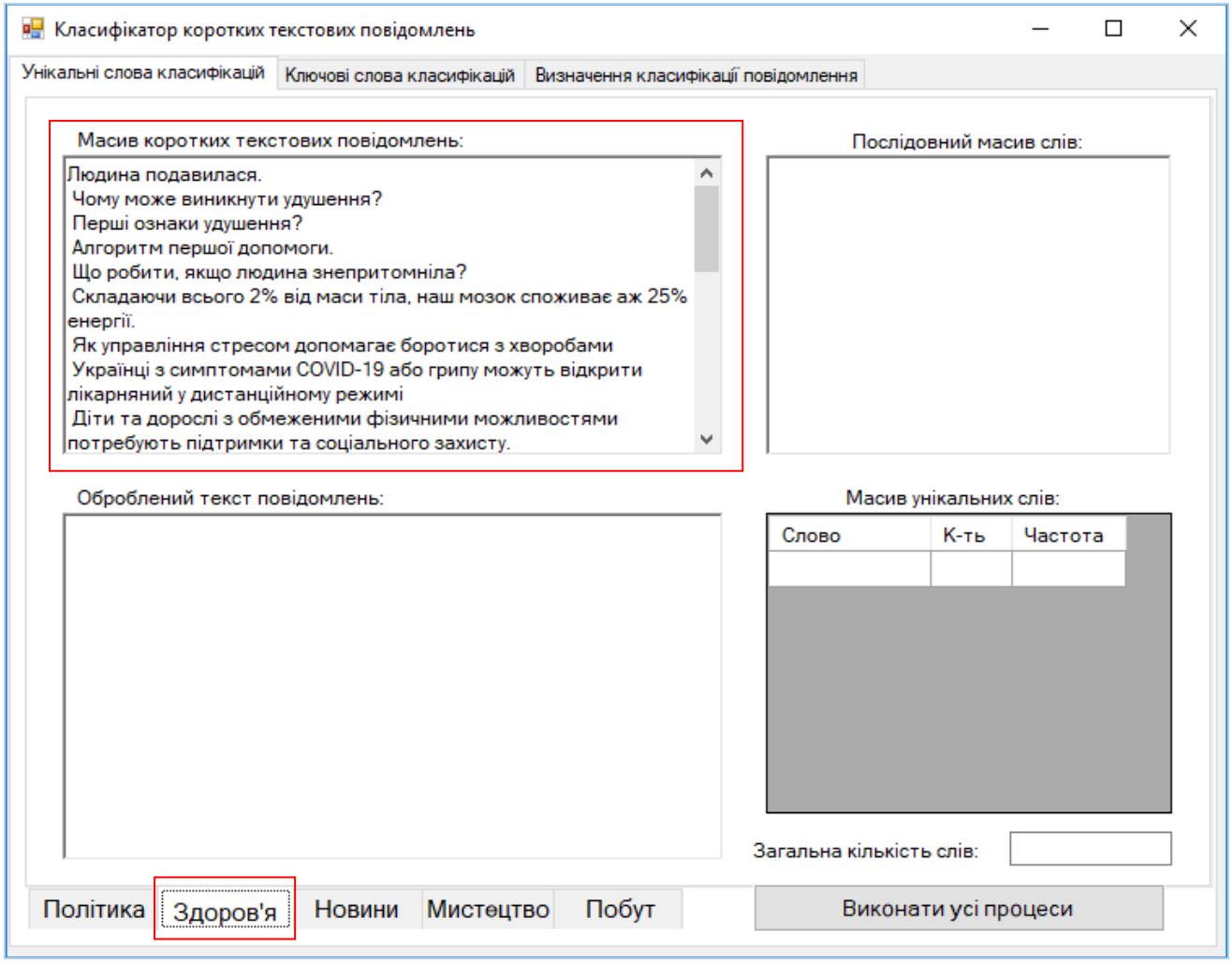


Рисунок 3.12 – Масив коротких текстових повідомлень тегу «Здоров’я»

Після натискання на кнопку «Виконати усі процеси», відбувається основна обробка тексту та визначення необхідний значень для унікальних слів текстів. Під час обробки тексту визначається послідовний масив слів (рисунок 3.13), масив унікальних слів (рисунок 3.14), загальна кількість слів та формується оброблений текст.

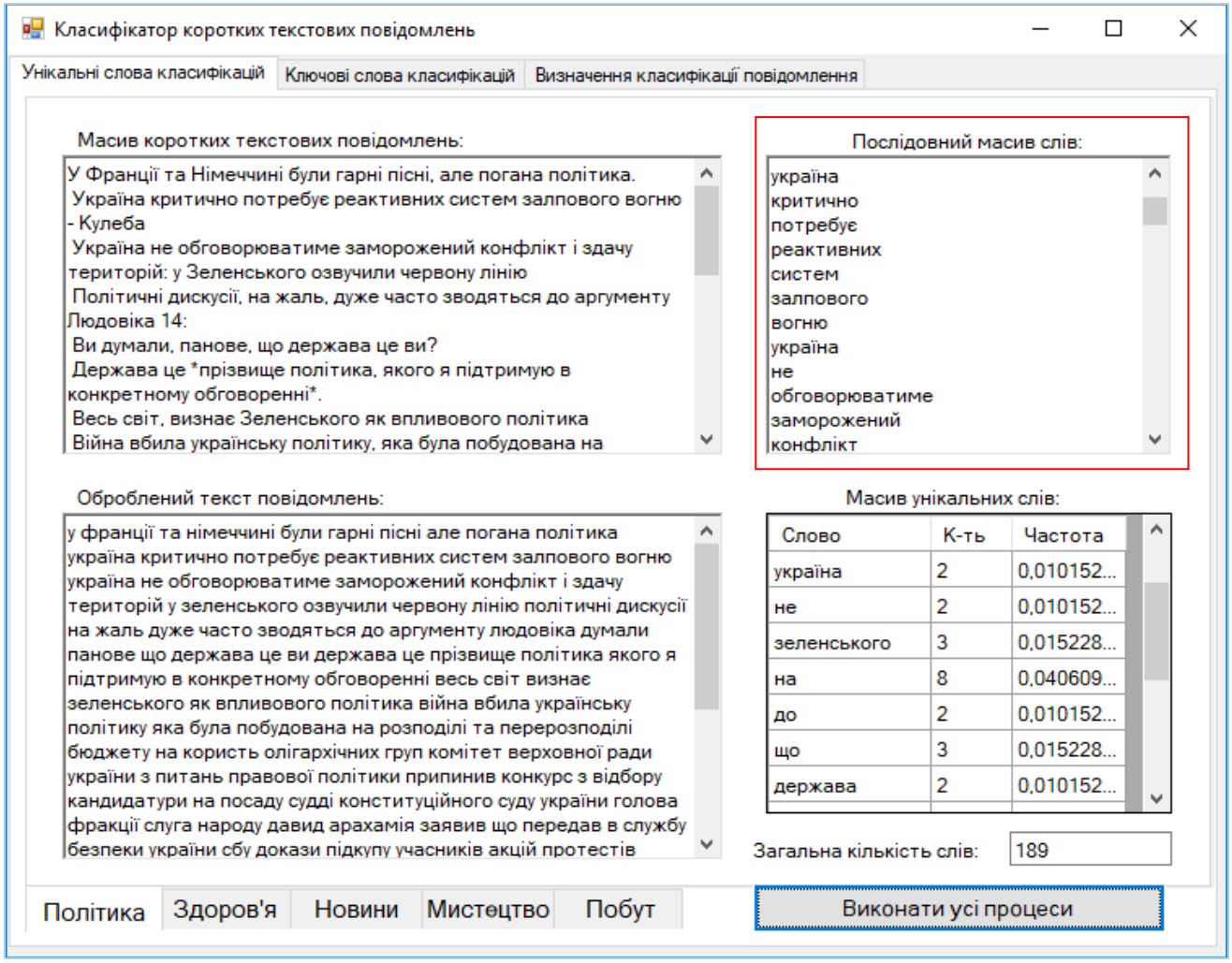


Рисунок 3.13 – Масив послідовних слів тексту

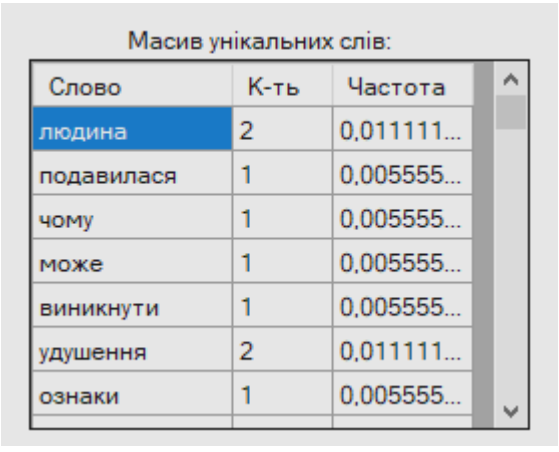


Рисунок 3.14 – Масив унікальних слів тексту

Наступна вкладка «Ключові слова класифікацій» проводить визначення ключових слів тексту, за допомогою альтернативних текстів (рисунок 3.15), які

беруться з бази даних інформаційної системи, та обрахунку значень TF-IDF для унікальних слів класифікацій.

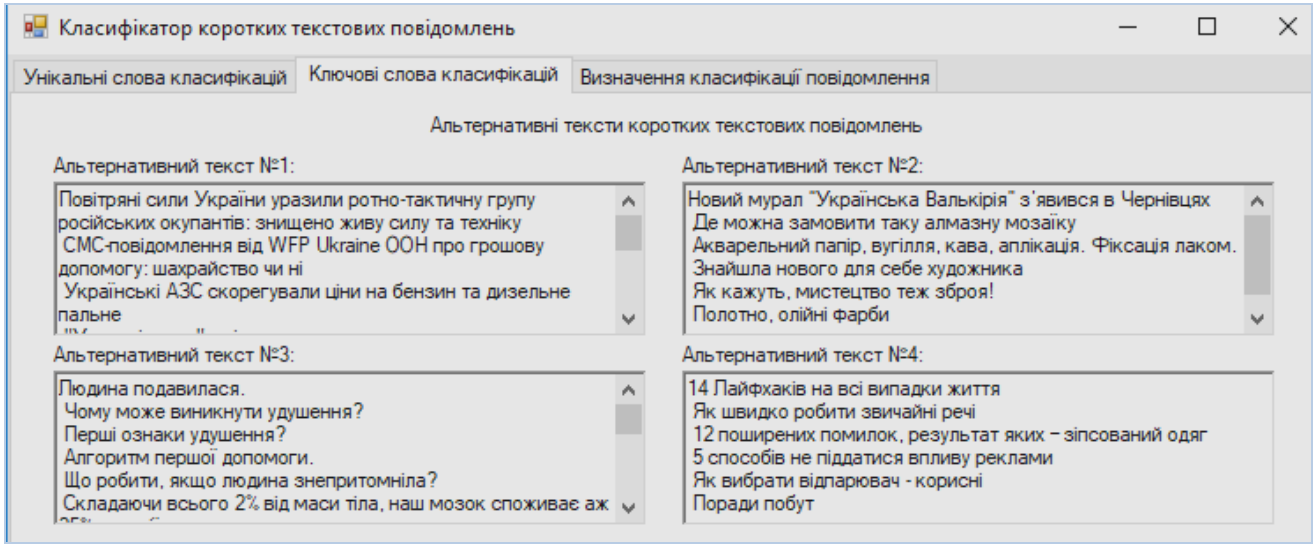


Рисунок 3.15 – Альтернативні тексти класифікації «Політика»

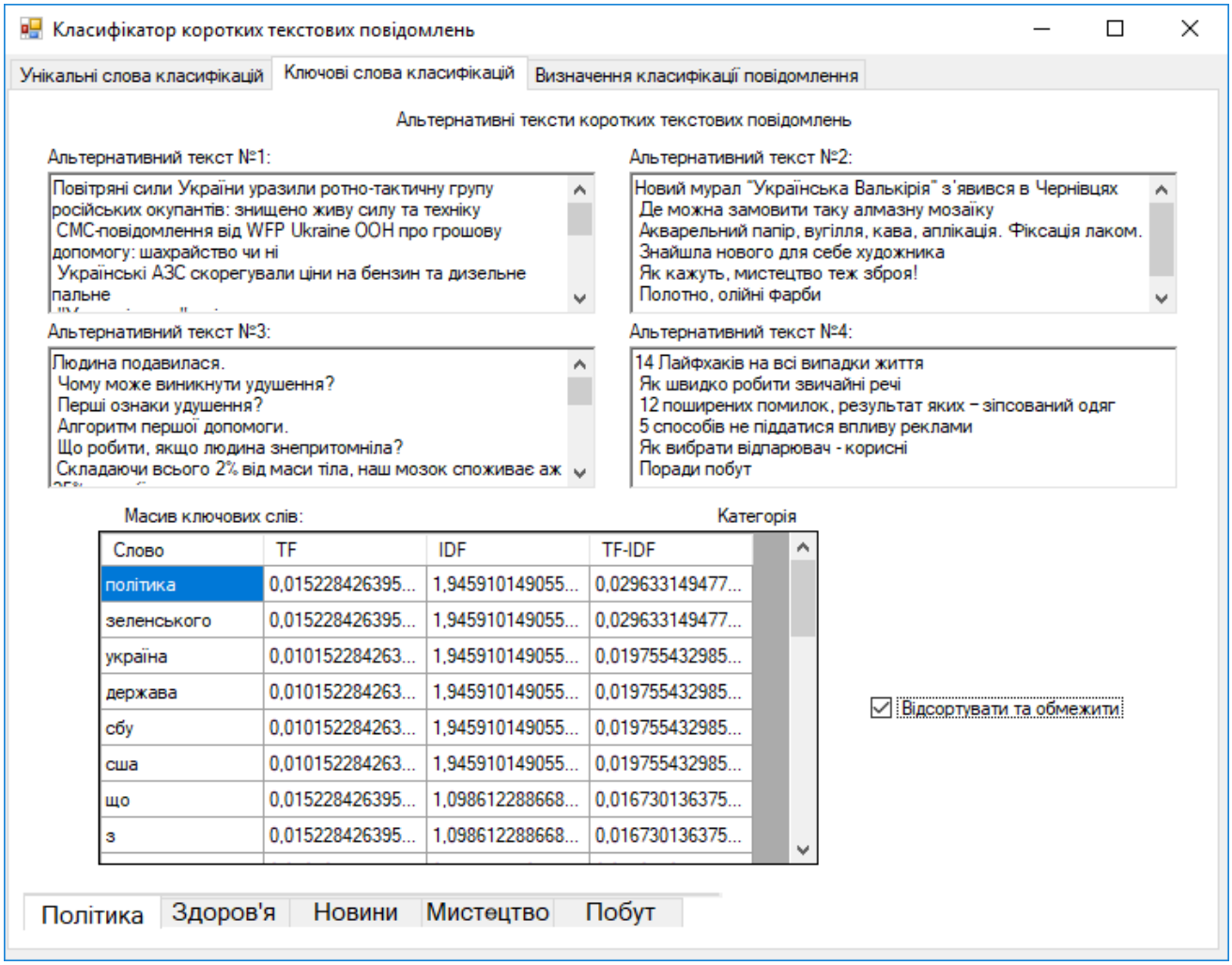


Рисунок 3.16 – Сортування та обмеження ключових слів

Після обрахування значень TF-IDF, по натисканню на відповідний чекбокс відбувається сортування значень та обмеження їх у кількості (рисунок 3.16)

Вкладка «Визначення класифікації повідомлення» дозволяє користувачу додавати тестовий текст повідомлення, або знаходити тег для тестового повідомлення з бази даних, обраховувати його приналежність до інших тегів.

Визначення приналежності унікальних слів тестового повідомлення до тегів зображено на рисунку 3.17. Результат приналежності тексту до тега зображено на рисунку 3.18.

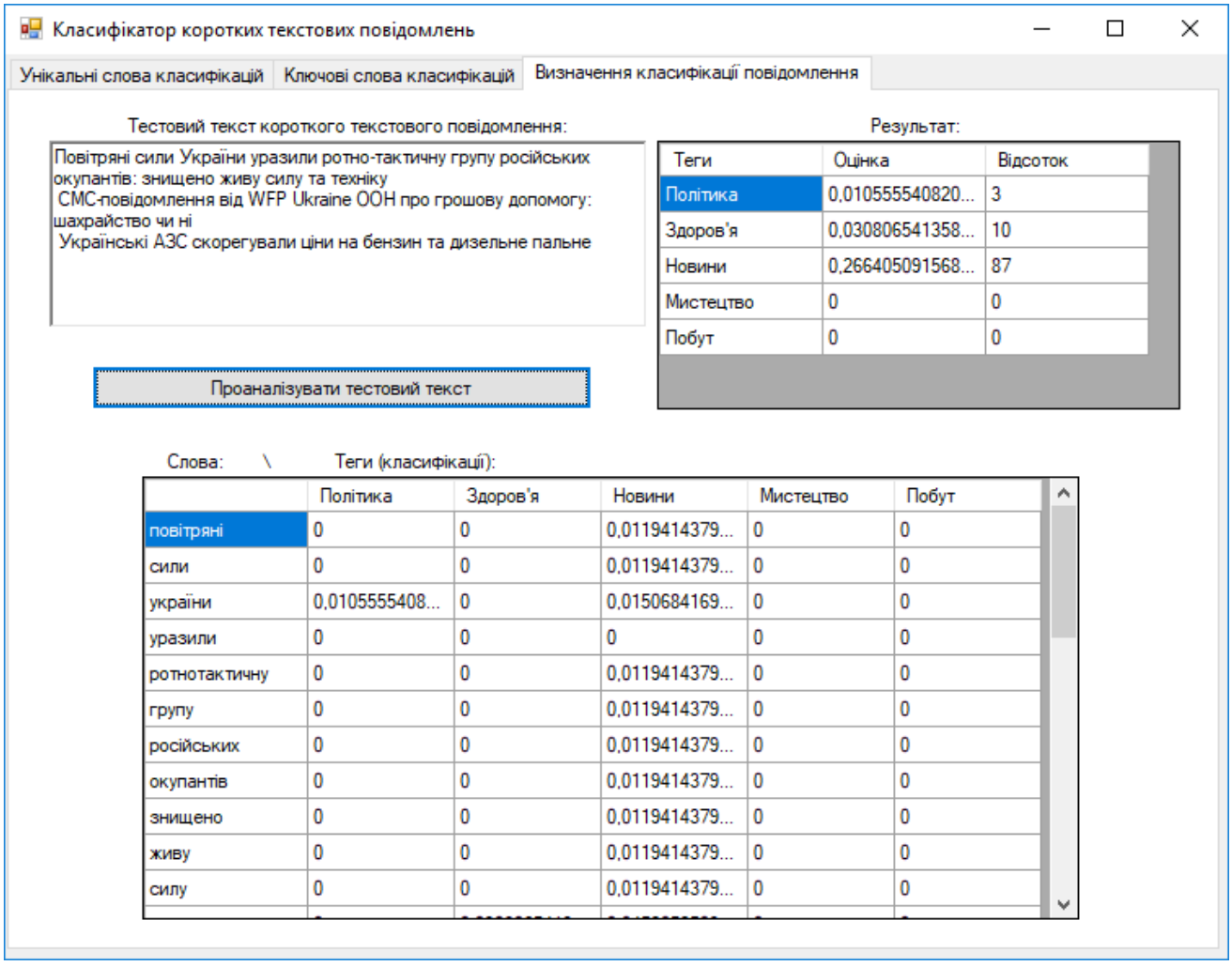


Рисунок 3.17 – Приналежність унікальних слів тестового повідомлення до тегів

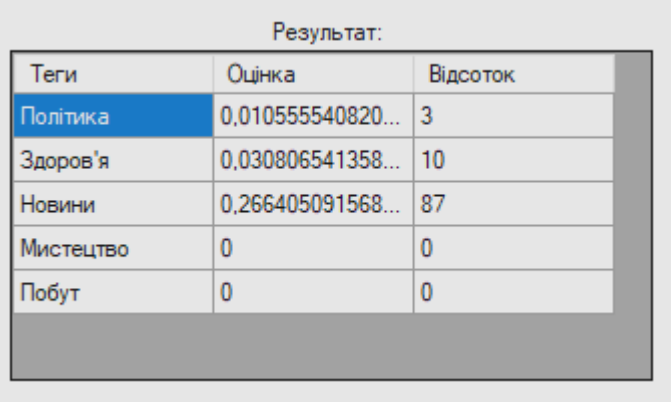


Рисунок 3.18 – Приналежність тексту тестового повідомлення до тегів

Отже, робота користувача з системою є досить проста та зрозуміла, що робить дану інформаційну систему зручною у користуванні.

3.5 Вимоги до розгортання інформаційної системи

Вимоги до апаратних засобів:

1.Процесор: AMD, Intel

2.Тактова частота процесора - 1.8 GHz +

3.Об’єм оперативної пам’яті - 1024 Mb +

4.Операційна система: Windows XP,7

Вимоги до програмних засобів:

– Фреймворк: .NET Framework 3.5;

– Visual Studio 2019.

Висновки

Кваліфікаційна робота бакалавра виконує задачу автоматизованої тематичної класифікації коротких текстових повідомлень. Були виконані розробка методу автоматизованої тематичної класифікації коротких текстових повідомлень й розробка відповідної інформаційної системи. При виконанні класифікації було використано метод класифікації, унікальні та ключові слова текстів класифікацій, альтернативні тексти та параметри семантичної важливості слів у текстах.

Під час роботи були поставлені та вирішені такі задачі:

1. Провести аналіз предметної області семантичного аналізу цифрових текстів.
2. Провести аналіз засобів спілкування в Інтернеті.
3. Огляд теоретичних підходів до розв’язання задач подібних до автоматизації формування тематичної класифікації коротких текстових повідомлень.
4. Обрати алгоритм початкової обробки коротких текстових повідолень.
5. Вдосконалити метод автоматизованої тематичної класифікації коротких текстових повідомлень.
6. Розробити інформаційну технологію автоматизованого пошуку та формування масиву ключових слів коротких текстових повідомлень.
7. Провести прикладне дослідження методу автоматизованої тематичної класифікації коротких текстових повідомлень і виконати аналіз результатів.

В результаті роботи було отримано інформаційну систему автоматизованої тематичної класифікації коротких текстових повідомлень, яка виконує наступні основні функції:

- розбивання тексту на слова та додаткові знаки, цифри;
- визначення загальних параметрів тексту: кількість слів, знаків;
- виділення з тексту лише слова, та зменшення регістру тексту;
- формування масиву унікальних слів;

– формування обробленого тексту;

– визначення позиції та частоти зустрічання кожного унікального слова;

– формування альтернативних текстів для кожного тегу;

– обрахунок значень TF-IDF для кожного унікального слова;

– сортування та обмеження значень TF-IDF;

– визначення ключових слів тексту;

– обробка тестового тексту короткого текстового повідомлення;

– визначення ключових слів тестового тексту, обрахунок їх частоти – CW;

– визначення приналежності ключових слів тестового тексту до кожного тегу.

– сортування слів тестового тексту по приналежності до тегів.

В майбутньому розроблену інформаційну систему можна автоматизувати для аналізу не лише по ключовим словам, але і по ключовим словосполученням.

Перелік посилань

1. Кілька найпопулярніших сфер використання інтернету речей URL: https://lpnu.ua/news/kilka-naipopuliarnishykh-sfer-vykorystannia-internetu-rechei

2. О. Мартиняк. Віртуальне спілкування: позитивні та негативні аспекти. 2019. № 2. С. 57–58. URL: http://elartu.tntu.edu.ua/bitstream/lib/30229/2/FVT_2019_Martyniak_O-Virtual_communication_positive_57-58.pdf.

3. Що таке соціальні мережі? URL: https://futurenow.com.ua/shho-take-sotsialni-merezhi-vydy-klasyfikatsiya-bezpeka/

4. Класифікація соціальних мереж. URL: https://core.ac.uk/download/pdf/84825408.pdf

5. Види соціальних мереж URL: https://uk.economy-pedia.com/11035023-types-of-social-networks

6. Значення слова месенджер. URL: https://slovotvir.org.ua/words/mesendzher

7. Найпопулярніші месенджери серед українців. URL: https://pingvin.pro/gadgets/article-gadget/najpopulyarnishi-mesendzhery-sered-ukrayintsiv.html

8. Viber. URL: https://viber.com

9. Facebook Messenger. URL: https://www.messenger.com/

10. Telegram. URL: https://play.google.com/store/apps/details?id=org.telegram.messenger&hl=uk&gl=US

11. Чати. URL: https://sites.google.com/site/sluzbamereziinternet/cati

12. Методи використання вебсайтів в інтегрованому просуванні бізнесу організацій URL: http://repository.kpi.kharkov.ua/bitstream/KhPI-Press/56313/4/EP_2021_2_Zaruba_Metody_vykorystannia.pdf

13. Мікроблогінг URL: https://uk.wikipedia.org/wiki/Мікроблогінг

14. Що таке мікроблогінг і які платформи найкращі? URL: https://ciksiti.com/uk/chapters/10205-what-is-microblogging-and-which-platforms-are-best

15. Що таке Твіттер, для чого він потрібен: Twitter як соціальна мережа. URL: http://smartandyoung.com.ua/shho-take-tvitter-dlja-chogo-vin-potriben-twitter

16. Класифікація. URL: http://sum.in.ua/s/klasyfikacija

17. Пам'ятка щодо тегування. URL: https://webometr.kpi.ua/files/about-tags.pdf

18. Комп’ютерна лінгвістика URL: https://esu.com.ua/search_articles.php?id=4396

19. Пушик Н.В. Комп’ютерна лінгвістика та «штучний інтелект». 2021. №2 (90). С. 151-155. URL: https://molodyivchenyi.ua/index.php/journal/article/view/390/379

20. О.В. Мазурець, О.Ю. Тимуш, А.П. Федорко. Інформаційна технологія тематичної класифікації текстових повідомлень. 2019. №5. С. 203-209. URL: http://elar.khnu.km.ua/jspui/bitstream/123456789/8714/1/21.pdf

21. Які бувають ключові слова і як їх гармонійно вписати в статтю? URL: https://textum.com.ua/blog/kakie-byvayut-klyuchevye-slova-i-kak-ih-garmonichno-vpisat-v-statyu/

22. Що таке ключові слова? URL: https://outsourcing.team/uk/blog/seo-prosuvannya/shho-take-klyuchovi-slova/

23. Словосполучення. URL: https://repetitor.org.ua/slovospoluchennya-2

24. Термін. URL: https://www.jnsm.com.ua/cgi-bin/u/book/sis.pl?Qry=%F2%E5%F0%EC%B3%ED&action=search

25. Модель «торба слів». URL: https://uk.wikipedia.org/wiki/Модель_«торба_слів»

26. О.В. Мазурець, О.В. Ковальчук, В.О. Слободзян. Використання спеціалізованих програмних розширень для автоматизації роботи з цифровими документами навчальних матеріалів . 2018. №1. С. 65-67. URL: http://lib.khnu.km.ua/pdf/visnyk_tup/2018/(257)2018-1-t.pdf

27. Токенізація. URL: https://www.researchgate.net/publication/325071336_TOKENIZACIA_AK_SPOSIB_OBROBKI_KORPUSNOGO_TEKSTU

28. Алгоритм токенізації та стемінгу для текстів українською мовою. URL: http://ekmair.ukma.edu.ua/bitstream/handle/123456789/12541/Hlybovets_Tochytskyi_Alhorytm_tokenizatsii.pdf

29. BM25. URL: https://seo.ru/seowiki/bm25-okapi-bm25/

30. TF-IDF. URL: https://seonomad.net/wiki/tf-idf

31. База даних. URL: http://apeps.kpi.ua/shco-take-basa-danykh

32. Семантичний Аналізатор Тексту. URL: https://site-analyzer.pro/uk/services-seo/text-semantic/

33. З.В. Дудар. Дослідження методів вилучення інформації з неструктурованих текстів на природній мові. URL: https://openarchive.nure.ua/bitstream/document/18869/1/2021_M_PI_Zagoruyko_AV.pdf

34. Twitter. URL: https://twitter.com/explore

35. Keyword Planner. URL: https://ads.google.com/home/tools/keyword-planner/

36. Serpstat. URL: https://serpstat.com/uk/

37. Мова та середовище програмування. URL: https://subject.com.ua/gdz/informatics/8klas_2/7.html

38. Середовище розробки Microsoft Visual Studio. URL: https://learn.ztu.edu.ua/mod/page/view.php?id=9974

39. Середовище створення програм Microsoft Visual Studio. URL: https://studfile.net/preview/5994722/page:7/

40. Розробка веб-застосунків. URL: https://webstudio2u.net/ua/site-develop/641-razrabotka-veb-prilozheniy.html

41. Мобільний застосунок. URL: https://buduysvoe.com/publications/mobilnyy-zastosunok-navishcho-y-komu-vin-potribnyy

42. Десктопні додатки. URL: https://www.quality-assurance-group.com/osoblyvosti-testuvannya-desktopnyh-dodatkiv-u-porivnyanni-z-web-ta-mobilnymy-dodatkamy/

43. C# — Преимущества и недостатки. URL: https://shwanoff.ru/plus-minus-c-sharp/

44. Обґрунтування вибору мови програмування. URL: https://studopedia.ru/15_59823_obruntuvannya-viboru-movi-programuvannya.html

45. Microsoft .NET Framework. URL: https://flexberry.github.io/ru/gbt_dotnet.html

46. Переваги та недоліки .NET. URL: https://dou.ua/lenta/articles/pros-and-cons-of-dotnet/

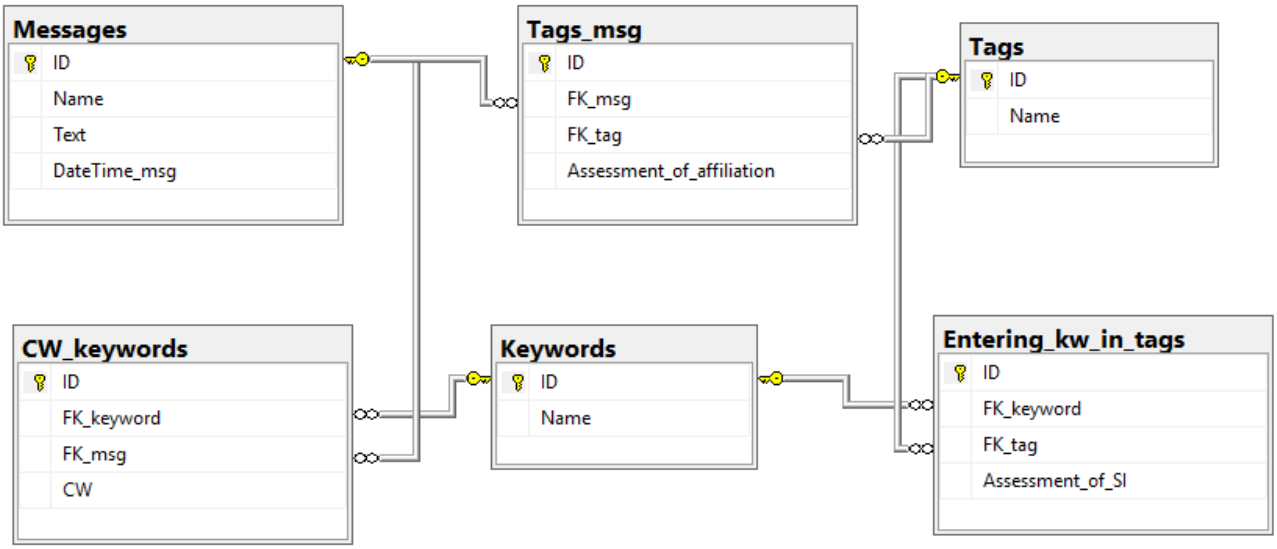
47. 10 причин перейти на Microsoft SQL Server 2019. URL: https://softline.ru/about/blog/10-prichin-pereyti-na-microsoft-sql-server-2019

48. Аналіз найактуальніших серверних систем управління базами даних. URL: http://vlp.com.ua/files/14_22.pdf

ДОДАТКИ

Додаток А

Структура бази даних інформаційної системи автоматизованої тематичної класифікації коротких текстових повідомлень

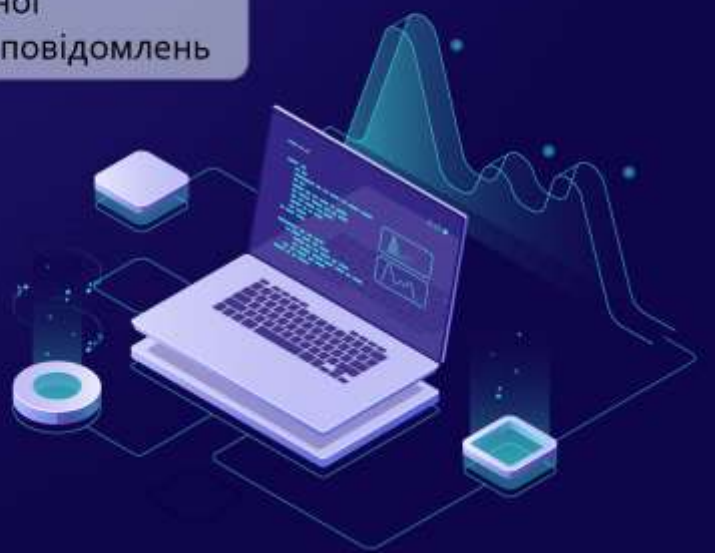


Додаток Б

Презентаційний матеріал

КВАЛІФІКАЦІЙНА РОБОТА БАКАЛАВРА

Метод автоматизованої тематичної класифікації коротких текстових повідомлень



Виконав:
студент 4 курсу, група КН-18-1
О.В. Здоровик

Керівник:
старший викладач кафедри КН
Т.К. Скрипник

Актуальність

Сучасний світ базується на використанні інтернету та інформаційних ресурсів, які допомагають людству полегшити своє життя.

Масштаби інформаційних потоків збільшуються, саме тому тексти необхідно класифікувати, відносити до категорій та певних класів, тобто, необхідно створити комплекс методів та алгоритмів, які працюючи разом будуть виконувати поставлену задачу

Завдання

Мета кваліфікаційної роботи бакалавра – розробка та прикладна програмна реалізація методу автоматизованої тематичної класифікації коротких текстових повідомлень.

Автоматизувати наступні функції:

- обробка тексту повідомлень;
- формування масиву унікальних слів;
- обрахунок значень TF-IDF для кожного унікального слова;
- сортування та обмеження значень TF-IDF;
- визначення ключових слів тексту;
- обробка тестового тексту короткого текстового повідомлення;
- визначення ключових слів тестового тексту, обрахунок їх частоти – CW;
- визначення тегу тестового тексту повідомлення.

Схема інформаційної технології

Вхідні дані:
- База текстів коротких текстових повідомлень;
- Тестовий текст короткого текстового повідомлення.

Крок 1:
Поелементна обробка тексту, визначення унікальних слів:
1. Розбивання тексту на слова та додаткові знаки.
2. Визначення загальних параметрів тексту (кількість слів, знаків).
3. Очищення тексту від додаткових знаків, символів, цифр.
4. Зменшення регістру тексту.
5. Формування масиву унікальних слів тексту.
6. Формування обробленого тексту.

Крок 2:
Підготовка текстів та пошук частоти зустрічей кожного унікального слова:
1. Пошук та запис позицій кожного унікального слова у тексті.
2. Визначення частоти зустрічання кожного унікального слова у тексті.
3. Формування альтернативних текстів для кожного тегу.

Крок 3:
Обрахунок значень семантичної важливості – TF, TF-IDF, визначення ключових слів:
1. Обрахунок значень TF для кожного унікального слова.
2. Обрахунок значень IDF для кожного унікального слова.
3. Обрахунок значень TF-IDF для кожного унікального слова.
4. Сортування та обмеження значень TF-IDF.
5. Визначення ключових слів.

Крок 4:
Визначення приналежності до тегу тестового тексту:
1. Виконання обробки тестового тексту (крок 1).
2. Визначення ключових слів тексту, обрахунок їх частоти - CW (крок 2).
3. Порівняння ключових слів тестового тексту та альтернативних текстів.
4. Визначення приналежності ключових слів до кожного тегу.
5. Сортування слів по приналежності до тегів.

Вихідні дані:
- Ключові слова тегів (класифікацій).
- Визначений тег до якого відноситься тестовий текст.

Діаграма класів інформаційної системи

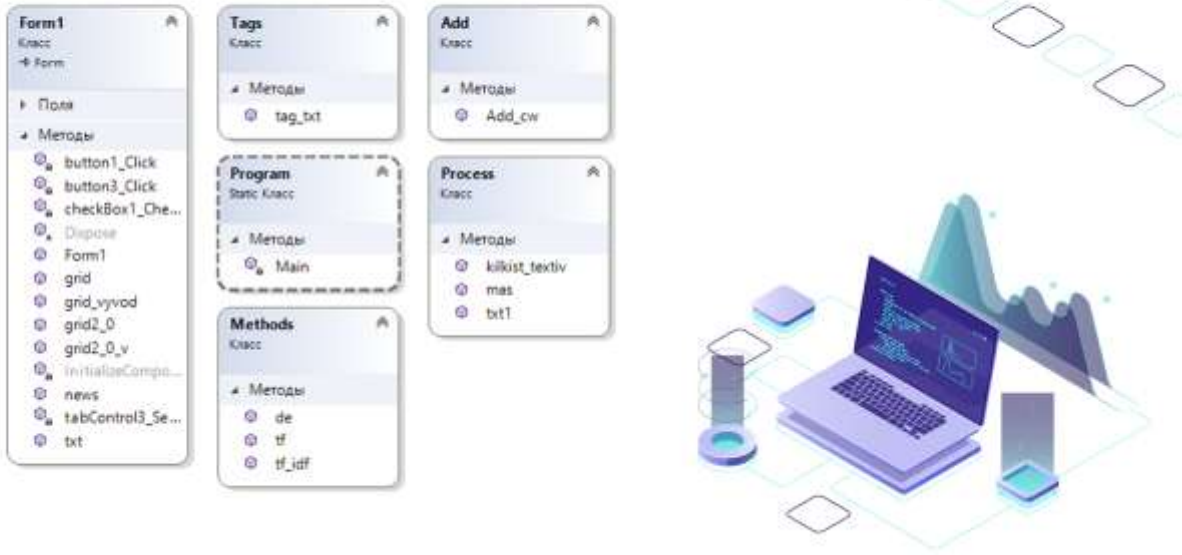
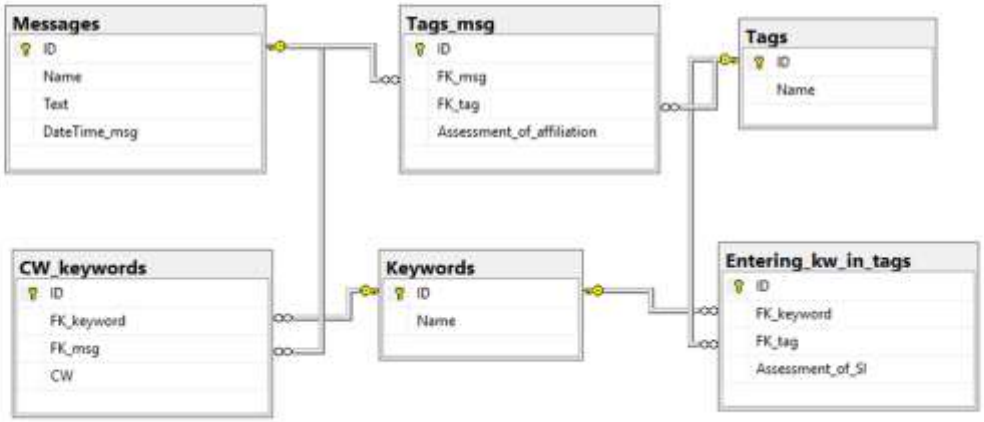
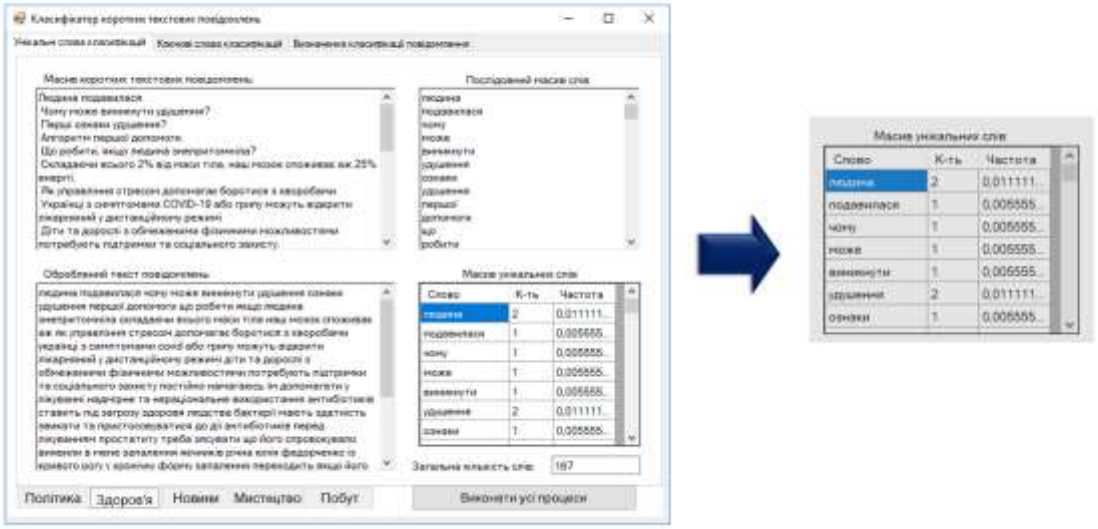


Схема бази даних

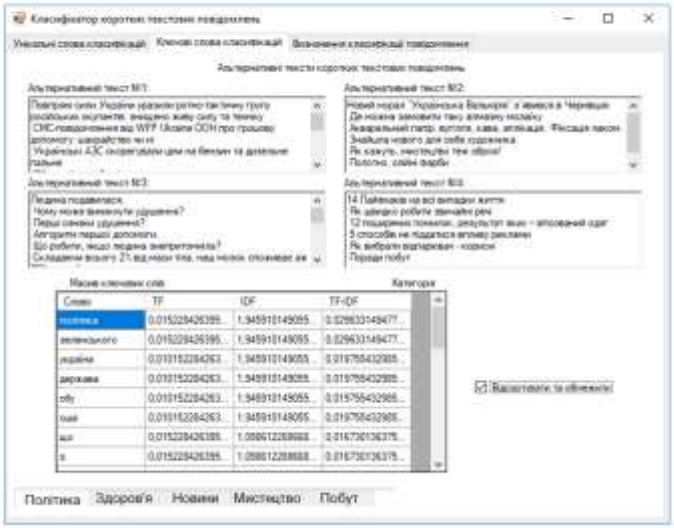
Для забезпечення збереження даних було розроблено базу даних автоматизованої інформаційної системи методу автоматизованої тематичної класифікації коротких текстових повідомлень



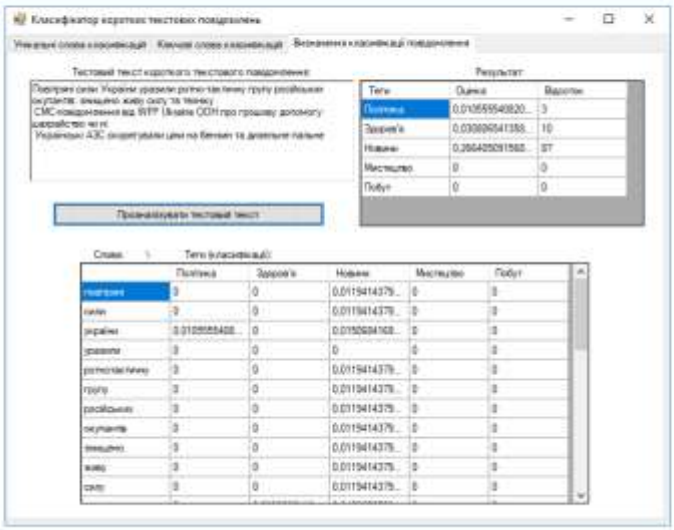
Визначення масиву унікальних слів



Визначення масиву ключових слів



Визначення тегу (класифікації) тестового тексту



Висновки

В результаті виконання кваліфікаційної роботи бакалавра було розроблено інформаційну систему автоматизованого методу тематичної класифікації коротких текстових повідомлень.

Під час створення інформаційної технології були поставлені та вирішені такі задачі:

- Провести аналіз предметної області семантичного аналізу шифрових текстів.
- Провести аналіз засобів спілкування в Інтернеті.
- Огляд теоретичних підходів до розв'язання подібних задач
- Обрати алгоритм початкової обробки коротких текстових повідомлень
- Вдосконалити метод автоматизованої тематичної класифікації
- Розробити інформаційну технологію автоматизованого пошуку та формування масиву ключових слів
- Провести прикладне дослідження методу автоматизованої тематичної класифікації

UNICHECK

Ім'я користувача:
Кафедра КН

ID перевірки:
1011582822

Дата перевірки:
15.06.2022 08:18:02 EEST

Тип перевірки:
Doc vs Internet + Library

Дата звіту:
15.06.2022 08:19:35 EEST

ID користувача:
100005671

Назва документа: **Здоровик_ЗАПИСКА_short**

Кількість сторінок: **59** Кількість слів: **8475** Кількість символів: **64614** Розмір файлу: **2.11 MB** ID файлу: **1011452325**

Виявлено модифікації тексту (можуть впливати на відсоток схожості)

8.87%
Схожість

Найбільша схожість: **2.78%** з джерелом з Бібліотеки (ID файлу: **1011254702**)

2.4% Джерела з Інтернету	171	Сторінка 61
7.73% Джерела з Бібліотеки	67	Сторінка 61

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.

Замінені символи	4
Підозріле форматування	29 сторінок

Wed Jun 15 07:28:37 EEST 2022, Петровський Сергій Степанович, Хмельницький національний університет, ХНУ

Anti-Plagiarism v-15.257

Максимальное совпадение с одним документом 7.0%

Словари проверки: en_US, ru_RU, ua_UA. **Ошибок в документах: 11%**

	Документ		Суммарное совпадение по Базе Данных	
ID: 105391 Название: КВАЛІФІКАЦІЙНА РОБОТА БАКАЛАВРА на тему Метод автоматизованої тематичної класифікації коротких текстових повідомлень Добавлено в БД: 2022-06-15 Авторы: О.В. Здоровик Руководители: Т.К. Скрипник Консультанты: Опоненты:	Символы	Лексемы	Символы	Лексемы
	52816	765	6554 (12%)	94 (12%)

Источник плагиата

ID	Описание	Наличие плагиата в документе	
		Символы	Лексемы

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: *Метод автоматизованої тематичної класифікації коротких текстових повідомлень*

Автор: *студент групи КН-18-1 Здоровик Олександр Васильович*

Спеціальність: *122 – Комп'ютерні науки*

Освітня програма: *освітньо-професійна*

Науковий керівник: *старший викладач кафедри КН Скрипник Тетяна Казимирівна*

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 Здоровика Олександра Васильовича, є законними і не є плагіатом, оскільки:

1) за програмою Anti-Plagiarism виявлені 12% запозичень вказують на документ автора роботи та містять його ж Звіт з практики.

2) За програмою UNICHECK виявлені 8,87%, які є фрагментарними, не більше 2,78% на джерело – містять поширені конструкції, загальновідомі терміни та визначення.

3) запозичення розміщені в розділах аналізу існуючих аналогів та прототипів, які не описують безпосередньо авторське дослідження і не стосуються результатів роботи.

Керівник роботи _____ *Тетяна СКРИПНИК*

Гарант ОП _____ *Олександр МАЗУРЕЦЬ*

Завідувач кафедри КН _____ *Олександр БАРМАК*

ХМЕЛЬНИЦЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ
МОН УКРАЇНИ

Кафедра комп'ютерних наук

РЕЦЕНЗІЯ

на кваліфікаційну роботу бакалавра

студента *гр. КН-18-1 Здоровика Олександра Васильовича*
за темою: *Метод автоматизованої тематичної класифікації коротких текстових повідомлень*

1. Актуальність обраної теми

Проблема обробки текстів, пошуку їх семантичних величин та значень залишається актуальною і досі. Вона ускладнюється ще й тим, що для якісної автоматизованої обробки інформації необхідно створити комплекс методів та алгоритмів, які працюючи разом будуть виконувати поставлену задачу. Пошук інформації відбувається за ключовими словами, які є у кожній статті, на кожному сайті або ж на сторінках соціальних мереж, де ключові слова в основному називаються – хештегами. Саме хештеги (теги) дозволяють класифікувати невелику або ж навпаки більш масштабну інформацію. Відповідно, метод автоматизованої тематичної класифікації коротких текстових повідомлень є актуальною задачею комп'ютерних наук.

2. Повнота розкриття мети та завдань роботи

Метою кваліфікаційної роботи бакалавра є розробка та прикладна програмна реалізація методу автоматизованої тематичної класифікації коротких текстових повідомлень. Для досягнення поставленої мети було вирішено наступні задачі: провести аналіз предметної області семантичного аналізу цифрових текстів, провести аналіз засобів спілкування в інтернеті, огляд теоретичних підходів до розв'язання задач подібних до автоматизації формування тематичної класифікації коротких текстових повідомлень, обрати алгоритм початкової обробки коротких текстових повідомлень, вдосконалити метод автоматизованої тематичної класифікації коротких текстових повідомлень, розробити інформаційну технологію автоматизованого пошуку та формування масиву ключових слів коротких текстових повідомлень, провести прикладне дослідження методу автоматизованої тематичної класифікації коротких текстових повідомлень і виконати аналіз результатів.

3. Зміст кожного розділу роботи

Перший розділ присвячений проведенню аналізу предметної області та визначенню основних параметрів для розв'язку поставленої задачі. Другий розділ присвячений проєктуванню функціональної структури інформаційної системи. Третій розділ присвячений програмній реалізації спроєктованої функціональної структури інформаційної системи.

4. Оцінка розробленої інформаційної системи, її практична цінність

Спроєктований програмний застосунок на основі методу автоматизованої тематичної класифікації коротких текстових повідомлень можна інтегрувати в установи та заклади, що аналізують ключові терміни в коротких текстових повідомленнях для подальшої роботи з ними.

5. Якість оформлення кваліфікаційної роботи бакалавра

Робота виконана на належному науково-методичному рівні та відповідає встановленим вимогам щодо оформлення такого роду праць.

6. Недоліки кваліфікаційної роботи бакалавра

У пояснювальній записці кваліфікаційної роботи містяться формули без відповідних позначень та пояснень та наявні випадки непослідовної нумерації посилань на Перелік джерел.

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 ***«задовільно».***

Рецензент Макарчишкін Р.А доц. каф. АКІТ

ХМЕЛЬНИЦЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ
МОН УКРАЇНИ

Кафедра комп'ютерних наук

ВІДГУК НАУКОВОГО КЕРІВНИКА

на кваліфікаційну роботу бакалавра

студента *ср. КН-18-1 Здоровика Олександра Васильовича*

за темою *Метод автоматизованої тематичної класифікації коротких текстових повідомлень*

1. Актуальність теми

У соціальних мережах користувачі можуть не лише обмінюватись повідомленнями, а і переглядати фото, відео та іншу інформацію про співрозмовника. Текстових повідомлень є досить велика кількість, і ця цифра збільшується з кожним днем, з кожною хвилиною, тому для зручності пошуку та спілкування буде актуально створити програму для автоматизованої тематичної класифікації коротких текстових повідомлень. Саме хештеги (теги) дозволяють класифікувати невелику або ж навпаки більш масштабну інформацію. Відповідно, метод автоматизованої тематичної класифікації коротких текстових повідомлень є актуальною задачею комп'ютерних наук.

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

Тема кваліфікаційної роботи відповідає предметній області спеціальності 122 Комп'ютерні науки та вимогам до кваліфікаційної роботи бакалавра, адже метою роботи є створення методу автоматизованої тематичної класифікації коротких текстових повідомлень. При вирішенні поставленої задачі використано математичні моделі, методи та алгоритми розв'язання теоретичних і прикладних задач, що виникають при розробці інформаційних технологій.

3. Професійні та особистісні якості бакалавра

Студент Здоровик О.В. під час роботи над кваліфікаційною роботою бакалавра та під час навчання продемонстрував достатній рівень знань та умінь за спеціальністю "Комп'ютерні науки", проявив себе відповідальним студентом. Опанував необхідні професійні навички за напрямком «Комп'ютерні науки».

4. Ступінь самостійності під час виконання кваліфікаційної роботи

Одержані в роботі результати є наслідком особистої діяльності студента, який самостійно виконував всі поставлені задачі.

5. Ступінь оволодіння методами дослідження

При реалізації кваліфікаційної роботи студент Здоровик О.В. показав достатній рівень компетентностей та володіння необхідними інструментами та обладнанням, методами, методиками та технологіями предметної області комп'ютерних наук.

6. Повнота та якість розкриття теми роботи

Усі поставлені вимоги до роботи виконані в повному обсязі, проведено аналіз актуальності та відомих досліджень в межах обраної теми, також реалізоване відповідне програмне забезпечення.

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

Викладення матеріалу логічне, послідовне та аргументоване. Мова і стиль викладення кваліфікаційної роботи відповідають стандартам, що забезпечує доступність сприймання матеріалу і відповідає вимогам до сучасних наукових робіт.

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

Реалізований програмний застосунок на базі методу автоматизованої тематичної класифікації коротких текстових повідомлень може бути використаний в установах та службах, що досліджують тексти для подальшою роботи з ними.

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

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

Керівник _____ *ст. викладач кафедри КН Скрипник Т.К.*