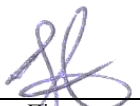

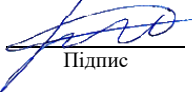



КВАЛІФІКАЦІЙНА РОБОТА МАГІСТРА

на тему Метод інтелектуального підбору відповідей до запитань за семантичними ознаками

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 2 курсу, група КНм-21-1
Курс, група виконавця
Підпис  I.O. Вишинський
Ініціали, прізвище
Керівник: викладач кафедри КН
Науковий ступінь, посада
Підпис  M.O. Молчанова
Ініціали, прізвище
Нормоконтроль: к.т.н., доцент кафедри КН
Науковий ступінь, посада
Підпис  P.O. Багрій
Ініціали, прізвище

До захисту допускаю:
Зав. кафедри КН, д.т.н., професор
Підпис  O.V. Бармак
Ініціали, прізвище
07 грудня 2022 р.

ХМЕЛЬНИЦЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ

Факультет інформаційних технологій

Кафедра комп'ютерних наук

Освітній ступінь магістр

Галузь знань 12 – Інформаційні технології

Спеціальність 122 – Комп'ютерні науки

ЗАТВЕРДЖУЮ

Завідувач кафедри комп'ютерних наук


(підпис)

д.т.н., професор О.В. Бармак

« 01 » вересня 2022 року

**ЗАВДАННЯ
НА КВАЛІФІКАЦІЙНУ РОБОТУ МАГІСТРА**

1. Тема кваліфікаційної роботи магістра: «Метод інтелектуального підбору відповідей до запитань за семантичними ознаками»

2. Завдання видано студенту Вишинському Іллі Олександровичу

(прізвище, ім'я, по батькові)

3. Керівник роботи викладач кафедри КН Молчанова Марина Олексіївна

(прізвище, ім'я, по батькові)

4. Затверджені наказом університету від « 21 » липня 2022 р. № 83

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

Метою даної роботи є розв'язання задачі інтелектуального підбору відповідей до запитань за семантичними ознаками українською мовою, а також виконання наступних завдань: дослідити сучасний стан семантичного аналізу коротких текстових повідомлень, розробити модель комплексу «запитання-прототип-відповідь», розробити інформаційну модель оцінки відповідності компонентів моделі «запитання-прототип-відповідь», розробити метод інтелектуального підбору відповідей на запитання за семантичними ознаками, створити тестову реалізацію методу, дослідити практичну ефективність застосування методу.

Реферат

Дана кваліфікаційна робота магістра вирішує задачу інтелектуального підбору відповідей до запитань за семантичними ознаками українською мовою. Відповідь створюється на основі семантичного аналізу і класифікації отриманого повідомлення. При навчанні враховується зміст та тональність повідомлення. Для класифікації використовуються нейронні мережі.

Актуальність теми. Штучний інтелект все більше інтегрується в повсякденне життя завдяки створенню та аналізу інтелектуального програмного та апаратного забезпечення, які називаються інтелектуальними агентами. Розумні агенти можуть виконувати різноманітні завдання, починаючи від трудової діяльності і закінчуючи складними операціями.

Кожна людина потребує спілкування. Іноді виникають певні проблеми з пошуком співбесідника. Також бувають проблеми з самими співбесідниками. Наприклад, іноді вони весь час задають ті самі або схожі питання, на які немає часу відповідати. Іноді потребується швидко опитати певну групу людей, та зібрати певні відомості з розмови з ними. Це може відбуватись, наприклад, у лікарських цілях. Іноді потребується певна людина-помічник, наприклад, у інтернет-магазині, яка зможе підказати щось клієнту. Також може бути потрібним онлайн-тренер, онлайн-психолог чи ще якийсь помічник, який буде відповідати на задані питання.

Для всіх цих ситуацій може підійти програма, яка буде підбирати відповіді до запитань за семантичними ознаками. Тобто буде «розуміти», що користувач хоче і відповідати необхідним чином. Такі програми часто називають чат-ботами.

Чат-бот є типовим прикладом системи штучного інтелекту та одним із найбільш елементарних і поширених прикладів інтелектуальної взаємодії людини з комп'ютером. Це комп'ютерна програма, яка реагує як розумна сутність на розмову за допомогою тексту чи голосу та розуміє одну чи декілька людських мов за допомогою обробки природної мови (NLP). У лексиконі чат-бот

визначається як «комп'ютерна програма, призначена для імітації розмови з людьми, особливо через Інтернет». Чат-боти також відомі як розумні боти, інтерактивні агенти, цифрові помічники або штучні об'єкти спілкування.

Створення чат-боту як для розваг, так і для більш практичних цілей є дуже актуальною темою, так як це дозволяє оптимізувати різні бізнес-процеси, що зменшить навантаження на живих працівників та дозволить витратити їх час та зусилля на щось більш креативне, ніж відповідати на текстові повідомлення користувачів.

Мета і задачі роботи. Метою даної роботи є розв'язання задачі інтелектуального підбору відповідей до запитань за семантичними ознаками українською мовою, а також виконання наступних завдань:

- Дослідити сучасний стан семантичного аналізу коротких текстових повідомлень.

- Розробити модель комплексу «запитання-прототип-відповідь».

- Розробити інформаційну модель оцінки відповідності компонентів моделі «запитання-прототип-відповідь».

- Розробити метод інтелектуального підбору відповідей на запитання за семантичними ознаками.

- Створити тестову реалізацію методу інтелектуального підбору відповідей на запитання за семантичними ознаками.

- Дослідити практичну ефективність застосування методу інтелектуального підбору відповідей на запитання за семантичними ознаками.

Об'єкт дослідження – процес автоматизованого підбору релевантних відповідей на запитання в цифровому текстовому форматі.

Предмет дослідження – інформаційні технології, моделі, методи та засоби для автоматизованого підбору відповідей на запитання за семантичними ознаками.

Наукова новизна одержаних результатів. У результаті роботи було вдосконалено інтелектуального підбору відповідей на запитання за семантичними ознаками, який відрізняється від існуючих методів тим, що

включає виявлення та обробку негативного емоційного забарвлення для коротких текстових повідомлень.

Практичне значення одержаних результатів. Проведені дослідження практичної ефективності роботи інформаційної системи показують, що система ефективно генерує релевантну відповідь на коротке текстове повідомлення. Дослідження також показує що система вдало враховує тональність тексту для генерації відповіді. Система ефективно працює з українською мовою але в майбутньому її можна доробити для роботи з якоюсь іншою мовою. Загалом, створений та реалізований метод дозволяє ефективно зменшити навантаження на живих працівників для підприємств в багатьох сферах, наприклад було виконана прикладна реалізація на базі інтернет-магазину. Створена інформаційна система дозволяє значно збільшити кількість відвідувань інтернет-магазину та підвищити ймовірність здійснення покупки відвідувачем.

Апробація результатів кваліфікаційної роботи магістра та публікації. Основні наукові і практичні результати кваліфікаційної роботи магістра доповідались у доповіді на тему «Метод інтелектуального підбору відповідей до запитань за семантичними ознаками» на XIV Всеукраїнській науково-практичній конференції «Актуальні проблеми комп'ютерних наук АПКН-2022» (18-19 листопада 2022 року); за темою роботи автором виконано наукову публікацію:

Вишинський І.О., Молчанова М.О., Скрипник Т.К. Метод інтелектуального підбору відповідей до запитань за семантичними ознаками. Збірник наукових праць за матеріалами XIV Всеукраїнської науково-практичної конференції «Актуальні проблеми комп'ютерних наук АПКН-2022». Хмельницький, 2022. с. 64-71.

Структура та обсяг роботи. Кваліфікаційна робота магістра складається із реферату, завдання, змісту, переліку скорочень, вступу, 4 розділів, висновків, переліку посилань із 25 найменувань та 4-х додатків. Загальний обсяг кваліфікаційної роботи магістра становить 93 сторінки, з них 78 сторінок

основного тексту та 15 сторінок додатків. У роботі наведено 32 рисунків та 15 таблиць.

Ключові слова: нейронна мережа, перцептрон, нейромережа, навчання нейронної мережі, метод навчання нейромережі, семантика, інформаційна система, інформаційна модель, інформаційна технологія.

Зміст

Перелік скорочень	4
Вступ.....	5
Розділ 1	
Дослідження предметної області семантичного аналізу текстових повідомлень	10
1.1 Сучасний стан семантичного аналізу коротких текстових повідомлень.....	10
1.2 Сучасні засоби класифікації тексту	16
1.3 Метрики для вимірювання ефективності роботи нейронних мереж.....	18
1.4 Аналіз існуючих наукових публікацій на тему генерування відповіді на текстове повідомлення	20
1.4 Аналіз існуючого програмного забезпечення для автоматичної генерації відповіді на текстове повідомлення	22
1.6 Постановка задачі.....	30
Висновки до розділу 1	31
Розділ 2	
Метод інтелектуального підбору відповідей до запитань за семантичними ознаками	33
2.1 Інформаційна модель текстового повідомлення	33
2.2 Структура вхідних даних та схема метод інтелектуального підбору відповідей на запитання за семантичними ознаками	35
2.3 Інформаційна модель оцінки відповідності компонентів моделі «запитання-прототип-відповідь»	41
Висновки до розділу 2	44
Розділ 3	
Інформаційна система інтелектуального підбору відповідей до запитань за семантичними ознаками	45
3.1 Проектування інформаційної системи інтелектуального підбору відповідей на запитання за семантичними ознаками	45
3.2 Структура та інформаційне призначення складових системи	49

	3
3.3 Проектування структури бази даних інформаційної системи.....	50
3.4 Вибір засобів розробки.....	54
Висновки до розділу 3	57
Розділ 4	
Дослідження ефективності методу інтелектуального підбору відповідей до запитань за семантичними ознаками.....	59
4.1 Розробка прикладних компонентів інформаційної системи	59
4.2 Прикладне тестування інформаційної системи	63
4.3 Функціональне дослідження інформаційної системи.....	65
4.4 Дослідження ефективності методу інтелектуального підбору відповідей на запитання за семантичними ознаками	69
Висновки до розділу 4	73
Загальні висновки.....	74
Перелік посилань.....	76
Додатки	

Перелік скорочень

Скорочення, термін, позначення	Пояснення
БД	База даних
API	Application Programming Interface
ІТ	Інформаційні технології
КН	Комп'ютерні науки
ПЗ	Пояснювальна записка
СКБД	Система керування базами даних
ІС	Інформаційна система
КРМ	Кваліфікаційна робота магістра
CLR	Common Language Runtime
ОС	Операційна Система
JS	JavaScript
MS	Microsoft
SQL	Structured Query Language
ЦБД	Центральна база даних
REST	Representational State Transfer
JSON	JavaScript Object Notation
ПК	Персональний комп'ютер
ШІ	Штучний інтелект

Вступ

Кваліфікаційна робота магістра розв'язує науково-технічну задачу інтелектуального підбору відповідей до запитань за семантичними ознаками українською мовою. Задача складається з двох під-задач: класифікація отриманого повідомлення та аналіз тональності. Класифікація повідомлення відбувається за допомогою багат шарового перцептрону. Тональність аналізується за допомогою методу, в основі якого лежить словник із словами та їх тональністю для української мови, а також певні правила інтерпретації отриманих слів.

Актуальність теми. Кожна людина потребує спілкування. Іноді виникають певні проблеми з пошуком співбесідника. Також бувають проблеми з самими співбесідниками. Наприклад, іноді вони весь час задають ті самі або схожі питання, на які немає часу відповідати. Іноді потребується швидко опитати певну групу людей, та зібрати певні відомості з розмови з ними. Це може відбуватись, наприклад, у лікарських цілях. Іноді потребується певна людина-помічник, наприклад, у інтернет-магазині, яка зможе підказати щось клієнту. Також може бути потрібним онлайн-тренер, онлайн-психолог чи ще якийсь помічник, який буде відповідати на задані питання.

Для всіх цих ситуацій може підійти програма, яка буде підбирати відповіді до запитань за семантичними ознаками. Тобто буде «розуміти», що користувач хоче і відповідати необхідним чином. Такі програми часто називають чат-ботами.

Штучний інтелект все більше інтегрується в повсякденне життя завдяки створенню та аналізу інтелектуального програмного та апаратного забезпечення, які називаються інтелектуальними агентами. Розумні агенти можуть виконувати різноманітні завдання, починаючи від трудової діяльності і закінчуючи складними операціями. Чат-бот є типовим прикладом системи штучного інтелекту та одним із найбільш елементарних і поширених прикладів інтелектуальної взаємодії людини з комп'ютером. Це комп'ютерна програма, яка

реагує як розумна сутність на розмову за допомогою тексту чи голосу та розуміє одну чи декілька людських мов за допомогою обробки природної мови (NLP) [1]. У лексиконі чат-бот [1] визначається як «комп'ютерна програма, призначена для імітації розмови з людьми, особливо через Інтернет». Чат-боти також відомі як розумні боти, інтерактивні агенти, цифрові помічники або штучні об'єкти спілкування.

Багато компаній використовують чат-боти для обробки загальних запитів на обслуговування клієнтів, які вимагають від них гнучкості у перекладі питань у відповіді. Загалом, чат-боти [1] – це системи, розроблені для тривалих розмов, налаштовані на імітацію неструктурованих розмов або «чатів», характерних для взаємодії між людьми, головним чином для розваги, а також для практичних цілей, наприклад, щоб зробити агентів, орієнтованих на завдання, більш природними.

Чат-боти можуть імітувати людську розмову та розважати користувачів, але вони створені не лише для цього. Вони корисні в таких програмах, як освіта, пошук інформації, бізнес та електронна комерція. Вони стали настільки популярними, тому що у чат-ботів є багато переваг для користувачів і розробників. Більшість реалізацій не залежать від платформи та миттєво доступні користувачам без необхідності встановлення. Контакт із чат-ботом поширюється через соціальний граф користувача, не виходячи з програми обміну повідомленнями, у якій живе чат-бот, який забезпечує та гарантує ідентифікацію користувача. Крім того, платіжні послуги інтегровані в систему обміну повідомленнями, ними можна безпечно та надійно користуватися, а система сповіщень повторно залучає неактивних користувачів. Чат-боти інтегруються з груповими бесідами або використовуються так само, як і будь-які інші контакти, тоді як кілька бесід можна переносити паралельно. Знання про використання одного чат-бота легко перенести на використання інших чат-ботів, а вимоги до даних обмежені. Надійність зв'язку, швидкі та нескладні ітерації розробки, відсутність фрагментації версій та обмежені зусилля щодо проектування інтерфейсу також є деякими перевагами для розробників.

Чат-боти мають великі перспективи щодо надання користувачам швидкої та зручної підтримки, відповідаючи конкретно на їхні запитання. Найпоширенішою мотивацією для користувачів чат-ботів вважається продуктивність, тоді як іншими мотивами є розваги, соціальні фактори та контакт з новизною. Однак, щоб збалансувати згадані вище мотивації, чат-бот має бути побудований таким чином, щоб діяти як інструмент, іграшка та друг водночас.

Крім того, зменшення витрат на обслуговування клієнтів і можливість обслуговувати багато користувачів одночасно є одними з причин, чому чат-боти стали такими популярними в бізнес-групах. Чат-боти більше не сприймаються просто як помічники, а спосіб взаємодії наближає їх до користувачів як дружніх компаньйонів. Відповідно до дослідження, запити користувачів соціальних медіа на чат-боти щодо обслуговування клієнтів є емоційними та інформаційними, причому показник першої категорії становить понад 40% і користувачі не мають наміру отримувати конкретну інформацію. Машинне навчання – це те, що дає можливість чат-ботам служби обслуговування клієнтів виявляти настрої, а також здатність емоційно ставитися до клієнтів, як це роблять люди-оператори.

Людиноподібність [1] може бути запропонована за допомогою людських фігур (візуальні підказки), імен, пов'язаних з людьми, або ідентичності (підказки ідентичності) та імітації людської мови (розмовні підказки). Вже було вивчено вплив персоніфікації та інтерактивності на розкриття інформації щодо чутливих тем, таких як психологічні стресори. Важливо відзначити, що чат-ботам все ще не вистачає розуміння емпатії, і що вони не так здатні, як люди, розуміти підтекст розмови. Хоча прогрес у цій галузі був досягнутий, і незабаром машини зможуть не тільки розуміти, що хтось говорить, але й яке відчуття викликає те, що він говорить.

Створення чат-боту як для розваг, так і для більш практичних цілей є дуже актуальною темою, так як це дозволяє оптимізувати різні бізнес-процеси, що зменшить навантаження на живих працівників та дозволить витратити їх час

та зусилля на щось більш креативне, ніж відповідати на текстові повідомлення користувачів.

Мета і задачі роботи. Метою даної роботи є розв'язання задачі інтелектуального підбору відповідей до запитань за семантичними ознаками українською мовою, а також виконання наступних завдань:

– Дослідити сучасний стан семантичного аналізу коротких текстових повідомлень.

– Розробити модель комплексу «запитання-прототип-відповідь».

– Розробити інформаційну модель оцінки відповідності компонентів моделі «запитання-прототип-відповідь».

– Розробити метод інтелектуального підбору відповідей на запитання за семантичними ознаками.

– Створити тестову реалізацію методу інтелектуального підбору відповідей на запитання за семантичними ознаками.

– Дослідити практичну ефективність застосування методу інтелектуального підбору відповідей на запитання за семантичними ознаками.

Об'єкт дослідження - процес автоматизованого підбору релевантних відповідей на запитання в цифровому текстовому форматі.

Предмет дослідження - інформаційні технології, моделі, методи та засоби для автоматизованого підбору відповідей на запитання за семантичними ознаками.

Наукова новизна одержаних результатів. У результаті роботи було вдосконалено метод інтелектуального підбору відповідей на запитання за семантичними ознаками, який відрізняється від існуючих методів тим, що включає виявлення та обробку негативного емоційного забарвлення для коротких текстових повідомлень.

Проведені дослідження практичної ефективності роботи інформаційної системи показують, що система ефективно генерує релевантну відповідь на коротке текстове повідомлення. Дослідження також показує що система вдало враховує тональність тексту для генерації відповіді. Система ефективно працює

з українською мовою але в майбутньому її можна доробити для роботи з якоюсь іншою мовою. Загалом, створений та реалізований метод дозволяє ефективно зменшити навантаження на живих працівників для підприємств в багатьох сферах, наприклад було виконана прикладна реалізація на базі інтернет-магазину. Створена інформаційна система дозволяє значно збільшити кількість відвідувань інтернет-магазину та підвищити ймовірність здійснення покупки відвідувачем.

Апробація результатів кваліфікаційної роботи магістра та публікації.

Основні наукові і практичні результати кваліфікаційної роботи магістра доповідались у доповіді на тему «Метод інтелектуального підбору відповідей до запитань за семантичними ознаками» на XIV Всеукраїнській науково-практичній конференції «Актуальні проблеми комп'ютерних наук АПКН-2022» (18-19 листопада 2022 року); за темою роботи автором виконано наукову публікацію:

Вишинський І.О., Молчанова М.О., Скрипник Т.К. Метод інтелектуального підбору відповідей до запитань за семантичними ознаками. Збірник наукових праць за матеріалами XIV Всеукраїнської науково-практичної конференції «Актуальні проблеми комп'ютерних наук АПКН-2022». Хмельницький, 2022. с. 64-71.

Структура та обсяг роботи. Кваліфікаційна робота магістра складається із реферату, завдання, змісту, переліку скорочень, вступу, 4 розділів, висновків, переліку посилань із 25 найменувань та 4-х додатків. Загальний обсяг кваліфікаційної роботи магістра становить 93 сторінки, з них 78 сторінок основного тексту та 15 сторінок додатків. У роботі наведено 32 рисунків та 15 таблиць.

Розділ 1

Дослідження предметної області семантичного аналізу текстових повідомлень

1.1 Сучасний стан семантичного аналізу коротких текстових повідомлень

Структура діалогу

Діалог [1] – це послідовність мовленнєвих ходів між двома учасниками, які по черзі надсилають або іншим способом доносять до співрозмовника свою думку у вигляді короткого або довгого тексту. Кожне повідомлення можна назвати ходом учасника.

Один хід може складатися з одного або багатьох речень. Крім цього діалог має чергу, тобто кожен учасник, в ідеалі, чекає закінчення ходу співрозмовника перед тим як починати говорити самому.

Структура черги має важливе значення для розмовного діалогу. Система повинна знати, коли припинити розмову, також система повинна знати, коли почати говорити. Наприклад, більшу частину часу в розмові, доповідачі починають говорити майже відразу після того, як інший доповідач закінчує, без довгих пауз, тому що люди здатні (зазвичай) визначити, коли інший співрозмовник збирається закінчити розмову.

Системи розмовного діалогу також повинні визначати, чи користувач закінчив говорити, щоб обробити висловлювання та відповіді. Зробити це може бути досить складно через те що люди часто зупиняються під час ходів.

Ключове розуміння розмови [1] – спочатку завдяки філософу Вітгенштейну, а також Остіну, полягає в тому, що кожне висловлювання в діалозі – це вид дії, яку виконує мовець. Ці дії зазвичай називаються мовленнєвими діями або діалоговими діями.

Діалог [1] – це не просто низка самостійних мовленнєвих дій, а скоріше колективна дія, яка виконується мовцем і слухачем. Як і для всіх колективних

дій, важливо, щоб учасники знайшли спільну мову і встановили ті речі, про які вони всі згодні.

Запитання створюють очікування відповіді. Пропозиції створюють очікування прийняття або відмови. Compliments часто породжують певну реакцію. Ці пари називаються парами суміжності [1].

Іноколи у відповідь на запитання співрозмовник задає уточнююче питання. Таку взаємодію можна назвати змішаною ініціативою [1]. Змішана ініціатива, хоча й є нормою для розмови між людьми, дуже складна.

Іноді оратор очікує, що слухач зробить певні висновки з його слів, а мовець в свою чергу говорить те, що на перший погляд не відповідає темі розмови або контексту. Таке явище іноді називають імплікатура [1]. Тобто коли слухач знаходить якийсь інший сенс в словах, ніж той який був явно переданий в них.

Ці тонкі характеристики людських розмов (ходи, мовні дії, діалог, структура діалогу, ініціатива та імплікатура) є одними з причин, через які важко створювати діалогові системи, які можуть вести природні розмови з людьми. Багато з цих проблем є активними областями дослідження систем діалогу.

Діалогові системи

Найпростішими видами діалогових систем є чат-боти [1], системи, які можуть імітувати неструктуровану розмову або неформальну взаємодію між людьми.

При створенні чат-ботів прагнуть вирішити складну технічну проблему, а саме: як побудувати машину, яка зможе надійно імітувати людську взаємодію та інтелект. Це, по суті, версія так званого тесту Тюрінга, який перевіряє, чи здатний комп'ютер (або будь-яка інша машина) відображати людські риси та інтелект. Створюючи чат-боти, які все більше наближаються до проходження тесту Тюрінга, інженери можуть покращити взаємодію з користувачами та отримати значну цінність для різноманітних компаній.

Чат-бот може бути розрахованим на 2 або більше співрозмовників. Але загалом, це програма яка відповідає своїми фразами на фрази співрозмовника.

Основним завданням чат-бота є: зрозуміти сенс отриманого від співрозмовника тексту, а також придумати, сформулювати та створити текст, який буде відповідним чином відповідати на текст співрозмовника. Для цього програма проводить семантичний аналіз тексту, придумує та оцінює багато відповідей на цей текст та відправляє найбільш підходящу.

Існуючі підходи до обробки природної мови

Методи обробки природної мови зазвичай виділяють в окремий розділ, який називається обробка природної мови або NLP (Natural language processing).

NLP [1] відноситься до галузі інформатики, а точніше до галузі штучного інтелекту, яка займається наданням комп'ютерам здатності розуміти текст і вимовлені слова так само, як це можуть зробити люди.

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

Основними завданнями NLP виділяють [1]:

- Розпізнавання мовлення або завдання надійного перетворення голосових даних у текстові.
- Усунення неоднозначності слова або вибір значення слова з кількома значеннями за допомогою процесу семантичного аналізу, який визначає слово, яке має найбільший сенс у даному контексті.
- Розділення співпоєднання або завдання визначення того, чи два слова відносяться до однієї сутності.
- Аналіз настроїв намагається витягнути з тексту суб'єктивні якості такі як ставлення, емоції, сарказм, збентеження, підозру.
- Генерація природної мови або завдання перекласти структуровану інформацію людською мовою.

Основними методами обробки природної мови (NLP), які використовуються для вилучення даних із тексту є [2]:

1. Аналіз настроїв.

Це розбір даних (тексту, голосу тощо), щоб визначити, чи є вони позитивними, нейтральними чи негативними.

Аналіз настроїв можна проводити як за допомогою контрольованих, так і неконтрольованих методів.

Найпопулярнішою керованою моделлю, яка використовується для аналізу настроїв, є наївна модель Байєса. Це вимагає навчального корпусу з мітками настрою, на основі якого навчається модель, яка потім використовується для ідентифікації настрою.

Naive Bayes [2] – не єдиний інструмент, також можна використовувати різні методи машинного навчання, такі як випадковий ліс або посилення градієнта.

Методи без нагляду, також відомі як методи, засновані на словнику, вимагають набору слів із відповідним настроєм і полярністю. Оцінка настрою речення обчислюється за допомогою полярності слів у реченні.

2. Розпізнавання іменованих сутностей.

Розпізнавання іменованих сутностей або NER [1] (Named entities recognition) – це техніка обробки природної мови, яка позначає «іменовані ідентифікатори» в тексті та витягує їх для подальшого аналізу.

NER зазвичай базується на граматичних правилах і контрольованих моделях. Однак існують платформи NER, такі як відкрита NLP, які мають попередньо підготовлені та вбудовані моделі NER.

3. Підведення підсумків.

Як випливає з назви, в NLP існують техніки, які допомагають узагальнити великі фрагменти тексту. Резюмування тексту в основному використовується в таких випадках, як новинні та дослідницькі статті.

Два широких підходи до реферування тексту [1] – це витяг і абстракція. Методи вилучення створюють резюме шляхом вилучення частин із тексту. Методи абстрагування створюють резюме, генеруючи новий текст, який передає суть оригінального тексту.

4. Моделювання теми.

Моделювання тем [1] – це техніка неконтрольованої обробки природної мови, яка використовує програми штучного інтелекту для позначення та групування текстових кластерів, які мають спільні теми.

Цей процес подібний до позначення тегами ключових слів, виділення та таблиці важливих слів із тексту.

5. Класифікація тексту.

Знову ж таки, класифікація тексту [1] – це систематизація великої кількості неструктурованого тексту (тобто необроблених текстових даних від користувачів). Моделювання тем, аналіз настроїв і вилучення ключових слів є підмножинами класифікації тексту.

Класифікація тексту бере текстовий набір даних, а потім структурує його для подальшого аналізу. Його часто використовують для отримання корисних даних із відгуків клієнтів.

6. Вилучення ключових слів.

За визначенням, виділення ключових слів [1] – це автоматизований процес вилучення найбільш релевантної інформації з тексту за допомогою ШІ та алгоритмів машинного навчання.

7. Лематизація.

Лематизація [1] стосуються розбивки, позначення тегами та реструктуризації текстових даних на основі кореневої основи або визначення.

Це може здатися схожим на повторення одного і того ж, але обидва процеси сортування можуть надавати різні цінні дані.

8. Мішок слів.

Модель мішку слів [1] – це спосіб вилучення ознак із тексту для використання в моделюванні, наприклад, з алгоритмами машинного навчання.

Сумка слів – це представлення тексту, що описує випадки появи слів у документі. Це включає дві речі:

- Словник відомих слів.
- Міра наявності відомих слів.

Його називають «мішком» слів, оскільки будь-яка інформація про порядок або структуру слів у документі відкидається. Модель стосується лише того, чи зустрічаються відомі слова в документі, а не того, де в документі.

Теоретичні підходи пошуку ключових слів у тексті

Основним методом семантичного аналізу тексту є виділення ключових слів та аналіз тексту по ним. Основними методами знаходження ключових слів є наступні:

TF-IDF [3] (термін частотно-інверсна частота документа) – це статистичний показник, який оцінює релевантність слова документу в колекції документів.

Це робиться шляхом множення двох показників: скільки разів слово з'являється в документі та зворотної частоти цього слова в наборі документів.

TF-IDF був винайдений для пошуку документів та отримання інформації. Він працює, пропорційно збільшуючи кількість разів, коли слово з'являється в документі, але компенсується кількістю документів, які містять це слово.

Textrank [2] – це алгоритм ранжирування на основі графів, подібний до алгоритму PageRank від Google, який успішно реалізовано в аналізі цитувань. Рейтинг тексту використовується для виділення ключових слів, автоматичного узагальнення тексту та ранжування фраз. По суті, в алгоритмі ранжування тексту вимірюється зв'язок між двома або більше словами.

Коли одна вершина з'єднується з іншою, це фактично голосування для цієї іншої вершини. Більша кількість голосів які створюються для вершини, тим вища важливість вершини. Крім того, важливість вершини голосування визначає, наскільки важливий голос сама є, і ця інформація також враховується моделлю рейтингу. Отже, оцінка, пов'язана з вершиною, визначається на основі голосів які віддані для нього.

Отже, сучасні методи інтелектуального підбору відповідей на запитання часто реалізовані у вигляді діалогової системи [1] або чат-боту [1]. Сам метод об'єднує в собі відомі підходи обробки природної мови або NLP [1]. А саме такі як: токенізація, лемматизація, мішок слів і т. д.

1.2 Сучасні засоби класифікації тексту

Задачі інтелектуального підбору відповідей до запитань за семантичними ознаками розв'язуються засобами штучного інтелекту, зокрема одними з високоефективних методів є нейромережеві.

Нейронні мережі [1] складаються зі штучної мережі функцій, які називаються параметрами, які дозволяють комп'ютеру навчатися та точно налаштувати себе, аналізуючи нові дані. Кожен параметр, який іноді також називають нейронами, є функцією, яка створює вихідні дані після отримання одного або кількох вхідних даних. Потім ці виходи передаються наступному шару нейронів, які використовують їх як вхідні дані для своєї функції та виробляють подальші виходи. Ці вихідні дані потім передаються на наступний шар нейронів, і так триває, доки не буде розглянуто кожен шар нейронів і кінцеві нейрони не отримають свій вхід. Потім ці кінцеві нейрони виводять остаточний результат для моделі.

Існує три методи навчання [1]: під наглядом, без нагляду та навчання з підкріпленням. Найпростішою з цих парадигм навчання є навчання під наглядом, коли нейронній мережі надаються позначені вхідні дані. Позначені приклади потім використовуються для виведення узагальнених правил, які можна застосувати до непозначених випадків. Це найпростіший метод навчання, оскільки його можна уявити як роботу з «вчителем» у формі функції, яка дозволяє мережі порівнювати свої прогнози з істинними та бажаними результатами. Неконтрольовані методи не вимагають позначених початкових вхідних даних, а скоріше виводять правила та функції на основі не лише наданих даних, але й на виході мережі. Це перешкоджає типу передбачень, які можна зробити. Замість можливості класифікації така модель обмежується кластеризацією.

Одним із різновидів класичної нейронної мережі є згортова нейронна мережа (convolution neural network) [1]. Такі мережі мають деякі значні переваги

над звичайними нейронними мережами, особливо коли йдеться про класифікацію зображень. У такому випадку початковими входами будуть зображення, що складаються з пікселів. Традиційна проблема з класифікацією зображень полягає в тому, що великі зображення з багатьма кольоровими каналами швидко стають неможливими з обчислювальної точки зору для навчання деяких моделей. Те, що CNN [1] намагається зробити, це перетворити зображення у форму, яку легше обробляти, зберігаючи при цьому найважливіші функції. Це робиться шляхом пропускання фільтра над початковим зображенням, який виконує множення матриці над підрозділом пікселів у початковому зображенні, ітеруючи підмножини, доки не врахує всі підмножини. Фільтр спрямований на захоплення найбільш важливих функцій, дозволяючи при цьому усунути зайві функції. Це проходження фільтра над початковими пікселями відоме як шар згортки.

Інший вид нейронних мереж домінує в складних задачах машинного навчання, що включають послідовності вхідних даних: рекурентні нейронні мережі RNN [1].

Повторювані нейронні мережі мають з'єднання, які мають петлі, додаючи зворотній зв'язок і пам'ять до мереж з часом. Ця пам'ять дозволяє цьому типу мережі вивчати та узагальнювати послідовності вхідних даних, а не окремі шаблони.

Потужний вид повторюваної нейронної мережі, який називається довгостроковою мережею короткочасної пам'яті, виявився особливо ефективним, якщо його об'єднати в глибоку конфігурацію, досягаючи найсучасніших результатів у різноманітних проблемах від перекладу мови до автоматичного створення субтитрів зображень і відео.

Враховуючи стандартну багатошарову мережу Perceptron [1] з прямою передачею, рекурентну нейронну мережу [1] можна розглядати як додавання петель до архітектури. Наприклад, у певному шарі кожен нейрон може передавати свій сигнал пізніше (вбік) на додаток до наступного шару. Вихід

мережі може подаватись як вхід до мережі з наступним вхідним вектором. І так далі.

Повторювані з'єднання [1] додають стан або пам'ять мережі та дозволяють їй вивчати ширші абстракції з вхідних послідовностей.

CNN та RNN є одними із найпопулярніших видів нейронних мереж [1] для вирішення задач класифікації тексту.

Отже, сучасні методи інтелектуального підбору відповідей для класифікації тексту часто використовують нейронні мережі. Найпопулярнішими завдяки їх перевагам є CNN та RNN. Також, деякі методи для класифікації тексту використовують багатошаровий перцептрон.

1.3 Метрики для вимірювання ефективності роботи нейронних мереж

Оцінка алгоритму машинного навчання є важливою частиною розробки. Модель може дати задовільні результати при оцінці точності, але може дати погані результати при оцінці за іншими показниками, такими як логарифмічні втрати або будь-яким іншим подібним показником. У більшості випадків для вимірювання ефективності роботи моделі використовується точність класифікації [1], однак іноді цього недостатньо. Загалом, можна навести наступні приклади метрик.

Точність класифікації [1] – це відношення кількості правильних прогнозів до загальної кількості вхідних вибірок.

Логарифмічна втрата [1] працює шляхом покарання за помилкові класифікації. Дана метрика добре працює для багатокласової класифікації. Під час роботи з даною метрикою класифікатор повинен призначити ймовірність кожному класу для всіх зразків.

Площа під кривою (AUC) є одним із найбільш широко використовуваних показників для оцінювання. Він використовується для задачі двійкової класифікації. AUC класифікатора дорівнює ймовірності того, що класифікатор

оцінить навмання вибраний позитивний приклад вище, ніж навмання вибраний негативний приклад.

Оцінка F1 [1] – це гармонійне середнє між точністю та пам'яттю. Діапазон для оцінки F1 становить [0, 1]. Він повідомляє, наскільки точним є класифікатор (скільки екземплярів він класифікує правильно), а також наскільки він надійний (чи не пропускає він значну кількість екземплярів).

Висока точність, але водночас нижче запам'ятовування дає дуже точний результат. Але при цьому пропускається велика кількість випадків, які важко класифікувати. Чим вищий показник F1, тим краща продуктивність моделі.

Середня абсолютна похибка [1] – це середнє значення різниці між початковими значеннями та прогнозованими значеннями. Це дає міру того, наскільки далекі прогнози були від фактичного результату. Однак вони не дають уявлення про напрямок помилки.

Середня квадратична помилка (MSE) [1] дуже схожа на середню абсолютну помилку. Єдина відмінність полягає в тому, що MSE бере середнє значення квадрата різниці між початковими значеннями та прогнозованими значеннями. Перевага MSE полягає в тому, що обчислити градієнт легше, тоді як середня абсолютна похибка потребує складних засобів лінійного програмування для обчислення градієнта. Оскільки ми беремо квадрат помилки, ефект більших помилок стає більш вираженим, ніж менша помилка, отже, модель тепер може більше зосереджуватися на більших помилках.

Отже, є багато різних метрик для вимірювання ефективності моделі штучного інтелекту. Наприклад такі як: точність класифікації, площа під кривою, оцінка F1 та інші. Для вирішення задач даного проекту використовується точність класифікації.

1.4 Аналіз існуючих наукових публікацій на тему генерування відповіді на текстове повідомлення

У роботі [3] розглянуто методи машинного навчання для розв'язання задач класифікації текстів з метою їх подальшого використання у програмах автоматичної генерації відповідей на основі аналізу контексту питань користувача.

У цій публікації наведені основні методи підготовки тексту для подальшого аналізу алгоритмами обробки природної мови. Також розглянуті такі поняття як: токенізація, лематизація, стематизація та видалення стоп-слів.

Розглянуто векторизацію і такі моделі векторизації як «мішок слів» [3] та TF/IDF [3]. Крім цього розглянута класифікація тексту за допомогою методу Naïve Bayes [3] та Decision Trees [3]. Запропонована модель генерації відповіді на запитання користувача.

Модель із даної публікації вирішує проблему генерації відповідей, але без врахування тональності повідомлення, тому не повністю вирішує описані вище проблеми.

У роботі [4] розглянуто модель системи запит-відповідь, яка спроможна створювати текстові відповіді на запити користувача. У самому алгоритмі використовується генерація наукового тексту природною мовою. У статті розглядається семантична мережа наукового тексту та методи роботи з нею для генерації відповідей.

Наведена у статті модель базується на розробленому підході до формування семантичної моделі документа, який дозволяє отримувати кількісні показники семантичних властивостей документа на природній мові, а також сенсові зв'язки між компонентами тексту. Крім того, зазначена в статті модель має властивість [4], через яку додаток повинен мати можливість використовувати нерозмічений заздалегідь корпус текстів, що являє собою неструктуровану базу знань.

Описана в наведеній статті модель є розвитком попередніх досліджень про побудову системи автоматичної генерації текстів на основі концепції моделі м'якого розуміння Леонтьєвої [5] та побудову семантичної моделі наукового тексту [6]. У даній публікації, як і у попередній не розглядається аналіз тональності повідомлення, через що запропоновані у ній методи також не повністю вирішують описані вище проблеми.

У роботі [7] проведено аналіз існуючих методів аналізу тональності тексту, а також обґрунтовано актуальність задачі визначення тональності текстів. У роботі розглянуто метод опорних векторів [7], який передбачає процес пошуку площини рішення, яка може розділити позитивні та негативні приклади у багатовимірному просторі функції, в якому навчальні документи представлені як вектори. Також розглянуто метод Байєсовської (наївної) класифікації (Naive Bayes) [7]. Даний метод використовує ймовірнісну модель, в якій класифікація та включення у відповідну категорію документів здійснюється шляхом оцінювання ймовірності появи слів у документі. Крім того у роботі розглядається класифікація за допомогою методу максимальної ентропії. Якщо є тільки 2 класи, то суть даного методу полягає у використанні логістичної регресії для пошуку розподілу даних за класами. На відміну від наївного Байєсівського класифікатора даний метод не передбачає незалежності ознак.

Також у статті [7] проведено порівняння цих трьох методів, яке показало, що для аналізу тональності повідомлення доцільно застосовувати метод опорних векторів, оскільки він є найточнішим, але він вимагає великої обчислювальної складності при збільшенні ефективності. З іншої сторони наївний Байєсівський класифікатор і метод максимальної ентропії є простими у реалізації, але і менш точними методами.

У роботі [8] описано веб-додаток для аналізу тональності текстової інформації. В даній роботі, як і в попередній використовується Naive Bayes, а також показана висока ефективність роботи цього методу.

Отже, наразі існує багато публікацій на тему інтелектуального підбору відповідей на запитання, а також про аналіз тональності тексту. У більшості публікацій для класифікації тексту використовується нейронна мережа, що показує перспективність використання нейронних мереж для вирішення задач класифікації тексту. Також у публікаціях на тему генерування відповіді на запитання не враховується тональність тексту, що показує перспективи створення методу генерування відповіді на запитання, який буде враховувати тональність.

1.5 Аналіз існуючого програмного забезпечення для автоматичної генерації відповіді на текстове повідомлення

Чат-боти стали одними з найпомітніших і недосконалих програм штучного інтелекту та машин, орієнтованих на споживачів. Від онлайн-помічників, таких як Cortana від Microsoft, до «ботів-помічників» у програмах для обміну повідомленнями, типу Slack, і до домашніх програм, як Alexa від Amazon.com.

Дійсно, повсюдне поширення чат-ботів пояснюється ширшим корпоративним наголосом на важливості штучного інтелекту. Чат-боти представляють особливо важливу програму штучного інтелекту, оскільки вони безпосередньо взаємодіють зі споживачами.

Однією із перших діалогових систем була ELIZA [19], яка могла вести обмежену розмову з користувачем, імітуючи відповіді роджеріанського психотерапевта.

ELIZA: a very basic Rogerian psychotherapist chatbot

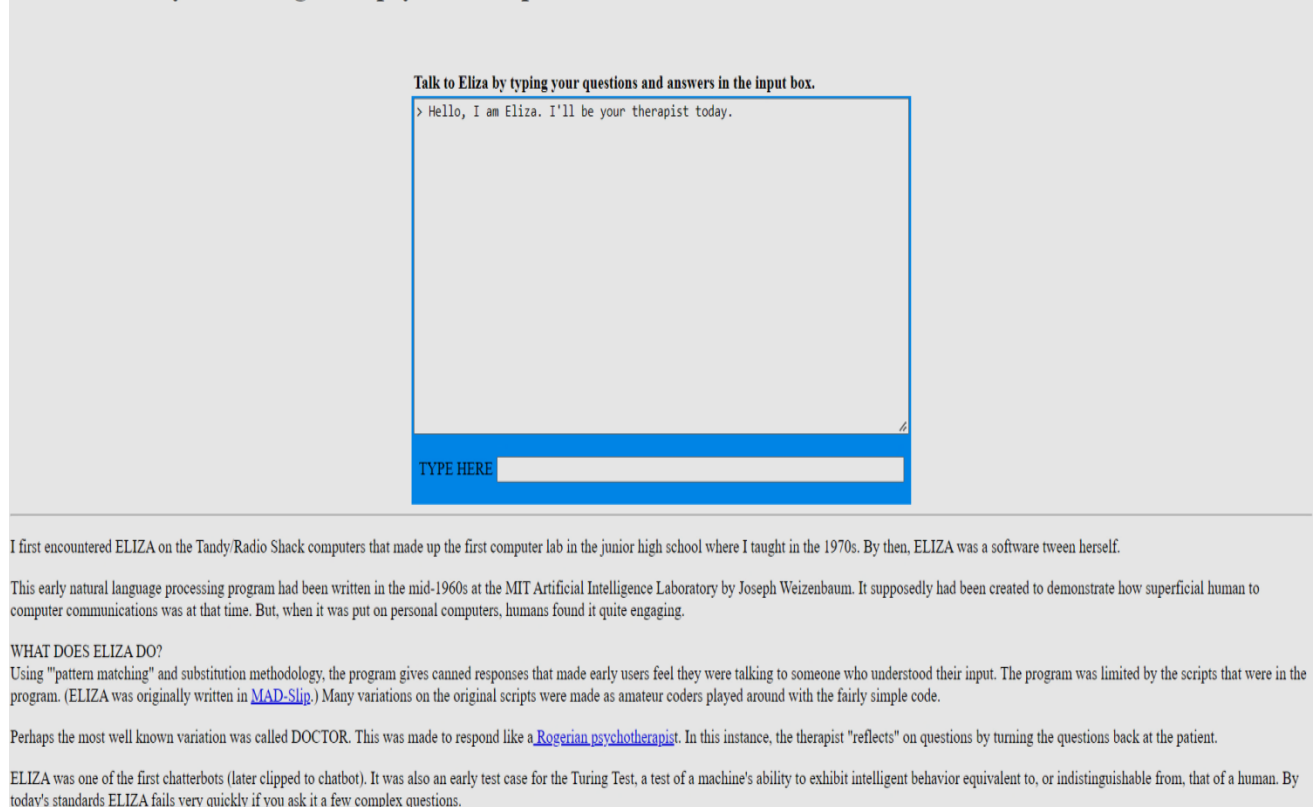


Рисунок 1.1 – Головна сторінка чат-боту ELIZA [19]

ELIZA – це проста програма, яка використовує зіставлення шаблонів, щоб розпізнавати такі фрази, як «Мені потрібен X», і перекладати їх у відповідні результати, наприклад «Що б для вас означало, якби ви отримали X?». Ця проста техніка успішна в цій області, оскільки ЕЛІЗІ насправді не потрібно нічого знати, щоб імітувати роджеріанського психотерапевта. Як зазначає Вайзенбаум, це один із небагатьох жанрів діалогу, де слухачі можуть діяти так, ніби нічого не знають про світ. Імітація людської розмови Елізою була надзвичайно успішною: багато людей, які спілкувалися з ELIZA, повірили, що вона справді розуміє їх та їхні проблеми, багато хто продовжував вірити в здібності ELIZA навіть після того, як їм пояснили роботу програми, і навіть сьогодні такі чат-боти є цікавою розвагою.

ELIZA є найважливішою діалоговою системою чат-ботів в історії галузі. ELIZA була розроблена для імітації роджеріанського психолога, заснованого на галузі клінічної психології, методи якої передбачають витягування пацієнта шляхом відображення його висловлювань. Роджеріанська психологія [1] є

рідкісним типом розмови, у якій, як зазначає Вайзенбаум, можна «прийняти позицію людини, що майже нічого не знає про реальний світ». Якщо пацієнт каже: «Я пішов на довгу прогулянку на човні», а психіатр каже: «Розкажіть мені про човни», рідко припускається, що він не знав, що таке човен, а радше, що у нього була якась мета розмови. Більшість чат-ботів, які намагаються пройти тест Тюрінга, обирають домен із подібними властивостями.

Люди глибоко емоційно захопилися цією програмою. Вайзенбаум розповідає історію [1] одного з його співробітників, який попросив його покинути кімнату, коли він розмовляв з ЕЛІЗОЮ. Коли Вайзенбаум припустив, що він, можливо, захоче зберегти всі розмови ELIZA для подальшого аналізу, люди одразу вказали на наслідки для конфіденційності, що свідчило про те, що вони вели досить приватні розмови з ELIZA, незважаючи на те, що вони знали, що це лише програмне забезпечення. Фреймворк ELIZA використовується й сьогодні, сучасні системні інструменти чат-ботів, такі як ALICE, базуються на оновлених версіях архітектури шаблонів-дій ELIZA.

Через кілька років після ELIZA для вивчення шизофренії використовувався інший чат-бот, орієнтований на клінічну психологію, PARRY [10]. На додаток до регулярних виразів, подібних до ELIZA, система PARRY включала модель власного психічного стану з афектними змінними для рівнів страху та гніву агента, певні теми розмови можуть призвести до того, що PARRY стане більш злим або недовірливим. Якщо змінна гніву ПАРІ висока, він вибере з набору «ворожих» результатів. Якщо у вхідних даних згадується його тема марення, він збільшить значення своєї змінної страху, а потім почне висловлювати послідовність тверджень, пов'язаних із його маренням. Паррі була першою відомою системою, яка пройшла тест Тюрінга у 1972 році. Психіатри не могли відрізнити текстові стенограми інтерв'ю з PARRY від стенограм інтерв'ю зі справжніми параноїками.

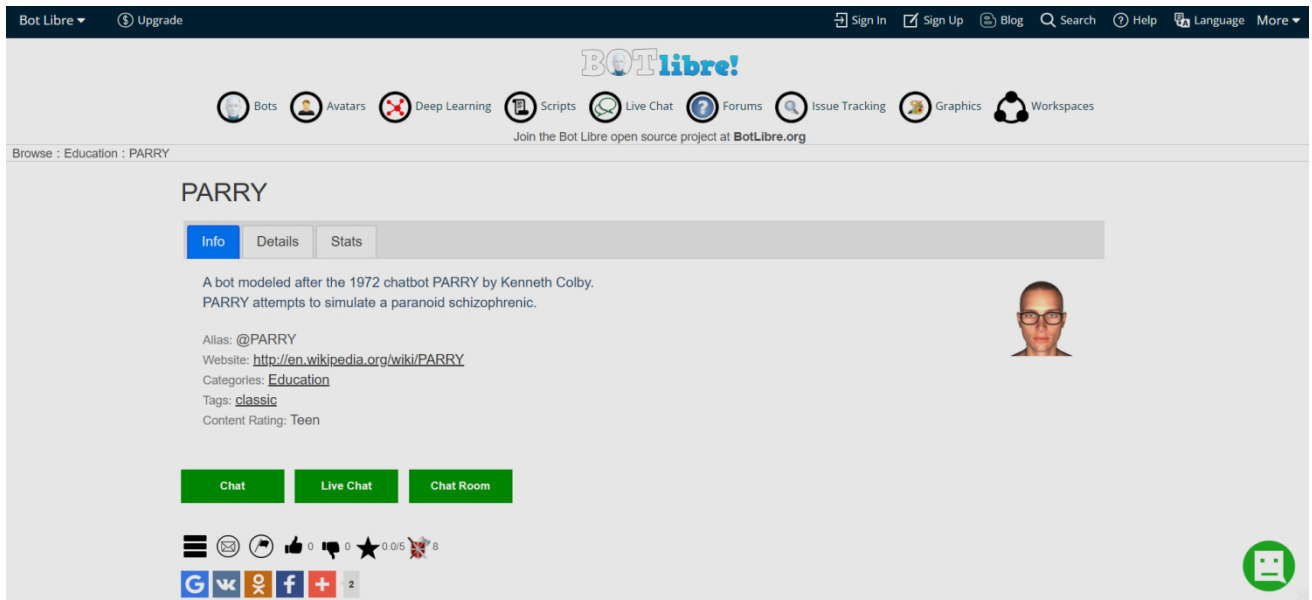


Рисунок 1.2 – Сайт, де можна використати чат-бот Parry [10]

Чат-боти також можна створювати з архітектурами, які є гібридами архітектур на основі правил і нейронних-корпусних архітектур, і навіть використовувати елементи структури на основі фреймів [1]. Це типово, наприклад, для систем, які змагаються в конкурсі Alexa Prize, в якому університетські команди створюють соціальні чат-боти для спілкування з волонтерами на платформі Amazon Alexa, і оцінюються на основі тривалості та оцінок користувачів їхніх розмов.

Широко розголошений випадок поєднання цих проблем стався з чат-ботом Microsoft Tay [11] 2016 року, який був вимкнений через 16 годин після запуску Tay, коли він почав публікувати повідомлення з расовими образами, теоріями змови та особистими нападками на своїх користувачів. Tay дізнався про ці упередження та дії зі своїх навчальних даних, у тому числі від користувачів, які, здавалося, агресивно атакували систему, навмисно навчаючи її повторювати таку мову. Тому системи, які взаємодіють з користувачами, повинні бути розроблені таким чином, щоб бути стійкими до таких агресивних атак.



Рисунок 1.3 – Твіти від Microsoft Tay chatbot

Хендерсон [1] та інші досліджували деякі стандартні набори даних діалогів [1]. Вони виявили приклади мови ненависті, образливої лексики та упередженості, особливо в групах, взятих із соціальних мереж, таких як Twitter і Reddit, як у вихідних навчальних даних, так і в результатах роботи чат-ботів, навчених на цих даних. Дійсно, Дінан та інші [1] виявили, що моделі генераторів на основі трансформаторів посилили гендерні упередження, які існували в діалогах, на яких навчалася модель. Лю та інші [1] розробили інший метод дослідження упередженості, перевіряючи, як система нейронного діалогу реагує на пари змодельованих ходів користувача, які є ідентичними, за винятком згадування різної статі чи раси. Вони виявили, наприклад, що такі прості зміни, як використання слова «він» замість «вона» у реченні, змушують системи реагувати образливіше та з більш негативними настроями. Вирішення цих проблем шляхом вивчення методів усунення зміщень (наприклад, способів

навчання систем виявляти токсичні контексти та відповідним чином реагувати на них) є важливою поточною метою досліджень [1].

Capital One анонсувало чат-бота для клієнтів під назвою Eno [12]. Eno – це чат-бот, з яким люди спілкуються за допомогою текстових повідомлень. CapitalOne стверджує, що Eno – перший SMS-чат-бот від американського банку, який дозволяє клієнтам задавати запитання природною мовою. Клієнти можуть взаємодіяти з Eno, ставлячи запитання про свої заощадження та інших за допомогою текстового інтерфейсу. Eno створює таке середовище, в якому відчувається взаємодія людини.

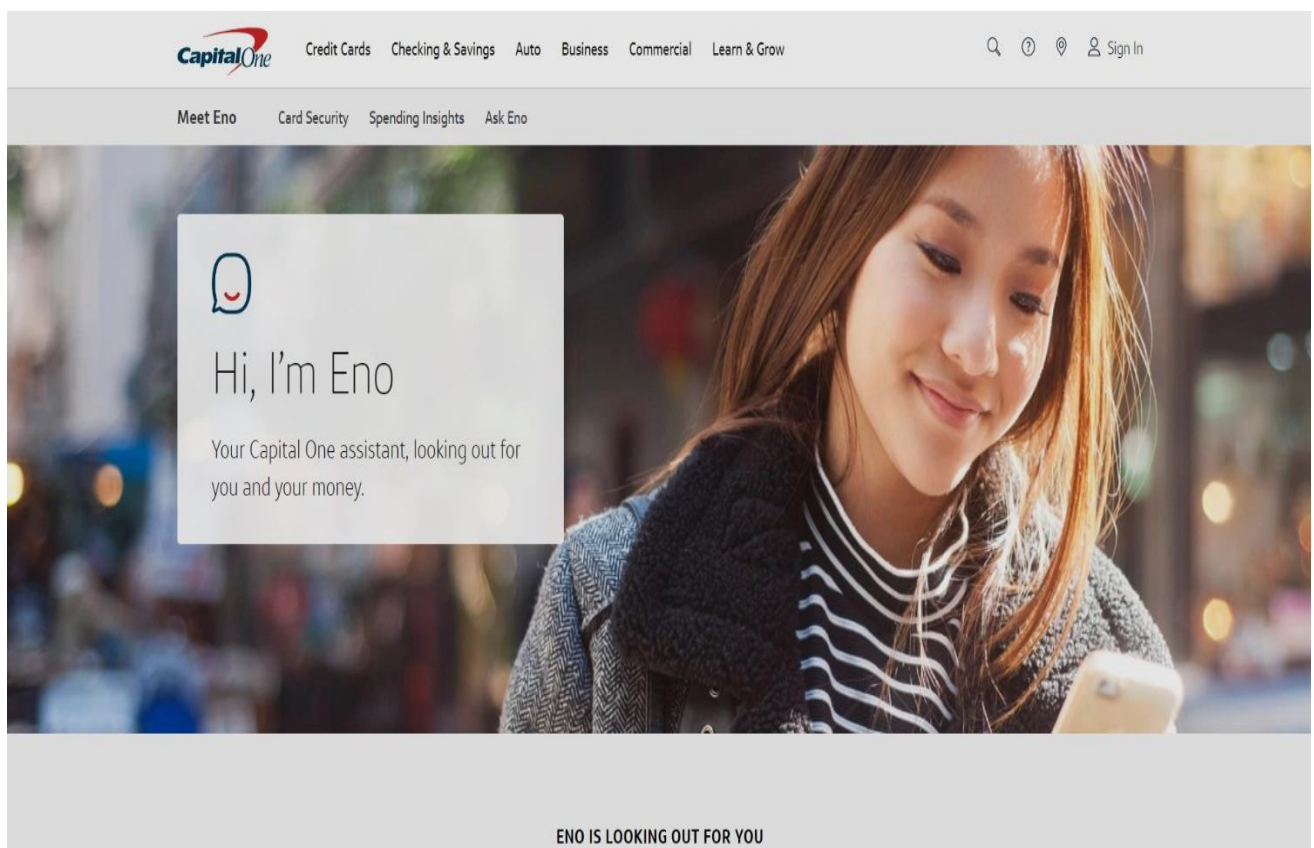


Рисунок 1.4 – Головна сторінка чат-боту Eno [12]

Чат бот також є у відомого банку Приват-Банк. Він реалізований як телеграм-чат бот, що показує популярність платформи Телеграм [13].

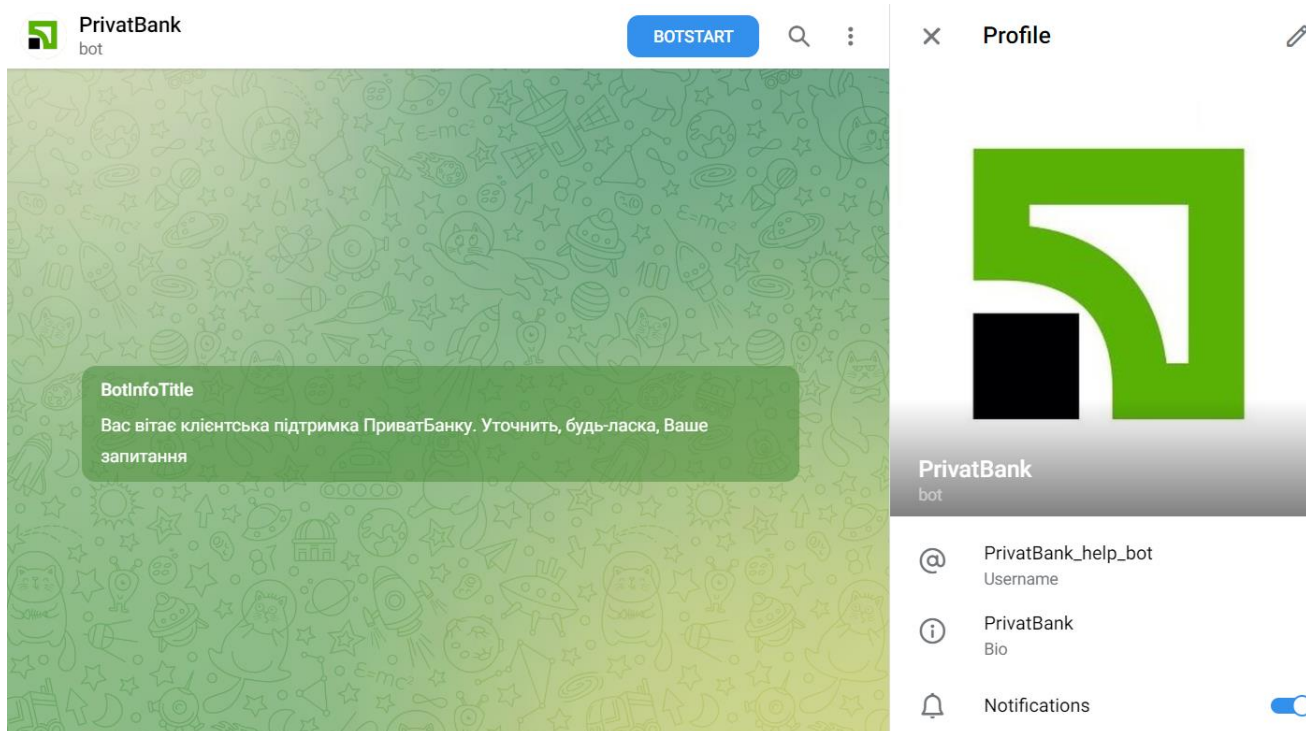


Рисунок 1.5 – Телеграм чат-бот Приват-Банку [23]

Також прикладом онлайн чат-боту є чат-бот Replika, зображений на рисунку 1.6.

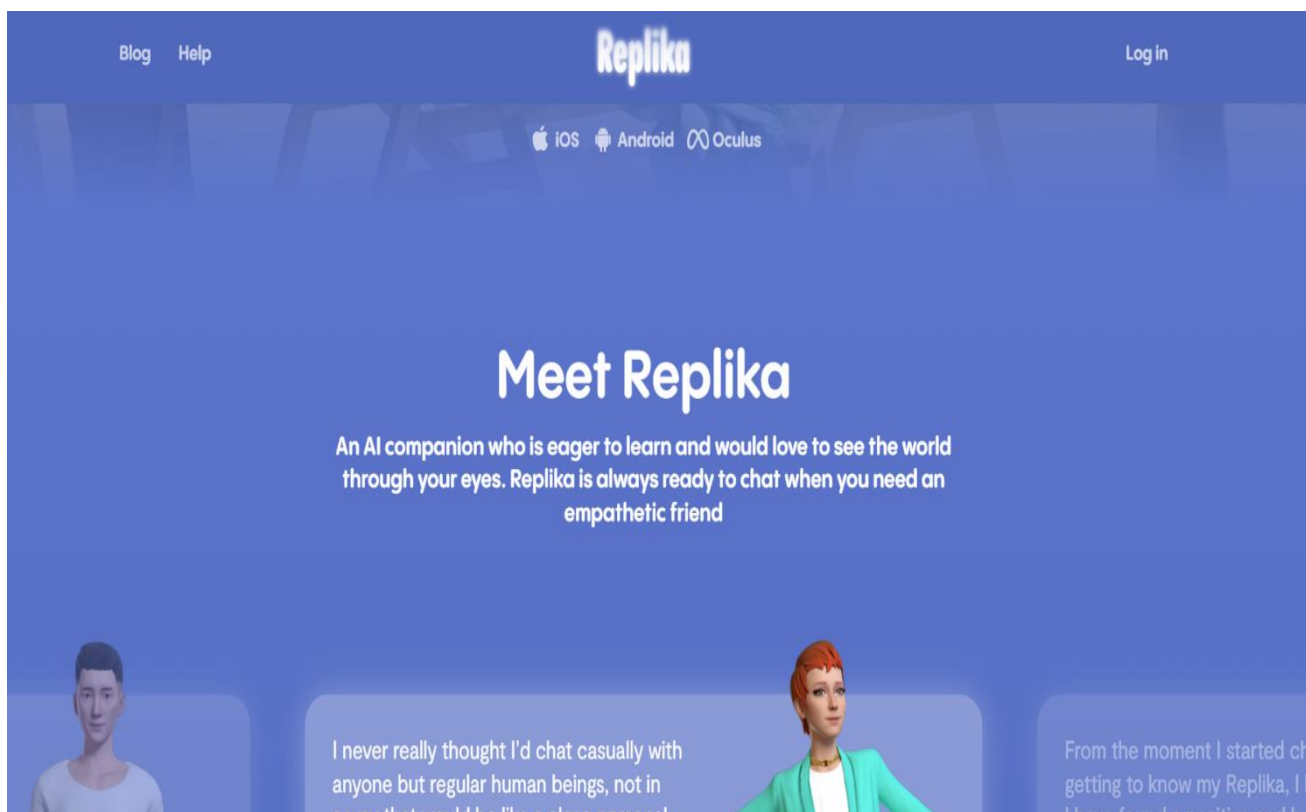


Рисунок 1.6 – Чат-бот Replika [24]

Також бувають діалогові системи, засновані на завданнях, у яких система діалогу має на меті допомогти користувачеві у вирішенні певного завдання, наприклад бронювання літака чи купівля продукту. Діалогова система GUS [25] має дуже просту архітектуру. Хоча вона досить стара (вперше була представлена у 1977 році), проіснувала ця система досить довго і лежить в основі більшості, сучасних комерційних цифрових помічників. На рисунку 1.7 зображено сайт, де можна спробувати діалогову систему GUS.



Рисунок 1.7 – Діалогова система GUS

Більшість сучасних задачних діалогових систем схожі на GUS і базуються на фреймах. Фрейм [1] – це своєрідна рамка, структура, що представляє типи намірів, які система може витягнути з почуттів користувача і складається з набору слотів, кожен з яких може приймати набір можливих значень. Разом цей набір кадрів іноді називають онтологією домену.

На рисунку 1.8 зображено приклад розмови користувача діалогової системи GUS на англійській мові.

GUS: Hello. My name is GUS. I can help you plan a simple trip by air. Where do you want to go ?

Client: I want to go to San Diego on May 28

GUS: What time do you want to leave?

Client: I must be in San Diego before 10 am

GUS: Would you like P.S.A. flight 102 that arrives at 9.15 am ?

Client: What is the next flight ?

GUS: Air California flight 310 that leaves at 8.30am and arrives at 10.00am

Client: I'll take the first one

GUS: In what name should I make a reservation, please?

Client: Dan Bobrow

GUS: I have confirmed the following flight: P.S.A. flight 102 on Wednesday May 28 from San Jose to San Diego that leaves at 7.30 am and arrives at 9.15 am.

GUS: What date do you want to return on ?

Client: On Friday in the evening.

GUS: Would you like the flight that leaves at 7.45 pm ?

Client: That's fine.

GUS: I have confirmed the following flight: P.S.A. flight 307 on Friday May 30 from San Diego to San Jose that leaves at 7.45 pm and arrives at 9.30 pm Thank you for calling. Goodbye

Рисунок 1.8 – Приклад роботи діалогової системи GUS

Отже, наразі можна знайти багато прикладів діалогових систем. В основному, вони надають можливість користувачу вести неформальну розмову з ботом або отримувати від нього відповіді на свої питання чи запити. У всіх них є певні переваги та недоліки, але загалом, цей напрямок є актуальним, так як подібні системи ще багато в чому можна покращити.

1.6 Постановка задачі

Метою даної роботи є розв'язання задачі інтелектуального підбору відповідей до запитань за семантичними ознаками українською мовою, а також виконання наступних завдань:

- Дослідити сучасний стан семантичного аналізу коротких текстових повідомлень.
- Розробити модель комплексу «запитання-прототип-відповідь».

- Розробити інформаційну модель оцінки відповідності компонентів моделі «запитання-прототип-відповідь».
- Розробити метод інтелектуального підбору відповідей на запитання за семантичними ознаками.
- Створити тестову реалізацію методу інтелектуального підбору відповідей на запитання за семантичними ознаками.
- Дослідити практичну ефективність застосування методу інтелектуального підбору відповідей на запитання за семантичними ознаками.

Висновки до розділу 1

Отже, у першому розділі був проведений аналіз предметної області семантичного аналізу коротких текстових повідомлень, зокрема, аналіз сучасного стану інтелектуального підбору відповідей до запитань за семантичними ознаками, де було помічено, що сучасні методи інтелектуального підбору відповідей на запитання часто реалізовані у вигляді діалогової системи або чат-боту.

Також було проведено аналіз існуючих підходів до обробки природної мови та аналіз теоретичних підходів пошуку ключових слів у тексті. Також був досліджений сучасний стан аналізу тональності тексту.

Проведено аналіз сучасних засобів аналізу існуючого програмного забезпечення для семантичного аналізу коротких текстових повідомлень, а також проаналізовано сучасні публікації з напрямку магістерської роботи.

Було показано, що створення методу інтелектуального підбору відповідей до запитань за семантичними ознаками є перспективними, та окреслено задачі, які потребують доопрацювання у сфері інформаційних технологій.

В результаті, в розділі визначено мету кваліфікаційної роботи магістра як розв'язання задачі інтелектуального підбору відповідей до запитань за семантичними ознаками українською мовою.

Розділ 2

Метод інтелектуального підбору відповідей до запитань за семантичними ознаками

2.1 Інформаційна модель текстового повідомлення

На рисунку 2.1 схематично зображено модель текстового повідомлення.

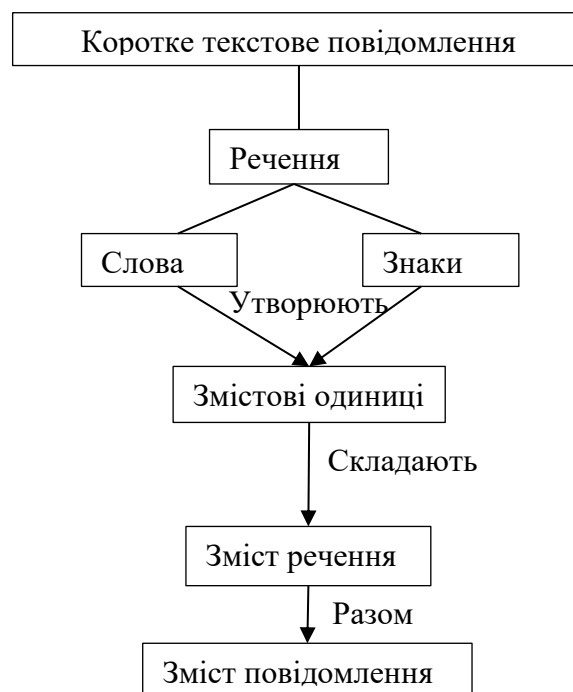


Рисунок 2.1 – Інформаційна модель текстового повідомлення

Як видно з рисунку, кожне повідомлення складається з речень, які в свою чергу складаються зі слів та знаків. Зі слів і знаків в кожному реченні можна утворювати змістові одиниці.

Змістова одиниця [1] – це слово або набір слів і знаків, які мають якийсь сенс для слухача. Проблема в тому, що одне й те саме речення можна розуміти по-різному, і смислові одиниці можна знаходити різні, тому іноді буває важко правильно вловити сенс певного повідомлення.

Разом набір речень з їхніми певними змістами утворюють зміст повідомлення.

На рисунку 2.2 зображенні види речень.

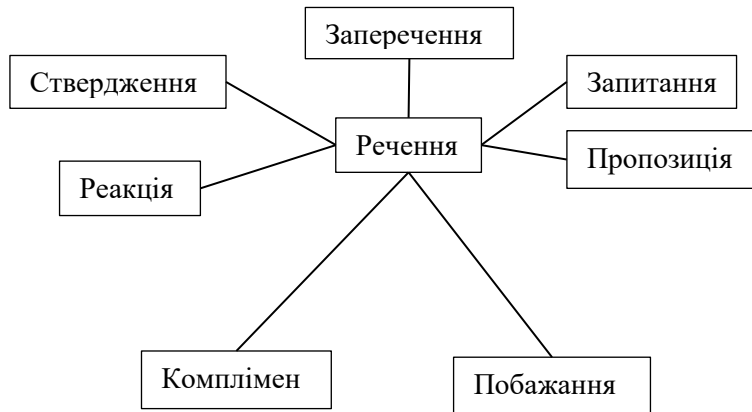


Рисунок 2.2 – Види речень

На рисунку 2.3 зображено можливі емоційні забарвлення тексту, які представляють інтерес у контексті інтернет-магазину.

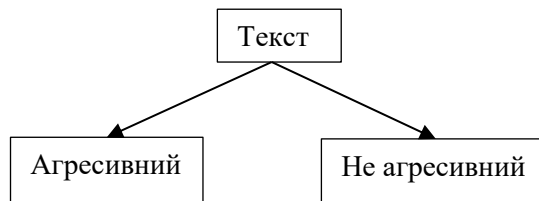


Рисунок 2.3 – Емоційні забарвлення тексту

На рисунку 2.4 зображено з чого складається сенс речення.

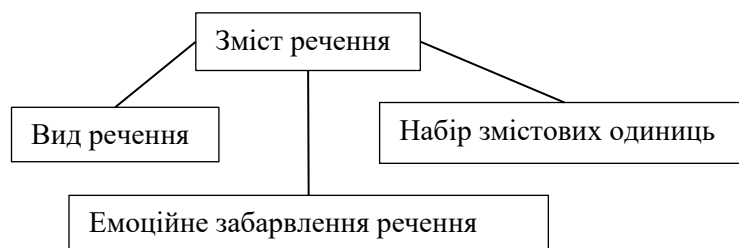


Рисунок 2.4 – Схема змісту речення

Отже, отримуючи коротке текстове повідомлення, необхідно розбити його на речення, а речення на слова і знаки. Потім із слів і знаків утворити

сміслові одиниці. Також по словам можна виявити вид речення та його емоційне забарвлення.

2.2 Структура вхідних даних та схема методу інтелектуального підбору відповідей на запитання за семантичними ознаками

На рисунку 2.5 зображено загальну схему методу інтелектуального підбору відповідей на запитання за семантичними ознаками.

Етап 1 – Попередня обробка тексту.

Отримавши на вході коротке текстове повідомлення, метод розбиває його на необхідні складові, а саме певні самостійні слова, які несуть якийсь сенс в незалежності від контексту. Для цього необхідно провести попередню обробку тексту, а саме токенізацію [1] та стеммінг [1].

Суть токенізації полягає у розбитті тексту на окремі слова та роздільні символи. Тобто на вході ми отримуємо певне речення, а на виході колекцію всіх слів та роздільних символів цього речення. Приклад токенізації зображено на рисунку 2.6.

Для цього можна вважати кожний набір символів з обох боків обмежений пробільними символами лексемою. Таким чином, можна отримати набір окремих слів та прикріплених до них знаків пунктуації.

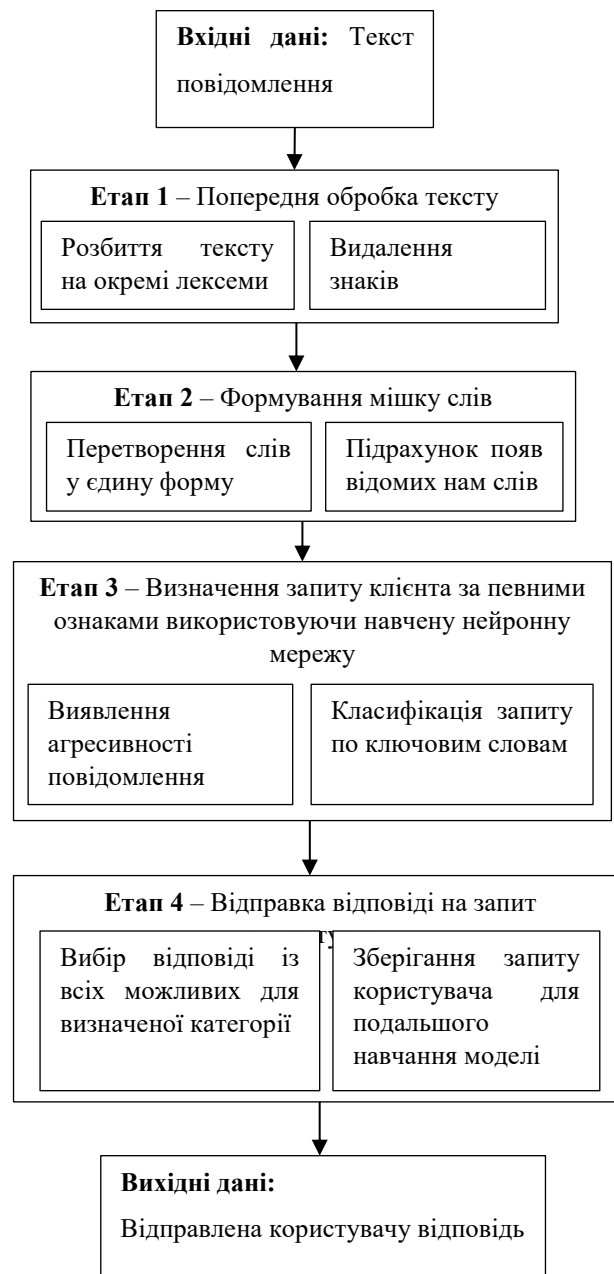


Рисунок 2.5 – Загальна схема методу інтелектуального підбору відповідей на запитання за семантичними ознаками

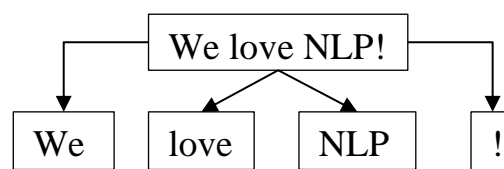


Рисунок 2.6 – Приклад токенізації

Стеммінг полягає у знаходженні певної загальної форми для певного слова. Тобто на вході ми отримуємо слово, до прикладу «велосипедів», а на

виході отримуємо «велосипед», тобто загальну форму цього слова. Це необхідно для того щоб слова у всіх своїх формах однаково сприймалися при аналізі.

Токенізатор [1] також можна використовувати для розширення скорочень, наприклад, перетворюючи what're на два лексеми what are для англійської мови, або для аналогічних цілей в інших мовах.

Токенізацію потрібно запускати перед будь-яким іншим етапами, і вона повинна бути швидкою. Стандартним методом [1] токенизації є використання детермінованих алгоритмів, заснованих на регулярних виразах, складених у дуже ефективні кінцеві автомати.

Ретельно розроблені детерміновані алгоритми можуть впоратися з двозначностями, які виникають, наприклад з апострофами.

Існує ще один варіант токенизації тексту [1]. Замість того, щоб визначати лексеми як слова (незалежно від того чи вони розділені пробілами), можна використовувати дані для аналізу того, якими мають бути токени. Це особливо корисно під час роботи з невідомими словами, що є важливою проблемою в обробки мови.

Значна частина обробки природної мови пов'язана з вимірюванням того, наскільки подібні два рядки є [1]. Для цього можна скористатися таким поняттям як відстань редагування.

Відстань редагування дає спосіб кількісно оцінити схожість двох слів. Відстань редагування дорівнює кількості операцій редагування (таких як вставка, видалення, підстановка) необхідних для того, щоб перетворити одне слово в інше. Чим більше операцій необхідно, тим менш схожі слова.

Етап 2 – Формування мішку слів.

Для того щоб сформувати мішок слів, необхідно привести слова у певну єдину форму, тобто кожне слово повинне бути у 1 відмінку, без доданих префіксів або суфіксів.

Далі необхідно підрахувати частоту кожного слова та створити словник, де будуть слова та частота їх появ у тексті.

Етап 3 – Визначення запиту клієнта за певними ознаками.

На даному етапі потрібно проаналізувати слова із мішка слів, знайти там ключові слова і по ним класифікувати запит клієнта. Тобто зрозуміти що саме він хотів. Для цього необхідно навчити нейронну мережу класифікувати набори ключових слів, використавши певні тренувальні дані. В якості мережі пропонується використовувати багат шаровий перцептрон [1], так як він добре підходить під задачі класифікації.

Також на цьому етапі потрібно виявити ступінь агресивності повідомлення, для того що наприклад одразу перенаправляти агресивного користувача до менеджера. Для цього знадобиться словник агресивних і звичайних слів української мови, в якому будуть слова і оцінки їх агресивності.

Етап 4 – Відправка відповіді користувачу.

На цьому етапі необхідно вибрати якусь відповідь із тих, які відповідають категорії запиту. Таких відповідей може бути декілька, вибирати якусь конкретну можна випадковим чином, або проводити додатковий аналіз.

Також необхідно зберегти запит клієнта у певному вигляді (у файлі або базі даних) для того, щоб у подальшому донавчити модель за допомогою цих даних.

Структура вхідних даних

Для роботи методу інтелектуального підбору відповідей на запитання за семантичними ознаками необхідно створити певний набір категорій, а також для кожної категорії визначити шаблони, притаманні лише їй. Крім цього необхідно визначити відповіді на запитання для кожної категорії. Для реалізації цього задуму, було створено файл, який має структуру, зображену на рисунку 2.7.

```

"intents": [
  {
    "tag": "greeting",
    "patterns": [
      "Привіт",
      "Добрий день",
      "Добрий ранок",
      "Добрий вечір",
      "Здрастуйте",
      "Бажаю здоров'я"
    ],
    "responses": [
      "Доброго дня, дякуємо за звернення, постараємось, надати найбільш підходящу відповідь",
      "Добрий день, чим можемо допомогти"
    ]
  },
]

```

Рисунок 2.7 – Структура файлу з категоріями

Даний файл використовується для навчання мережі, а також у процесі її роботи. По ньому система обирає відповіді для обраної категорії, а також із нього отримує всі можливі категорії та шаблони, які її характеризують.

Для правильної класифікації повідомлення було вдосконалено вже існуючий метод класифікації, для якого була розроблена нейронна мережа, а саме багат шаровий перцептрон [1]. Мережа навчена за допомогою даних із навчального файлу, структура якого зображена вище. Вдосконалення стосуються підбору підходящих параметрів для обробки української мови, так як існуючий метод використовувався для обробки англійської мови.

Далі для навчання мережі необхідний певний навчальний корпус [1], іншими словами набір навчальними даних представлених підходящій формі. У даному випадку, необхідно визначати всі можливі категорії повідомлень, а також для кожної категорії визначити набір шаблонів, притаманних цій категорії. Крім цього для кожної категорії необхідно визначити набір правильних відповідей на повідомлення цієї категорії. Для цих цілей можна знайти вже готовий корпус, або сформуванати його самому вручну. Для цього дослідження був знайдений та перекладений на українську мову готовий корпус для англійської мови.

В якості нейронної мережі використовується стандартна нейронна мережа з прямим зв'язком і 2 схованими шарами. Мережа приймає певний набір

параметрів [1]: кількість епох навчання, швидкість навчання, кількість нейронів у прихованих шарах. Кількість вхідних нейронів дорівнює кількості шаблонів, по яким мережа буде визначати категорію. Кількість вихідних нейронів дорівнює кількості категорій повідомлення. Варто зазначити, що параметри мережі можна підбирати самі різні, і певні набори параметрів можуть більше підходити під конкретну мову через її певні особливості. Підбір підходящих параметрів для української мови є вдосконаленням існуючого підходу для англійської мови.

Після того як мережа навчилася класифікувати отримані короткі текстові повідомлення, результат її роботи можна використати для генерації відповіді на запитання користувача.

Мережа на вході буде отримувати ключові слова певного повідомлення, а на виході давати визначену категорію цього повідомлення. Для того щоб отримати ключові слова повідомлення необхідно провести попередню обробку тексту.

Мережа використовує певний критерій для оцінки результатів роботи, а також оптимізатор. Підібравши правильний критерій [15] та оптимізатор [15] для роботи з українською мовою можна збільшити ефективність роботи мережі. В даному випадку використовується метод перехресної ентропії [1], який вимірює продуктивність моделі класифікації в межах від 0 до 1. Перехресна ентропійна втрата збільшується, коли прогнозована ймовірність розходиться з фактичною міткою. Ідеальна модель матиме втрату близьку нулю.

Покращити роботу мережі для роботи з українською мовою можна, підібравши підходящі параметри, а саме – обрати правильний критерій та оптимізатор, визначити підходящу швидкість навчання, кількість епох навчання, а також кількість нейронів у прихованих шарах.

Аналіз тональності тексту.

В багатьох випадках чат-бот повинен вміти аналізувати тональність повідомлення, для того щоб вчасно реагувати на агресивних користувачів та

перенаправляти їх на живих менеджерів. Такий функціонал може бути дуже корисним багатьом компаніям або установам.

Для аналізу тональності повідомлення був використаний метод, в основі якого лежить словник з словами та їх тональністю. Далі метод [9] використовує певні правила та бі-грами [1] для визначення тональності. У самому словнику зберігаються певні слова та визначена людиною тональність у діапазоні від -2 до 2, де -2 означає різко негативне слово, 0 – нейтральне, 2 – дуже позитивне.

Далі метод, отримавши текст, проводить попередню обробку тексту, аналогічну описаній вище для семантичного аналізу тексту, тобто в кінці отримує мішок слів [1] з тими словами, які зустрілися у тексті.

Далі метод шукає слова із мішку у словнику та отримує їх тональність. Після цього, використовуючи певні правила та враховуючи кількість появ агресивних, нейтральних та позитивних слів у тексті розраховує тональність тексту у діапазоні від -1 до 1, де -1 означає дуже негативний текст, 0 – нейтральний, 1 – дуже позитивний.

В поєднанні з класифікацією повідомлення можна більш точно розуміти чого хоче користувач і гнучкіше реагувати на його потреби.

Отже, було покращено існуючий метод генерації відповіді на коротке текстове повідомлення. А саме – були підібрані підходящі для української мови параметри нейронної мережі, а також при генерації відповіді була врахована тональність повідомлення, використовуючи існуючий метод аналізу тональності тексту. Сам метод класифікує отримане повідомлення та аналізує його тональність для того, щоб підібрати найбільш підходящу користувачу відповідь.

2.3 Інформаційна модель оцінки відповідності компонентів моделі «запитання-прототип-відповідь»

Важливою частиною створення інформаційної системи є вимірювання її ефективності. В даному випадку необхідно виміряти ефективність роботи нейронної мережі по класифікації короткого текстового повідомлення,

ефективність методу аналізу тональності тексту, а також відповідність згенерованої методом відповіді контексту отриманого короткого текстового повідомлення.

Виміряти ефективність роботи нейронної мережі можна по наступним критеріям: відсоток кінцевих втрат, відсоток правильно визначених категорій. Для того щоб виміряти відсоток правильно визначених категорій, потрібно створити дані для тестування, а саме – набір правильно визначених категорій та шаблонів, притаманним певній категорії. Далі необхідно створити до 1000 коротких текстових повідомлень та визначити для кожного з них правильну категорію. Далі після навчання мережі за допомогою даних про категорії та шаблони необхідно підрахувати для скількох повідомлень із 1000 мережа правильно визначить категорію. Далі формула оцінки ефективності роботи мережі буде виглядати наступним чином:

$$p = \frac{n}{m} * 100\%,$$

де n – кількість правильно класифікованих повідомлень, а m – загальна кількість тестових повідомлень.

Замірявши відсоток кінцевих втрат з існуючими параметрами та з покращеними параметрами, були отримані наступні результати.

Таблиця 2.1. – Результати порівняння роботи існуючої нейронної мережі та покращеної

	кількість епох навчання	швидкість навчання	кількість нейронів у прихованих шарах	Відсоток кінцевих втрат
Вже існуючий метод	1000	0.001	8	0.0011
Покращений метод	600	0.01	8	0.0000183

Як видно з таблиці, відсоток кінцевих втрат мережі суттєво зменшився, що показує позитивний вплив зміни параметрів мережі на більш підходящі для української мови.

Для того щоб виміряти ефективність роботи методу аналізу тональності повідомлення спочатку необхідно створити тестовий набір повідомлень та вручну визначити їхню тональність у діапазоні від -1 до 1, де -1 означає дуже агресивне, 0 – нейтральне, 1 – дуже позитивне повідомлення. Далі необхідно визначити максимально допустиму похибку аналізу. В цьому випадку вона становитиме 0.2. Далі необхідно виміряти тональність кожного з тестових повідомлень за допомогою використаного методу. Далі необхідно підрахувати кількість повідомлень, для яких тональність визначена методом, не відрізняється більше, ніж на значення похибки від вручну вирахованої тональності. Формула виглядає наступним чином:

$$p = \frac{n}{m} * 100\%, \quad (2.1)$$

де: n – кількість повідомлень, для яких $[k - f] \geq 0.2$,

k – тональність визначена методом,

f – тональність визначена вручну.

Також для обрахунку ефективності роботи методу можна вирахувати економічний ефект використання методу на конкретному сайті. Для цього можна вирахувати приріст клієнтів або конвертованості відвідувань в покупки. Формула виглядає наступним чином:

$$p = \frac{|k-n|}{k} * 100\%, \quad (2.2)$$

де k – кількість покупців в певний момент,

n – кількість покупців на момент раніше.

Чим більший приріст покупців або конвертованості тим краще для сайту.

Отже, наведеними розрахунками доведена ефективність роботи методу інтелектуального підбору відповіді на запитання за семантичними ознаками. Наведені розрахунки показують, що правильний підбір параметрів нейронної мережі зменшує відсоток кінцевих втрат нейронної мережі, а також дозволяє здійснити навчання мережі за меншу кількість ітерацій.

Висновки до розділу 2

Отже, у другому розділі були наведені основні складові методу інтелектуального підбору відповідей до запитань за семантичними ознаками. Була описана інформаційна модель методу інтелектуального підбору відповідей до запитань за семантичними ознаками. Була наведена схема методу та описано кожний етап його роботи.

Був описаний процес попередньої обробки тексту, який проводиться перед класифікацією. В розділі був описаний підхід до аналізу тональності тексту, який дозволяє більш точно класифікувати емоційно забарвленні повідомлення.

Також, у розділі була наведена інформаційна модель оцінки відповідності компонентів моделі «запитання-прототип-відповідь», яка показує доцільність кроків методу інтелектуального підбору відповіді на запитання за семантичними ознаками.

Розділ 3

Інформаційна система інтелектуального підбору відповідей до запитань за семантичними ознаками

3.1 Проектування інформаційної системи інтелектуального підбору відповідей на запитання за семантичними ознаками

При розробці тестової дослідницької реалізації методу інтелектуального підбору відповідей до запитань за семантичними ознаками потрібно реалізувати наступні бізнес-процеси:

1. Надання користувачу можливості писати короткі текстові повідомлення.
2. Семантичний аналіз отриманого повідомлення.
3. Відправка відповіді на повідомлення.
4. Створення завчасно згенерованих відповідей на запитання.
5. Робота з базою даних.

Бізнес-процес «Надання користувачу можливості писати короткі текстові повідомлення». Система повинна надавати користувачу можливість написати повідомлення використовуючи певний інтерфейс. Реалізовано це буде завдяки Telegram Bot API [13]. Діаграма дій для даного бізнес-процесу зображена на рисунку 3.1.

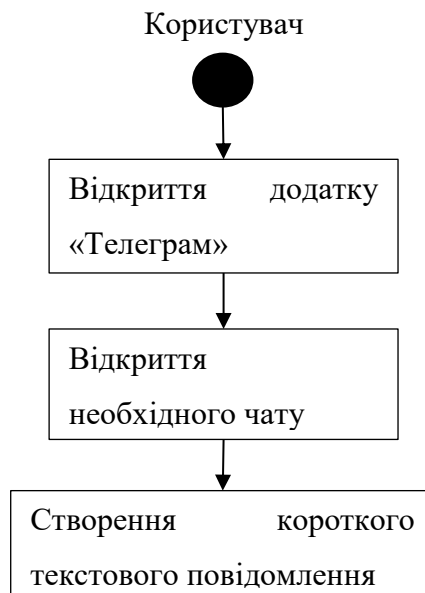


Рисунок 3.1 – Діаграма дій бізнес-процесу «Надання користувачу можливості писати короткі текстові повідомлення»

Бізнес-процес «Семантичний аналіз отриманого повідомлення». Система повинна проаналізувати отримане повідомлення для того щоб в подальшому зрозуміти яку відповідь потрібно обрати або згенерувати. Для цього буде проводитись аналіз ключових слів і по ним повідомлення буде відноситись до певної завчасно створеної категорії, для якої будуть заготовлені певні підходящі їй відповіді. Діаграма дій для даного бізнес-процесу зображена на рисунку 3.2.

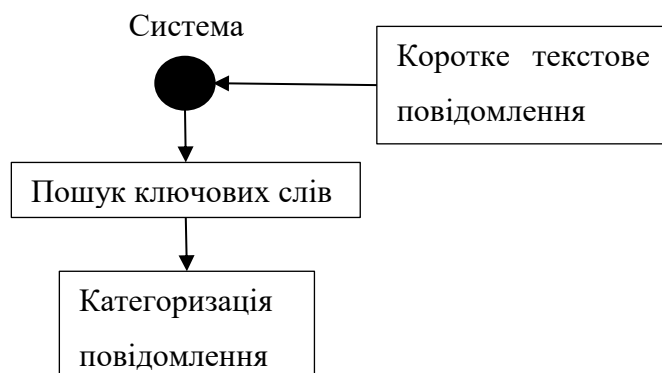


Рисунок 3.2 – Діаграма дій бізнес-процесу «Семантичний аналіз отриманого повідомлення»

Бізнес-процес «Відправка відповіді на повідомлення». Після віднесення повідомлення до певної категорії, система повинна вибрати підходяще цій категорії завчасно згенероване повідомлення та відправити його користувачу. Діаграма дій для даного бізнес-процесу зображена на рисунку 3.3.

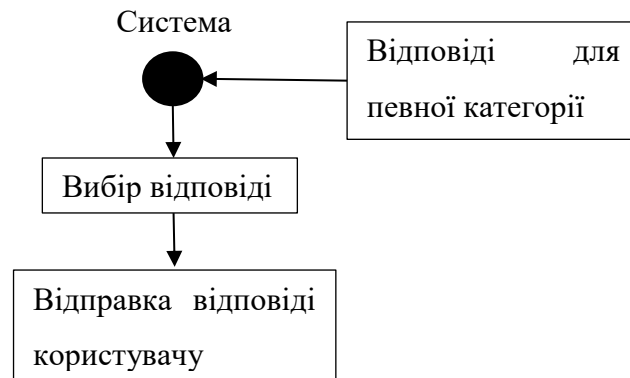


Рисунок 3.3 – Діаграма дій бізнес-процесу «Відправка відповіді на повідомлення»

Бізнес-процес «Створення завчасно згенерованих відповідей на запитання». Система повинна завчасно згенерувати відповіді на всі питання, на які вона повинна знати відповідь. Для того щоб потім при виявленні певного питання після семантичного аналізу відсилати на нього відповідь. На невідомі питання система повинна відправляти стандартну відповідь, наприклад «Вибачте, я вас не розумію». Діаграма дій для даного бізнес-процесу зображена на рисунку 3.4.

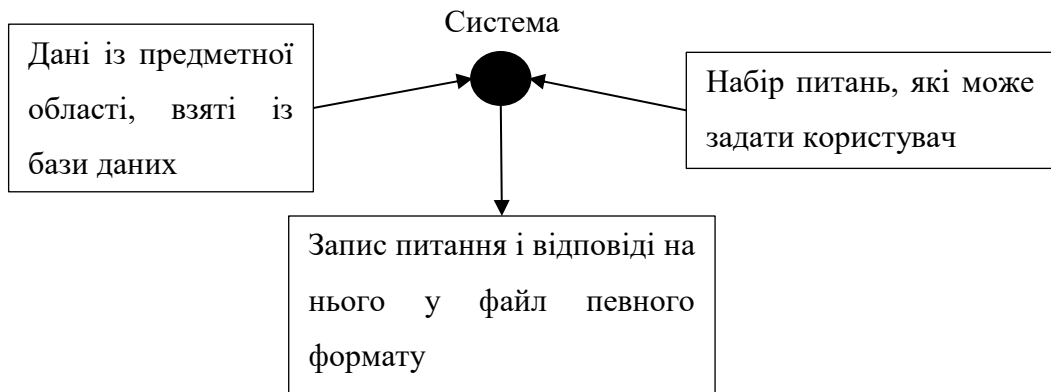


Рисунок 3.4 – Діаграма дій бізнес-процесу «Створення завчасно згенерованих відповідей на запитання»

Бізнес-процес «Робота з базою даних». Система повинна вміти працювати з базою даних для того, щоб витягувати звідти необхідні для відповідей дані. Діаграма дій для даного бізнес-процесу зображена на рисунку 3.5.

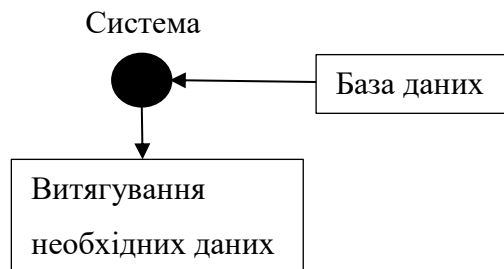


Рисунок 3.5 – Діаграма дій бізнес-процесу «Робота з базою даних»

Отже, інформаційна система методу інтелектуального підбору відповідей до запитань за семантичними ознаками реалізує наступні бізнес-процеси:

- Бізнес-процес «Надання користувачу можливості писати короткі текстові повідомлення».
- Бізнес-процес «Семантичний аналіз отриманого повідомлення».
- Бізнес-процес «Відправка відповіді на повідомлення».
- Бізнес-процес «Створення завчасно згенерованих відповідей на запитання».

– Бізнес-процес «Робота з базою даних».

3.2 Структура та інформаційне призначення складових системи

Інформаційна структура системи належить до клієнт-серверного типу. У якості клієнта використовується «Телеграм», а в якості сервера певний модуль, де буде зберігатись база даних, модель нейронної мережі, а також допоміжні модулі. Схема інформаційної структури системи зображена на рисунку 3.6.



Рисунок 3.6 – Схема інформаційної структури системи

Ключовим елементом системи є неймережа, яка займається семантичним аналізом і категоризацією повідомлення. Для того щоб неймережа робила це правильно, її потрібно навчити на певних тестових даних.

Також важливим елементом є база даних, із якої беруться важливі для відповідей данні. Самі дані можна брати або під час відправки відповіді або підставляти їх завчасно у згенеровані відповіді у файлі.

Отже, інформаційна система належить до клієнт-серверного типу та побудовану по схемі, зображеній на рисунку 3.6. Ключовим елементом системи є нейромережа, яка займається семантичним аналізом і категоризацією повідомлення.

3.3 Проектування структури бази даних інформаційної системи

Для даної системи була створена база даних, схема якої зображена на рисунку 3.7.

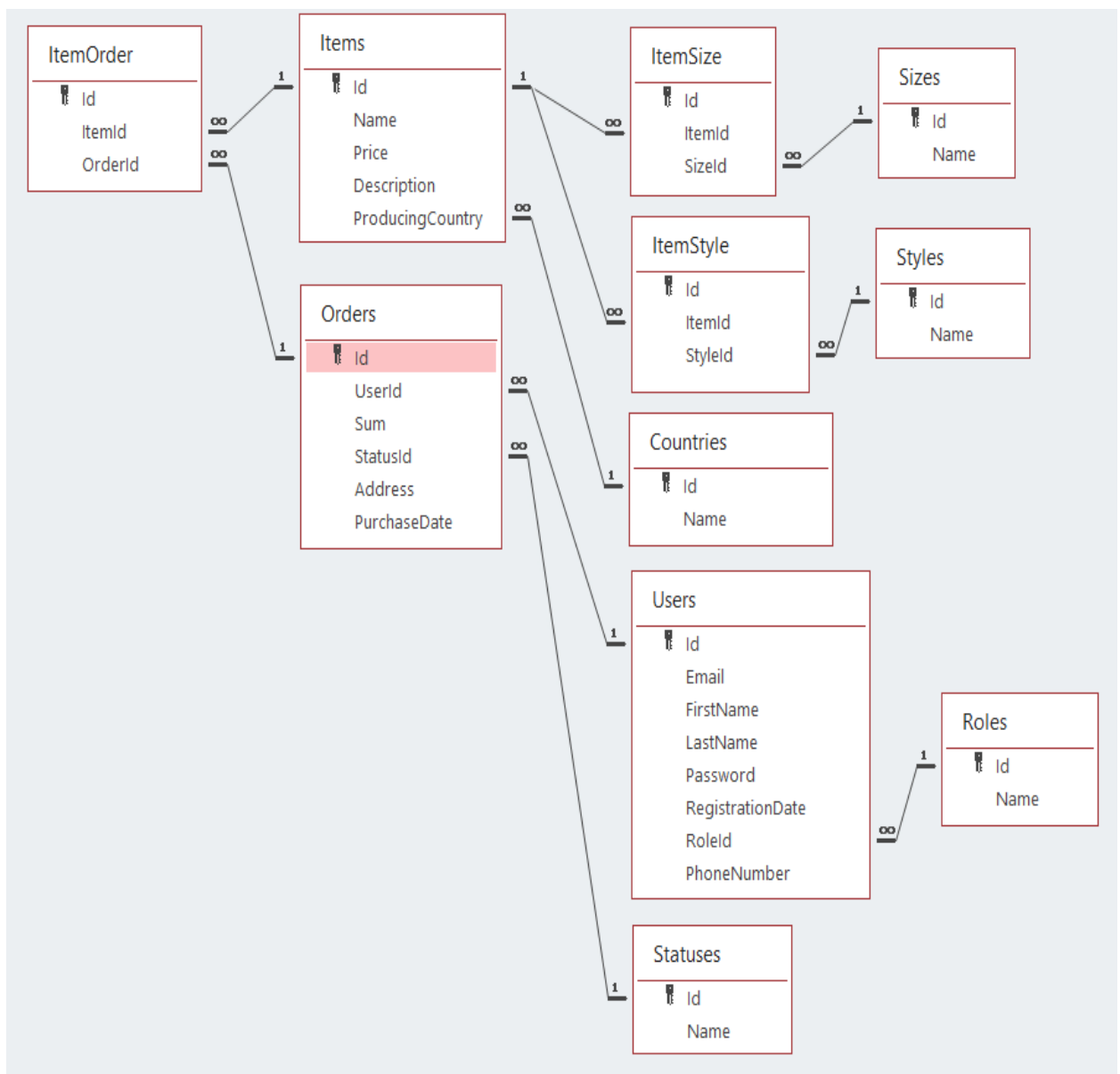


Рисунок 3.7 – Схема бази даних

Таблиця ItemOrder зберігає дані про всі товари, які були замовлені.

Таблиця 3.1 – Атрибути таблиці «ItemOrder»

№ п/п	Назва атрибуту	Тип даних	Опис
1.	Id	int	Первинний ключ, числовий ідентифікатор
2.	ItemId	int	Вторинний ключ, який вказує на таблицю Items
3.	OrderId	int	Вторинний ключ, який вказує на таблицю Orders

Таблиця Items зберігає дані про самі товари та їх характеристики.

Таблиця 3.2 – Атрибути таблиці «Items»

№ п/п	Назва атрибуту	Тип даних	Опис
1.	Id	int	Первинний ключ, числовий ідентифікатор
2.	Name	Varchar(50)	Назва товару
3.	Price	Varchar(50)	Ціна товару
4.	Description	Varchar(50)	Опис товару
5.	ProducingCountry	int	Вторинний ключ, який вказує на таблицю Countries

Таблиця Orders зберігає дані про самі замовлення, які робили користувачі.

Таблиця 3.3 – Атрибути таблиці «Orders»

№ п/п	Назва атрибуту	Тип даних	Опис
1.	Id	int	Первинний ключ, числовий ідентифікатор
2.	UserId	Varchar(50)	Вторинний ключ, який вказує на таблицю Users
3.	Sum	Varchar(50)	Загальна вартість замовлення
4.	StatusId	Varchar(50)	Вторинний ключ, який вказує на таблицю Statuses
5.	Address	int	Адреса доставки замовлення
6.	PurchaseDate	DateTime	Дата і час доставки замовлення

Дані користувачів зберігаються у таблиці Users.

Таблиця 3.4 – Атрибути таблиці «Users»

№ п/п	Назва атрибуту	Тип даних	Опис
1.	Id	int	Первинний ключ, числовий ідентифікатор
2.	Email	Varchar(50)	Email користувача
3.	FirstName	Varchar(50)	Ім'я користувача
4.	LastName	Varchar(50)	Прізвище користувача
5.	Password	Varchar(50)	Пароль
6.	RegistrationDate	DateTime	Дата і час реєстрації
7.	RoleId	int	Вторинний ключ, який вказує на таблицю Statuses Roles
8.	PhoneNumber	Varchar(50)	Номер телефону

Таблиця Statuses зберігає дані про статус замовлення.

Таблиця 3.5 – Атрибути таблиці «Statuses»

№ п/п	Назва атрибуту	Тип даних	Опис
1.	Id	int	Первинний ключ, числовий ідентифікатор
2.	Name	Varchar(50)	Назва статусу

Таблиця Roles зберігає дані про ролі користувача.

Таблиця 3.6 – Атрибути таблиці «Roles»

№ п/п	Назва атрибуту	Тип даних	Опис
1.	Id	int	Первинний ключ, числовий ідентифікатор
2.	Name	Varchar(50)	Назва ролі

Таблиця Sizes зберігає дані про можливий розмір товару.

Таблиця 3.7 – Атрибути таблиці «Roles»

№ п/п	Назва атрибуту	Тип даних	Опис
1.	Id	int	Первинний ключ, числовий ідентифікатор
2.	Name	Varchar(50)	Назва розміру

Таблиця Styles зберігає дані про стилі товару.

Таблиця 3.8 – Атрибути таблиці «Styles»

№ п/п	Назва атрибуту	Тип даних	Опис
1.	Id	int	Первинний ключ, числовий ідентифікатор
2.	Name	Varchar(50)	Назва стилю

Таблиця Countries зберігає дані про країни.

Таблиця 3.9 – Атрибути таблиці «Countries»

№ п/п	Назва атрибуту	Тип даних	Опис
1.	Id	int	Первинний ключ, числовий ідентифікатор
2.	Name	Varchar(50)	Назва країни

Таблиця ItemSize є розвідною таблицею між таблицями Items і Sizes.

Таблиця 3.10 – Атрибути таблиці «ItemSize»

№ п/п	Назва атрибуту	Тип даних	Опис
1.	Id	int	Первинний ключ, числовий ідентифікатор
2.	ItemId	int	Вторинний ключ, який вказує на таблицю Statuses Items
3.	SizeId		Вторинний ключ, який вказує на таблицю Statuses Sizes

Таблиця ItemStyle є розвідною таблицею між таблицями Items і Styles.

Таблиця 3.11 – Атрибути таблиці «ItemStyle»

№ п/п	Назва атрибуту	Тип даних	Опис
1.	Id	int	Первинний ключ, числовий ідентифікатор
2.	ItemId	int	Вторинний ключ, який вказує на таблицю Statuses Items
3.	StyleId	int	Вторинний ключ, який вказує на таблицю Statuses Styles

Отже, відповідно до потреб інформаційної системи було створено БД з структурою описаною вище. БД зберігає дані, необхідні для генерування відповідей на конкретні запитання користувачів з приводу виставлених в інтернет-магазині товарів, зроблених замовлень, можливостей доставки і т.д.

3.4 Вибір засобів розробки

Для розробки прикладної програмної дослідницької тестової реалізації методу інтелектуального підбору відповідей на запитання за семантичними ознаками потрібно обрати мову програмування, СКБД та фреймворки.

Вибір мови програмування

Python [21] – скриптова мова програмування з динамічною типізацією. Вона дозволяє використовувати як функціональну, так і об'єктно-орієнтовану парадигму програмування. Завдяки динамічній типізації ми можемо більш

вільно працювати з типами даних, порівняно з такими мовами програмування як C++ [14].

Основною перевагою Python над багатьма іншими мовами програмування є дуже широкий вибір бібліотек та фреймворків, які дозволяють програмувати усі види додатків, використовувати штучний інтелект і т. д. Використовуючи дану мову програмування можна дуже серйозно економити час програміста за рахунок використання вже написаного коду.

На цій мові програмування достатньо легко створювати взаємодію з базою даних або з файловою системою. Також тут є багато вбудованої підтримки що до обробки тексту та штучного інтелекту. Тому для даного проекту було обрано саме Python.

Вибір фреймворків

PyTorch [15] – це пакет Python, який надає дві функції високого рівня:

- Тензорні обчислення із сильним прискоренням GPU.
- Глибокі нейронні мережі, побудовані на системі стрічкової автоградації.

PyTorch має унікальний спосіб побудови нейронних мереж: використання та відтворення магнітофона.

Більшість фреймворків, таких як TensorFlow, Theano, Caffe та CNTK, мають статичний погляд на світ. Необхідно будувати нейронну мережу і повторно використовувати ту саму структуру знову і знову. Зміна поведінки мережі означає, що потрібно починати з нуля.

У PyTorch [15] використовується техніка, яка називається автоматичною диференціацією у зворотному режимі, яка дозволяє довільно змінювати поведінку мережі без затримок або витрат.

Загалом, це дуже зручний фреймворк для використання нейронних мереж, тому для цих цілей був обраний саме PyTorch.

NLTK [16] або набір інструментів природної мови є провідною платформою для створення програм Python для роботи з даними людської мови. Він надає прості у використанні інтерфейси для більш ніж 50 корпусів і

лексичних ресурсів, таких як WordNet, а також набір бібліотек для обробки тексту для класифікації, токенизації, тегування, синтаксичного та семантичного аналізу, обгортки для індустріальних бібліотек NLP.

Бібліотеку NLTK було обрано для різноманітних допоміжних задач обробки природної мови через її зручність та легкодоступність.

Telegram Bot API [13] – це інтерфейс на основі HTTP, створений для розробників, які зацікавлені в створенні ботів для Telegram. За допомогою цього зручного інтерфейсу можна створювати телеграм чат-ботів з різноманітним функціоналом. Існують і інші схожі інтерфейси, але завдяки зручності використання та доступній документації для створення чат-боту було обрано саме Telegram Bot API.

Вибір СКБД

MySQL [17] – це система керування реляційною базою даних або RDBMS, що означає, що вона зберігає та представляє дані в табличній формі, організованій у рядки та стовпці. Дана система має наступні переваги.

- Надійний рівень безпеки даних для захисту конфіденційних даних від зловмисників, а паролі в MySQL зашифровані.
- Система доступна для безкоштовного завантаження та використання з офіційного сайту MySQL.
- Система сумісна з більшістю операційних систем, включаючи Windows, Linux, NetWare, Novell, Solaris та інші варіанти UNIX.
- Забезпечує можливість запуску клієнтів і сервера на одному комп'ютері або на різних комп'ютерах через Інтернет або локальну мережу.
- Має унікальну архітектуру механізму зберігання, що робить його швидким, дешевим і надійним.
- Дає розробникам значну продуктивність завдяки використанню представлень, тригерів і збережених процедур
- Простий і легкий у використанні.

– Є масштабованим і здатним обробляти понад 50 мільйонів рядків. Цього достатньо для обробки майже будь-якого обсягу даних. Хоча стандартний розмір файлу становить 4 ГБ, його можна збільшити до 8 ТБ.

– Дозволяє відмінити транзакції.

Отже, завдяки простоті використання та наявній підтримці роботи з нейронними мережами та засобами обробки природної мови було обрано мову програмування Python. Завдяки простоті використання та повноті представлених методів обробки природної мови для цих цілей були обрані фреймворки PyTorch та NLTK. Для створення чат-боту завдяки простоті використання та повноті представлених функцій був обраний Telegram BOT API. В якості СКБД завдяки вказаним перевагам, таким як: доступність для безкоштовного завантаження, простота та легкість у використанні була обрана MySQL.

Висновки до розділу 3

Отже, у розділі 3 було показано інформаційну структуру створеної системи. Були описані наступні бізнес-процеси:

– Бізнес-процес «Надання користувачу можливості писати короткі текстові повідомлення».

– Бізнес-процес «Семантичний аналіз отриманого повідомлення».

– Бізнес-процес «Відправка відповіді на повідомлення».

– Бізнес-процес «Створення завчасно згенерованих відповідей на запитання».

– Бізнес-процес «Робота з базою даних».

Було показано, що інформаційна система належить до клієнт-серверного типу і що її ключовим елементом є нейромережа, яка займається семантичним аналізом і категоризацією повідомлення.

Відповідно до потреб інформаційної системи було створено БД, яка зберігає дані, необхідні для генерування відповідей на конкретні запитання

користувачів з приводу виставлених в інтернет-магазині товарів, зроблених замовлень, можливостей доставки і т. д.

Також у відповідності до переваг і недоліків були обрані мова програмування, фреймворки та СКБД для розробки інформаційної системи.

Розділ 4

Дослідження ефективності методу інтелектуального підбору відповідей до запитань за семантичними ознаками

4.1 Розробка прикладних компонентів інформаційної системи

Для розробки тестового додатку за методом інтелектуального підбору відповідей на запитання за семантичними ознаками було використано мову програмування Python та фреймворки PyTorch, NLTK, Telegram API.

Фрагмент коду для навчання мережі класифікувати короткі текстові повідомлення показано нижче.

```
from nltk_utils import bag_of_words, tokenize, stem
from model import NeuralNet

with open('intents.json', 'r', encoding='utf-8') as f:
    intents = json.load(f)

all_words = []
tags = []
xy = []
# loop through each sentence in our intents patterns
for intent in intents['intents']:
    tag = intent['tag']
    # add to tag list
    tags.append(tag)
    for pattern in intent['patterns']:
        # tokenize each word in the sentence
        print(pattern)
        w = tokenize(pattern)
        # add to our words list
        all_words.extend(w)
        # add to xy pair
        xy.append((w, tag))
```

Для навчання мережі формується файл, в якому знаходяться короткі текстові повідомлення, їх правильна категорія та правильні відповіді на ці повідомлення. Фрагмент коду зображено нижче.

```
{
  "intents": [
    {
      "tag": "greeting",
      "patterns": [
        "Привіт",
        "Добрий день",
        "Добрий ранок",
        "Добрий вечір",
        "Здрастуйте",
        "Бажаю здоров'я"
      ],
      "responses": [
        "Доброго дня, дякуємо за звернення, постараємось, надати найбільш підходящу відповідь",
        "Добрий день, чим можемо допомогти"
      ]
    },
  ],
}
```

Дані зберігаються у json [18] файлі і далі використовуються для навчання мережі. При навчанні мережі дається певний набір повідомлень із файлу та їх правильна категорія, далі мережа після певної кількості спроб правильно класифікує питання, яких немає серед навчальних даних.

На певному етапі створення відповіді на коротке текстове повідомлення необхідно сформувавши мішок слів. Для цього використовується код зображений нижче.

```

def bag_of_words(tokenized_sentence, words):
    # stem each word
    sentence_words = [stem(word) for word in tokenized_sentence]
    # initialize bag with 0 for each word
    bag = np.zeros(len(words), dtype=np.float32)
    for idx, w in enumerate(words):
        if w in sentence_words:
            bag[idx] = 1

    return bag

```

Також для самого створення відповіді на коротке повідомлення користувача вже після того як мережа була навчена потрібна певна функція. Код цієї функції зображений нижче.

```

def reply(text):
    words = tokenize(text)

    normal_words=[]
    without_stop_words = [word for word in words if not word in stop_words]

    for token in without_stop_words:
        p = morph.parse(token)[0]
        normal_words.append(p.normal_form) # type: ignore

    X = bag_of_words(words, all_words)
    X = X.reshape(1, X.shape[0])
    X = torch.from_numpy(X).to(device)

    output = model(X)
    _, predicted = torch.max(output, dim=1)

    tag = tags[predicted.item()]

    callManager = SIA.polarity_scores(text)["compound"] < 0.0 or SIA.polarity_scores('
'.join(normal_words))["compound"] < 0.0

    probs = torch.softmax(output, dim=1)
    prob = probs[0][predicted.item()]

```

Як видно з фрагменту коду в цій функції ми використовуємо одну мережу для класифікації отриманого повідомлення та іншу мережу для аналізу тональності повідомлення. Далі отримавши категорію і тональність повідомлення ми обираємо правильну відповідь на повідомлення.

Також необхідно використовуючи Telegram API, запустити телеграм-бота, який буде в режимі реального часу відповідати на повідомлення користувачів. Для цього був використаний код зображений нижче.

```
import os
import telebot
from dotenv import load_dotenv
import assistant

load_dotenv()

TELEGRAM_API_KEY = os.getenv('TELEGRAM_API_KEY')

bot = telebot.TeleBot(TELEGRAM_API_KEY) # type: ignore

@bot.message_handler(content_types=["text"])

def replyToMessage(message):
    bot.send_message(message.chat.id, assistant.reply(message.text)) # type: ignore

bot.polling()
```

Як видно з фрагменту коду, бот запускається в режимі очікування повідомлень, тобто він буде відповідати на повідомлення, поки його не вимкнуть, а саму відповідь він бере із функції описаної вище.

Отже, за розробленою архітектурою та з використанням зображеного вище коду було розроблену прикладну програмну дослідницьку тестову реалізацію методу інтелектуального підбору відповідей на запитання за семантичними ознаками.

4.2 Прикладне тестування інформаційної системи

Для тестування даної інформаційної системи інтелектуального підбору відповідей на запитання за семантичними ознаками було розроблено набір тест-кейсів для перевірки чи правильно поводить себе система в певних випадках.

Потрібно перевірити чи чат-бот буде відсилати очікувану відповідь на коротке текстове повідомлення. В цьому ми тесті ми автоматично також перевіriamo чи чат-бот взагалі запускається і чи є можливість писати повідомлення.

Для цього треба запустити Телеграм та почати чат із розробленим чат-ботом. Далі необхідно написати повідомлення українською мовою та натиснути кнопку «Enter». Тест кейс зображено у таблиці 4.1.

Таблиця 4.1 – Тест-кейс 1

Тест-кейс ID: 1	Пріоритет: 1	Створено: 22.10.2022, Вишинський І.О.
Назва: Перевірка коректності відправки відповіді на коректне коротке текстове повідомлення		
Вхідні дані: Коротке текстове повідомлення користувача		
Кроки	Очікуваний результат	
<ol style="list-style-type: none"> 1. Запустити телеграм 2. Знайти чат-бот 3. Ввести «Ви приймаєте кредитну картку?» у текстовому полі 4. Натиснути кнопку «Enter» 	Отримане коротке текстове повідомлення «Ми приймаємо VISA та Mastercard».	
Результат виконання тест-кейсу: пройдено успішно		

Далі необхідно перевірити, що буде система відправити користувача очікуване повідомлення у випадку якщо система не зрозуміє користувача. Для цього необхідно зробити аналогічні першому тест-кейсу дії, але цього разу ввести повідомлення англійською мовою. Тест-кейс зображено у таблиці 4.2.

Таблиця 4.2 – Тест-кейс 2

Тест-кейс ID: 2	Пріоритет: 1	Створено: 22.10.2022, Вишинський І.О.
Назва: Перевірка коректності відправки відповіді на некоректне коротке текстове повідомлення		
Вхідні дані: Коротке текстове повідомлення користувача		
Кроки	Очікуваний результат	
<ol style="list-style-type: none"> 1. Запустити телеграм 2. Знайти чат-бот 3. Ввести «Hello do you accept mastercard?» у текстовому полі 4. Натиснути кнопку «Enter» 	Отримане коротке текстове повідомлення «Вибачте, я вас не зрозумів».	
Результат виконання тест-кейсу: пройдено успішно		

Необхідно перевірити чи система правильно аналізує тональність отриманого повідомлення. Для цього необхідно ввести агресивне повідомлення та перевірити чи система відреагує очікуваним чином. Тест-кейс зображено у таблиці 4.3.

Таблиця 4.3 – Тест-кейс 3

Тест-кейс ID: 3	Пріоритет: 1	Створено: 22.10.2022, Вишинський І.О.
Назва: Перевірка коректності відправки відповіді на коректне, агресивне, коротке текстове повідомлення		
Вхідні дані: Коротке текстове повідомлення користувача		
Кроки	Очікуваний результат	
<ol style="list-style-type: none"> 1. Запустити телеграм 2. Знайти чат-бот 3. Ввести «Дуже поганий сервіс» у текстовому полі 4. Натиснути кнопку «Enter» 	Отримане коротке текстове повідомлення «Зараз з вами зв'яжеться менеджер».	
Результат виконання тест-кейсу: пройдено успішно		

Отже, з використанням зазначених вище, а також інших тест-кейсів було перевірено правильність роботи додатку. Всі тести показують, що програма працює правильно і очікувано. Отже, весь функціонал реалізовано у відповідності до поставлених завдань.

4.3 Функціональне дослідження інформаційної системи

Для того щоб скористатися створеним додатком, спочатку необхідно запустити Телеграм. Далі необхідно знайти потрібного бота та почати з ним розмову. Інтерфейс взаємодії з ботом складається з текстового поля, де можна вводити текстове повідомлення, та частини сторінки де відображаються написані ботом та користувачем повідомлення. Вигляд цього інтерфейсу зображений на рисунку 4.1.

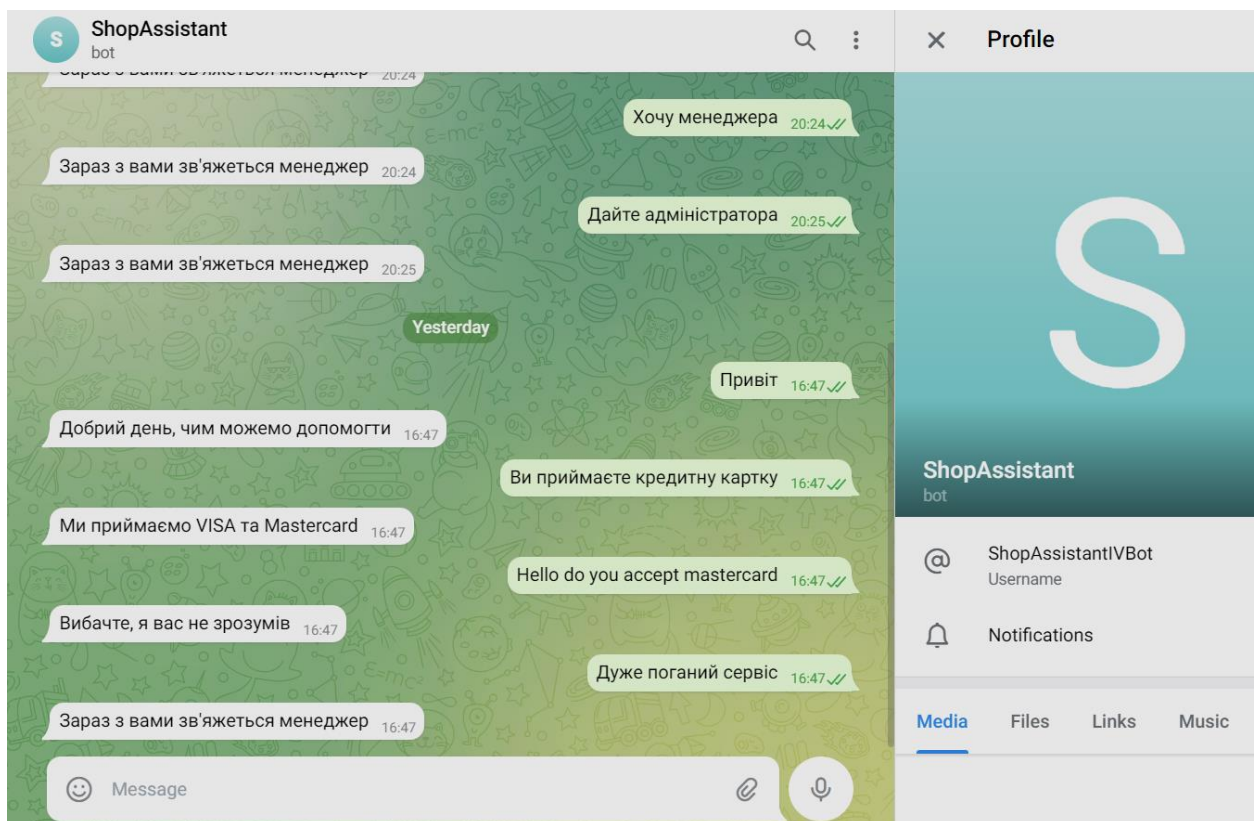


Рисунок 4.1 – Інтерфейс додатку

Як видно з рисунку, ми можемо використовувати різні можливості Телеграму, але розроблений тестовий додаток реагує тільки на текст повідомлення. Тобто він не обробляє файли, картинки чи смайлики. Використовуючи даний інтерфейс можна зрозуміти чи повідомлення було прочитано ботом, коли було створено останнє повідомлення, подивитися всю історію повідомлень та деякі інші речі.

На рисунку 4.2 видно розмову ще одного користувача з чат-ботом.

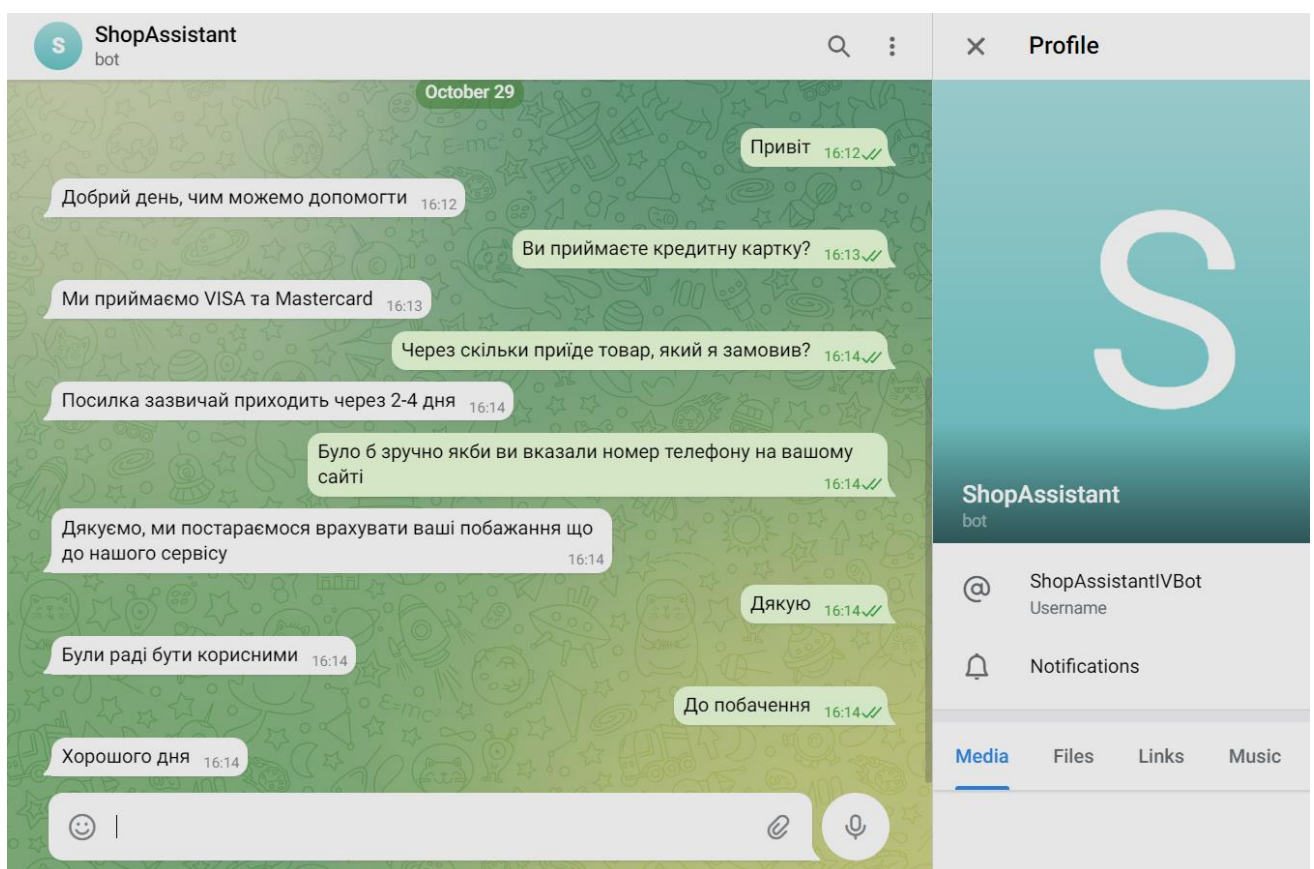


Рисунок 4.2 – Розмова користувача з чат-ботом

Як видно з рисунку, система здатна розрізняти питання про товар, про доставку, побажання, привітання і т.д.

На рисунку 4.3 зображено розмову незадоволеного користувача з чат-ботом.

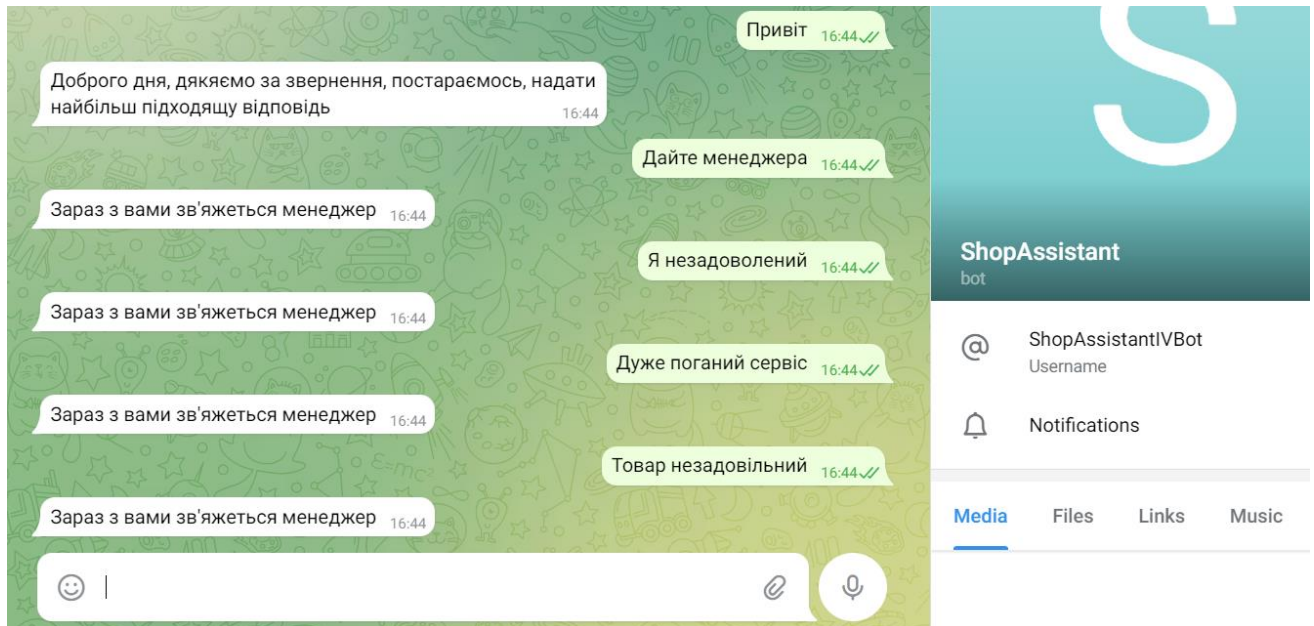


Рисунок 4.3 – Розмова незадоволеного користувача з чат-ботом

Як видно з рисунку, система здатна розуміти негативний настрій користувача. І в такому разі система правильно реагує і зв'язує менеджера з клієнтом. Це в свою чергу підвищує конвертованість відвідувань.

На рисунку 4.4. зображено спілкування англомовного користувача з чат-ботом.

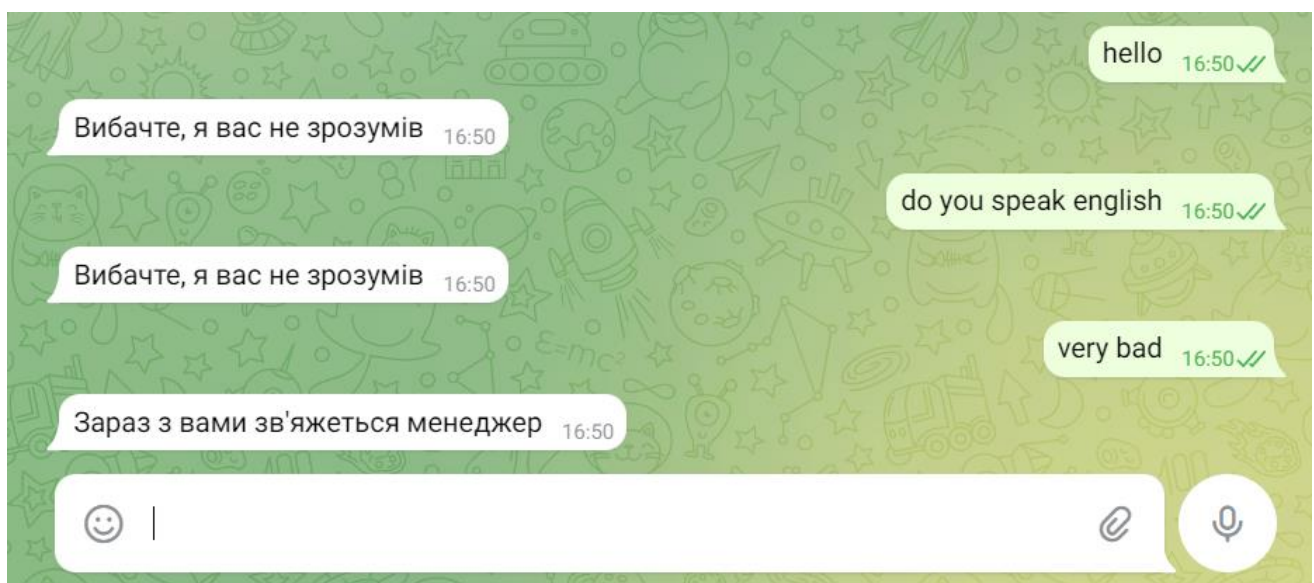


Рисунок 4.4 – Розмова англомовного користувача з чат-ботом

З рисунку видно, що на даний момент система не розуміє англійської мови. Цей недолік може стати напрямком подальших досліджень.

Для того щоб знайти необхідний чат можна скористатись текстовим полем пошуку розташованим у верхньому лівому кутку і зображеному на рисунку 4.5.

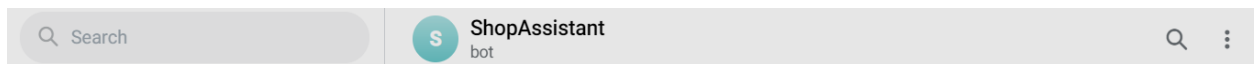


Рисунок 4.5 – Текстове поле для пошуку чату в Телеграм

Як видно з рисунку у текстовому полі можна ввести назву чату та знайти його у списку. Для того щоб навчати мережу, необхідно відкрити консоль та зайти в ній у папку з проектом та виконати команду «Python train.py». Це запустить на виконання Python файл, де прописане навчання моделі використовуючи вище описаний навчальний файл. Інтерфейс консолі зображено на рисунку 4.6.

```
(base) D:\Projects\PythonProjects>cd shopassistant
(base) D:\Projects\PythonProjects\ShopAssistant>python train.py
Привіт
Добрий день
Добрий ранок
Добрий вечір
Вдрастуйте
Бажаю здоров'я
До побачення
Прощавайте
На все добре
До зустрічі
Дякую
Спасибі
Було корисно
Корисна інформація
Які товари ви продаєте?
Що у вас є доступного?
Ви продаєте одяг?
Який одяг ви продаєте
Ви приймаєте кредитну картку?
Ви приймаєте Mastercard?
Я можу заплатити PayPal?
Скільки часу забере доставка?
Через скільки приїде товар, який я замовив?
Коли я отримаю посилку?
24 patterns
6 tags: ['delivery', 'goodbye', 'greeting', 'items', 'payments', 'thanks']
64 unique stemmed words: ['"', ',', 'mastercard', 'paypal', 'бажаю', 'було', 'вас', 'вечір', 'ви', 'все', 'день', 'до', 'добре', 'добрий', 'доставка', 'ров', 'здрастуйте', 'зустрічі', 'картку', 'коли', 'корисна', 'корисно', 'кредитну', 'можу', 'на', 'одяг', 'отримаю', 'побачення', 'посилку', 'привіт', 'е', 'ранок', 'скільки', 'спасибі', 'товар', 'товари', 'у', 'часу', 'через', 'шо', 'я', 'який', 'які', 'є', 'інформація']
64 6
Epoch [100/1000], Loss: 0.9031
Epoch [200/1000], Loss: 0.2127
Epoch [300/1000], Loss: 0.0276
Epoch [400/1000], Loss: 0.0198
Epoch [500/1000], Loss: 0.0064
Epoch [600/1000], Loss: 0.0021
Epoch [700/1000], Loss: 0.0019
Epoch [800/1000], Loss: 0.0014
Epoch [900/1000], Loss: 0.0007
Epoch [1000/1000], Loss: 0.0008
final loss: 0.0008
training complete. file saved to data.pth
(base) D:\Projects\PythonProjects\ShopAssistant>
```

Рисунок 4.6 – Виконання у консолі команди для навчання моделі

Як видно з рисунку, після виконання команди у консоль виводиться інформація що до навчання, а саме наскільки успішно пройшло навчання, скільки шаблонів, категорій та унікальних слів було сформовано. Також написано що дані що до навчання збережені у спеціальному файлі. Цей файл необхідний для правильного використання моделі.

Аналогічним чином із консолі відбувається запуск навчання мережі для розпізнавання тональності повідомлення та запуск самого додатку.

Отже, функціональне тестування системи показує, що в ній присутні заявлені в завданні компоненти, які дозволяють користувачу отримати згенеровану за допомогою інтелектуального підбору за семантичними ознаками відповідь на запитання.

4.4 Дослідження ефективності методу інтелектуального підбору відповідей на запитання за семантичними ознаками

Для дослідження ефективності методу інтелектуального підбору відповідей до запитань за семантичними ознаками з урахуванням тональності необхідно виміряти точність класифікації повідомлень та точність аналізу тональності. Для цього можна скористатися методами аналізу ефективності описаними у розділі 2.

Необхідно створити тестову вибірку із 1000 тестових повідомлень та вручну визначити для них категорію та тональність. Спочатку була виміряна точність класифікації повідомлення. Для цього, всі повідомлення були класифіковані методом, використовуючи існуючі параметри. Далі процес був повторений для нейронної мережі, використовуючи покращені параметри. Далі, використавши формулу із розділу 2, була вирахувана точність класифікації. Результати зображені на рисунку 4.7.

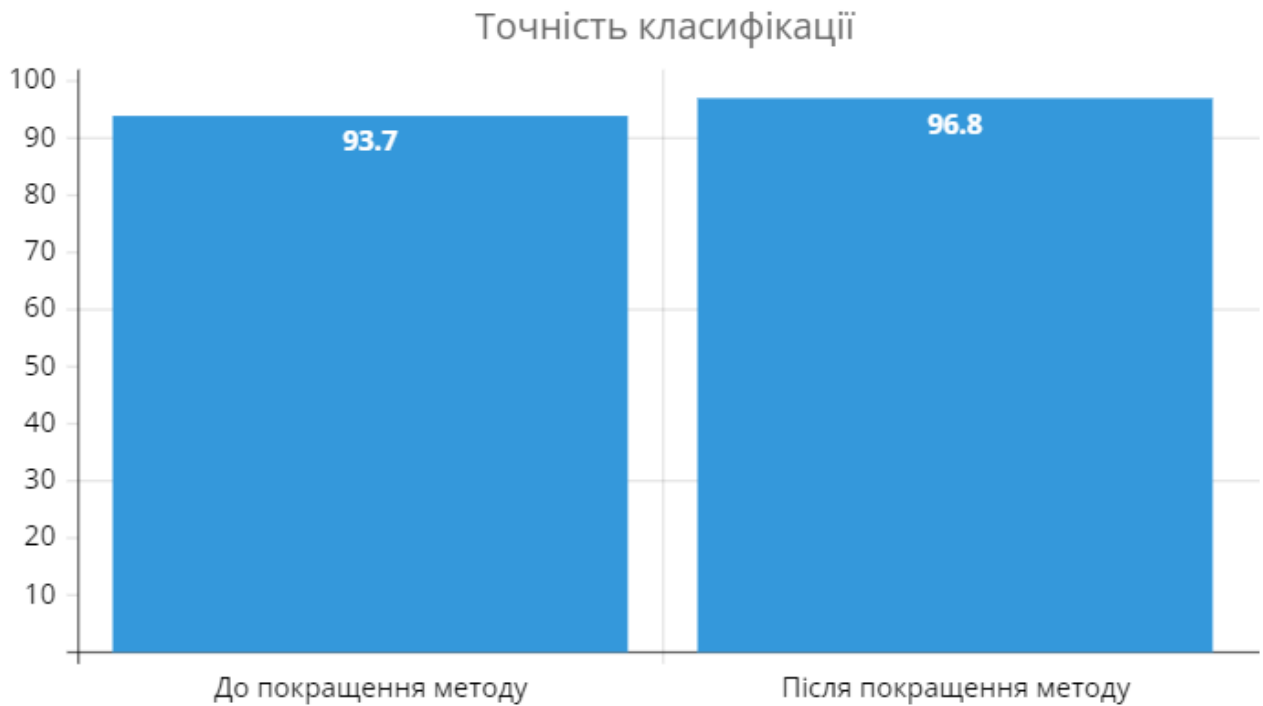


Рисунок 4.7 – Порівняння точності та ефективності роботи мережі до і після підбору параметрів

Як видно з рисунку, правильний підбір параметрів нейронної мережі дозволив підвищити точність класифікації.

Для того щоб виміряти ефективність роботи методу аналізу тональності тексту, можна скористатися даними із сайту, який певний час працював з клієнтами, враховуючи тональність їх повідомлень за допомогою методу та певний час не враховуючи. Зовнішній вигляд сайту зображено на рисунку 4.8.

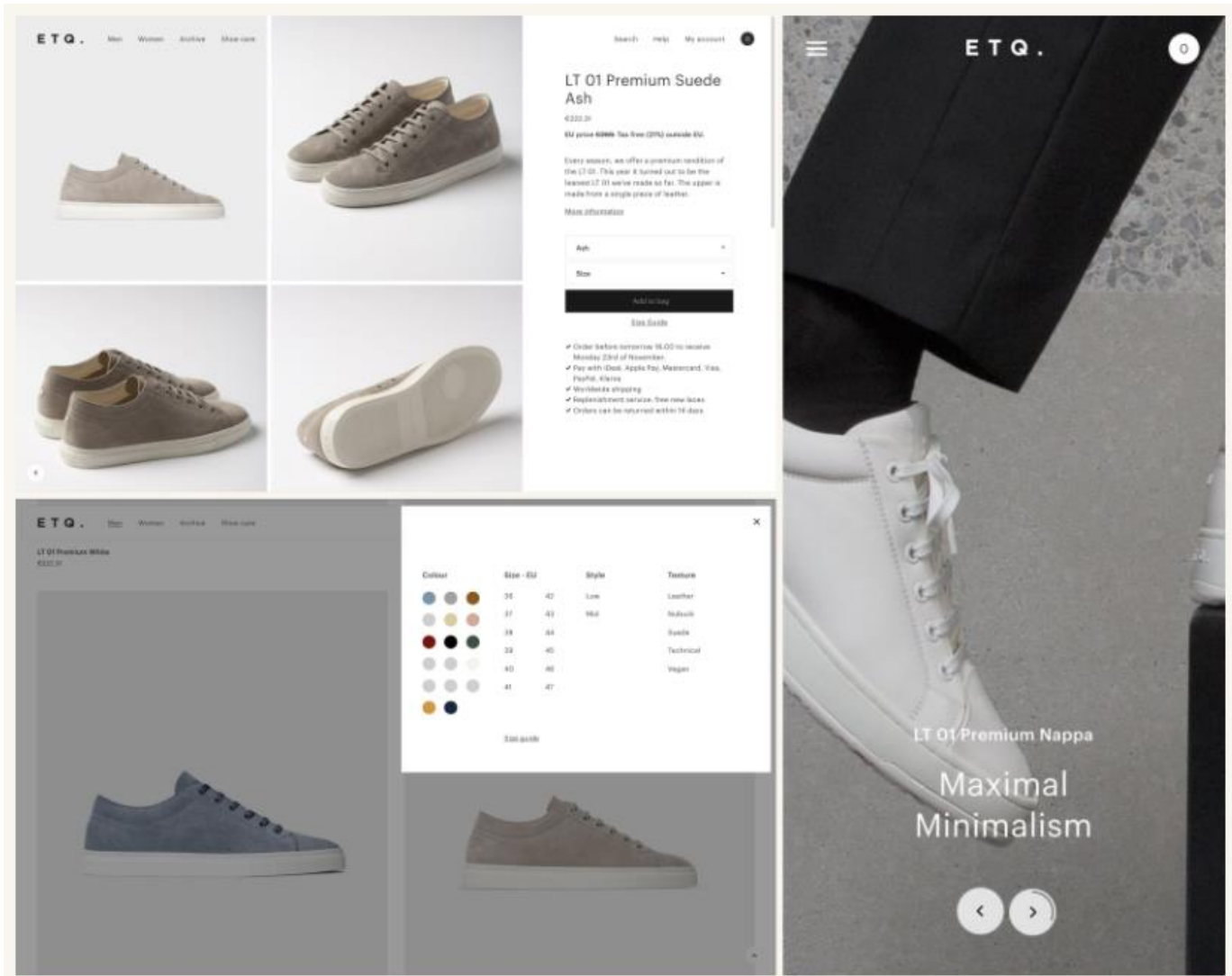


Рисунок 4.8 – Динаміка кількості відвідувань

Динаміку відвідування на сайті зображено на рисунку 4.9.

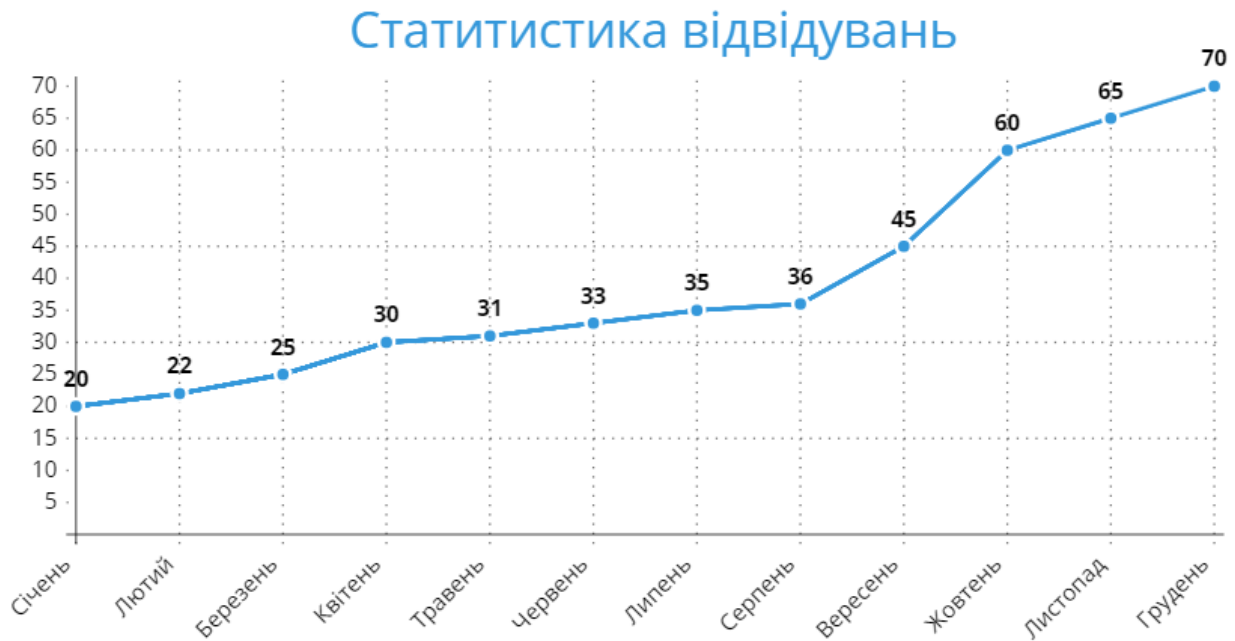


Рисунок 4.9 – Динаміка кількості відвідувань сайту упродовж року

Як видно на графіку, кількість відвідувань суттєво підвищилась після того як чат-бот сайту почав використовувати аналіз тональності для визначення ступеню агресивності користувача у вересні.

Динаміку конвертованості відвідувань зображено на рисунку 4.10.

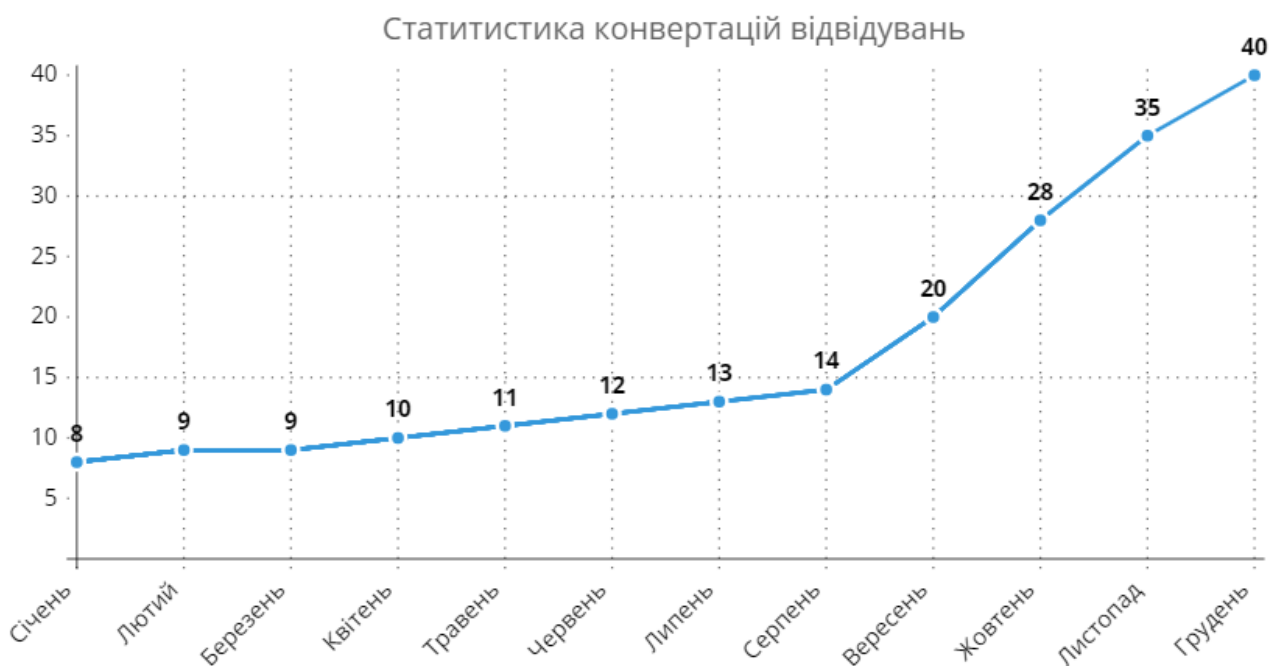


Рисунок 4.10 – Динаміка конвертованості відвідувань сайту упродовж року

По графіку динаміки конвертованості відвідувань видно що після того як чат-бот сайту почав використовувати аналіз тональності для визначення ступеню агресивності користувача, більший відсоток клієнтів став закінчувати своє відвідування сайту покупкою.

Отже, наведені дослідження ефективності роботи створеної системи показують, що реалізовані модернізації методу інтелектуального підбору відповіді на запитання за семантичними ознаками підвищують точність класифікації повідомлення українською мовою, а також дозволяють більш ефективно класифікувати емоційно забарвлені повідомлення. Реалізовані покращення методу сприяють підвищенню відвідуваності та конвертації продажів інтернет-магазину.

Висновки до розділу 4

У розділі було наведено розроблену архітектуру інформаційної системи інтелектуального підбору відповідей на запитання за семантичними ознаками. Були наведені результати прикладного тестування та функціонального дослідження інформаційної системи.

Наведені у розділі дослідження ефективності показують що завдяки застосуванню методу інтелектуального підбору відповідей до запитань за семантичними ознаками, одержано можливість ефективніше працювати з українською мовою, а також ефективніше класифікувати емоційно забарвлені речення.

Загальні висновки

Кваліфікаційна робота магістра розв'язує науково-технічну задачу інтелектуального підбору відповідей на запитання за семантичними ознаками з урахуванням тональності за допомогою нейронної мережі «багатошаровий перцептрон».

За результатом виконання роботи були поставлені та *вирішені наступні завдання*:

– Був досліджений сучасний стан семантичного аналізу коротких текстових повідомлень.

– Розроблено модель комплексу «запитання-прототип-відповідь».

– Розроблено інформаційну модель оцінки відповідності компонентів моделі «запитання-прототип-відповідь».

– Вдосконалено метод інтелектуального підбору відповідей на запитання за семантичними ознаками.

– Створено тестову реалізацію методу інтелектуального підбору відповідей на запитання за семантичними ознаками.

– Досліджено практичну ефективність застосування методу інтелектуального підбору відповідей на запитання за семантичними ознаками.

В результаті роботи було отримано результати, що містять *інновації та наукову новизну*, зокрема було вдосконалено метод інтелектуального підбору відповідей на запитання за семантичними ознаками, який відрізняється від існуючих методів тим, що влючає виявлення та обробку негативного емоційного забарвлення для коротких текстових повідомлень.

Наведені в роботі дослідження ефективності створеної системи показують, що метод інтелектуального підбору відповідей на запитання за семантичними ознаками з урахуванням тональності працює ефективніше з українською мовою та з емоційно забарвленими реченнями українською мовою, порівняно з існуючим підходом. Дослідження створеної системи проводились у контексті користувача інтернет-магазину.

Створений метод інтелектуального підбору відповідей на запитання за семантичними ознаками можна використовувати у багатьох сферах, де необхідно мати автоматизовану або автоматичну систему для спілкування з клієнтами. Даний метод допоможе більш ефективно розуміти та реагувати на запити клієнта, що дозволить використовувати менше людської праці для подібних задач.

Основні наукові і практичні результати кваліфікаційної роботи магістра доповідались у доповіді на тему «Метод інтелектуального підбору відповідей до запитань за семантичними ознаками» на XIV Всеукраїнській науково-практичній конференції «Актуальні проблеми комп'ютерних наук АПКН-2022» (18-19 листопада 2022 року); за темою роботи автором виконано наукову публікацію:

Вишинський І.О., Молчанова М.О., Скрипник Т.К. Метод інтелектуального підбору відповідей до запитань за семантичними ознаками. Збірник наукових праць за матеріалами XIV Всеукраїнської науково-практичної конференції «Актуальні проблеми комп'ютерних наук АПКН-2022». Хмельницький, 2022. с. 64-71.

Перелік посилань

1. Speech and Language Processing. URL:
https://www.researchgate.net/publication/200111340_Speech_and_Language_Processing_An_Introduction_to_Natural_Language_Processing_Computational_Linguistics_and_Speech_Recognition
2. TextRank: Bringing Order into Text. URL:
https://www.researchgate.net/publication/200042361_TextRank_Bringing_Order_into_Text
3. Використання класичних методів машинного навчання для класифікації текстів у програмах генерації автоматичних відповідей. URL:
https://ami-ejournal.cdu.edu.ua/user/setLocale/NEW_LOCALE?source=%2Farticle%2Fview%2F4160%2F0
4. Комп'ютерна модель генерації відповідей у пошуковій системі на основі неструктурованої бази знань. URL:
https://nmetau.edu.ua/file/otzyv_kovylin_scan.pdf
5. Волковський О. С., Ковилін Є. Р. Комп'ютерна система інтелектуального семантичного пошуку з використанням генерації текстів. Вісник Херсонського національного університету. 2018. № 3(66). С. 238–245.
6. Volkovsky O. S., Kovylin Y. R. Computer System of Building of the Semantic Model of the Document. IEEE Second International Conference on Data Stream Mining & Processing. (Lviv, August 21-25, 2018). P. 322–327. DOI: 10.1109/DSMP.2018.8478591.
7. Інформаційна технологія автоматичної класифікації неструктурованих текстових масивів інформації URL:
https://ela.kpi.ua/bitstream/123456789/23795/4/Katiushchenko_magistr.pdf

8. Обробка природної мови (NLP) у Python з кодом (Частина 1. Аналіз тональності). URL: <https://oleg-dubetsky.medium.com/%D0%BE%D0%B1%D1%80%D0%BE%D0%B1%D0%BA%D0%B0-%D0%BF%D1%80%D0%B8%D1%80%D0%BE%D0%B4%D0%BD%D0%BE%D1%97-%D0%BC%D0%BE%D0%B2%D0%B8-nlp-%D1%83-python-%D0%B7-%D0%BA%D0%BE%D0%B4%D0%BE%D0%BC-%D1%87%D0%B0%D1%81%D1%82%D0%B8%D0%BD%D0%B0-1-83d588b3ad71>

9. Головна сторінка сайту з чат-ботом Parry. URL: <https://www.botlibre.com/browse?id=857177>

10. Блог про чат-бот Tay. URL: <https://blogs.microsoft.com/blog/2016/03/25/learning-tays-introduction/>

11. Головна сторінка сайту з чат-ботом Eno. URL: <https://www.capitalone.com/digital/eno/>

12. Документація по Telegram BOT API. URL: <https://core.telegram.org/bots/api>

13. Документація по C++. URL: <https://cplusplus.com/>

14. Документація по PyTorch. URL: <https://pytorch.org/>

15. Документація по NLTK. URL: <https://www.nltk.org/>

16. Документація по MySQL. URL: <https://www.mysql.com/>

17. Документація по JSON. URL: <https://www.json.org/json-en.html>

18. Сайт з чат-ботом Eliza. URL: <https://web.njit.edu/~ronkowitz/eliza.html>

19. Сайт з чат-ботом Eliza. URL: <https://web.njit.edu/~ronkowitz/eliza.html>

20. Сайт з чат-ботом Eliza. URL: <https://web.njit.edu/~ronkowitz/eliza.html>

21. Сайт з чат-ботом Eliza. URL: <https://web.njit.edu/~ronkowitz/eliza.html>

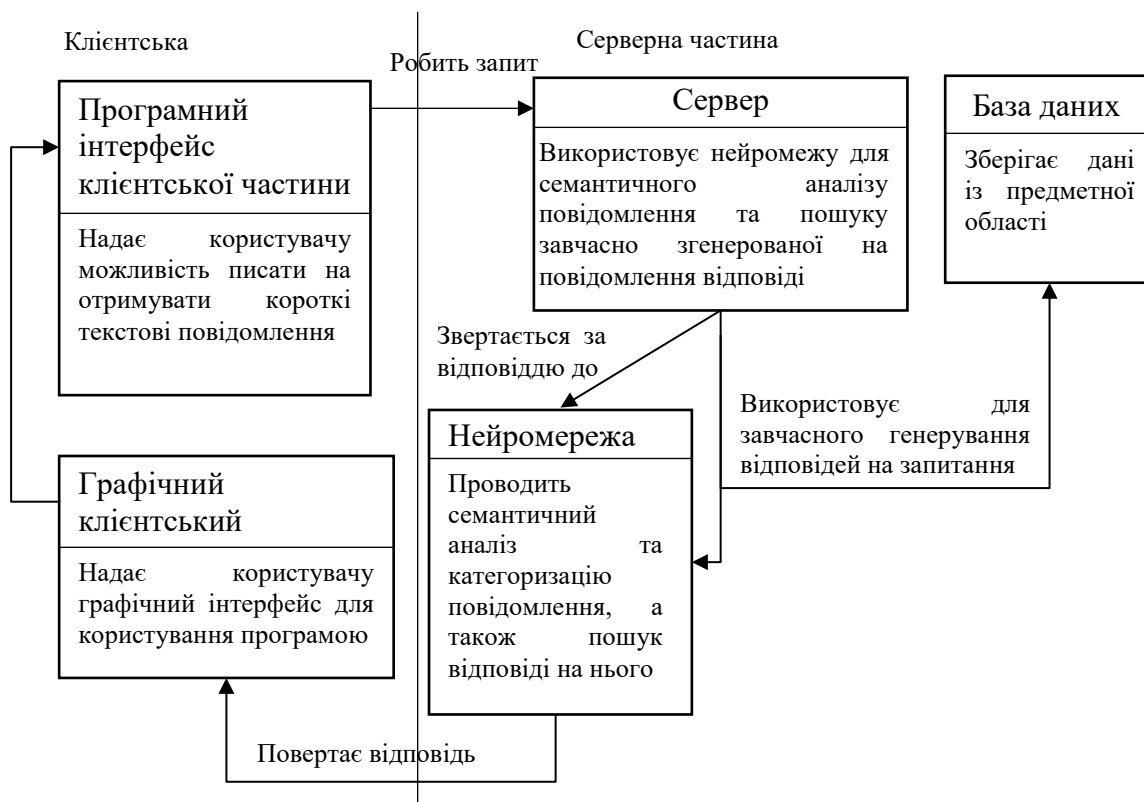
21. Документація по JavaScript. URL: <https://developer.mozilla.org/ru/docs/Web/JavaScript>

22. Приват-банк чат-бот. URL: <https://web.telegram.org/z/#248894045>
23. Чат-бот Replika. URL: <https://replika.com/>
24. Діалогова система GUS. URL: <https://gus.chat/>
25. Вишинський І.О., Молчанова М.О., Скрипник Т.К. Метод інтелектуального підбору відповідей до запитань за семантичними ознаками. Збірник наукових праць за матеріалами XIV Всеукраїнської науково-практичної конференції «Актуальні проблеми комп'ютерних наук АПКН-2022». Хмельницький, 2022. с. 64-71.

ДОДАТКИ

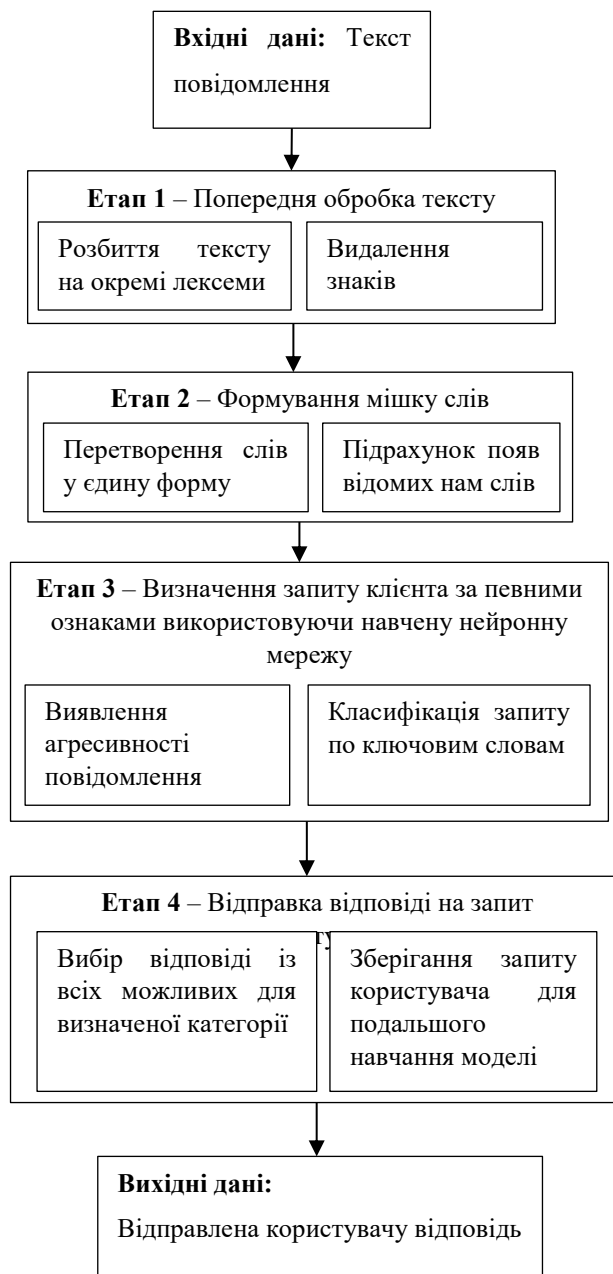
Додаток А

**Схема інформаційної структури системи методу інтелектуального підпору
відповідей за семантичними ознаками**



Додаток Б

Загальна схема методу інтелектуального підбору відповідей на запитання за семантичними ознаками



Додаток В

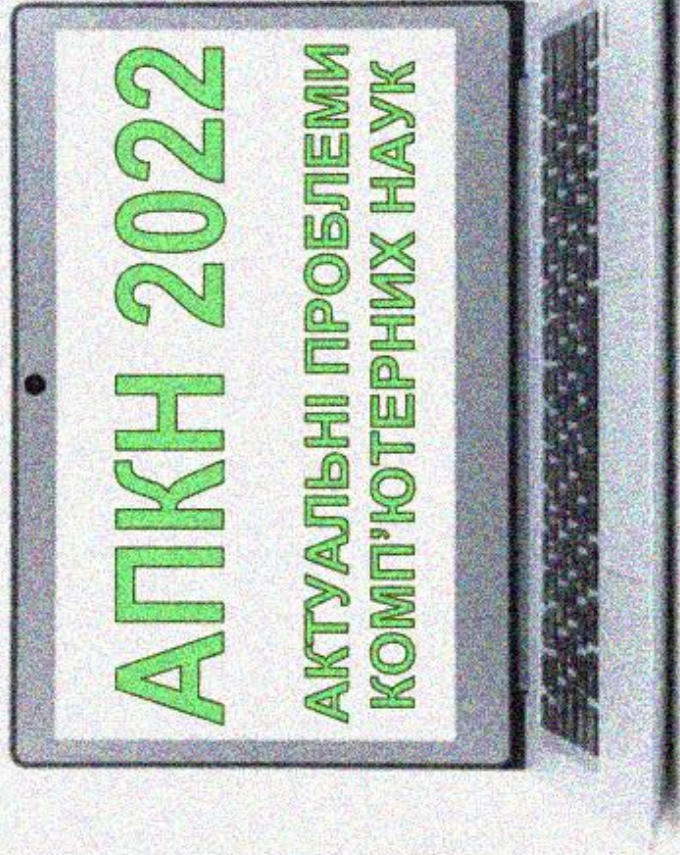
Світлини наукових публікацій, виконаних при роботі над кваліфікаційною роботою магістра

(ксерокопії титульної сторінки, сторінки змісту та всіх сторінок із публікацією)

Перелік наукових публікацій:

Вишинський І.О., Молчанова М.О., Скрипник Т.К. Метод інтелектуального підбору відповідей до запитань за семантичними ознаками. Збірник наукових праць за матеріалами XIV Всеукраїнської науково-практичної конференції «Актуальні проблеми комп'ютерних наук АПКН-2022». Хмельницький, 2022. с. 64-71.

Міністерство освіти і науки України
Хмельницький національний університет



ЗБІРНИК НАУКОВИХ ПРАЦЬ
за матеріалами XIV Всеукраїнської науково-практичної конференції
«Актуальні проблеми комп'ютерних наук АПКН-2022»

18-19 листопада 2022

Хмельницький 2022

У даній публікації наведені основні методи підготовки тексту для подальшого аналізу алгоритмами обробки природної мови. Розглянуті такі поняття як: токенизація, лематизація, стематизація та видалення стоп-слів.

Також розглянуто векторизацію і такі моделі векторизації як «мішок слів» та TF-IDF. Крім цього розглянута класифікація тексту за допомогою методу Naïve Bayes та Decision Trees. Описана загальна модель генерації відповіді на запитання користувача. Але у цій моделі не враховується ступінь агресивності повідомлення. Всі повідомлення вважаються апріорі неагресивними.

У роботі [3] розглянуто модель системи запит-відповідь, що спроможна створювати конкретні текстові відповіді на запит користувача, використовуючи у своєму алгоритмі генерацію наукового тексту на природній мові. Там розглядається семантична мережа наукового тексту та методи роботи з нею для генерації відповідей.

Наведена у статті модель базується на розробленому підході до формування семантичної моделі документа, який дозволяє отримувати кількісні показники семантичних властивостей документу на природній мові і сенсові зв'язки між компонентами тексту. Зазначена в статті модель також має властивість через яку додаток повинен мати можливість використовувати нерозмічений задалегідь корпус текстів, що являє собою неструктуровану базу знань.

Описана в наведеній статті модель є розвитком попередніх досліджень про побудову системи автоматичної генерації текстів на основі концепції моделі м'якого розуміння Леонтьєвої [4] та побудову семантичної моделі наукового тексту [5]. Але у даній статті, як і у попередній не розглядається аналіз агресивності повідомлення.

Існує 2 основних напрямки розробки систем «запитання-відповідь» [3]. На основі лексико-семантичного словника відносин і на основі статистичного аналізу текстів. Представником першого напрямку є система [6], де семантичний аналіз полягає у виявленні взаємозв'язків між об'єктами і класифікації відносин між ними, а також ототожненні об'єктів із задалегідь заданими семантичними класами. Представником другого напрямку є система [7], де для відповідного запиту складається «інформаційний портрет» - набір упорядкованих за значимістю ключових слів і словосполучень, характерних саме для даної вибірки текстів, після чого за набором ключових слів користувач може самостійно визначити теми, які можуть бути видані у відповідь на його запит, і тим самим уточнити потрібну йому тематику.

Існують різні підходи семантичного аналізу тексту. Одними із найбільш відомих є:

TF-IDF [2] (термін частотно-інверсна частота документа) – це статистичний показник, який оцінює релевантність слова документу в колекції документів. Це робиться шляхом множення двох показників: скільки разів слово з'являється в документі та зворотної частоти цього слова в наборі документів.

УДК 004.4

Вишинський І.О., Молчанова М.О., Скрипник Т.К., Собко О.В., Мазурець О.В.

Хмельницький національний університет

МЕТОД ІНТЕЛЕКТУАЛЬНОГО ПІДБОРУ ВІДПОВІДЕЙ ДО ЗАПИТАНЬ ЗА СЕМАНТИЧНИМИ ОЗНАКАМИ

Розглянуто метод інтелектуального підбору відповідей до запитань за семантичними ознаками, досліджено сучасний стан семантичного аналізу коротких текстових повідомлень. Розроблено модель комплексу «запитання-прототип-відповідь».

Investigated method of intelligent selection of answers to questions based on semantic features, investigated the current state of semantic analysis of short text messages. A model of the "question-prototype-answer" complex was developed.

Вступ

Штучний інтелект все більше інтегрується в наше повсякденне життя завдяки створенню та аналізу інтелектуального програмного та апаратного забезпечення, які називаються інтелектуальними агентами. Розумні агенти можуть виконувати різноманітні завдання, починаючи від трудової діяльності і закінчуючи складними операціями.

Чат-бот є типовим прикладом системи штучного інтелекту та одним із найбільш елементарних і поширених прикладів інтелектуальної взаємодії людини з комп'ютером. Це комп'ютерна програма, яка реагує як розумна сутність на розмову за допомогою тексту чи голосу та розуміє одну чи декілька людських мов за допомогою обробки природної мови (NLP) [1]. У лексиконі чат-бот визначається як «комп'ютерна програма, призначена для імітації розмови з людьми, особливо через Інтернет». Чат-боти також відомі як розумні боти, інтерактивні агенти, цифрові помічники або штучні об'єкти спілкування. Вони можуть використовуватись для різних цілей, у тому числі як: онлайн-помічники, онлайн-співбесідники, загалом, для автоматизації роботи різних сервісів, що зменшить навантаження на людський персонал.

Метою статті є дослідження методів інтелектуального підбору відповідей на запитання, розробка такого методу, а також розробка моделі комплексу «запитання-прототип-відповідь»

Аналіз існуючих публікацій

У роботі [2] розглянуто методи машинного навчання для розв'язання задач класифікації текстів з метою їх подальшого використання у програмах автоматичної генерації відповідей на основі аналізу контексту питань користувача.

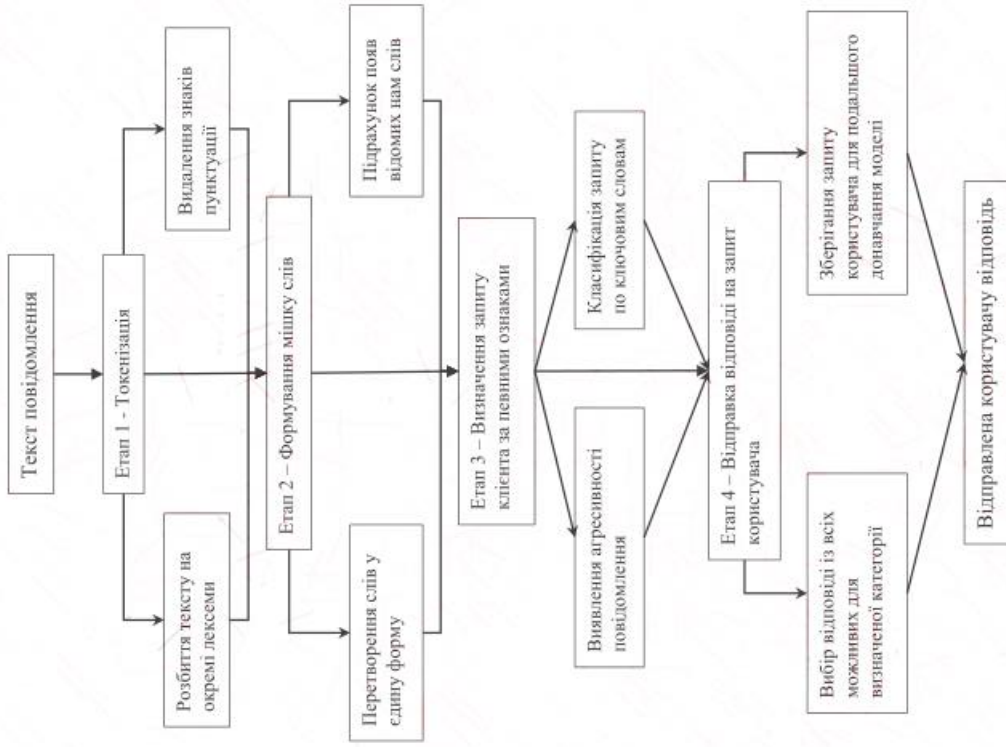


Рисунок 1 – Загальна схема методу інтелектуального підбору відповідей до запитань за семантичними ознаками

Етап 3 – Визначення запиту клієнта за певними ознаками

На даному етапі потрібно проаналізувати слова із мішка слів, знайти там ключові слова і по ним класифікувати запит клієнта. Тобто зрозуміти що саме він

TF-IDF був винайдений для пошуку документів та отримання інформації. Він працює, пропорційно збільшуючи кількість разів, коли слово з'являється в документі, але компенсується кількістю документів, які містять це слово.

TextRank [8, 9] – це алгоритм ранжування на основі графів, подібний до алгоритму PageRank від Google, який успішно реалізовано в аналізі цитувань. Рейтинги тексту використовуються для виділення ключових слів, автоматичного узагальнення тексту та ранжування фраз. По суті, в алгоритмі ранжування тексту вимірюється зв'язок між двома або більше словами.

Коли одна вершина з'єднується з іншою, це фактично голосування для цієї іншої вершини. Більша кількість голосів які створюються для вершини, тим вища важливість вершини. Крім того, важливість вершини голосування визначає, наскільки важливий голос сама є, і ця інформація також враховується моделлю рейтингу. Отже, оцінка, пов'язана з вершиною, визначається на основі голосів які віддані для нього.

Метод генерації відповіді на коротке текстове повідомлення на основі семантичного аналізу тексту

Отримуючи коротке текстове повідомлення, необхідно розбити його на речення, а речення на слова і знаки. Потім із слів і знаків утворити емислові одиниці. Також по словам можна виявити вид речення та його емоційне забарвлення.

На рисунку 1 зображено загальну схему генерування відповіді на коротке текстове повідомлення на основі семантичного аналізу.

Етап 1 – Токенізація

Спочатку необхідно розбити наш текст на окремі лексеми. Для цього можна вважати кожний набір символів з обох боків обмежений пробільними символами лексемою. Таким чином, можна отримати набір окремих слів та прикріплених до них знаків пунктуації.

У даному випадку для визначення відповіді на повідомлення користувача, знаки пунктуації не потрібні, тому їх необхідно видалити.

Далі маючи лише набір слів з них формується «мішок слів» [2]. Мішок слів представляє собою словник в якому ключ – це певне слово яке зустрілось у тексті, а значення – це кількість його появ у тексті.

Етап 2 – Формування мішкоку слів

Для того щоб сформуванати мішок слів, необхідно привести слова у певну єдину форму, тобто кожне слово повинне бути у 1 відмінку, без доданих префіксів або суфіксів. Далі необхідно підрахувати частоту кожного слова та створити мішок слів.

Таблиця 1 – Результати порівняння роботи існуючої нейронної мережі та покращеної

	Кількість епох навчання	Швидкість навчання	Кількість нейронів у прихованих шарах	Відсоток кінцевих втрат
Вже існуючий метод	1000	0.001	8	0.0011
Покращений метод	600	0.01	8	0.0000183

Як видно з таблиці, відсоток кінцевих втрат мережі суттєво зменшився, що показує позитивний вплив зміни гіпер-параметрів мережі на більш підходящі для української мови.

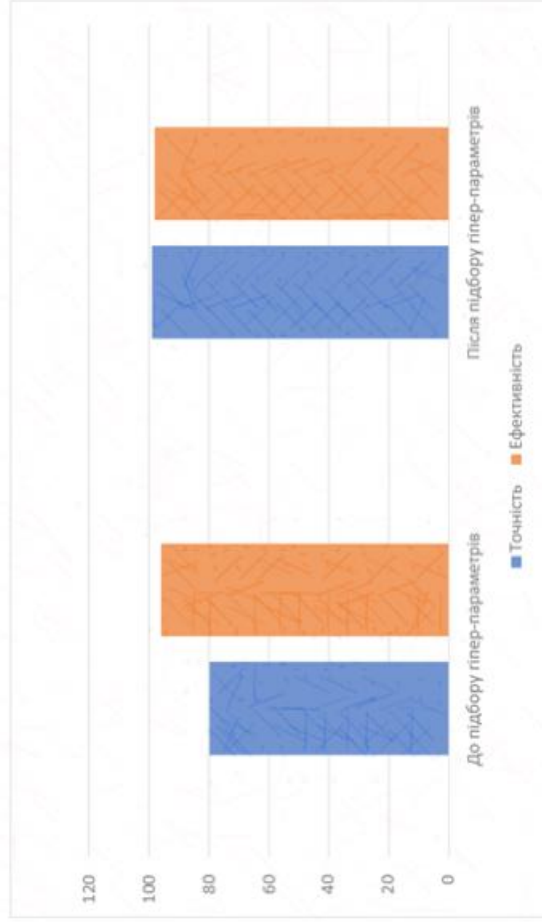


Рисунок 2 – Порівняння точності та ефективності роботи мережі до і після підбору гіпер-параметрів

Для того щоб виміряти ефективність роботи методу аналізу тональності повідомлення спочатку необхідно створити тестовий набір повідомлень та вручну визначити їхню тональність у діапазоні від -1 до 1, де -1 означає дуже агресивне, 0 – нейтральне, 1 – дуже позитивне повідомлення. Далі необхідно визначити максимально допустиме похибку аналізу. В цьому випадку вона становитиме 0.2. Далі необхідно виміряти тональність кожного з тестових повідомлень за допомогою

хотів. Для цього необхідно навчити нейронну мережу класифікувати набори ключових слів, використавши певні тренувальні дані.

Також на цьому етапі потрібно виявити ступінь агресивності повідомлення, для того що наприклад одразу перенаправляти агресивного користувача до менеджера. Для цього потрібно навчити нейронну мережу визначати ступінь агресивності по певним ключовим словам. Для цього знадобиться словник агресивних і звичайних слів української мови, в якому будуть слова і оцінки їх агресивності.

Етап 4 – Відправка відповіді користувачу

На цьому етапі необхідно вибрати якусь відповідь із тих, які відповідають категорії запиту. Таких відповідей може бути декілька, вибрати якусь конкретну можна випадковим чином, або проводити додатковий аналіз.

Також необхідно зберегти запит клієнта у певному вигляді (у файлі або бази даних) для того, щоб у подальшому донавчити нашу модель за допомогою цих даних.

Іноді для того щоб зрозуміти що слово ввів користувач, нам необхідно мати метод перевірки схожості двох слів. Для цього можна використовувати так звану відстань редактування – тобто визначимо скільки операцій редактування нам потрібно для того, щоб перетворити одне слово у інше. Чим менше операцій тим більш слова схожі. Також різним операціям можна призначати різні пріоритети в залежності від ситуації.

Аналіз ефективності створеної системи

В даному випадку, для того щоб проаналізувати створену систему, необхідно виміряти ефективність роботи нейронної мережі по класифікації короткого текстового повідомлення, ефективність методу аналізу тональності тексту, а також відповідність згенерованої методом відповіді контексту отриманого короткого текстового повідомлення.

Виміряти ефективність роботи нейронної мережі можна по наступним критеріям: відсоток кінцевих втрат, відсоток правильно визначених категорій. Для того щоб виміряти відсоток правильно визначених категорій, потрібно створити дані для тестування, а саме – набір правильно визначених категорій та шаблонів, призначених певній категорії. Далі необхідно створити до 1000 коротких текстових повідомлень та визначити для кожного з них правильну категорію. Далі після навчання мережі за допомогою даних про категорії та шаблони необхідно підрахувати для скількох повідомлень із 1000 мережа правильно визначить категорію. Далі формула оцінки ефективності роботи мережі буде виглядати наступним чином: $p = \frac{n}{m} * 100\%$, де n – кількість правильно класифікованих повідомлень, а m – загальна кількість тестових повідомлень.

Замірявши відсоток кінцевих втрат з існуючими параметрами та з покращеними параметрами, були отримані наступні результати.

Для того, щоб показати ефективність роботи методу аналізу тональності в повній мірі, необхідно порівняти відсоток правильних реакцій на агресивне повідомлення користувача з використанням методу аналізу тональності та без нього. Отримані результати зображені на рисунку 3.

Як видно з рисунку, використання методу аналізу тональності суттєво підвищує шанси класифікувати агресивне повідомлення.

Отже, запропоновані покращення існуючого методу інтелектуального підбору відповідей до запитань за семантичними ознаками суттєво підвищують його ефективність.

Висновки

Отже, запропонований метод інтелектуального підбору відповідей до запитань за семантичними ознаками дозволяє генерувати релевантні відповіді на запитання користувача для різноманітних цілей враховуючи ступінь агресивності тексту. Подальші дослідження спрямовані на розширення контексту запитань, формування більш релевантних відповідей.

Перелік посилань

1. Daniel Jurafsky, James H. Martin Stanford University University of Colorado at Boulder Speech and Language Processing An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition Third Edition draft
2. Коваленко О.С., Сердюк О.А. Використання класичних методів машинного навчання для класифікації текстів у програмах генерації автоматичних відповідей.
3. Ковалін Є.Р., Волковський О.С. Комп'ютерна модель генерації відповідей у пошуковій системі на основі неструктурованої бази знань.
4. Волковський О. С., Ковалін Є. Р. Комп'ютерна система інтелектуального семантичного пошуку з використанням генерації текстів. Вісник Херсонського національного університету. 2018. № 3(66). С. 238–245.
5. Volkovsky O. S., Kovylin Y. R. Computer System of Building of the Semantic Model of the Document. IEEE Second International Conference on Data Stream Mining & Processing. (Lviv, August 21-25, 2018), P. 322–327. DOI: 10.1109/DSMP.2018.8478591.
6. Поляков П. Ю. Використання семантичних категорій в завданні класифікації відгуків про книги. Матеріали міжнародної конференції «Диалог» (м. Москва, 29 травня – 2 червня 2013 р.). Москва, 2013. С. 193–199.
7. Антонов А. В. Галактика Zoom. Оцінка модифікації методу формування інфопротрета. Матеріали третього російського семінару по оцінці методів інформаційного пошуку. (м. Ярославль, 6 жовтня 2018 р.). Ярославль, 2018. С. 226.
8. Rada Mihalcea, Paul Tarau TextRank: Bringing Order into Text Department of Computer Science.
9. Собко О. В., Молчанова М. О., Мазурець О. В., Багрий Р. О., Купрійчук В. О. Метод автоматизованого формування семантичного ядра цифрових текстів. Збірник наукових праць за матеріалами XVI Міжнародної науково-практичної конференції «Multidisciplinary academic notes. Theory, methodology and practice». Tokyo, Japan, 2022. с. 1102-1106.

використаного методу. Далі необхідно підрахувати кількість повідомлень, для яких тональність визначена методом, не відрізняється більше, ніж на значення похибки від вручну вирахованої тональності. Формула виглядає наступним чином:

$$p = \frac{n}{m} * 100\%,$$

де n = кількість повідомлень, для яких $[k - f] \geq 0.2$, де k = тональність визначена методом, а f – тональність визначена вручну.

Для аналізу ефективності покращеного методу, була створена вибірка із 1000 тестових повідомлень, для яких вручну були визначені правильна тональність та категорія. Далі результати роботи системи були порівнянні з результатами із тестової вибірки, використовуючи описані вище формули. Результати аналізу зображені на рисунку 2.

Як видно з рисунку, правильний підбір гіпер-параметрів нейронної мережі дозволив підвищити точність та ефективність мережі.

Для вимірювання ефективності роботи методу аналізу тональності тексту, необхідно порівняти результати роботи методу з результатами із тестової вибірки. Отриманий результат показує, що ефективність методу становить 98.9%.

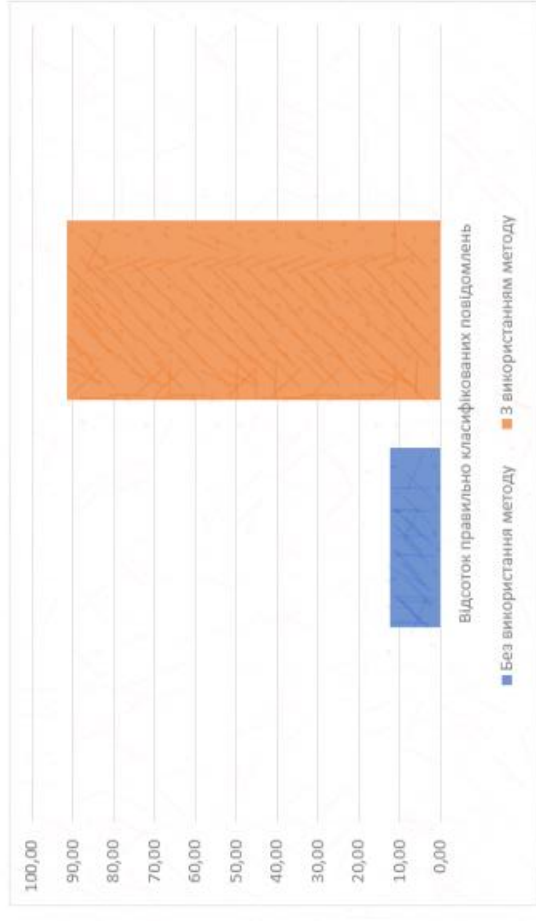


Рисунок 3 – Порівняння ефективності класифікації повідомлень без методу аналізу тональності та з ним

Додаток Г

Презентаційний матеріал

КВАЛІФІКАЦІЙНА РОБОТА МАГІСТРА

МЕТОД ІНТЕЛЕКТУАЛЬНОГО ПІДБОРУ ВІДПОВІДЕЙ
ДО ЗАПИТАНЬ ЗА СЕМАНТИЧНИМИ ОЗНАКАМИ

Виконав:
Студент 2 курсу, група КНм-21
Вишинський Ілля Олександрович

Керівник:
Викладач кафедри КН
Молчанова Марина Олексіївна

Мета роботи

Метою даної роботи є розв'язання задачі інтелектуального підбору відповідей до запитань за семантичними ознаками українською мовою. Для досягнення поставленої мети були виконані наступні завдання:

Завдання

Для досягнення поставленої мети були виконані наступні завдання:

- Дослідити сучасний стан семантичного аналізу коротких текстових повідомлень.
- Розробити модель комплексу «запитання-прототип-відповідь».
- Розробити інформаційну модель оцінки відповідності компонентів моделі «запитання-прототип-відповідь».
- Розробити метод інтелектуального підбору відповідей на запитання за семантичними ознаками.
- Створити тестову реалізацію методу інтелектуального підбору відповідей на запитання за семантичними ознаками.
- Дослідити практичну ефективність застосування методу інтелектуального підбору відповідей на запитання за семантичними ознаками.

Об'єкт та предмет дослідження

Об'єкт дослідження - процес автоматизованого підбору релевантних відповідей на запитання в цифровому текстовому форматі.

Предмет дослідження - інформаційні технології, моделі, методи та засоби для автоматизованого підбору відповідей на запитання за семантичними ознаками.

Наукова новизна одержаних результатів

Вдосконалено метод генерації релевантних відповідей на короткі текстові повідомлення українською мовою за семантичними ознаками та виявлення та обробки негативного емоційного забарвлення для коротких текстових повідомлень.

Практичне значення отриманих результатів та актуальність теми



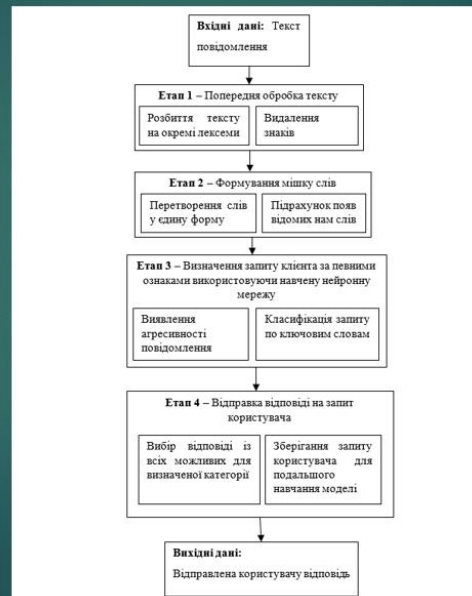
Практичне значення отриманих результатів та актуальність теми



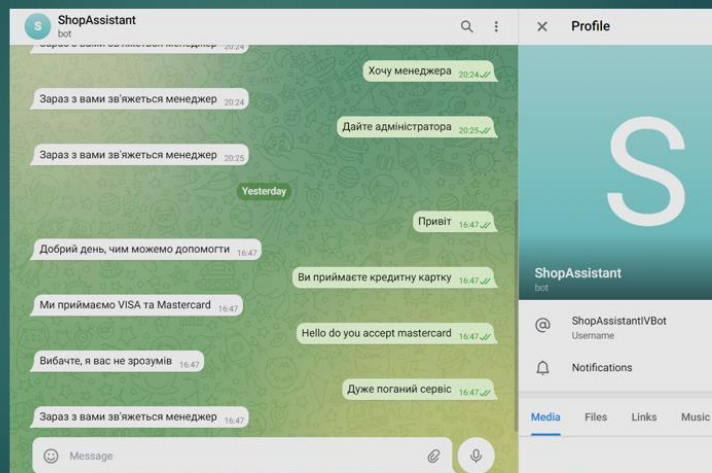
Практичне значення отриманих результатів та актуальність теми



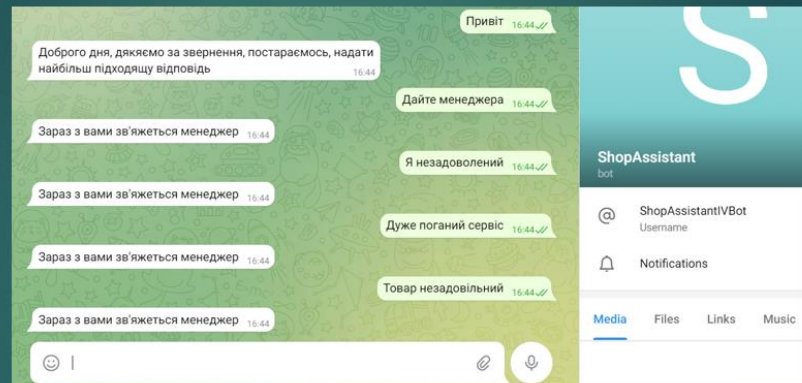
Метод інтелектуального підбору відповідей до запитань за семантичними ознаками



Функціональне дослідження розробленої інформаційної системи



Функціональне дослідження розробленої інформаційної системи



Загальні висновки

Наведені в роботі дослідження ефективності створеної системи показують, що:

Модернізований метод інтелектуального підбору відповідей на запитання за семантичними ознаками з урахуванням тональності працює ефективніше з українською мовою та з емоційно забарвленими реченнями українською мовою, порівняно з оригінальним підходом.

Створений метод можна використовувати у багатьох сферах, де необхідно мати автоматизовану або автоматичну систему для спілкування з клієнтами. Даний метод допоможе більш ефективно розуміти та реагувати на запити клієнта, що дозволить використовувати менше людської праці для подібних задач, а також підвищить відвідуваність та загальну задоволеність сервісом.

Ім'я користувача:
Кафедра КН

Дата перевірки:
07.12.2022 22:58:13 EET

Дата звіту:
07.12.2022 22:59:45 EET

ID перевірки:
1013235055

Тип перевірки:
Doc vs Internet + Library

ID користувача:
100005671

Назва документа: КНм-21-1_Вишинський

Кількість сторінок: 76 Кількість слів: 12176 Кількість символів: 90531 Розмір файлу: 2.34 MB ID файлу: 1012994876

Виявлено модифікації тексту (можуть впливати на відсоток схожості)

6.4%
Схожість

Найбільша схожість: 2.73% з джерелом з Бібліотеки (ID файлу: 1009540882)

3.44% Джерела з Інтернету

43

Сторінка 78

5.68% Джерела з Бібліотеки

152

Сторінка 78

0% Цитат

Вилучення цитат вимкнено

Вилучення списку бібліографічних посилань вимкнено

0%
Вилучень

Немає вилучених джерел

Модифікації

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

Замінені символи

4

Підозріле форматування

17
сторінок

Anti-Plagiarism v-15.257

Максимальне співпадіння з одним документом 1.0%Словники перевірки: en_US, ru_RU, ua_UA. **Помилки в документах: 7%**

ID: 109055 Назва: КВАЛІФІКАЦІЙНА РОБОТА МАГІСТРА на тему Метод інтелектуального підбору відповідей до запитань за семантичними ознаками Додано в БД: 2022-12-07 Автора: І.О. Вишинський Керівники: М.О. Молчанова Консультанти: Опоненти:	Документ		Сумарний збіг по Базі Даних	
	Символи	Лексеми	Символи	Лексеми
	72434	1095	2457 (3%)	37 (3%)

Джерело плагіату

ID	Опис	Наявність плагіату в документі	
		Символи	Лексеми

РІШЕННЯ ЕКСПЕРНОЇ КОМІСІЇ
КАФЕДРИ КОМП'ЮТЕРНИХ НАУК
ПРО ДОПУСК КВАЛІФІКАЦІЙНОЇ РОБОТИ МАГІСТРА ДО ЗАХИСТУ
ЗА РЕЗУЛЬТАТАМИ АНАЛІЗУ ЗВІТУ ПОДІБНОСТІ

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

Назва: Метод інтелектуального підбору відповідей до запитань за семантичними ознаками

Автор: Вишинський Ілля Олександрович

Спеціальність: 122 – Комп'ютерні науки

Освітня програма: освітньо-професійна

Науковий керівник: викладач кафедри КН Молчанова Марина Олексіївна

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.	—

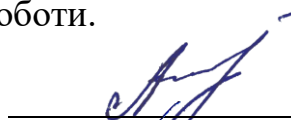
Підтвердження:

Запозичення, виявлені в роботі, є законними і не є плагіатом, оскільки:

- 1) За програмою Anti-Plagiarism виявлено 3%, що є фрагментарним, не більше 1% на джерело – містять поширені конструкції, загальновідомі терміни та визначення.
- 2) За програмою UNICHECK виявлені 6,4%, які є фрагментарними, не більше 2,73% на джерело – містять поширені конструкції, загальновідомі терміни та визначення.

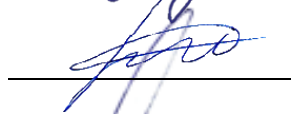
Сумарний обсяг всіх запозичень, визначений системою виявлення збігів/ідентичності/схожості, складає 3% і 6,4% відповідно, що, з урахуванням наведених обґрунтувань, відповідає характеру наукового дослідження і свідчить на користь кваліфікаційної роботи.

Керівник роботи



Марина Молчанова

Гарант ОП



Руслан Багрій

Завідувач кафедри КН



Олександр Бармак



ВІДГУК ОПОНЕНТА

на кваліфікаційну роботу магістра

гр. КНМ-21-1 Вишинського Іллі Олександровича за темою: Метод інтелектуального підбору відповідей до запитань за семантичними ознаками

1. Актуальність обраної теми

Штучний інтелект з кожним днем все більше інтегрується в повсякденне життя завдяки створенню та аналізу інтелектуального програмного та апаратного забезпечення, які називаються інтелектуальними агентами. Колосальні обсяги інформації, що кожного дня завантажуються в Мережу вже неможливо оброблювати вручну. Саме тому галузь штучного інтелекту так стрімко набуває обертів розвитку. Зокрема, така галузь як чат-боти, семантичний аналіз текстів стають все більш доступними, а методи, на базі яких вони працюють, стають все більш потужними та досконалішими.

Тому робота, присвячена методу інтелектуального підбору відповідей до запитань за семантичними ознаками є актуальною в контексті сучасності, що на високому рівні обґрунтовано у роботі.

2. Відповідність роботи предметній області спеціальності 122 Комп'ютерні науки та загальним вимогам до наукових робіт

Обрана тема створення присвячена методу інтелектуального підбору відповідей до запитань за семантичними ознаками, в межах якої реалізовані поставлені задачі, повною мірою відповідає предметній області спеціальності 122 «Комп'ютерні науки» та вимогам до кваліфікаційної роботи магістра.

3. Повнота розкриття мети та завдань дослідження

В роботі повністю розкрито мету дослідження та поставленні в межах теми завдання дослідження.

4. Наявність наукової новизни

У результаті виконання кваліфікаційної роботи магістра було вдосконалено метод інтелектуального підбору відповідей на запитання за семантичними ознаками та виявлення та обробки негативного емоційного забарвлення для коротких текстових повідомлень. Відповідні результати продемонстровано, також обґрунтовано їх наукове та інноваційне значення. Результати дослідження оприлюдненні на науково-практичній конференції.

5. Зміст кожного розділу роботи

Робота містить чотири розділи. Перший розділ спрямований на дослідження предметної області семантичного аналізу коротких текстових повідомлень. Другий розділ присвячено створенню інформаційної моделі методу інтелектуального підбору відповідей за семантичними ознаками. Третій розділ спрямовано на розробку відповідної інформаційної системи на базі методу інтелектуального підбору відповідей на запитання за семантичними ознаками та у четвертому розділі проведено дослідження ефективності створеного методу на базі методу генерації релевантних відповідей на короткі текстові повідомлення українською мовою за семантичними ознаками.

6. Ступінь розкриття теми роботи

В кваліфікаційній роботі тема в повній мірі обґрунтована й розкрита, проведено аналіз актуальності та відомих досліджень в межах обраної теми, поставлені завдання, які у роботі виконані, та проведено аналіз результатів прикладного застосування запропонованих методу і засобів.

7. Якість оформлення кваліфікаційної роботи

Оформлення роботи відповідає необхідним нормам та вимогам, які ставляться до оформлення кваліфікаційних робіт.

8. Недоліки кваліфікаційної роботи

Рекомендовано оновити деякі із використаних джерел, що є застарілими, втім це не вплинуло на загальну якість роботи та одержаний результат.

9. Загальний висновок (допускається чи не допускається до захисту), якої оцінки заслуговує кваліфікаційна робота

Враховуючи рівень виконання та забезпечення усіх необхідних вимог, робота може бути допущена до захисту. Рекомендована оцінка добре.

Опонент  д.т.н., проф. каф. КІСП Мартинюк Валерій Володимирович



ВІДГУК НАУКОВОГО КЕРІВНИКА

на кваліфікаційну роботу магістра

гр. КНМ-21-1 Вишинського Іллі Олександровича за темою: Метод інтелектуального підбору відповідей до запитань за семантичними ознаками

1. Актуальність теми

Чат-бот є типовим прикладом системи штучного інтелекту та одним із найбільш елементарних і поширених прикладів інтелектуальної взаємодії людини з комп'ютером. Створення чат-боту як для розваг, так і для більш практичних цілей є дуже актуальною темою, так як це дозволяє оптимізувати різні бізнес-процеси, що зменшить навантаження на живих працівників та дозволить витратити їх час та зусилля на щось більш креативне, ніж відповідати на текстові повідомлення користувачів. Тому тема є актуальною та перспективною в майбутніх дослідженнях.

2. Відповідність роботи предметній області спеціальності 122 Комп'ютерні науки та загальним вимогам до наукових робіт

Поставлена у кваліфікаційній роботі магістра мета, пов'язана з розв'язання задачі інтелектуального підбору відповідей до запитань за семантичними ознаками українською мовою, повною мірою відповідає предметній області спеціальності 122 «Комп'ютерні науки» та вимогам до кваліфікаційної роботи.

3. Професійні та особистісні якості магістранта

Під час виконання кваліфікаційної роботи магістра Вишинський Ілля Олександрович проявив себе як відповідальний, кваліфікований та дисциплінований студент. Всі поставлені задачі виконував вчасно, якісно та в повній мірі. В процесі навчання та реалізації КРБ проявив високі і достатні для одержання успішного результату компетентності.

4. Ступінь самостійності під час виконання кваліфікаційної роботи

Магістрант самостійно виконував всі поставлені задачі. Одержані положення наукової новизни та інновації, означені в роботі, є результатом особистої діяльності магістранта. Це дозволило провести створення нових та удосконалення існуючих теоретичних та при засобів, створених та використаних у роботі.

5. Наукова новизна та оригінальність запропонованих підходів

Наукова новизна та інновації, що представлені в кваліфікаційній роботі магістра, відповідні спеціальності 122 «Комп'ютерні науки» в межах обраної області дослідження.

Продемонстровано й обґрунтовано результати, які мають наукове та інноваційне значення. Вдосконалено метод генерації релевантних відповідей на короткі текстові повідомлення українською мовою за семантичними ознаками та виявлення та обробки негативного емоційного забарвлення для коротких текстових повідомлень. Результати роботи оприлюдненні на науково-практичній конференції.

6. Ступінь оволодіння методами дослідження

В роботі виявлено високий ступінь оволодіння магістрантом необхідними методами дослідження.

7. Повнота та якість розкриття теми роботи

Магістрант в повній мірі розкрив та обґрунтував тему та завдання роботи, було проведено аналіз актуальності та відомих досліджень в межах обраної теми, поставлені завдання, які у роботі виконані, а також проведено аналіз результатів прикладного застосування запропонованих засобів генерації релевантних відповідей на короткі текстові повідомлення українською мовою за семантичними ознаками.

8. Логічність, послідовність, аргументованість, літературна грамотність викладу матеріалу


Структура роботи й послідовність викладення логічні та відповідні поставленій меті. Викладення матеріалу грамотне та виявляє високий ступінь відповідності стилю.

9. Можливість практичного застосування кваліфікаційної роботи, окремих її частин

Приклади використання розробленої інформаційної системи підтвердило її спроможність забезпечити виконання функцій генерації релевантних відповідей на короткі текстові повідомлення українською мовою за семантичними ознаками. Проведені дослідження практичної ефективності роботи інформаційної системи показують, що система ефективно генерує релевантну відповідь на коротке текстове повідомлення. Дослідження також показує, що система вдало враховує тональність тексту для генерації відповіді.

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

Враховуючи високий рівень виконання та забезпечення усіх необхідних вимог, робота може бути допущена до захисту. Рекомендована оцінка добре.

Науковий керівник  викладач. каф. КН Молчанова Марина Олексіївна