Шифр: **Сентимент Код**

АНАЛІЗ ТОНАЛЬНОСТІ КОМЕНТАРІВ ІНТЕРНЕТ-МАГАЗИНУ

ЗМІСТ

ВСТУП

Сучасний розвиток інформаційних технологій та електронної комунікації створює безмежні можливості для збору та аналізу великої кількості даних, які мають значення для різних галузей, включаючи електронну торгівлю, медіа, соціальні мережі, і багато інших. Одним із важливих завдань в цьому контексті є аналіз тональності коментарів та відгуків, залишених користувачами в мережі. Відгуки клієнтів відображають їхні особисті враження від придбаного товару чи отриманих послуг. Зрозуміння сутності цих вражень та виявлення тональності відгуків (позитивної, негативної або нейтральної) має велике значення для бізнесу, оскільки впливає на репутацію компанії, прийняття стратегічних рішень та покращення якості послуг. Системи аналізу настроїв, які базуються на методах обробки природної мови (Natural Language Processing - NLP), стали невід'ємною частиною багатьох сучасних досліджень і практичних застосувань.

Не дивлячись на те, що наразі на ринку існує велика кількість інструментів та інтелектуальних систем для здійснення аналізу настроїв, є сфери, в яких застосування універсальних прийомів оброблення природної мови може не привести до бажаного ефекту. Більшість існуючих систем орієнтовані на англомовне середовище, тому їх складно використовувати для аналізу текстів українською мовою. Тому створення інструментів аналізу настроїв для інтернет-магазину є актуальним.

Об’єкт дослідження: застосування аналізу тональності коментарів для інтернет-магазину.

Предметом дослідження є розробка інструментарію аналізу настроїв покупців інтернет-магазину.

Мета і завдання. Метою дослідження є проєктування інструментів аналізу тональності коментарів інтернет-магазину на основі технологій оброблення природної мови.

Для досягнення мети необхідно виконати такі завдання.

1. Дослідити теоретичні відомості щодо процесів оброблення природної мови.

2. Здійснити аналіз найбільш поширених інструментів аналізу настроїв та сфери його використання.

3. Дослідити існуючі моделі векторного представлення слів.

4. Виконати аналіз настроїв за відгуками та коментарями до товару клієнтів інтернет-магазину.

Матеріал дослідження. Робота виконана на основі наукових публікацій та публікацій з мережі інтернет

Наукова новизна. Запропонована система аналізу настроїв покупців інтернет-магазину на основі аналізу коментарів, яка використовує словник української мови, і дозволить здійснювати інструменти аналізу настроїв в процесі діяльності українського бізнесу.

Практичне значення. Практичне значення полягає у розробленні системи аналізу настроїв покупців інтернет-магазину «Розетка», яка може бути корисною для будь-якого інтернет-магазину, який надає можливість покупцю залишати коментарі.

Структура і обсяг роботи. Робота складається за вступу, трьох розділів, висновків, списку використаних джерел і додатку з програмним кодом. Перший розділ включає теоретичні відомості про технологію оброблення природної мови, аналіз існуючих інтелектуальних системи аналізу настроїв та аналіз існуючих областей використання технології аналізу настроїв. Другий розділ присвячений дослідженню алгоритму обробки текстів та існуючих моделей векторного представлення слів. Третій розділ – практична реалізація проведеного дослідження, представлена у вигляді запропонованої ситсеми аналізу настроїв на прикладі інтернет-магазину «Розетка».

Загальний обсяг роботи становить 35 сторінок друкованого тексту, 26 рисунків на 14 сторінках, 1 додаток на 2 сторінках. Список використаних джерел налічує 19 найменування.

РОЗДІЛ 1 ТЕОРЕТИЧНІ АСПЕКТИ ВИКОРИСТАННЯ ІНТЕЛЕКТУАЛЬНИХ СИСТЕМ АНАЛІЗУ НАСТРОЇВ

1.1 Теоретичні основи обробки природної мови та аналізу настроїв

Сучасні інформаційні технології та зростання обсягів даних сприяють широкому застосуванню засобів штучного інтелекту в різних галузях, таких як медицина, транспорт, фінанси, бізнес та інші. Одним із важливих аспектів розвитку штучного інтелекту є обробка природної мови (Natural Language Processing, NLP), технологія, що дозволяє комп'ютерам розуміти та аналізувати людську мову. NLP встановлює комунікаційний зв'язок між людьми та комп'ютерами, відкриваючи можливості для вирішення завдань, які раніше були виконувані виключно людьми.

Natural Language Processing (NLP) — це форма штучного інтелекту (AI), яка зосереджується на способах взаємодії комп’ютерів і людей за допомогою людської мови. Методи НЛП допомагають комп’ютерам аналізувати, розуміти та реагувати на нас, використовуючи наші природні способи спілкування: мову та письмовий текст [1]. NLP включає в себе міждисциплінарний підхід, який поєднує комп'ютерні науки, штучний інтелект і обчислювальну лінгвістику. Головною метою NLP є створення засобів для безпосередньої взаємодії між комп'ютерами та людьми за допомогою природної мови [2].

Область обробки природної мови (NLP) представляє значущий інтерес в сучасних інформаційних технологіях і має широкий спектр застосувань, починаючи від аналізу текстової інформації і закінчуючи розробкою систем автоматичного перекладу. В рамках NLP існують можливості для вирішення різноманітних завдань [3], включаючи:

- пошук інформації;
- автоматизований переклад текстів;
- перевірка грамотності текстів;
- розпізнавання мовлення та пошук відповідей;

- створення ботів та особистих асистентів;
- класифікація текстів;
- автоматичне створення відгуків та резюме;
- голосове управління;
- сумаризація – пошук головних фактів і переказ змісту тексту;
- аналіз настроїв;
- показ відповідної онлайн реклами (пошук схожого контексту);
- прогнозування тощо.

Для виконання цих завдань, необхідно мати вміння аналізувати та розуміти мову. І саме це досягається завдяки NLP, тобто технології, що дозволяє встановлювати зв'язок між людьми та комп'ютерами через мову [3].

Однією з ключових областей застосування NLP є sentiment analysis, що полягає у визначенні емоцій та настроїв людей. Ця область має важливе значення в різних галузях, таких як бізнес, маркетинг, наука та інші. Аналіз настроїв допомагає визначати попит на продукти та послуги, досліджувати поведінку та думки клієнтів, підтримувати репутацію бренду, покращувати комунікацію зі споживачами та багато інше.

Sentiment analysis (SA) – аналіз настроїв, (аналіз тональності, аналіз думок) – представляє собою галузь дослідження, яка зосереджена на вивченні думок, емоцій, ставлень, і вражень, які виражаються щодо об'єктів, таких як товари, послуги, організації, особистості, проблеми, події, теми та їхні атрибути [4,5].

Аналіз настроїв є важливою складовою обробки природної мови та ставить своєю метою виділення ознак, які виражають емоційне ставлення, і визначення полярності текстового документа як "позитивного," "негативного" або іноді "нейтрального". При аналізі настроїв для текстових відгуків фіксується думка, судження або емоції автора стосовно об'єктів, згаданих у тексті. Зокрема, аналіз настроїв допомагає встановити, які погляди та переконання мають інші люди щодо різних об'єктів, таких як товари, послуги та підприємства, з метою забезпечення обґрунтованих рішень.

Аналіз настрою - це процес класифікації тексту, адже його виконання включає в себе кроки, подібні до тих, що використовуються у тематичній класифікації. Втім, основна відмінність між аналізом настроїв і традиційною тематичною класифікацією полягає в тому, що аналіз настроїв вимагає ідентифікації більш тонких характеристик, таких як вибір слів і структура речень, щоб точно визначити настрої частини тексту. Аналіз настрою потребує навіть більш тонких функцій, таких як слова та фрази, наповнені емоціями, щоб точно визначити конкретні емоції, виражені в тексті. [6].

Основною метою аналізу настрою є розуміння поглядів та думок людей з метою сприяння розвитку бізнесу. Цей вид аналізу фокусується не лише на визначенні полярності тексту (тобто, чи він містить позитивні, негативні або нейтральні відгуки), але також на виявленні виражених емоцій, таких як радість, сум, агресія і т. д. Для досягнення цієї мети використовуються різні методи обробки природної мови, включаючи алгоритми на основі правил, автоматичні та гібридні підходи [7].

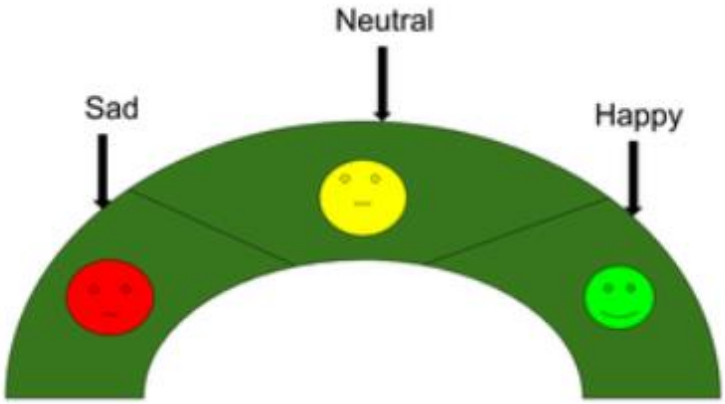

Рисунок 1.1 – Візуалізація Sentiment analysis [7]

Аналіз настроїв може бути класифікованим на кілька типів, включаючи:

Точний аналіз настроїв: Цей підхід базується на полярності тексту та включає такі категорії як дуже позитивний, позитивний, нейтральний, негативний або дуже негативний настрій. Оцінка зазвичай проводиться на шкалі від 1 до 5, де 1 відповідає дуже негативному настрою, а 5 - дуже позитивному.

Виявлення емоцій: Цей метод визначає емоційний стан тексту та може розпізнати емоції, такі як щастя, сум, гнів, смуток, радість, приємність та інші. Це часто відомо як лексиконний метод аналізу настроїв.

Аналіз настрою на основі аспектів: Цей підхід фокусується на аналізі настрою відносно конкретних аспектів. Наприклад, якщо користувач оцінює функціональність мобільного телефону, він може аналізувати такі аспекти, як акумулятор, екран та якість камери, і потім оцінювати настрій на основі цих аспектів.

Аналіз багатомовних настроїв: Цей підхід враховує різні мови, де класифікація може бути проведена в позитивному, негативному та нейтральному контексті. Проведення такого аналізу може бути важливим завданням через різноманітність мов та культур.

1.2 Аналіз існуючих інтелектуальних систем аналізу настроїв

Інструмент аналізу настроїв — це програмне забезпечення штучного інтелекту, яке автоматично аналізує текстові дані, щоб допомогти швидко зрозуміти, як клієнти ставляться до бренду, продукту чи послуги [8].

Інструменти аналізу настроїв використовуються для проведення ретельного аналізу текстового матеріалу за допомогою методів машинного навчання та обробки природної мови. Важливим аспектом їх роботи є кількість текстів, які були проаналізовані, оскільки це впливає на точність отриманих результатів.

Завдяки інструментам аналізу настроїв, компанії можуть в режимі реального часу відстежувати відношення клієнтів до свого бренду, продукту або послуги. Аналіз виражених в Інтернеті думок і відгуків може служити важливою інформацією для прийняття обґрунтованих бізнес-рішень.

Моделі аналізу настроїв здатні ефективно класифікувати настрої та надавати корисну інформацію, яку можна використовувати у різних підрозділах компанії.

На сучасному ринку інструментів для аналізу настроїв існує широкий вибір пропозицій. Наприклад, у продуктах компанії Microsoft, зокрема у системі Microsoft Dynamics 365 Customer Insights, є можливість використання інструментів аналізу настроїв для обробки відгуків клієнтів [9]. Цей сервіс дозволяє налаштовувати функції для кожного ідентифікатора клієнта. Для цього використовуються дві моделі обробки природної мови (NLP):

– перша модель надає оцінку настроїв кожному коментарю відгуку.

– друга модель пов'язує кожний відгук із всіма відповідними бізнес-аспектами.

Ці моделі навчаються на основі публічних даних, отриманих із різних джерел, включаючи соціальні мережі, роздрібну торгівлю, ресторани, споживчі товари та автомобільну промисловість (рисунок 1.2).

Рисунок 1.2 – Результати аналізу настроїв у відгуках клієнтів у Microsoft Dynamics 365 [9]

Ще однією популярною платформою для моніторингу медіа та аналізу настроїв є Brand24. Цей інструмент призначений для відстеження інформації в Інтернеті та соціальних мережах, і він включає в себе функцію аналізу глобального настрою. Brand24 проводить моніторинг всіх основних платформ соціальних мереж, а також враховує блоги, форуми, новинні веб-сайти, подкасти і інформаційні бюлетені[10]. Інтегрована система аналізу настроїв використовується для оцінки відгуків та публікацій у соціальних мережах. Brand24 проводить збір коментарів в реальному часі та надає інструменти для аналізу ЗМІ. Використання алгоритмів машинного навчання та обробки природної мови дозволяє аналізувати тексти у реальному часі (див. рис. 1.3).

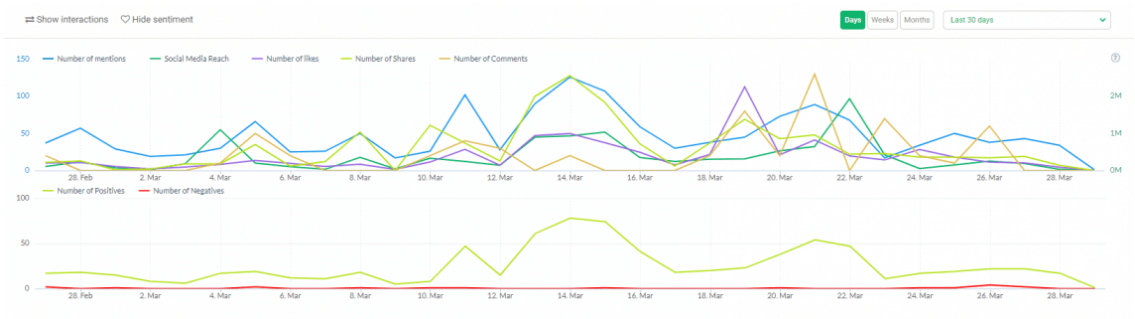

Рисунок 1.3 – Результати аналізу настроїв у відгуках клієнтів у Brand24[10]

Google Alerts [11] – це зручний і простий інструмент для моніторингу брендів, який може бути особливо корисним для нових компаній або малих підприємств. Однією з найбільших переваг цього інструменту є те, що він доступний абсолютно безкоштовно.

Робочий принцип Google Alerts полягає в тому, що користувач визначає ключові слова, які цікавлять його для моніторингу в інтернеті. Кожного разу, коли ці ключові слова зустрічаються в онлайн-середовищі, користувач отримує сповіщення на свою електронну пошту.

Проте, варто відзначити, що Google Alerts має свої обмеження та недоліки, серед яких можна виділити наступні:

– затримка у надходженні результатів або навіть їх відсутність у деяких випадках.
– обмежений обсяг джерел, які індексує інструмент.
– не завжди ефективна робота фільтрів для вибірки релевантних результатів.
– можливість отримання спаму в результатах моніторингу.
– іноді він може пропустити вміст, який вже існує в google і вже індексований пошуковою системою.

Clarabridge [10, 12] представляє собою інструмент аналізу настроїв, який входить у склад рішення Customer Experience Management. Це комплексне рішення включає в себе дві важливі компоненти: CX Analytics і CX Social.

Під час використання системи, для індексації настроїв, зібраних з різних джерел, використовується 11-бальна шкала. Під час оцінювання окремих фрагментів тексту беруть до уваги такі параметри, як граматика, контекст, галузь, та джерело походження інформації. Цей інструмент аналізу настроїв ідеально підходить для збору відгуків від клієнтів та для пошуку в них позитивних, негативних або нейтральних вражень.

Repustate [13] представляє собою інструмент для аналізу настроїв, який надає текстовий аналіз для компаній 17 мовами. Перед проведенням фактичного аналізу, цей інструмент використовує процес, відомий як частинне мовне тегування, який ґрунтується на розбитті тексту на граматичні частини. Після завершення цього етапу стає легше визначити, які фрази є найбільш цікавими для аналізу настроїв. Необхідно підкреслити, що цей інструмент акцентує увагу на інших факторах, таких як лематизація, попередня полярність і так далі.

У Repustate також вбудована система розпізнавання іменованих об'єктів (NER) з розширеним семантичним пошуком для ідентифікації брендів і бізнес-суб'єктів у вхідних даних. Незалежно від того, наскільки неправильно написано слово, модель природної мови (NLP) відновить ім'я в правильному написанні та забезпечить точний пошук імен, транслітерацію і перевірку ідентичності. Це дозволяє отримувати високоточні ранжирувані результати, враховуючи мовні, фонетичні і специфічні культурні варіації імен.

Інструмент аналізу настрою OpenText є важливою частиною платформи аналізу вмісту OpenText, яка дозволяє визначати та оцінювати суб'єктивні моделі і вирази настрою, які містяться у текстовому контенті [14]. Цей інструмент функціонує на трьох рівнях аналізу: рівень теми, рівень окремих речень та рівень цілих документів.

Аналіз здійснюється у такий спосіб, щоб визначити, чи частини тексту є фактичними або суб'єктивними. Щодо суб'єктивного аналізу, інструмент також призначений для визначення, чи виражені думки в окремих частинах тексту є позитивними, негативними, змішаними або нейтральними. OpenText підтримує

п'ять основних мов: англійську, німецьку, французьку, іспанську та португальську.

1.3 Аналіз областей використання систем аналізу настроїв

У сучасному інформаційному суспільстві системи аналізу настроїв знаходять широке застосування в різних галузях, включаючи науку, технології, медицину, бізнес та суспільство загалом. До найбільш поширених галузей використання систем аналізу настроїв відносять наступні.

Медицина та психологія. Системи аналізу настроїв широко використовуються у медицині та психології для діагностики та моніторингу психічного стану пацієнтів. Вони можуть допомагати виявляти ознаки депресії, тривожності та інших психічних розладів за текстовими виразами та відгуками пацієнтів.

Соціальні мережі і маркетинг. Багато компаній та маркетологів використовують системи аналізу настроїв для відслідковування реакцій споживачів на їх продукти та послуги через відгуки в соціальних мережах і відгуки на веб-сайтах. Це дозволяє підприємствам адаптувати свої стратегії маркетингу та взаємодії з клієнтами.

Фінансовий аналіз. У фінансовому секторі аналіз настроїв може бути корисним для прогнозування ринкових тенденцій та ризиків. Трейдери та інвестори можуть використовувати дані з соціальних мереж та новинних порталів для прийняття рішень щодо інвестицій.

Безпека. Аналіз настроїв також може бути важливим інструментом для правоохоронних органів та агентств національної безпеки для виявлення загроз громадській безпеці. Вони можуть моніторити соціальні мережі та медіа для виявлення потенційних загроз та кризових ситуацій.

Наукові дослідження. Системи аналізу настроїв також знаходять застосування в наукових дослідженнях. Дослідники можуть використовувати їх для аналізу громадської думки та реакції на події та тенденції у реальному часі.

Кліматичні дослідження. В останні роки системи аналізу настроїв також використовуються для вивчення реакції громадськості на кліматичні зміни та інші питання, пов’язані із природнім середовищем. Це може допомогти визначити градус глобального обурення і підтримки конкретних кліматичних ініціатив.

Освіта та особистий розвиток. Додатки та платформи з системами аналізу настроїв можуть бути корисними для освіти та особистого розвитку. Вони допомагають аналізувати та поліпшувати мовні та комунікативні навички студентів та користувачів.

Проте, проведені дослідження показали, що найчастіше системи аналізу настроїв використовують в **бізнесі та торгівлі**. Для власника бренду важливо розуміти сучасні тенденції попиту його товару.

З іншого боку, сучасні інформаційні технології дають можливість застосовувати системи аналізу настроїв завдяки можливостям збирати персональні дані про клієнтів та коментарі. В будь-якому інтернет-магазині є можливість залишити коментар про товар чи послугу, що є ключовим елементом подальшого аналізу.

Висновки до розділу 1

Обробка природної мови є актуальним і перспективним напрямом дослідження. Основними напрямками використання NLP є: автоматизований переклад текстів; перевірка грамотності текстів; створення ботів та особистих асистентів; класифікація текстів; аналіз настроїв тощо.

Аналіз настроїв дозволяють виявити ставлення клієнтів до бренду, продукту чи послуги. Існує багато систем аналізу настроїв, переважна більшість – закордонні розробки. Найчастіше аналіз настроїв використовують у маркетингу і торгівлі.

РОЗДІЛ 2. МОДЕЛІ ВЕКТОРНОГО ПРЕДСТАВЛЕННЯ СЛІВ

2.1 Алгоритм обробки текстів

Важливим компонентом у проведенні аналізу настроїв є підготовка та обробка текстових даних перед подальшим аналізом [15]. Цей процес включає наступні кроки:

1. Очищення тексту. Перший крок - це очищення тексту від зайвих елементів. Це включає видалення неалфавітних символів, тегів HTML, URL-адрес, знаків пунктуації, пробілів та інших елементів розмітки.

2. Сегментація та токенізація. Наступний крок - це сегментація та токенізація тексту. Під час токенізації текст розбивається на окремі слова або лексеми. Зазвичай всі слова перетворюються на нижній регістр для однорідності.

3. Лематизація та стеммінг. Для обробки різних граматичних форм слів і слів з однаковим коренем використовують лематизацію та стеммінг. Лематизація спрямована на приведення слова до його нормальної форми, тоді як стеммінг обрізає слова до їх основи шляхом відкидання закінчень або суфіксів.

4. Визначення незалежних від контексту ознак. Для кожної лексеми визначаються ознаки, які не залежать від суміжних елементів.

5. Видалення стоп-слів. Стоп-слова - це часті слова, які не несуть значущої інформації в тексті. Вони можуть бути видалені, оскільки вони можуть призвести до зайвого шуму при аналізі.

6. Перетворення тексту в векторне представлення. Для подальшого аналізу текст перетворюється в векторне представлення, яке відображає слова, які зустрічаються в схожих або ідентичних контекстах. Цей метод, відомий як векторне представлення, дозволяє представляти слова у вигляді числових векторів і є основною технікою для багатьох завдань обробки природної мови. Для цього найчастіше використовують метод word embedding [16], який перетворює слова в числові вектори, що дозволяє комп'ютеру працювати з ними.

Ембеддінг (word embedding) – представляє собою метод, що дозволяє асоціювати слова або фрази з числовими векторами. У цьому процесі використовуються математичні моделі, зокрема числові вектори, які представляють собою рядки в матриці, яка відома як "слово-контекст". Ця матриця будується на основі великого обсягу даних з різних датасетів. Контекстом для кожного слова може бути безпосередньо сусіднє слово або слова, які входять в одну семантичну або синтаксичну конструкцію разом з цим словом.

Зазвичай ці матриці містять частоти взаємопояви слова в даному контексті. Проте, зростає популярність використання коефіцієнта позитивної попарної взаємної інформації (Positive Pointwise Mutual Information, PPMI), який дозволяє визначити, наскільки ймовірною є поява слова в певному контексті в порівнянні з випадковою подією. Це допомагає зрозуміти семантичні відношення між словами та допомагає класифікувати слова за їх схожістю або синонімічністю.

Цей метод виявляється корисним у багатьох задачах, включаючи машинний переклад. Використовуючи векторне представлення тексту, можна легко здійснювати переклад тексту з однієї мови на іншу. Це спрощує завдання пошуку інформації в мережі, оскільки пошук може бути здійснений у всіх мовах, використовуючи схожий підхід.

2.2 Моделі перетворення слів у вектори

NLP включає в себе розмаїтні моделі для перетворення слів у вектори. Наведемо основні з них та опишемо деякі з них.

Bag-of-Words (BoW) - модель, в якій кожне слово розглядається індивідуально і отримує свій унікальний індекс. Текст подається у вигляді вектора, де кожний елемент відповідає кількості разів, які дане слово зустрічається в тексті. З іншими словами, ця модель створює матрицю входжень для речення або документа, і вона ігнорує граматику та порядок слів. Частоти, з

якими слова зустрічаються, потім використовуються як ознаки для подальшого навчання моделей.

Основна ідея BoW полягає в припущенні, що схожі документи мають схожий зміст. Отже, на основі змісту тексту ми можемо дізнатися дещо про значення документа.

Проте, не дивлячись на свою простоту та легкість розуміння, цей підхід має серйозний недолік. При кодуванні BoW використовується словник слів, і кожен текст представляється вектором, довжина якого відповідає розміру словника. Якщо слово присутнє в тексті, відповідним елементом вектора буде кількість разів, які дане слово зустрічається в тексті. Якщо користуватися векторами із великим словником, розмірність простору ознак також зростає, і вона може становити десятки чи навіть сотні тисяч. Ця розмірність зростає разом із збільшенням обсягу словника.

Continuous Bag of Words (CBOW). В даній моделі береться багато різних пропозицій з корпусу. Кожного разу, коли алгоритм бачить певне слово, він бере слова, які знаходяться поруч з цим словом, і намагається передбачити саме це слово, як центральне слово, на основі інших слів у контексті. Таким чином, для кожного такого контексту отримується один приклад даних для навчання нейронної мережі. Після тренування нейромережа створює вкладення (embedding) для кожного слова в корпусі. Якщо нейромережу навчати на великій кількості речень і слів з подібним контекстом, слова, які зазвичай зустрічаються в однаковому контексті, отримають схожі векторні представлення.

Word2Vec - модель використовує нейронні мережі для цього завдання і ґрунтується на ідеї, що слова, які зустрічаються в схожих контекстах, мають схожі значення.

Word2Vec приймає великий корпус (corpus) тексту, де кожне слово вже представлено у вигляді вектора. Алгоритм моделі пройде кожну позицію t тексту, вважаючи її центральним словом c та контекстним словом o. Використовуючи схожість між векторами слів для c і o, модель розраховує

ймовірність появи ***o*** у контексті слова ***c***. Потім вектори слів регулюються так, щоб максимізувати цю ймовірність.

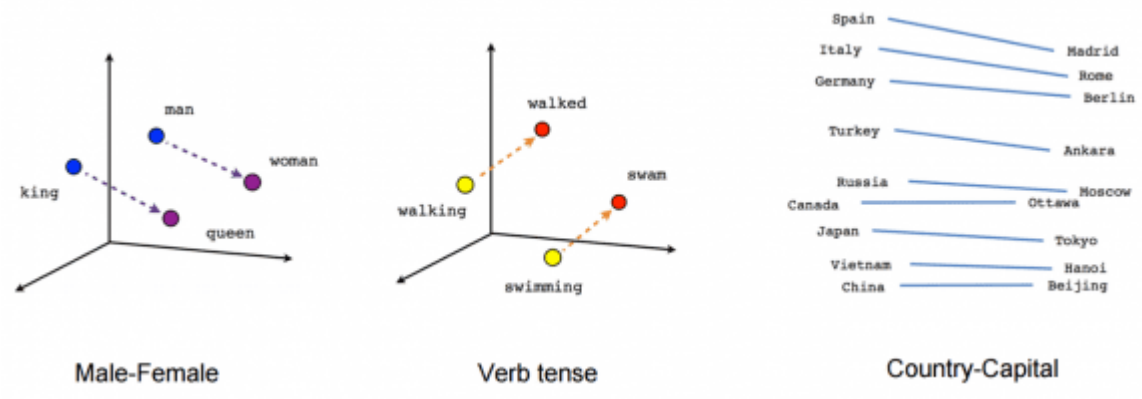

Рисунок 2.1 – Візуалізація моделі Word2Vec

Word2vec може бути представлений двома типами моделей Skip-Gram та Continuous Bag of Words (CBOW) (рис. 2.5):

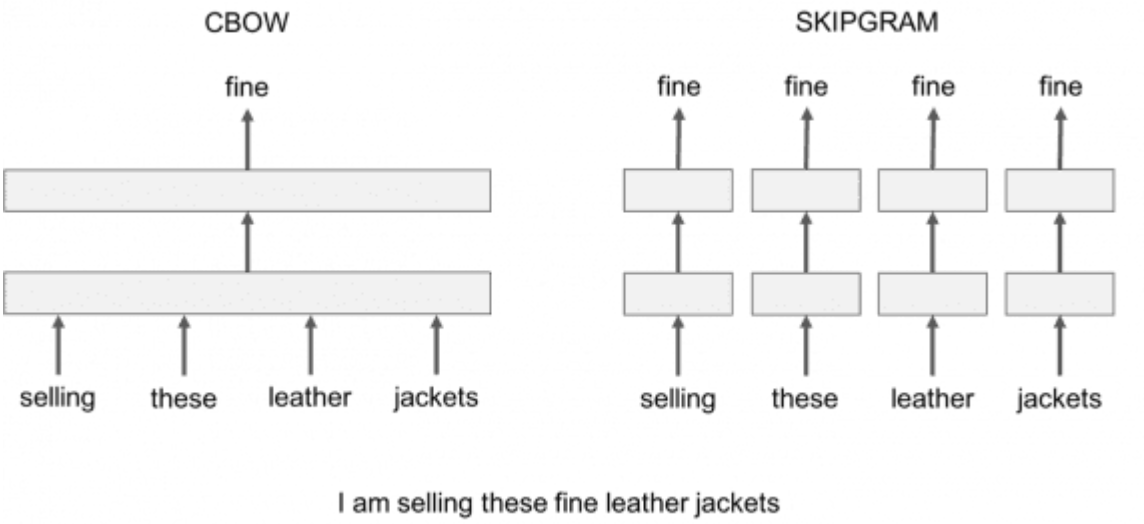

Рисунок 2.2 – Моделі Word2vec

Skip-Gram. В цій моделі розглядається контекстне вікно, яке включає послідовні слова. Потім одне слово в цьому контекстному вікні пропускається, і нейронна мережа навчається передбачати це пропущене слово на основі інших слів у вікні. Іншими словами, модель намагається зрозуміти, які слова часто зустрічаються разом в тексті. Якщо два слова мають схожий контекст і зазвичай зустрічаються поруч у корпусі, ці слова будуть мати схожі векторні представлення.

GloVe - модель, яка використовує глобальну статистику спільного входження слів у текстах для створення векторів слів. Ця модель враховує частоту спільного вживання слів і базується на статистичних даних.

FastText - модель, яка використовує нейронні мережі, але працює зі словами, розглядаючи їх як набори символів, включаючи n-грами. Це дозволяє моделі враховувати морфологічну та семантичну інформацію про слова.

ELMo - модель, яка генерує вектори слів, враховуючи контекст, в якому вони використовуються у тексті. Вона здатна враховувати значення слова, залежно від його контексту.

BERT - модель, яка використовує трансформерну архітектуру та навчається на великих корпусах текстів. Вона дозволяє розуміти семантичні відносини між словами та фразами у тексті, а також вивчає контекст слів.

N-gram. - це інший підхід для аналізу тексту, в якому розглядаються групи слів різної довжини, такі як уніграми (одне слово), бі-грами (послідовність двох слів), триграми (три слова) і так далі. Число "N" вказує на кількість слів у групі. Модель N-грам використовує лише ті групи слів, які мають відповідні дані в наборі даних, і дозволяє аналізувати спільне вживання групованих слів в текстах.

Висновки до розділу 2

Визначені основні етапи обробки природної мови, які включають: очищення текст, сегментація та токенізація, лематизація та стеммінг, визначення незалежних від контексту ознак: видалення стоп-слів, перетворення тексту в векторне представлення. Сучасні моделі векторного представлення слів наступні: Bag-of-Words, Word2Vec, Skip-Gram, GloVe, FastText, ELMo, BERT, N-gram. Даний список не є вичерпним.

РОЗДІЛ 3. РОЗРОБКА ІНСТРУМЕНТУ АНАЛІЗУ НАСТРОЇВ ПОКУПЦІВ ІНТЕРНЕТ-МАГАЗИНУ

3.1. Опис предметної області

У даному дослідженні ми використовуємо сентимент-аналіз для аналізу коментарів в інтернет-магазині "Розетка" з метою надання клієнтам інструменту, який допомагає при виборі товарів. При купівлі товарів у магазині, клієнти часто зіштовхуються з великим асортиментом і різнорідністю брендів і моделей. Щоб зробити правильний вибір, важливо мати доступ до інформації про якість та властивості товарів, а також думку інших клієнтів, які вже купували той чи інший продукт. Сентимент-аналіз допомагає класифікувати відгуки клієнтів на позитивні, негативні та нейтральні, що робить процес вибору товару більш швидким. Під час дослідження, ми показуємо, як сентимент аналіз може бути використаний для знаходження та відбору найкращих товарів, що допомагає клієнтам зробити свідомий вибір, а також рекомендуємо покупцям спираючись на емоційний фідбек інших користувачів.

3.2. Отримання даних

Обробка текстової інформації включає кілька етапів, які можна описати наступним чином:

- збір даних.
- підготовка даних для обробки.
- проведення обробки тексту з використанням методів NLP.
- аналіз результатів отриманих після обробки тексту.

Для проведення аналізу коментарів та відгуків до товарів інтернет магазину, необхідно створити корпус даних і виконати аналіз настроїв (sentiment analysis). Процес створення корпусу даних та виконання аналізу настроїв клієнтів інтернет-магазину «Розетка» показаний на рисунку 3.1. Для цих цілей

використовується мова програмування Python та відповідні бібліотеки для роботи з текстовою інформацією.

Рисунок 3.1 – Етапи здійснення аналізу настроїв клієнтів інтернет-магазину

Результати парсингу будуть зберігатися у форматі JSON і включатимуть такі дані:

– автор коментаря.
– дата коментаря.
– коментар.
– посилання на коментар.
– об'єкт коментаря, що описує, на який об'єкт спрямований коментар або відгук.

Отримані в результаті парсингу дані, необхідно підготувати для подальшої обробки. Цей етап називається передобробкою тексту (text preprocessing).

Text preprocessing – підготовка і очищення даних для обробки. Цей процес включає наступні етапи:

1. Очищення тексту від неалфавітних символів, різних тегів, URL-адрес, знаків пунктуації, пробілів та інших елементів розмітки.

2. Виключення стоп-слів.
3. Перетворення тексту на нижній регістр.
4. Токенізація тексту – розбиття тексту на лексеми.
5. Лематизація тексту – зведення форми слова до її нормальної (словникової) форми.
6. Здіснення sentiment analysis.

Розглянемо процес обробки даних на прикладі коментарів, отриманих з веб-сайту rozetka.com.ua [17]. Щоб отримати ці коментарі, перше, що необхідно зробити, - це проаналізувати структуру вказаного веб-сайту та визначити можливість отримання даних через API (інтерфейс програмування застосунків) [18]. API можна розглядати як спосіб для двох або більше комп'ютерних програм спілкуватися одна з одною.

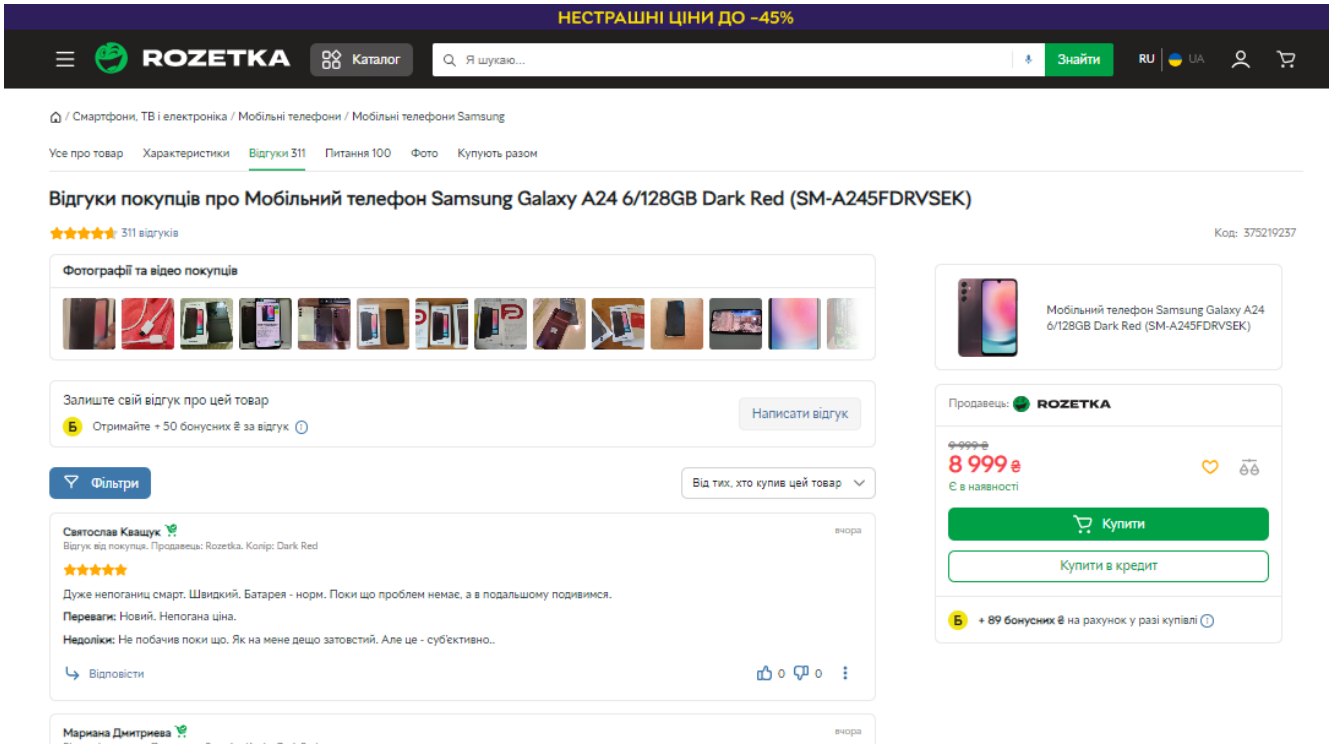

Рисунок 3.2 – Коментарі про товар з сайту rozetka.com.ua для парсингу і аналізу настроїв

На підставі аналізу веб-сайту rozetka.com.ua, було встановлено, що весь набір необхідної інформації для проведення аналізу настроїв може бути

здобутий з цього ресурсу. Результати процесу парсингу, описані у форматі JSON, представлено на рисунку 3.3.


```
{
"name": "Овчаренко Сергей",
"date_com": "18.10.2023",
"description": "Такий собі телефон, ні хороший, ні поганий, краще б Xiaomi redmi note 12s узяв...",
"Positiv": "Яркий екран, непогана камера",
"Negativ": "Не вартуе своїх грошей, ні чохла,ні зарядного пристрою в комплекті немае",
"title": "Відгуки покупців про Мобільний телефон Samsung Galaxy A24 6/128GB Dark Red (SM-A245FDRVSEK)",
"link": "https://rozetka.com.ua/ua/samsung-sm-a245fdrvsek/p375219237/comments/"
},
```


Рисунок 3.3 – Приклад спарсеного коментаря

Наступні кроки передбачають підготовку текстового матеріалу для подальшої обробки. В першу чергу, необхідно провести очищення інформації від символів, що не входять до кириличного алфавіту і перетворення тексту на нижній регістр.

Для цього можна використовувати стандартну бібліотеку регулярних виразів re (regular expressions) мови програмування Python, яка надає всі необхідні інструменти для цього. Програмний код показаний в додатку А.

На рисунку 3.4 зображена функція cl_text, що замінює не кириличні символи на пробіли, і перетворює текст на нижній регістр.

```python
import re

def cl_text(text: str) -> str:
    cl_text = re.sub(r'[^а-яА-ЯґҐєЄіІїЇўЎ\s]', ' ', text).lower()
    return ' '.join(cl_text.split())
```

Рисунок 3.4 – Функція cl_text

```
Це мій перший Samsung і від нього одне розчарування. У Xiaomi і то менше глюків. Не рекомендую людям зі слабкими нервами
==========
це мій перший і від нього одне розчарування у і то менше глюків не рекомендую людям зі слабкими нервами
```

Рисунок 3.5 – Результати роботи функції cl_text

Наступним кроком підготовки тексту є видалення стоп-слів. Для цього буде використовуватись розроблений Сергієм Курпієнко список стоп-слів української мови. Список знаходиться на ресурсі https://github.com/skupriienko/Ukrainian-Stopwords. Список містить 1983 слова.

Рисунок 3.6 – Перелік стоп-слів української мови

```python
# Читання вмісту файлу з стоп-словами
with open('/content/stopwords_ua_list.txt', 'r', encoding='utf-8') as file:
    stop_words = file.read()

# Перетворення в списку потрібний формат
stop_words = stop_words.strip('[]').replace("'", "").split(', ')

words = text.split()  # Розділення тексту на слова (токенізація)

# Очистка тексту від стоп-слів
filtered_words = [word for word in words if word.lower() not in stop_words]

# Крок 4: З'єднання слів назад
clean_text = " ".join(filtered_words)
```

Рисунок 3.7 – Код для видалення стоп-слів

```
Це мій перший Samsung і від нього одне розчарування. У Xiaomi і то менше глюків. Не рекомендую людям зі слабкими нервами
==========
розчарування глюків не рекомендую людям слабкими нервами
```

Рисунок 3.8 – Результат процесу видалення стоп слів

```python
def lemmatize_text(text: str) -> str:
    morph = pymorphy2.MorphAnalyzer(lang='uk')
    lemmas = [morph.parse(word)[0].normal_form for word in text.split()]
    return ' '.join(lemmas)
```

Рисунок 3.9 – Функція для лематизації тексту

```
Це мій перший Samsung і від нього одне розчарування. У Xiaomi і то менше глюків. Не рекомендую людям зі слабкими нервами
============================
розчарування глюк не рекомендувати люди слабкий нерв
```

Рисунок 3.10 – Результат лематизаці тексту

3.3. Виконання сентимент аналізу

Виконаємо сентимент аналіз на прикладі коментарів до мобільного телефону Samsung Galaxy A24 в інтернет магазині Rozetka. На малюнку 3.11 представлений вміст файлу, який готовий для виконання аналізу настроїв. Даний файл надається у форматі .csv і містить 15 коментарів.

```
N,text
1,телефон ні хороший ні поганий узяти
2,непоганиц смарта швидкий батарея норма проблема а подальший подивитися
3,купувати мама телефон хороший дизайн подобатися колір батерея тримати
4,купувати дитина зручний гра тян гарно тримати батерея
5,хороший середнячок гріш якісний екран камера працювати доволі швидко місяць користування не заглючити ся
6,покупка задоволений гріш
7,задоволений телефон а не вчора почати глючити періодично переставати працювати кнопка додому зліва кнопк
8,розчарування глюк не рекомендувати люди слабкий нерв
9,зручний користування яскравий колір гарний камера швидкий доставка
10,гарний телефон хороший камера батарея достатньо пам ять
11,рекомендувати
12,захват прекрасний телефон
13,хороший телефон працювати рекомендувати
14,гарний телефон ціна
15,цілком задоволений швидко працювати якість зображення суперова
```

Рисунок 3.11 – Зміст файлу для проведення сентимент-аналіз

В минулому розділі ми підготували дані здійснивши передобробку тексту (text preprocessing), отже тепер їх можна використовувати для розрахунку тональності кожного коментаря. Наш підхід базується на використанні лексичного методу, який використовує лексичні ресурси, що містять слова та словосполучення, які часто асоціюються з конкретними емоціями або настроями. Ідея полягає в тому, щоб оцінити загальний настрій коментаря на основі тональності кожного вжитого слова.

Для визначення тональності кожного слова ми використовуємо Словник тональностей української мови, розроблений Сергієм Шеховцевим та іншими. Дані для цього словника отримано за допомогою усереднення оцінок декількох експертів та ML-моделі з використанням векторів слів word2vec та lex2vec, а

також незначною пост-обробкою людиною. Цей словник та його розширення доступні на ресурсі https://github.com/lang-uk/tone-dict-uk.

1	аборигенний	0	-0,25	0,433012701892219
2	аборт	-1	-1	0,816496580927726
3	абсолютний	0	0,333333333333333	0,471404520791032
4	абстрактний	0	-0,111111111111111	0,87488976377909
5	абсурд	-1	-1	0
6	абсурдний	-1	-1	0
7	абсурдно	-1	-1	0
8	абхаз	0	0	0
9	авантюра	0	-0,333333333333333	0,471404520791032
10	авантюрний	0	-0,333333333333333	0,471404520791032

Рисунок 3.12 – Фрагмент українського тонального словника [19]

У файлі з назвою "tone-dict-uk.tsv" міститься словник, який складається з 3442 слів української мови, і кожне слово має відмінну від нейтральної тональність, яка виражена числами у діапазоні від -2 до 2. Формат даних у цьому файлі представлений наступним чином:

– слово в базовій граматичній формі.

– дискретна тональність слова, яка може приймати одне із значень з діапазону: -2, -1, 0, 1, 2.

Код для завантаження словника тонів і виконання сентимент аналізу зображений на рисунках 3.13 – 3.14.

```
# завантажуємо словник тонів
url = 'https://raw.githubusercontent.com/lang-uk/tone-dict-uk/master/tone-dict-uk.tsv'
r = requests.get(url)
with open('tone-dict-uk.tsv', 'wb') as f:
    f.write(r.content)
d = {}
with open('tone-dict-uk.tsv', 'r') as csv_file:
    for row in csv.reader(csv_file, delimiter='\t'):
        d[row[0]] = float(row[1])
```

Рисунок 3.13 – Завантаження словника тонів

```python
# ініціалізуємо SentimentIntensityAnalyzer з оновленим словником тонів
SIA = SentimentIntensityAnalyzer()
SIA.lexicon.update(d)

# функція для отримання оцінки настроїв
def get_sentiment(text):
    return SIA.polarity_scores(text)['compound']

# додаємо стовпець з оцінкою настроїв
data['sentiment'] = data['text'].apply(get_sentiment)

# виводимо результат
print(data[['N', 'text', 'sentiment']])
```

Рисунок 3.14 – Сентимент-аналіз та виведення результатів

Отриманий результат відображений на рисунку 3.13.

	N	text	sentiment
0	1	телефон ні хороший ні поганий узяти	0.0000
1	2	непоганиц смарта швидкий батарея норма проблем...	0.0000
2	3	купувати мама телефон хороший дизайн подобатис...	0.4588
3	4	купувати дитина зручний гра тян гарно тримати ...	0.2500
4	5	хороший середнячок гріш якісний екран камера п...	0.6124
5	6	покупка задоволений гріш	-0.2500
6	7	задоволений телефон а не вчора почати глючити ...	0.0000
7	8	розчарування глюк не рекомендувати люди слабки...	-0.2500
8	9	перестати ласка писати коментар камера нормаль...	0.0000
9	10	гарний телефон хороший камера батарея достатнь...	0.6124
10	11	нарікання працювати відмінно користуватися лег...	0.0000
11	12	працювати купити самсунг самсунг крутяк	0.0000
12	13	хороший телефон працювати рекомендувати	0.4588
13	14	гарний телефон ціна	0.4588
14	15	цілком задоволений швидко працювати якість зоб...	0.0000

Рисунок 3.15 – Результат оцінювання коментарів

3.4. Візуалізація отриманих результатів

Для візуалізації результатів скористаємось бібліотекою Matplotlib для Python.

На рисунку 3.16 – 3.18 зображений розподіл оцінок настрою у вигляді кругової діаграми (рис. 3.16), гістограми (рис. 3.17) та лінійної діаграми (рис. 3.18), які отримані в результаті здійсненого sentiment analysis, описаного в п.3.3.

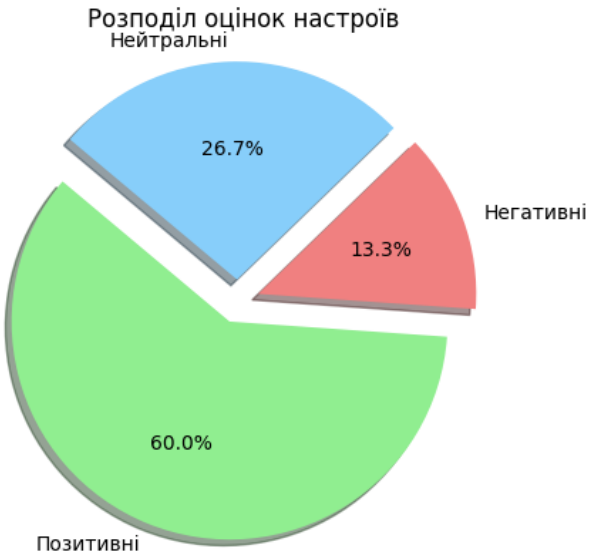

Рисунок 3.16 – Візуалізація результатів в вигляді кругової діаграми

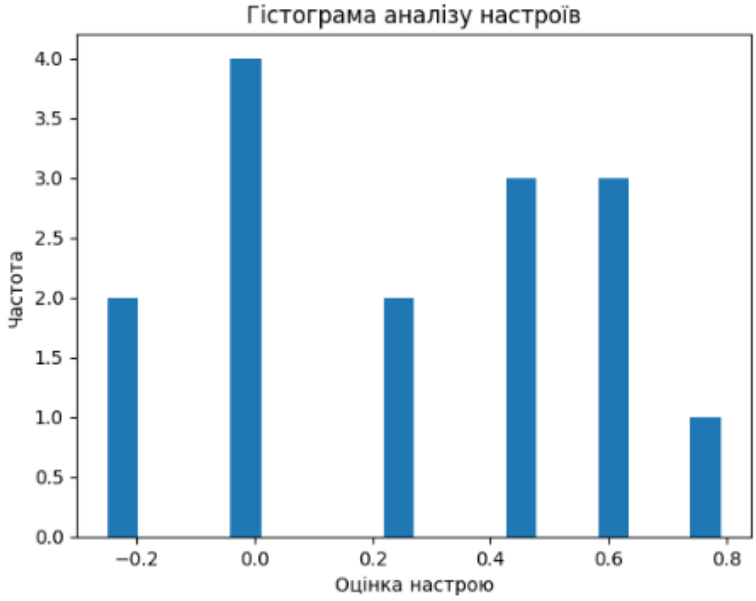

Рисунок 3.17 – Візуалізація результатів в вигляді гістограми

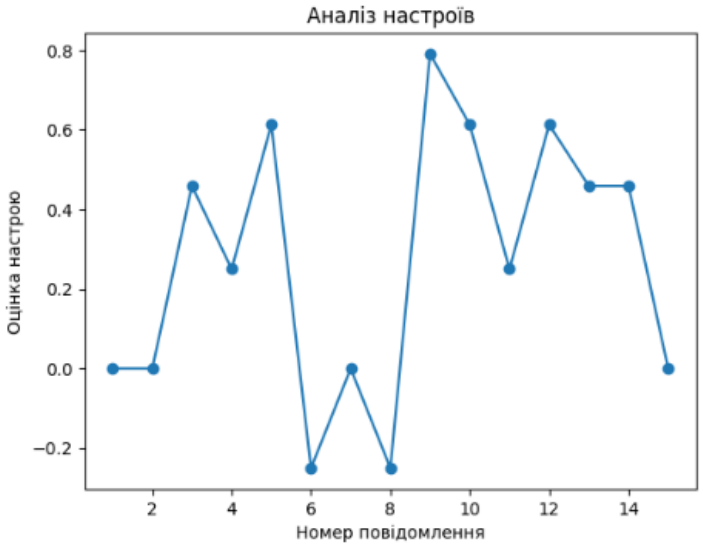

Рисунок 3.18 – Візуалізація результатів в вигляді лінійної діаграми

Далі, для створення візуалізації результатів можна використати метод побудови хмари слів. Цей метод графічно відображає список слів у текстовому корпусі, виділяючи найбільш значущі серед них. Існує багато інструментів для створення хмар слів, але більшість з них працюють із необробленим текстом.

Однак, оскільки наша інтелектуальна система для оцінки настроїв проводить попередню передобробку тексту, хмара слів будується на основі початкової форми кожного слова.

На рисунку 3.19 – 3.21 показані хмари слів для позитивних коментарів (рис. 3.19), позитивних коментарів (рис. 3.20) та усього корпусу слів (рис. 3.21).

Рисунок 3.19 – Візуалізація позитивних коментарів в вигляді хмари слів

Рисунок 3.20 – Візуалізація негативних коментарів в вигляді хмари слів

Рисунок 3.21 – Візуалізація всіх коментарів в вигляді хмари слів

Висновки до розділу 3

Розроблено систему аналізу настроїв для виявлення тональності коментарів інтернет-магазину «Розетка». Для створення ситсеми було проведено аналіз предметної області та отримано глибше розуміння основних аспектів предметної області, її особливостей та основних викликів. Також було розглянуто процес збору та підготовки даних для подальшого аналізу. Реалізація системи виконана мовою програмування Python.

Виконавши сентимент аналіз, ми отримали важливі відомості щодо настроїв та емоцій в текстах. Використовуючи графіки та хмари слів, ми зробили ці результати більш доступними та зрозумілим.

ВИСНОВКИ

Проведено аналіз основних теоретичних відомостей щодо технології оброблення природної мови. Виявлено, що обробка природної мови (NLP) є одним з ключовим напрямків у розвитку штучного інтелекту і знаходить широке застосування в таких галузях, як медицина, бізнес, маркетинг та інші. Однією з важливих областей NLP є аналіз настроїв, який дозволяє визначати емоційний стан та ставлення людей до різних об'єктів і має значуще значення для прийняття обґрунтованих управлінських рішень.

Здійснено аналіз найбільш поширених інструментів аналізу настроїв та сфери його використання. Було виявлено, що на сьогоднішній день існують багато систем аналізу настроїв, переважна більшість – закордонні розробки. Найчастіше аналіз настроїв використовують у маркетингу і торгівлі.

Досліджено існуючі моделі векторного представлення слів. З'ясовано, що основні етапи обробки природної мови включають: очищення текст, сегментація та токенізація, лематизація та стеммінг, визначення незалежних від контексту ознак: видалення стоп-слів, перетворення тексту в векторне представлення. Сучасні моделі векторного представлення слів наступні: Bag-of-Words, Word2Vec, Skip-Gram, GloVe, FastText, ELMo, BERT, N-gram.

Розроблено система аналізу настроїв та виконано аналіз настроїв за відгуками та коментарями до товару клієнтів інтернет-магазину. В результаті цього аналізу було виявлено, що використання технологій оброблення природної мови та sentiment analysis дозволяє ефективного визначити настрої клієнтів стосовно товарів, що дозволяє приймати обґрунтовані рішення при виборі товарів.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. What is natural language processing (NLP)? [Електронний ресурс]. – 2023. – Режим доступу до ресурсу: https://www.elastic.co/what-is/natural-language-processing

2. Yemelianova, O. V., and O. O. Kuksenko. "NATURAL LANGUAGE PROCESSING AS AN ASPECT OF MODERN TECHNOLOGIES DEVELOPMENT." Publishing House “Baltija Publishing” (2023). Режим доступу до ресурсу: https://essuir.sumdu.edu.ua/bitstream-download/123456789/90744/3/Yemelianova_natural_language.pdf;jsessionid=8D4ECC6C0485CE447C77D15BC254DA2E

3. NLP-технології розпізнавання людського мовлення. Можливості та сфери застосування [Електронний ресурс]. – 2019. – Режим доступу до ресурсу: https://evergreens.com.ua/ua/articles/natural-language-processing.html

4. Liu, B. (2012). Sentiment Analysis: A Fascinating Problem. In: Sentiment Analysis and Opinion Mining. Synthesis Lectures on Human Language Technologies. Springer, Cham. https://doi.org/10.1007/978-3-031-02145-9_1

5. Zhang, Lei and B. Liu. “Sentiment Analysis and Opinion Mining.” Encyclopedia of Machine Learning and Data Mining (2012).

6. Deep Learning for Sentiment Analysis: A Survey by Chenghua Lin and Yulan He, 2016. https://arxiv.org/pdf/1801.07883

7. What is sentiment analysis: principles, benefits and tools [Електронний ресурс]. – 2023. – Режим доступу до ресурсу: https://solutionshub.epam.com/blog/post/what-is-sentiment-analysis

8. 15 of The Best Sentiment Analysis Tools [2023] [Електронний ресурс] – Режим доступу до ресурсу: https://monkeylearn.com/blog/sentiment-analysis-tools/

9. Аналіз настроїв у відгуках клієнтів (ознайомлювальна версія) [Електронний ресурс]. – 2023. – Режим доступу до ресурсу: https://learn.microsoft.com/uk-ua/dynamics365/customer-insights/sentiment-analysis.

10. Guide to Brand Monitoring + TOOLS [2023 update] [Електронний ресурс] – Режим доступу до ресурсу: https://brand24.com/blog/brand-monitoring-tools/#1Brand24

11. Сповіщення Слідкуйте за новим цікавим вмістом в Інтернеті [Електронний ресурс]. – 2023. – Режим доступу до ресурсу: https://www.google.com/alerts

12. The gold standard in XM- now with the world's most powerful conversational analytics. The ultimate platform for collecting, understanding, and taking action on all forms of experience data [Електронний ресурс]. – 2023. – Режим доступу до ресурсу: https://www.qualtrics.com/clarabridge/?vid=clarabridge_redirect.

13. Instant customer and employee sentiments in any language [Електронний ресурс]. – 2023. – Режим доступу до ресурсу: https://www.repustate.com/.

14. Magellan Text Mining. What is text mining and content analytics? [Електронний ресурс]. – 2023. – Режим доступу до ресурсу: https://www.opentext.com.au/products-and-solutions/products/ai-and-analytics/opentext-magellan/magellan-text-mining.

15. Effectively Pre-processing the Text Data Part 1: Text Cleaning [Електронний ресурс]. – 2019. – Режим доступу до ресурсу: https://towardsdatascience.com/effectively-pre-processing-the-text-data-part-1-text-cleaning-9ecae119cb3e

16. Natural Language Processing: History, Evolution, Application and Future Work / Prashant Johri, Sunil K. Khatri, Ahmad T. Al-Taani [et al.] // Proceedings of 3rd International Conference on Computing Informatics and Networks. 2021. P. 365–375.

17. Відгуки покупців про Мобільний телефон ZTE Blade A31 Plus 1/32 GB Grey [Електронний ресурс]. – 2023. – Режим доступу до ресурсу: https://rozetka.com.ua/ua/zte-899612/p334610482/comments/

18. API [Електронний ресурс]. – 2023. – Режим доступу до ресурсу: https://en.wikipedia.org/wiki/API

19. Український тональний словник [Електронний ресурс]. – 2023. – Режим доступу до ресурсу: https://github.com/lang-uk/tone-dict-uk.

ДОДАТОК А

Програмний код. Модуль Preprocessing.

```
import re
import pymorphy2

text = "..."

def cl_text(text: str) -> str:
    cl_text = re.sub(r'[^a-яА-ЯґҐєЄіІїЇўЎ\s]', ' ', text).lower()
    return ' '.join(cl_text.split())

def lemmatize_text(text: str) -> str:
    morph = pymorphy2.MorphAnalyzer(lang='uk')
    lemmas = [morph.parse(word)[0].normal_form for word in text.split()]
    return ' '.join(lemmas)

text = cl_text(text)

# Читання вмісту файлу з стоп-словами
with open('/content/stopwords_ua_list.txt', 'r', encoding='utf-8') as file:
    stop_words = file.read()

# Перетворення в списку потрібний формат
stop_words = stop_words.strip('[]').replace("'", "").split(', ')

words = text.split()  # Розділення тексту на слова (токенізація)

# Очистка тексту від стоп-слів
filtered_words = [word for word in words if word.lower() not in stop_words]

# Крок 4: З'єднання слів назад
clean_text = " ".join(filtered_words)

lem_text = lemmatize_text(clean_text)

print(lem_text)
```

Програмний код. Модуль Sentiment analysis.

```
import pandas as pd
import requests
import csv
from nltk.sentiment.vader import SentimentIntensityAnalyzer
import matplotlib.pyplot as plt
import seaborn as sns
import nltk
from wordcloud import WordCloud
```

```
nltk.download('vader_lexicon')
# завантажуємо дані з файлу
data = pd.read_csv('/content/prepocessed.csv', encoding='cp1251')

# завантажуємо словник тонів
url = 'https://raw.githubusercontent.com/lang-uk/tone-dict-uk/master/tone-
dict-uk.tsv'
r = requests.get(url)
with open('tone-dict-uk.tsv', 'wb') as f:
    f.write(r.content)
d = {}
with open('tone-dict-uk.tsv', 'r') as csv_file:
    for row in csv.reader(csv_file, delimiter='\t'):
        d[row[0]] = float(row[1])

# ініціалізуємо SentimentIntensityAnalyzer
SIA = SentimentIntensityAnalyzer()
SIA.lexicon.update(d)

# функція для отримання оцінки настроїв
def get_sentiment(text):
    return SIA.polarity_scores(text)['compound']

# додаємо стовпець з оцінкою настроїв
data['sentiment'] = data['text'].apply(get_sentiment)

# виводимо результат
print(data[['N', 'text', 'sentiment']])


positive_count = len(data[data['sentiment'] > 0])
negative_count = len(data[data['sentiment'] < 0])
neutral_count = len(data[data['sentiment'] == 0])

# Дані для кругової діаграми
labels = ['Позитивні', 'Негативні', 'Нейтральні']
sizes = [positive_count, negative_count, neutral_count]
colors = ['lightgreen', 'lightcoral', 'lightskyblue']
explode = (0.1, 0.1, 0.1)  # Виокремлюємо кілька сегментів

# Побудова кругової діаграми
plt.pie(sizes, explode=explode, labels=labels, colors=colors,
autopct='%1.1f%%', shadow=True, startangle=140)
plt.axis('equal')  # Забезпечує кругову форму діаграми
plt.title('Розподіл оцінок настроїв')

# Відображення кругової діаграми на екрані
plt.show()
```

АНОТАЦІЯ

У даній науковій роботі під шифром «Сентимент код» розглядається проблема аналізу тональності коментарів до товарів в інтернет-магазині.

Актуальність дослідження обумовлена зростанням популярності онлайн-торгівлі та важливістю збереження репутації підприємств, які продають товари через Інтернет.

Метою дослідження є проєктування інструментів аналізу тональності коментарів інтернет-магазину на основі технологій оброблення природної мови.

Завдання наукової роботи: 1. Дослідити теоретичні відомості щодо процесів оброблення природної мови. 2. Здійснити аналіз найбільш поширених інструментів аналізу настроїв та сфери його використання. 3. Дослідити існуючі моделі векторного представлення слів. 4. Виконати аналіз настроїв за відгуками та коментарями до товару клієнтів інтернет-магазину.

Робота складається за вступу, трьох розділів, висновків, списку використаних джерел і додатку з програмним кодом. Перший розділ включає теоретичні відомості про технологію оброблення природної мови, аналіз існуючих інтелектуальних системи аналізу настроїв та аналіз існуючих областей використання технології аналізу настроїв. Другий розділ присвячений дослідженню алгоритму обробки текстів та існуючих моделей векторного представлення слів. Третій розділ – практична реалізація проведеного дослідження, прдставлена у вигляді запропонованої ситсеми аналізу настроїв на прикладі інтернет-магазину «Розетка».

Практичне значення полягає у розробленні системи аналізу настроїв покупців інтернет-магазину «Розетка», яка може бути корисною для будь-якого інтернет-магазину, який надає можливість покупцю залишати коментарі.

Загальний обсяг роботи становить 35 сторінок друкованого тексту, 26 рисунків на 14 сторінках, 1 додаток на 2 сторінках. Список використаних джерел налічує 19 найменування.

Ключові слова: аналіз тональності, інтернет-магазин, відгуки, природна мова, машинне навчання.