МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
Західноукраїнський національний університет
Факультет комп'ютерних інформаційних технологій
Кафедра інформаційно- обчислювальних систем і управління

КОВАЛЬЧУК Наталя Богданівна

Інтелектуальне моделювання тем відгуків в електронній комерції/Intelligent Modeling of Review Topics in E-commerce

спеціальність: 122 - Комп'ютерні науки
освітньо-професійна програма - Комп’ютерні науки

Кваліфікаційна робота

Виконала студентка групи КН-42
Н. Б. Ковальчук

_____

Науковий керівник:
к.т.н., доцент, Х. В.
Ліп'яніна-Гончаренко

_____

Кваліфікаційну роботу
допущено до захисту:

"___" _____ 20___ р.

Завідувач кафедри
_____ **М. П. Комар**

ТЕРНОПІЛЬ - 2024

Факультет комп'ютерних інформаційних технологій
Кафедра інформаційно-обчислювальних систем і управління
Освітній ступінь «бакалавр»
спеціальність 122 – Комп'ютерні науки
освітньо-професійна програма – Комп'ютерні науки

ЗАТВЕРДЖУЮ
Завідувач кафедри
_____ М.П. Комар
«____» _____2023 р.

ЗАВДАННЯ
НА КВАЛІФІКАЦІЙНУ РОБОТУ СТУДЕНТУ
КОВАЛЬЧУК Наталі Богданівні
(прізвище, ім'я, по батькові)

1. Тема кваліфікаційної роботи: Інтелектуальне моделювання тем відгуків в електронній комерції / Intelligent Modeling of Review Topics in E-commerce
керівник роботи Ліп'яніна-Гончаренко Христина Володимирівна, к.т.н., доцент
(прізвище, ім'я, по батькові, науковий ступінь, вчене звання)
затверджені наказом по університету від 12 грудня 2023 р. № 753.

2. Строк подання студентом закінченої кваліфікаційної роботи 15 травня 2024 р.

3. Вихідні дані до кваліфікаційної роботи: завдання на кваліфікаційну роботу студента, наукові статті, технічна література.

4. Основні питання, які потрібно розробити:
- Вивчити ключові поняття та термінологію, які застосовуються в сфері електронної комерції.
- Провести огляд сучасних підходів до аналізу текстових даних у електронній комерції.
- Проаналізувати існуючі інструменти та алгоритми для попередньої обробки даних.
- Розробити методику моделювання тем відгуків користувачів з використанням алгоритмів LDA, BERTopic та LSA.
- Спроектувати та реалізувати алгоритм моделювання тем відгуків в електронній комерції.
- Провести експериментальну перевірку розробленого алгоритму.
- Використовувати комплекс метрик для оцінювання якості роботи моделі.
- Розробити рекомендації щодо використання результатів аналізу текстових даних в стратегіях управління та маркетингу в електронній комерції.

5. Перелік графічного матеріалу в роботі:
- Алгоритм моделювання тем відгуків в електронній комерції
- графіки з результатами експериментальних досліджень.

6. Консультанти розділів кваліфікаційної роботи

Розділ	Прізвище, ініціали та посада консультанта	Підпис, дата	
		Завдання видав	Завдання прийняв

7. Дата видачі завдання 12 грудня 2023 р.

КАЛЕНДАРНИЙ ПЛАН

№ з/п	Назва етапів кваліфікаційної роботи	Строк виконання етапів кваліфікаційної роботи	Примітка
1	Затвердження теми кваліфікаційної роботи, ознайомлення з літературними джерелами та складання плану роботи.	до 01.01. 2024 р.	
2	Написання 1 розділу кваліфікаційної роботи	до 01.03. 2024 р.	
3	Написання 2 розділу кваліфікаційної роботи	до 01.04.2024 р.	
4	Написання 3 розділу кваліфікаційної роботи	до 01.05. 2024 р.	
5	Представлення попереднього варіанту кваліфікаційної роботи, перевірка та внесення змін керівником	до 15.05.2024 р.	
6	Опрацювання зауважень та представлення завершеного варіанту кваліфікаційної роботи. Підготовка супроводжуючих документів.	до 20.05.2024 р.	
7	Перевірка кваліфікаційної роботи на оригінальність тексту у системі «Unicheck».	до 10.06.2024 р.	
8	Оформлення кваліфікаційної роботи та отримання допуску до захисту	до 14.06.2024 р.	
9	Подання кваліфікаційної роботи до захисту на засіданні атестаційної комісії.	до 14.06. 2024 р.	

Студент_____ Н. Б. Ковальчук
(підпис) (прізвище та ініціали)

Керівник кваліфікаційної роботи_____ Х.В. Ліп’яніна-Гончаренко
(підпис) (прізвище та ініціали)

АНОТАЦІЯ

Кваліфікаційна робота на тему «Інтелектуальне моделювання тем відгуків в електронній комерції» на здобуття освітнього ступеня «бакалавр» зі спеціальності 122 «Комп'ютерні науки» освітньої програми «Комп'ютерні науки» написана обсягом в 35 сторінок і містить 8 ілюстрацій, 5 таблиць та 14 використаних джерел.

Метою роботи є розробка та впровадження комплексного підходу до аналізу текстових даних в електронній комерції з використанням методів обробки природної мови (NLP) та машинного навчання.

Методами розроблення обрано метод аналізу (для дослідження існуючих підходів до аналізу тексту), метод синтезу (для поєднання переваг існуючих методів), методи моделювання (для представлення та дослідження процесів аналізу тексту), метод порівняльного аналізу (для оцінювання адекватності моделі).

Внаслідок виконання роботи обґрунтовано раціональний підхід до розроблення моделей аналізу текстових даних в електронній комерції та розроблено програмний засіб, який дозволяє створювати і досліджувати моделі аналізу тексту.

Результати дослідження можуть бути використані в науково-дослідних установах і підрозділах підприємств, що займаються розробленням моделей аналізу тексту.

Ключові слова: ЕЛЕКТРОННА КОМЕРЦІЯ, АНАЛІЗ ТЕКСТУ, ОБРОБКА ПРИРОДНОЇ МОВИ, МАШИННЕ НАВЧАННЯ, ТЕМАТИЧНЕ МОДЕЛЮВАННЯ.

ANNOTATION

Qualification work on the topic «Intelligent modeling of review topics in e-commerce» for Bachelor's degree on speciality 122 «Computer Science» educational and professional program «Computer Science» is written on 35 pages and it contains 8 figures, 5 tables, and 14 sources.

The purpose of the work is to develop and implement a comprehensive approach to analyzing textual data in e-commerce using natural language processing (NLP) methods and machine learning.

Research methods include analysis (to study existing approaches to text analysis), synthesis (to combine the advantages of existing methods), modeling (to represent and study text analysis processes), and comparative analysis (to evaluate the adequacy of the model).

As a result of the work, a rational approach to the development of text data analysis models in e-commerce was substantiated, and a software tool was developed that allows creating and researching text analysis models.

The research results can be used in research institutions and enterprise departments involved in the development of text analysis models.

Keywords: E-COMMERCE, TEXT ANALYSIS, NATURAL LANGUAGE PROCESSING, MACHINE LEARNING, TOPIC MODELING.

ЗМІСТ

ВСТУП

Актуальність теми дослідження аналізу текстових даних у сфері електронної комерції обумовлена стрімким розвитком цифрових технологій та постійним зростанням обсягів цифрового контенту, що генерується користувачами. В умовах високої конкуренції на ринку електронної комерції, здатність компаній ефективно аналізувати великі обсяги текстових відгуків клієнтів стає критично важливою для розуміння потреб споживачів, оптимізації асортименту товарів, та підвищення якості обслуговування. Такий аналіз дозволяє не тільки виявляти та реагувати на негативні відгуки, але й відкриває можливості для адаптації маркетингових стратегій та покращення репутації бренду.

З іншого боку, прогрес у галузі машинного навчання та обробки природної мови створює передумови для розвитку ефективних методів аналізу текстових даних, які можуть автоматизувати процес виявлення тем та сентименту у відгуках користувачів. Втім, попри наявність численних інструментів аналізу, багато компаній все ще стикаються з викликами, пов'язаними з обробкою неструктурованих текстових даних, що підкреслює потребу в подальшому дослідженні та розробці нових підходів і алгоритмів. Таким чином, тема дослідження є вкрай актуальною та має важливе значення для розвитку ефективних інструментів електронної комерції, спрямованих на оптимізацію взаємодії з клієнтами та підвищення конкурентоспроможності бізнесу.

Метою роботи є розробка та впровадження комплексного підходу до аналізу текстових даних в електронній комерції з використанням методів обробки природної мови (NLP) та машинного навчання. Цей підхід передбачає автоматизацію процесів збору, очищення, аналізу та інтерпретації великих обсягів текстових даних з метою ідентифікації ключових трендів, виявлення

настроїв користувачів, класифікації відгуків та екстракції корисної інформації для підтримки прийняття рішень в бізнесі.

Для досягнення поставленої мети було визначено та послідовно вирішено наступні завдання:

1. Вивчити ключові поняття та термінологію, які застосовуються в сфері електронної комерції.
2. Провести огляд сучасних підходів до аналізу текстових даних у електронній комерції.
3. Проаналізувати існуючі інструменти та алгоритми для попередньої обробки даних.
4. Розробити методику моделювання тем відгуків користувачів з використанням алгоритмів LDA, BERTopic та LSA.
5. Спроектувати та реалізувати алгоритм моделювання тем відгуків в електронній комерції.
6. Провести експериментальну перевірку розробленого алгоритму.
7. Використовувати комплекс метрик для оцінювання якості роботи моделі.
8. Розробити рекомендації щодо використання результатів аналізу текстових даних в стратегіях управління та маркетингу в електронній комерції.

Об'єктом дослідження є процес аналізу текстових даних в електронній комерції, включаючи збір, обробку, та інтерпретацію відгуків користувачів, а також розробку та застосування методів машинного навчання та обробки природної мови для класифікації текстів, виявлення емоційного забарвлення та тематичних категорій.

Предметом дослідження є методи та алгоритми машинного навчання та обробки природної мови, призначені для аналізу текстових даних у контексті електронної комерції. Це включає розгляд різних підходів до попередньої

обробки даних, моделювання тем, класифікації емоційних станів та визначення ключових характеристик продуктів чи послуг на основі відгуків користувачів.

Методи дослідження включають застосування статистичного аналізу, теорії інформації та алгоритмічних процедур обробки даних для виявлення тенденцій, закономірностей та структур великих обсягів текстових даних. Використовуються методи машинного навчання, зокрема навчання з підкріпленням, навчання без вчителя та навчання з вчителем, для класифікації текстових даних, визначення сентименту, та тематичного моделювання. Також застосовуються методи обробки природної мови (NLP) для попередньої обробки текстів, токенізації, лематизації, та видалення стоп-слів, що покращує якість та точність аналізу.

Практичне значення дослідження полягає в розробці та впровадженні ефективних інструментів для аналізу текстових даних в контексті електронної комерції, що дозволяє підвищити якість взаємодії з клієнтами та оптимізувати маркетингові стратегії.

Структура та обсяг кваліфікаційної роботи. Робота складається зі вступу, трьох розділів, висновків та списку використаних джерел. Загальний обсяг роботи становить 35 сторінок комп'ютерного тексту, включаючи 8 рисунків та 5 таблиць. Список використаних джерел містить 14 найменувань і займає 2 сторінки.

Апробація результатів дослідження. Основні теоретичні положення роботи й практичні результати дослідження доповідалися й обговорювалися V Всеукраїнської мультидисциплінарної студентської наукової конференції «Розвиток сучасної науки: актуальні питання теорії та практики», яка відбулася 19 квітня 2024 року у місті Тернопіль, Україна.

1 АНАЛІЗ ПРЕДМЕТНОЇ ОБЛАСТІ І ПОСТАНОВКА ЗАДАЧІ ДОСЛІДЖЕННЯ

1.1 Визначення ключових понять та термінології в електронній комерції

У сучасному світі електронної комерції важливим аспектом є розуміння ключових понять та термінології, що лежать в основі цієї галузі. Електронна комерція, або е-комерція, визначається як процес купівлі та продажу товарів або послуг через Інтернет. Цей тип комерції охоплює широкий спектр діяльності, включаючи онлайн-торгівлю, електронні платежі, онлайн-аукціони та інтернет-банкінг. Розвиток цифрових технологій та інтернету зробили електронну комерцію невід'ємною частиною глобальної економіки.

Для успішної діяльності в сфері електронної комерції важливо розуміти такі поняття, як "клієнтський досвід" (Customer Experience, CX), який включає всі аспекти взаємодії клієнта з компанією та її продуктами чи послугами. Це означає, що успіх у електронній комерції не лише про продаж товарів, але й про створення позитивного досвіду для користувачів, який спонукає їх повертатися.

Структура електронної комерції може бути розглянута через декілька ключових компонентів, що разом формують її основу:

1. Вебсайти та онлайн-магазини - це цифрові вітрини, де товари та послуги представлені для широкої аудиторії. Вони включають інформацію про продукт, ціни, описи, відгуки клієнтів та можливості покупки.
2. Платіжні системи - набір інструментів та сервісів, що забезпечують обробку онлайн-платежів. Включають кредитні та дебетові карти, електронні гаманці, банківські перекази та інші методи платежу.
3. Логістика та доставка - системи управління ланцюгом постачання та доставки товарів до кінцевого споживача. Включає у себе складську логістику, упаковку, відправлення та слідкування за статусом доставки.

4. Сервіси підтримки клієнтів - механізми забезпечення взаємодії з клієнтами, включаючи служби підтримки через електронну пошту, чати, гарячі лінії та FAQ-секції для допомоги та консультацій.
5. Маркетинг та реклама - стратегії та інструменти просування продуктів та послуг в інтернеті, включаючи пошукову оптимізацію (SEO), контекстну та банерну рекламу, email-маркетинг, SMM (маркетинг у соціальних медіа) та інші канали.
6. Аналітика даних - інструменти та методи збору, обробки та аналізу даних про поведінку користувачів, ефективність рекламних кампаній, продажі та інші ключові показники ефективності бізнесу.
7. Інтеграція з третіми сторонами - з'єднання електронної комерційної платформи з зовнішніми сервісами та інструментами, такими як CRM-системи, платформи автоматизації маркетингу, ERP-системи та інші.

Ці компоненти разом створюють екосистему електронної комерції, забезпечуючи компаніям можливість ефективно торгувати в інтернеті та надавати клієнтам високоякісні сервіси.

Ще одним ключовим поняттям є "оптимізація під пошукові системи" (Search Engine Optimization, SEO), що відіграє критичну роль у просуванні онлайн-бізнесу. SEO означає адаптацію вебсайту таким чином, щоб забезпечити йому високі позиції у результатах пошукових систем за релевантними запитами. Це допомагає залучити більше потенційних покупців та підвищує видимість бізнесу в інтернеті.

"Аналіз великих даних" (Big Data Analytics) стає все важливішим в електронній комерції, оскільки компанії збирають величезні обсяги даних про клієнтів та їх поведінку. Використання аналітичних інструментів для обробки цих даних дозволяє краще розуміти потреби клієнтів, оптимізувати маркетингові стратегії та покращити продукти чи послуги.

Наостанок, "маркетплейси" (Marketplaces) як Amazon або eBay, є центральними елементами в екосистемі електронної комерції, пропонуючи платформи, де продавці можуть пропонувати свої товари широкій аудиторії. Розуміння механізмів роботи цих платформ та стратегій взаємодії з ними є ключовим для успішної торгівлі в онлайн-просторі.

1.2 Огляд сучасних підходів до аналізу текстових даних у електронній комерції

Аналіз текстових даних у електронній комерції стає дедалі актуальнішим завдяки стрімкому зростанню обсягів інформації, що генерують користувачі. В останні роки значною мірою розширилися можливості збору та обробки великих обсягів невпорядкованих текстових даних, що дозволяє компаніям краще розуміти потреби та відгуки своїх клієнтів. Сучасні підходи до аналізу текстових даних зосереджуються на виявленні тенденцій, сентименту, тематичному моделюванні та розпізнаванні інтенцій користувачів, що відкриває нові можливості для покращення продукції, маркетингових стратегій та обслуговування клієнтів.

Технології обробки природної мови (NLP) і машинного навчання лягли в основу більшості сучасних методів аналізу тексту. Використання алгоритмів NLP дозволяє автоматично виявляти ключові слова, фрази та концепти, витягати корисну інформацію із тексту та аналізувати емоційне забарвлення (сентимент) відгуків. Методи машинного навчання, зокрема навчання з підкріпленням та глибоке навчання, забезпечують здатність систем адаптуватися до змінних умов і покращувати якість аналізу з часом, враховуючи нові дані.

Одним з ключових напрямків є сентимент-аналіз, який визначає емоційний відтінок тексту та допомагає розпізнати позитивні, негативні чи нейтральні відгуки. Це особливо цінно у контексті електронної комерції, де

відгуки клієнтів мають велике значення для репутації товарів і брендів. Автоматизація процесу аналізу сентименту дозволяє швидко обробляти великі обсяги даних і отримувати цінні інсайти щодо сприйняття продукції користувачами.

Тематичне моделювання є ще одним ефективним інструментом для аналізу текстових даних, який дозволяє ідентифікувати та класифікувати великі обсяги тексту за тематикою без необхідності попереднього маркування. Моделі, такі як Latent Dirichlet Allocation (LDA) та BERTopic, використовуються для виявлення прихованих тематичних структур у текстових даних, що допомагає краще зрозуміти контекст відгуків та виявити загальні тренди серед думок клієнтів.

Таблиця 1.1 представляє типи підходів до аналізу текстових даних у електронній комерції, їх перевагами та недоліками.

Таблиця 1.1 - Перевагами та недоліками підходів до аналізу текстових даних у електронній комерції

Тип підходу	**Переваги**	**Недоліки**
Частотний аналіз [1]	Простота у використанні та розумінні; Швидкість обробки даних	Не враховує контекст слова; Може ігнорувати семантику
Сентимент-аналіз [2, 3, 4, 5]	Можливість оцінки емоційного забарвлення текстів; Підходить для автоматизації відгуків клієнтів	Залежить від якості навчальних даних; Може бути упередженим
Класифікація тексту [6, 7]	Висока точність розпізнавання тематики тексту при наявності достатньої кількості даних для навчання	Потребує великої кількості ручно анотованих даних для навчання; Складності з іронією та двозначністю
Тематичне моделювання [8, 9, 10]	Автоматичне виявлення тем у великих текстових колекціях; Виявлення прихованих зв'язків між текстами	Висока складність моделей; Необхідність у великій кількості обчислювальних ресурсів
Екстракція ентитетів [11, 12]	Ідентифікація конкретних осіб, місць, організацій у текстах; Полегшення пошуку інформації	Складність у визначенні зв'язків між ентитетами; Може потребувати додаткової обробки для визначення зв'язків

Завдяки постійному розвитку технологій обробки природної мови та машинного навчання, аналіз текстових даних у електронній комерції стає все більш гнучким і точним. Це дозволяє компаніям не тільки виявляти та вирішувати проблеми у відносинах з клієнтами, але й прогнозувати майбутні тенденції, адаптуватися до змін у споживацьких перевагах та формувати більш ефективні стратегії взаємодії.

1.3 Вибір перспективного шляху і постановка задачі дослідження

Актуальність аналізу текстових даних в електронній комерції в сучасному цифровому світі неможливо переоцінити. Завдяки широкому поширенню інтернету та зростаючій доступності цифрових технологій, обсяги генерованих користувачами даних в інтернет-магазинах стрімко зростають. Відгуки клієнтів, описи товарів, форуми для споживачів – все це містить безцінну інформацію, яка може бути використана для покращення продуктів, сервісів, маркетингових стратегій та клієнтського досвіду. Але для того, щоб цінність цих даних була в повній мірі реалізована, необхідно ефективно їх аналізувати та інтерпретувати, що робить аналіз текстових даних ключовим інструментом для кожного онлайн-бізнесу.

З розвитком технологій штучного інтелекту та машинного навчання з'явилась можливість автоматизувати процеси збору, обробки та аналізу великих обсягів текстової інформації. Використання таких методів як обробка природної мови (NLP), семантичний аналіз, тематичне моделювання дозволяє виявляти тренди, визначати настрої споживачів, класифікувати відгуки за темами, та витягувати корисну інформацію, яка може слугувати основою для стратегічних бізнес-рішень. Такий підхід дозволяє компаніям бути більш гнучкими та адаптивними до змін на ринку, а також краще розуміти потреби та бажання своїх клієнтів.

На додаток до комерційної цінності, аналіз текстових даних має значний вплив на покращення користувацького досвіду. Здатність швидко виявляти та реагувати на негативні відгуки, визначати згадки про дефекти товарів або проблеми з обслуговуванням, надає компаніям можливість оперативно вирішувати проблеми та покращувати якість своїх послуг. Відповідно, інвестиції в розвиток інструментарію для аналізу текстових даних не тільки сприяють збільшенню прибутків шляхом оптимізації продуктового портфоліо та маркетингових стратегій, але й зміцнюють лояльність клієнтів за рахунок покращення якості обслуговування та відповідності товарів очікуванням споживачів.

Метою роботи є розробка та впровадження комплексного підходу до аналізу текстових даних в електронній комерції з використанням методів обробки природної мови (NLP) та машинного навчання. Цей підхід передбачає автоматизацію процесів збору, очищення, аналізу та інтерпретації великих обсягів текстових даних з метою ідентифікації ключових трендів, виявлення настроїв користувачів, класифікації відгуків та екстракції корисної інформації для підтримки прийняття рішень в бізнесі.

1 Для досягнення поставленої мети було визначено та послідовно вирішено наступні завдання:
2 Вивчити ключові поняття та термінологію, які застосовуються в сфері електронної комерції, для забезпечення правильного розуміння контексту дослідження.
3 Провести огляд сучасних підходів до аналізу текстових даних у електронній комерції, зокрема методів обробки природної мови та машинного навчання.
4 Проаналізувати існуючі інструменти та алгоритми для попередньої обробки даних, включаючи очищення, нормалізацію та структурування текстових даних.

5 Розробити методику моделювання тем відгуків користувачів з використанням алгоритмів LDA, BERTopic та LSA для виявлення основних тематичних кластерів.

6 Спроектувати та реалізувати алгоритм моделювання тем відгуків в електронній комерції, забезпечуючи його інтеграцію з сучасними технологіями обробки тексту.

7 Провести експериментальну перевірку розробленого алгоритму на реальних наборах даних з метою оцінки його ефективності та точності виявлення тем.

8 Використовувати комплекс метрик для оцінювання якості роботи моделі, включаючи точність, повноту, F1-середнє та інші показники для аналізу результативності моделювання.

9 Розробити рекомендації щодо використання результатів аналізу текстових даних в стратегіях управління та маркетингу в електронній комерції, а також для покращення взаємодії з клієнтами.

Завдання 1-3 мають бути розглянуті в теоретичних розділах дипломної роботи, а завдання 4-7 – у практичних розділах, що охоплюють розробку, оцінку та тестування алгоритму.

2 АЛГОРИТМІЧНЕ ТА ІНФОРМАЦІЙНЕ ЗАБЕЗПЕЧЕННЯ МОДЕЛЮВАННЯ ТЕМ ВІДГУКІВ В ЕЛЕКТРОННІЙ КОМЕРЦІЇ

2.1 Попередня обробка даних

Попередня обробка текстових даних є критичним етапом в обробці природної мови (NLP), який передує моделюванню тем або будь-якому іншому задуму аналізу тексту. Цей процес включає серію трансформацій, які призначені для нормалізації тексту, зменшення його розмірності та виділення суттєвих лінгвістичних одиниць. Метою попередньої обробки є перетворення "сирого" тексту в структуровану форму, що піддається аналізу, при цьому зберігаючи важливу інформацію та знижуючи вплив неінформативних елементів.

Неінформативні символи, такі як додаткові пробіли, табуляції або переноси рядків, можуть спотворити лінгвістичну структуру тексту та вплинути на якість подальшого аналізу. Видалення зайвих пробілів включає застосування регулярних виразів або вбудованих функцій мови програмування для нормалізації пробільних символів.

В контексті аналізу тексту, пропущені значення в полях, що містять текстові дані, можуть вплинути на результати аналізу. Заповнення пропущених значень порожніми рядками ("") дозволяє зберегти структурну цілісність датасету, спрощуючи обробку тексту на подальших етапах.

Нормалізація лексичних одиниць тексту, включаючи виправлення помилок друку та невідповідностей у назвах категорій, є важливою задачею для забезпечення консистентності даних. Це може включати мапінг синонімічних назв до єдиного стандарту або автоматизоване виправлення помилок за допомогою алгоритмів розпізнавання та корекції тексту.

Токенізація полягає в розбитті тексту на мінімальні семантичні одиниці (токени), зазвичай слова або фрази. Математично це можна представити як

перетворення текстової стрічки T на послідовність токенів $\{t_1, t_2, \ldots, t_n\}$. Видалення стоп-слів (слів, що не несуть значної семантичної ваги, як "і", "в", "на") дозволяє знизити розмірність текстових даних та зосередитись на важливих лексичних одиницях.

Лематизація - це процес перетворення слова на його базову, або лему, форму з урахуванням його морфологічних характеристик. Математично, лематизація кожного токена t_i в T може бути представлена як функція $L(t_i)$, що відображає t_i на відповідну лему l_i. Цей процес допомагає зменшити варіативність форм слова в тексті, спрощуючи подальший аналіз.

Загалом, попередня обробка тексту створює основу для ефективного тематичного моделювання та інших завдань NLP, зменшуючи вплив шуму в даних та підвищуючи якість отриманих моделей.

2.2 Моделювання тем

2.2.1 Модель LDA

Модель Latent Dirichlet Allocation (LDA) є байєсовською генеративною статистичною моделлю, яка дозволяє виявити латентні теми у великій колекції текстових документів. LDA припускає, що кожен документ може бути представлений як суміш тем, де кожна тема характеризується розподілом слів.

Нехай D - множина документів, W - множина слів у корпусі, і T - множина тем. Для кожного документа $d \in D$ та кожного слова $w \in W$ в документі, LDA моделює ймовірність $P(w|d)$ як:

$$P(w \mid d) = \sum_{t \in T} P(w \mid t) P(t \mid d)$$

де:

- $P(w|t)$ - ймовірність слова w з'явитися в темі t,
- $P(t|d)$ - ймовірність теми t в документі d.

Припущення:

1. Розподіл Діріхле для тем у документах: Для кожного документа d, розподіл тем $P(t|d)$ є випадковою величиною, що підкоряється розподілу Діріхле з параметром α, який є гіперпараметром моделі.

$$P(t \mid d) \sim Dirichlet(\alpha)$$

2. Розподіл Діріхле для слів у темах: Аналогічно, для кожної теми t, розподіл слів $P(w|t)$ підкоряється розподілу Діріхле з параметром β, що також є гіперпараметром моделі.

$$P(w \mid t) \sim Dirichlet(\beta)$$

Алгоритм:

1. Ініціалізація: Випадково ініціалізується розподіл тем для кожного слова в кожному документі на основі α та β.
2. Оновлення за Гіббса: Використовуючи зразкування за Гіббса, оновлюються присвоєння тем словам в документах, максимізуючи сумісну ймовірність моделі.

Після збігу моделі, $P(w|t)$ та $P(t|d)$ можуть бути використані для інтерпретації тем в документах. Теми можуть бути представлені через найбільш ймовірні слова, а документи - через найбільш ймовірні теми.

Модель LDA надає потужний інструмент для розуміння складної структури латентних тем у великих текстових колекціях. Через її байєсову природу та гнучкість у моделюванні, LDA знайшла застосування у широкому спектрі доменів від аналізу соціальних медіа до автоматичного резюмування текстів.

2.2.2 Модель BERTopic

Модель BERTopic є сучасним методом для виявлення тематики в текстових даних, який базується на використанні передових трансформерних

мереж, таких як BERT (Bidirectional Encoder Representations from Transformers). Відрізняючись від традиційних підходів тематичного моделювання, як-от LDA, BERTopic використовує глибоке навчання та векторні представлення слів (embeddings) для кращого збереження семантичних відношень між словами та підвищення якості визначення тем.

Основні концепції:

- Трансформерні мережі: BERT та інші трансформерні моделі здатні забезпечувати контекстно залежні векторні представлення для слів у великих текстових корпусах, враховуючи двонаправлені контексти (вліво та вправо від даного слова).
- Векторні представлення (Embeddings): Слова та документи представляються як вектори в багатовимірному просторі, де семантично схожі одиниці розташовані ближче одна до одної.

Нехай $w_1, w_2, \ldots, w_n$ - послідовність слів у документі, а $e(w_i)$ - векторне представлення слова w_i отримане за допомогою BERT. Тоді кожен документ D можна представити як вектор $E(D) = f(e(w_1), e(w_2), \ldots, e(w_n))$, де f - агрегуюча функція, що об'єднує вектори слів у вектор документа. Зазвичай для f використовують середнє значення або SVD (метод головних компонент) для зниження розмірності.

Після отримання векторних представлень для всіх документів використовуються методи зниження розмірності (наприклад, UMAP) та кластеризації (наприклад, HDBSCAN), щоб групувати документи зі схожим семантичним змістом.

Нехай $V = \{E(D_1), E(D_2), \ldots, E(D_m)\}$ - множина векторів документів. Застосування UMAP до V перетворює кожен вектор $E(D_i)$ в новий вектор $U(D_i)$ у просторі зниженої розмірності. Потім, використовуючи HDBSCAN на множині $\{U(D_1), U(D_2), \ldots, U(D_m)\}$, отримуємо кластери $\{C_1, C_2, \ldots, C_k\}$, де кожен кластер C_j відповідає певній темі в даних.

Кожен кластер C_j представляє тему, і для інтерпретації теми аналізуються найбільш репрезентативні слова та документи цього кластера. Репрезентативні слова можуть бути визначені через аналіз частотності слова в кластері порівняно з його частотності в інших кластерах.

Модель BERTopic відкриває нові можливості для тематичного моделювання, забезпечуючи глибше розуміння семантичних структур у текстових даних. Використання трансформерних моделей для генерації векторних представлень слів та документів разом із застосуванням сучасних методів зниження розмірності та кластеризації дозволяє ефективно виявляти та інтерпретувати латентні теми.

2.2.3 Модель LSA

Латентний семантичний аналіз (LSA) є методом зниження розмірності та виявлення латентної семантики у колекції текстових документів за допомогою статистичного моделювання. Центральним елементом LSA є декомпозиція матриці термін-документ (term-document matrix), що представляє відносини між словами (термінами) та документами у вигляді векторів у багатовимірному просторі.

Нехай A означає матрицю термін-документ розмірності $m \times n$, де m - кількість унікальних термінів у корпусі, а n - кількість документів. Елемент aij матриці A відображає вагу (частоту) терміну i у документі j.

LSA використовує сингулярний розклад (Singular Value Decomposition, SVD) матриці A, розкладаючи її на три матриці:

$$A = U\Sigma V^T$$

де:

- U - ортогональна матриця розмірності $m \times m$, стовпці якої (ліві сингулярні вектори) відображають терміни у семантичних розмірах,

- Σ - діагональна матриця розмірності $m \times n$, елементи діагоналі якої (сингулярні значення) представляють важливість кожного семантичного розміру,
- V^T - транспонована ортогональна матриця розмірності $n \times n$, стовпці якої (праві сингулярні вектори) відображають документи у семантичних розмірах.

Для зниження розмірності матриці A та виділення латентних семантичних розмірів, k найбільших сингулярних значень у матриці Σ зберігаються, а решта обнуляються. Це відповідає взяттю k-рангового наближення вихідної матриці A, що мінімізує помилку відтворення у нормі Фробеніуса:

$$A_k = U_k \Sigma_k V_k^T$$

де U_k, Σ_k і V_k^T - матриці, отримані в результаті видалення стовпців/рядків, що відповідають видаленим сингулярним значенням.

Латентні семантичні розміри, визначені стовпцями матриці U_k, представляють собою суміші термінів, що відображають ключові концепції або теми у корпусі. Аналогічно, документи, представлені стовпцями матриці V_k^T, розташовуються у просторі цих тем, дозволяючи оцінити їх семантичну близькість та відносини.

Латентний семантичний аналіз дозволяє розкрити приховану семантичну структуру у великих текстових корпусах, забезпечуючи інструмент для зниження розмірності, кластеризації та пошуку за семантикою. LSA застосовується у широкому спектрі задач NLP, включаючи пошук інформації, автоматичне резюмування та аналіз сентименту.

2.3 Алгоритм моделювання тем відгуків в електронній комерції

Далі розглянемо по-кроковий алгоритм (рис.2.1), що описує процес впровадження NLP тематичного моделювання для аналізу відгуків клієнтів в електронній комерції, зокрема для бренду жіночого одягу. Метою є витяг тематичних груп, пов'язаних з перевагами та проблемами клієнтів, такими як якість продукції, час доставки, невідповідності розміру та кольору.

По-кроковий алгоритм:

1. Завантаження даних
   1.1. Імпортуйте необхідні бібліотеки: numpy, pandas, matplotlib.pyplot.
   1.2. Завантажте набір даних з відгуками про жіночий електронний комерційний одяг за допомогою pandas.
   1.3. Відобразіть перші кілька рядків, щоб зрозуміти структуру даних.
2. Огляд даних
   2.1. Використовуйте метод .info(), щоб отримати огляд набору даних, включаючи назви стовпців та типи даних.
   2.2. Проаналізуйте розподіл відгуків за різними рейтингами, віком та кількістю позитивних відгуків, використовуючи гістограми.
3. Попередня обробка даних
   3.1. Очистіть і попередньо обробіть текстові дані:
   3.2. Видаліть зайві пробіли.
   3.3. Заповніть пропущені значення порожніми рядками.
   3.4. Виправте помилки друку та невідповідності в назвах категорій.
   3.5. Токенізуйте текст відгуків і видаліть стоп-слова.
   3.6. За бажанням, проведіть лематизацію для зведення слів до їх базових форм.
4. Моделювання тем
   4.1. Модель LDA
      4.1.1. Реалізуйте модель Latent Dirichlet Allocation (LDA) за допомогою бібліотек, таких як gensim.

4.1.2. Створіть словник і корпус, необхідні для LDA.

4.1.3. Навчіть модель LDA на корпусі.

4.1.4. Оцініть модель, використовуючи когерентність тем, щоб визначити оптимальну кількість тем.

4.2.Модель BERTopic

4.2.1. Встановіть та імпортуйте бібліотеку BERTopic.

4.2.2. Застосуйте модель BERTopic до очищених текстів відгуків.

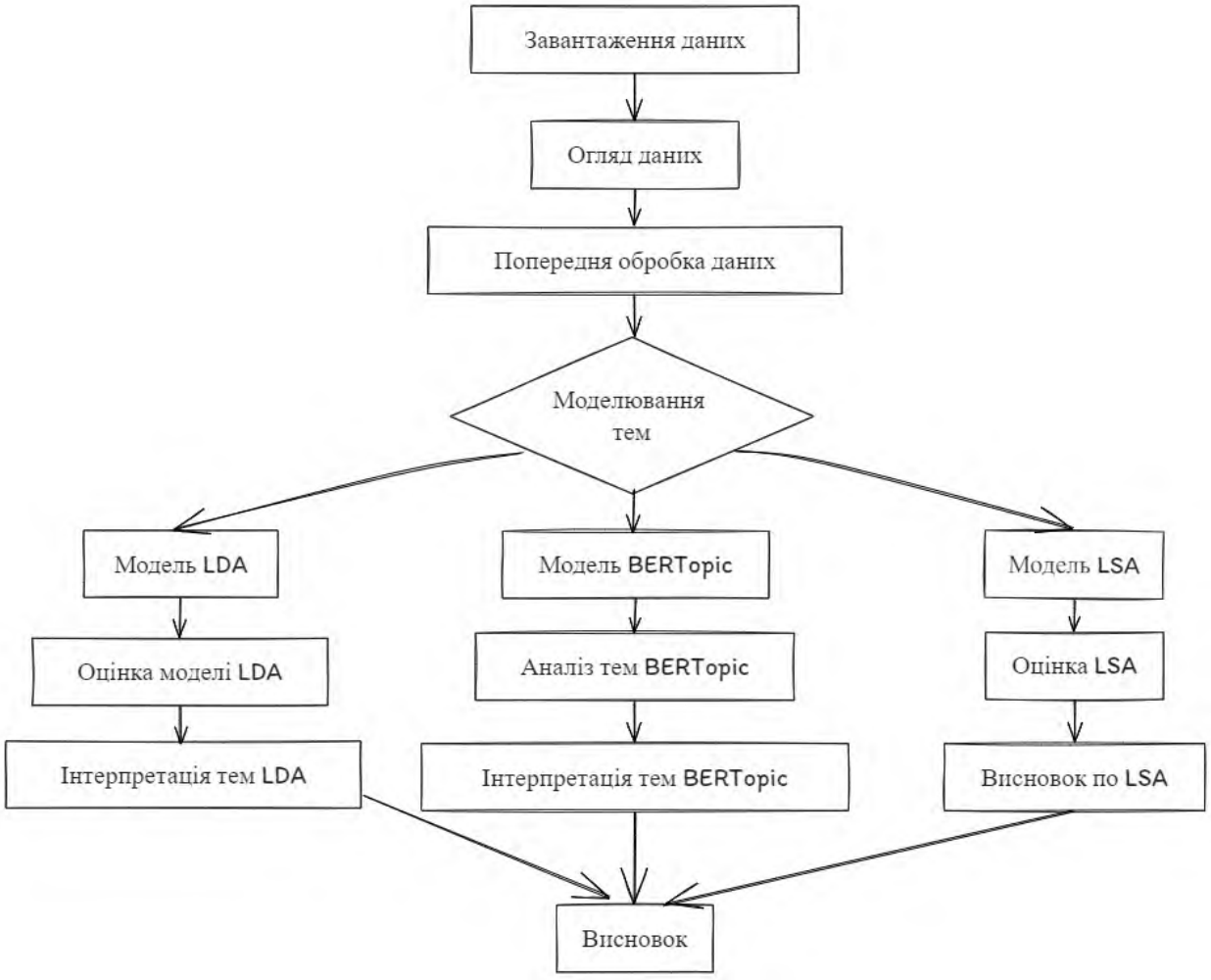


Рисунок 2.1 - Алгоритм моделювання тем відгуків в електронній комерції

4.2.3. Проаналізуйте теми для розуміння загальних тем у відгуках клієнтів.

4.3.Латентний семантичний аналіз

4.3.1. Використовуйте векторизацію TF-IDF для перетворення текстових даних на матрицю особливостей TF-IDF.

4.3.2. Застосуйте скорочений сингулярний розклад (Truncated SVD) для зниження розмірності.

4.3.3. Досліджуйте пояснювальну здатність, щоб вирішити кількість тем.

5. Висновок

Цей алгоритм надає всеосяжний підхід до аналізу відгуків клієнтів через тематичне моделювання, пропонуючи інсайти щодо особливостей продукції, які сприяють задоволенню клієнтів, і сфер для поліпшення.

3 ПРОГРАМНО-ТЕХНОЛОГІЧНЕ ЗАБЕЗПЕЧЕННЯ

3.1 Попередня обробка даних

Для розробленого (див.додаток А) алгоритму обрано набір даних (рис.3.1), що представляє собою колекцію відгуків клієнтів з жіночого електронного комерційного вебсайту одягу, який охоплює 23,000 відгуків і оцінок. Це реальні комерційні дані, які були анонімізовані; зокрема, згадки про компанію у тексті відгуків і заголовках були замінені на "рітейлер". Дані містять десять характеристик, включаючи унікальний ідентифікатор одягу, вік рецензента, заголовок та текст відгуку, рейтинг продукту, індикатор рекомендації продукту, кількість позитивних відгуків, а також категоріальні назви відділу, департаменту та класу продукту. Ці характеристики створюють сприятливе середовище для глибокого аналізу тексту відгуків на основі їхніх багатовимірних аспектів.

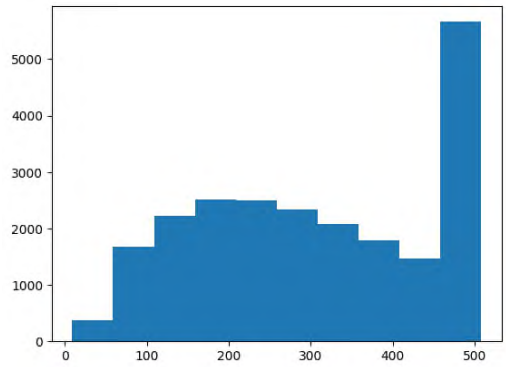
```
reviews_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 23486 entries, 0 to 23485
Data columns (total 11 columns):
 #   Column                   Non-Null Count  Dtype 
---  ------                   --------------  ----- 
 0   Unnamed: 0               23486 non-null  int64 
 1   Clothing ID              23486 non-null  int64 
 2   Age                      23486 non-null  int64 
 3   Title                    19676 non-null  object
 4   Review Text              22641 non-null  object
 5   Rating                   23486 non-null  int64 
 6   Recommended IND          23486 non-null  int64 
 7   Positive Feedback Count  23486 non-null  int64 
 8   Division Name            23472 non-null  object
 9   Department Name          23472 non-null  object
 10  Class Name               23472 non-null  object
dtypes: int64(6), object(5)
memory usage: 2.0+ MB
```

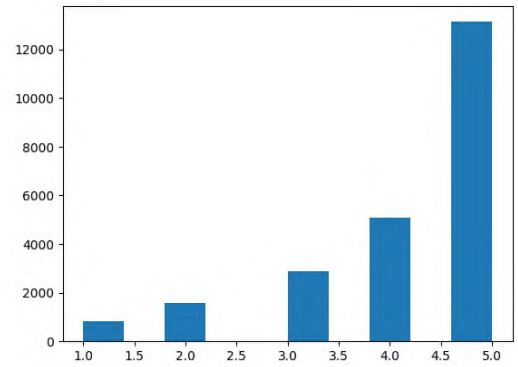
Рисунок 3.1 – Структура набору даних

На основі чотирьох гістограм, що відображають різні аспекти набору даних відгуків з жіночого електронного комерції одягу, можна зробити

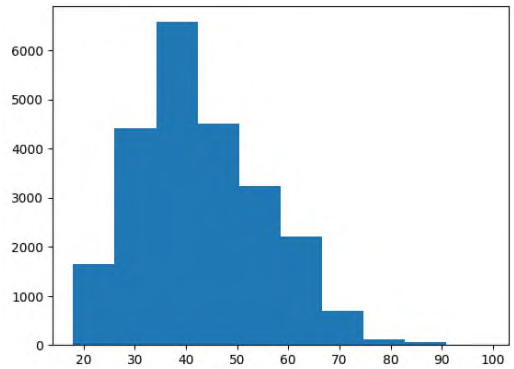
наступні спостереження. Перший рисунок (рис.3.2А) представляє гістограму довжини текстів відгуків, де видно, що більшість відгуків мають довжину близько 500 символів, з великим піком на правому краї, що вказує на обмеження довжини відгуку. На другому рисунку (рис.3.2Б) зображено розподіл рейтингів продуктів, з чітким домінуванням високих оцінок, особливо оцінки "5", що свідчить про загалом позитивний настрій відгуків. Третій рисунок (рис.3.2В) показує розподіл віку респондентів, що має форму дзвона та підкреслює основну демографічну групу клієнтів у віковому діапазоні від 30 до 40 років. Нарешті, четвертий рисунок (рис.3.2Г) демонструє розподіл кількості позитивних відгуків на кожен відгук, де більшість відгуків мають низький позитивний відгуковий рахунок, що вказує на те, що менше відгуків знаходять корисними велика кількість інших клієнтів.



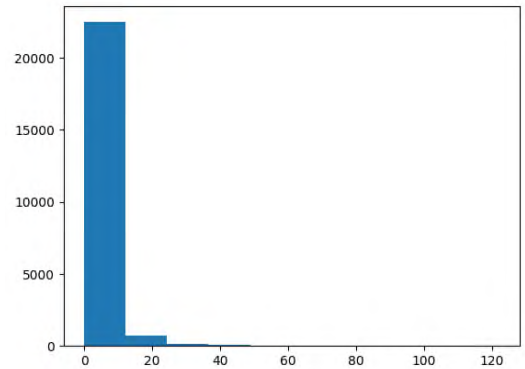
А. Гістограма довжини відгуків



Б. Гістограма рейтингів продуктів



В. Гістограма віку клієнтів



Г. Гістограма кількості позитивних відгуків

Рисунок 3.2 – Гістограми спостереження щодо відгуків клієнтів на жіночий одяг у електронній комерції

У процесі попередньої обробки даного набору даних з відгуками про жіночий одяг в електронній комерції, були ретельно очищені та нормалізовані текстові поля (рис.3.3). Це включало видалення зайвих пробілів, заповнення пропущених значень пустими рядками, корекцію помилок у назвах категорій та заміну слів для забезпечення консистентності. Потім було проведено токенізацію тексту відгуків з подальшим видаленням стоп-слів, використовуючи корпус Brown для визначення суттєвих термінів та відкиданням останніх слів, якщо вони не входили до цього корпусу. Це дозволило скоротити вплив несуттєвих токенів, які можуть спотворити семантичну вагу відгуків та завадити точному аналізу сентименту чи класифікації тем.

```
20081
[bought, dress, daughter, wear, rehearsal, dinner, could, n't, believe, beautifully, fit, ., fi
tted, top, ,, falling, gorgeous, flare, ., hit, knee, 's, 5, '', 6, ', ., classic, ,, elegant,
,, femme, fatale, look, work, range, occasions, .]
22163              [got, top, mail, today, ,, love, ., bought, white, small, ,, fits, picture
d, model, ., fabric, says, 's, cotton, ,, 's, substantial, fabric, feels, soft, silk, ., bodic
e, lined, ,, bra, lines, show, ., tiered, part, tank, lined, really, sheer, see-through, ., ar
m, holes, large, ,, sometimes, happens, wear, sleeveless, tops, ., think, 'll, wearing, one, lo
t, summer, ,, lot, summers]
1321
[great, top, !, !, love, fit, ,, pairs, great, dark, denim, pearls, !, !, ordered, 00, 32c, cu
p, 24-25, waist]
3112
[36dd, ,, 10/12, tops, ,, 12/14, bottoms, soft, warm, sweater, ,, boucle, body, almost, feels,
like, comfy, pajamas, !, lovely, heathered, pasted, colors, well, ., found, large, big, -, slee
ves, fit, fine, tighter, knit, ,, body, sagged, oddly, ., medium, best, fit, .]
8843     [review, titled, '', disappointed, '', could, n't, said, better, ., dresses, last, yea
r, perfect, ., probably, wear, least, week, ., excited, see, vibrant, blue, color, fully, expec
ted, quality, last, year, ., 'm, generally, xs, dresses, however, xxs, last, year, 's, version,
,, xs, one, lifeless, ., 's, heavy, ,, hot, ,, n't, nice, shape, last, year, ., may, cut, integ
rated, lining, lighten]
Name: Review Text, dtype: object
```

Рисунок 3.3 – Приклад очищених даних

Подальше лематизування (Рис.3.4) та обробка біграм (Рис.3.5) дозволили уточнити контекст та збагатити семантичне значення даних. Використовуючи лематизацію, кожне слово було зведено до своєї базової форми, що дозволяє краще зіставити семантично подібні терміни, не дивлячись на різноманітність їх морфологічних варіацій. Внаслідок цього, суттєві словосполучення (біграми), які з'являлись у тексті відгуків не менше

двадцяти разів, були додані до корпусу як окремі токени, що дозволяє підсилити аналіз та інтерпретацію тематичних зв'язків. Ця обробка створює міцну основу для подальшого моделювання тем та аналізу сентиментів, забезпечуючи більш чітке та глибоке розуміння взаємозв'язків у даних.

```
[('absolutely', 'RB'),
 ('wonderful', 'JJ'),
 ('-', ':'),
 ('silky', 'NN'),
 ('sexy', 'NN'),
 ('comfortable', 'JJ')]
```

Рисунок 3.4 – Лематизований текст

```
Number of unique tokens: 2451
Number of documents: 23486
```

Рисунок 3.5 – Кількість токенів

3.2 Моделювання тем відгуків в електронній комерції

3.2.1 Моделювання LDA

У рамках навчання моделі LDA для аналізу відгуків про жіночий одяг, була реалізована послідовність кроків для оптимізації кількості тем. Задано параметри навчання, включаючи розмір блоку даних, кількість проходів, ітерацій, та встановлено гіперпараметри моделі для автоматичного налаштування (α та β). Словникова база та корпус були створені з використанням попередньо обробленого тексту, відібраного на основі лематизації та видалення стоп-слів. Модель LDA була навчена з різною кількістю тем, в діапазоні від 2 до 11, де кожна модель була оцінена за когерентністю, що є показником її семантичної згуртованості.

Результати, зібрані в таблиці 3.1 когерентності, показують, що модель з сімома темами має найвищий рівень середньої когерентності, що свідчить про оптимальну збалансованість між кількістю тем та їх чіткістю (Coherence Score: 0.4969). Ці результати відображено на рисунку 3.5, який показує зв'язок між кількістю тем та когерентністю моделі.

Таблиця 3.1 - Когерентність тем моделі LDA в залежності від кількості тем

Num Topics	**Coherence Score**
2	0.412829
3	0.444483
4	0.469944
5	0.490289
6	0.485298
7	0.496909
8	0.481525
9	0.481995
10	0.478313
11	0.449600

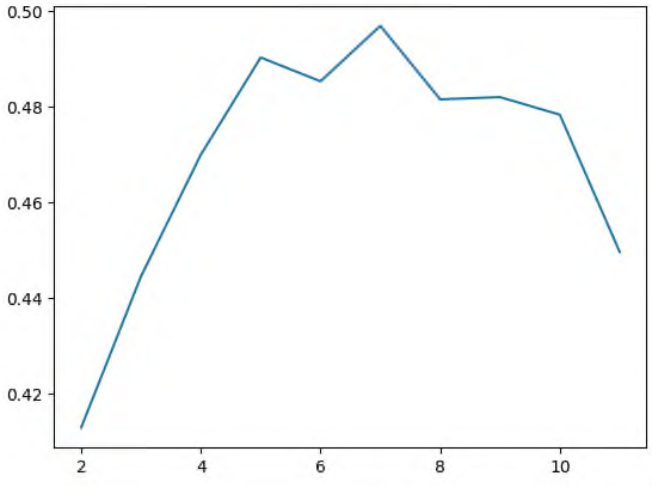


Рисунок 3.5 – Когерентність тем моделі LDA в залежності від кількості тем

Подальший детальний аналіз п'яти найкращих тем виявив їх ключові слова та вагу кожного слова у темі (табл 3.1). Наприклад, для теми з найвищим показником когерентності характерні такі слова як "довжина", "сукня", "талія", "пошиття" та інші, що відображають концептуальні особливості одягу. Наведені дані можуть бути корисними для розробників продукції та маркетологів для покращення стратегій виробництва та продажу продукції.

Таблиця 3.1 - Ключові слова та вагу кожного слова у темі

Topic	Word	Score	Coherence
1	!	0.086650	-1.825931
1	love	0.059440	-1.825931
1	wear	0.038725	-1.825931
1	great	0.036067	-1.825931
1	color	0.035360	-1.825931
2	look	0.082328	-2.124476
2	like	0.054783	-2.124476
2	would	0.042401	-2.124476
2	n't	0.031806	-2.124476
2	much	0.030735	-2.124476
3	top	0.056499	-2.219336
3	's	0.040548	-2.219336
3	nice	0.023566	-2.219336
3	little	0.022703	-2.219336
3	shirt	0.020255	-2.219336
4	size	0.087410	-2.291281
4	fit	0.055365	-2.291281
4	'm	0.041396	-2.291281
4	small	0.039109	-2.291281
4	order	0.036993	-2.291281
5	dress	0.150361	-2.369985
5	length	0.030628	-2.369985
5	petite	0.029935	-2.369985
5	waist	0.029432	-2.369985
5	skirt	0.027111	-2.369985

3.2.2 Моделювання BERTopic

Під час встановлення пакету BERTopic, що є розширенням для Python, яке забезпечує можливість моделювання тем на основі трансформаторів, спостерігається серія залежних завантажень та підготовка необхідних компонентів. Важливою частиною процесу є компіляція бібліотеки hdbscan та підготовка даних моделі sentence-transformers, яка включає різні конфігураційні файли та тренувальні скрипти. Завершальним етапом є ініціалізація моделі BERTopic та її застосування до набору текстових даних для виявлення тематичних структур з подальшим аналізом виразності та згуртованості обраних тем.

Застосування моделі BERTopic для аналізу текстових відгуків про жіночий одяг електронної комерції виявило, що вміст деяких тем містить неінформативні елементи (наприклад, тема 0 містить переважно пусті рядки з наближеним значенням 1e-05). Проте, аналіз конкретної теми (наприклад, тема 5) підкреслює зосередження на базових білих футболках (tee, tees, basic, white), з описом їхньої зручності, стилю та якості матеріалу. Аналіз репрезентативності документів у цій темі вказує на високу ймовірність приналежності відгуків до визначеної теми, що демонструється у таблиці "Document Info" для теми 5, де представлені зразки репрезентативних відгуків.

Таблиця 3.2 - Аномальний вихід моделі теми

Word	Value
0	0.00001
1	0.00001
2	0.00001
3	0.00001
4	0.00001
5	0.00001
6	0.00001
7	0.00001
8	0.00001
9	0.00001

В рамках подальшого дослідження, яке включає аналіз тем з допомогою моделі BERTopic (див.таблиця Б.1), було встановлено, що існує різноманітність сентиментів в рамках однієї теми. Деякі рецензенти високо оцінюють атрибути продукту, такі як пасування футболки, тоді як інші висловлюють критику. Для глибшого аналізу було запропоновано сегментувати відгуки за сентиментом перед класифікацією за темами, щоб визначити, які аспекти тем викликають позитивні чи негативні емоції серед користувачів, що може вказувати на різні спектри думок щодо даної теми. Ця інформація може бути корисною для вдосконалення продукції та маркетингових стратегій.

3.2.3 Моделювання LSA

В рамках дослідження Latent Semantic Analysis (LSA) застосовувалась методика сингулярного розкладу (SVD) для досягнення декомпозиції текстових даних відгуків споживачів на сайті електронної комерції з одягом для жінок. Цей процес дозволяє зменшити розмірність простору даних, зберігаючи при цьому основну інформацію про семантичні зв'язки між словами у текстах. Використання TfidfVectorizer забезпечило вагову оцінку важливості слова у документі відносно всієї колекції документів, дозволяючи сконцентруватися на значущих словах, що сприяє ефективнішій обробці даних перед застосуванням LSA.

При аналізі варіації об'ясненої дисперсії в залежності від кількості компонентів було встановлено, що навіть з великою кількістю компонентів пояснена дисперсія залишається низькою. Це підкреслює складність і багатогранність семантичного простору текстових відгуків та обмеженість LSA при роботі з дуже різноманітними і розрізненими даними. Графік (рис.3.6) "explained variance by n-components" ілюструє цей висновок, демонструючи, що для значного зростання пояснювальної здатності моделі необхідно використовувати велику кількість тем, що вимагає великих обчислювальних ресурсів та часу, роблячи LSA непрактичним для цього набору даних.

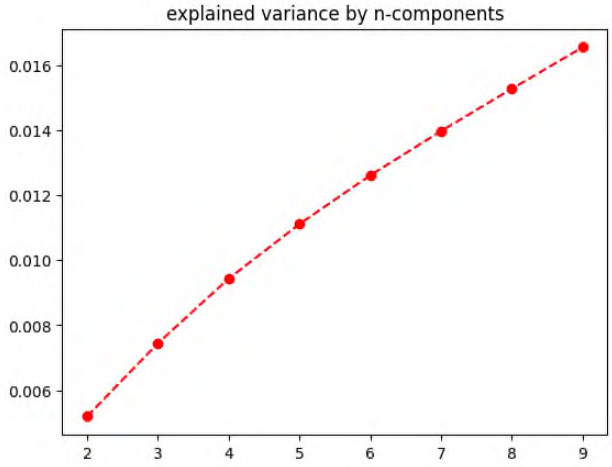


Рисунок 3.6 – Пояснена дисперсія за n-компонентами

3.3 Ефективність розробленого модуля

Отримані результати дослідження відгуків клієнтів про жіночий одяг в електронній комерції, використовуючи методики моделювання тем LDA, BERTopic, та аналізу LSA, відкривають нові горизонти для підтримки прийняття рішень у сфері маркетингу та розробки продукції. Аналіз виявив основні теми та сентименти, присутні у відгуках споживачів, що може слугувати важливою інформацією для розуміння переваг та недоліків продуктів з точки зору кінцевих користувачів. Це знання може бути використано для оптимізації асортименту продукції, покращення якості товарів та розробки більш цілеспрямованих маркетингових кампаній. Зокрема, зібрані дані про високу когерентність та популярність певних тем серед клієнтів можуть спонукати до впровадження нових дизайнів або модифікації існуючих продуктів, щоб краще задовольнити потреби ринку. Аналіз сентиментів дає змогу ідентифікувати ключові аспекти продуктів, які викликають позитивні або негативні емоції серед покупців, дозволяючи компаніям зосередити увагу на покращенні клієнтського досвіду.

Таблиця 3.3 відображає зібрані результати з ефективності різних моделей аналізу тем відгуків клієнтів.

Таблиця 3.3 - Ефективність різних моделей аналізу тем відгуків клієнтів

Модель	Середня когерентність	Об'яснена дисперсія	Репрезентативність документів	Загальна оцінка ефективності
LDA	0.4969	не застосовно	висока для теми 5	висока
BERTopic	не застосовно	не застосовно	змішана для теми 5	середня
LSA	не застосовно	низька для великої кількості компонентів	не застосовно	низька

Таблиця 3.3 наглядно демонструє, що модель LDA показала найвищу ефективність серед вибраних методів для аналізу відгуків, враховуючи її високу середню когерентність та здатність до репрезентативного відображення документів за темами. Натомість, BERTopic і LSA мають певні обмеження в контексті даного дослідження, особливо щодо об'ясненої дисперсії в LSA, що робить їх менш придатними для комплексного аналізу текстових даних.

Отже, використання подібних аналітичних методик може значно підвищити ефективність прийняття рішень у сфері роздрібної торгівлі одягом, сприяючи росту задоволеності клієнтів та збільшенню продажів.

ВИСНОВКИ

У світлі широкого спектру підходів до аналізу текстових даних в електронній комерції, зважаючи на їх переваги та недоліки, стає очевидним, що комплексне застосування цих методів може значно підвищити ефективність взаємодії компаній з клієнтами та оптимізувати маркетингові стратегії. Аналіз текстових даних дозволяє глибше зрозуміти потреби та переваги споживачів, адаптувати продукцію та послуги до вимог ринку, а також прогнозувати майбутні тенденції споживачів. Важливість цього аналізу не може бути переоцінена, особливо у контексті постійної еволюції цифрової економіки та змінних моделей споживання. Таким чином, глибоке розуміння та вміле застосування сучасних методів аналізу текстових даних стають ключовими факторами успіху в сфері електронної комерції.

Використання алгоритмічного та інформаційного забезпечення для моделювання тем відгуків в електронній комерції виявляється надзвичайно цінним для розуміння споживацької думки та оптимізації бізнес-стратегій. Попередня обробка даних, що забезпечує якісну підготовку тексту для аналізу, разом із застосуванням передових методів тематичного моделювання, таких як LDA, BERTopic, і LSA, дозволяє виявляти глибокі та релевантні інсайти з неструктурованих текстових даних. Цей підхід не тільки підвищує розуміння клієнтських потреб і переваг, але й сприяє розвитку продуктів і послуг, що краще відповідають очікуванням споживачів, водночас підвищуючи конкурентоспроможність бізнесу на ринку.

Реалізація програмно-технологічного забезпечення для аналізу тем відгуків в електронній комерції демонструє значний потенціал у збагаченні стратегій рішень та оптимізації взаємодії з клієнтами. Через детальну попередню обробку даних та застосування різноманітних методів моделювання, зокрема LDA, BERTopic, та LSA, стає можливим глибше

розуміння відгуків споживачів, виділення ключових тем та сентиментів. Такий підхід не лише відкриває шлях для вдосконалення продуктів і сервісів, але й надає компаніям цінні інсайти для формування більш ефективних маркетингових кампаній. Отримані результати підтверджують, що інтегрований аналіз текстових відгуків може стати ключовим фактором успіху в електронній комерції, сприяючи підвищенню конкурентоспроможності та задоволеності клієнтів.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. STHAL, Fabrice, and Enrico RUBIOLA. "Instrumentation Temps-Fréquence-Hautes fréquences et mesure de bruit de phase." (2010).
2. Vinodhini, G., & Chandrasekaran, R. M. (2012). Sentiment analysis and opinion mining: a survey. *International Journal*, *2*(6), 282-292.
3. Saberi, B., & Saad, S. (2017). Sentiment analysis or opinion mining: A review. *Int. J. Adv. Sci. Eng. Inf. Technol*, *7*(5), 1660-1666.
4. Liu, B. (2022). *Sentiment analysis and opinion mining*. Springer Nature.
5. Hutto, C., & Gilbert, E. (2014, May). Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the international AAAI conference on web and social media* (Vol. 8, No. 1, pp. 216-225).
6. Zhang, X., Zhao, J., & LeCun, Y. (2015). Character-level convolutional networks for text classification. *Advances in neural information processing systems*, *28*.
7. Joulin, A., Grave, E., Bojanowski, P., & Mikolov, T. (2016). Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.
8. Churchill, R., & Singh, L. (2022). The evolution of topic modeling. *ACM Computing Surveys*, *54*(10s), 1-35.
9. Kherwa, P., & Bansal, P. (2019). Topic modeling: a comprehensive review. *EAI Endorsed transactions on scalable information systems*, *7*(24).
10. Liu, L., Tang, L., Dong, W., Yao, S., & Zhou, W. (2016). An overview of topic modeling and its current applications in bioinformatics. *SpringerPlus*, *5*, 1-22.
11. Mansouri, A., Affendey, L. S., & Mamat, A. (2008). Named entity recognition approaches. *International Journal of Computer Science and Network Security*, *8*(2), 339-344.

12. Goyal, A., Gupta, V., & Kumar, M. (2018). Recent named entity recognition and classification techniques: a systematic review. *Computer Science Review*, *29*, 21-43.
13. Загальні методичні рекомендації з підготовки, оформлення, захисту та оцінювання кваліфікаційних робіт здобувачів вищої освіти першого (бакалаврського) і другого (магістерського) рівнів. Тернопіль: ЗУНУ, 2024. 83 с.
14. Комар М.П., Саченко А.О., Васильків Н.М., Гладій Г.М., Коваль В.С., Ліп'яніна-Гончаренко Х.В. Методичні рекомендації до виконання кваліфікаційної роботи з освітньо-професійної програми «Комп'ютерні науки» спеціальності 122 «Комп'ютерні науки» за першим (бакалаврським) рівнем вищої освіти. Тернопіль: ЗУНУ, 2024. 52 с.

ДОДАТОК А

КОД МОДЕЛЮВАННЯ ТЕМ ВІДГУКІВ В ЕЛЕКТРОННІЙ КОМЕРЦІЇ

```
# %% [code] {"execution":{"iopub.status.busy":"2023-12-07T06:43:57.97007Z","iopub.execute_input":"2023-12-07T06:43:57.970407Z","iopub.status.idle":"2023-12-07T06:43:58.238568Z","shell.execute_reply.started":"2023-12-07T06:43:57.970383Z","shell.execute_reply":"2023-12-07T06:43:58.237962Z"}}
import numpy as np
import pandas as pd # data processing, CSV file I/O (e.g. pd.read_csv)
import matplotlib.pyplot as plt

import os
# for dirname, _, filenames in os.walk('/kaggle/input'):
#     for filename in filenames:
#         print(os.path.join(dirname, filename))

# %% [code] {"execution":{"iopub.status.busy":"2023-12-06T07:43:23.983201Z","iopub.execute_input":"2023-12-06T07:43:23.983761Z","iopub.status.idle":"2023-12-06T07:43:23.99009Z","shell.execute_reply.started":"2023-12-06T07:43:23.983724Z","shell.execute_reply":"2023-12-06T07:43:23.988571Z"}}
# import logging
# import sys
# logger = logging.getLogger()
# logging.basicConfig(format='%(asctime)s : %(levelname)s : %(message)s', level=logging.ERROR)
# logger.addHandler(logging.StreamHandler(sys.stdout))

# %% [code] {"execution":{"iopub.status.busy":"2023-12-06T07:43:23.992513Z","iopub.execute_input":"2023-12-06T07:43:23.993311Z","iopub.status.idle":"2023-12-06T07:43:24.350965Z","shell.execute_reply.started":"2023-12-06T07:43:23.993253Z","shell.execute_reply":"2023-12-06T07:43:24.350043Z"}}
reviews_df = pd.read_csv('/kaggle/input/womens-ecommerce-clothing-reviews/Womens Clothing E-Commerce Reviews.csv')
reviews_df.head(5)

# %% [markdown]
# # Data Overview <a class="anchor" id="data-overview"></a>

# %% [code] {"execution":{"iopub.status.busy":"2023-12-06T07:43:24.352891Z","iopub.execute_input":"2023-12-06T07:43:24.353854Z","iopub.status.idle":"2023-12-06T07:43:24.40044Z","shell.execute_reply.started":"2023-12-06T07:43:24.353818Z","shell.execute_reply":"2023-12-06T07:43:24.399051Z"}}
reviews_df.info()

# %% [code] {"execution":{"iopub.status.busy":"2023-12-06T07:43:24.743315Z","iopub.execute_input":"2023-12-06T07:43:24.744128Z","iopub.status.idle":"2023-12-06T07:43:24.765508Z","shell.execute_reply.started":"2023-12-06T07:43:24.744093Z","shell.execute_reply":"2023-12-06T07:43:24.764345Z"}}
reviews_df['Title'].value_counts()

# %% [code] {"execution":{"iopub.status.busy":"2023-12-06T07:43:29.395297Z","iopub.execute_input":"2023-12-06T07:43:29.395734Z","iopub.status.idle":"2023-12-06T07:43:29.406025Z","shell.execute_reply.started":"2023-12-06T07:43:29.395704Z","shell.execute_reply":"2023-12-06T07:43:29.404985Z"}}
pd.options.display.max_colwidth = None
reviews_df['Review Text'].sample(1)

# %% [markdown]
# We found that text would get cut off if the character length exceeded 500.

# %% [code] {"execution":{"iopub.status.busy":"2023-12-06T07:48:38.648965Z","iopub.execute_input":"2023-12-06T07:48:38.649436Z","iopub.status.idle":"2023-12-06T07:48:39.452884Z","shell.execute_reply.started":"2023-12-06T07:48:38.649403Z","shell.execute_reply":"2023-12-06T07:48:39.451579Z"}}
plt.hist(reviews_df['Review Text'].str.len())
plt.show()

# %% [code] {"execution":{"iopub.status.busy":"2023-12-06T07:43:32.475562Z","iopub.execute_input":"2023-12-06T07:43:32.476163Z","iopub.status.idle":"2023-12-06T07:43:32.510465Z","shell.execute_reply.started":"2023-12-06T07:43:32.476122Z","shell.execute_reply":"2023-12-06T07:43:32.509147Z"}}
reviews_df[reviews_df['Review Text'].str.len() == 500]['Review Text']
```

```
# %% [code] {"execution":{"iopub.status.busy":"2023-12-06T07:47:21.843381Z","iopub.execute_input":"2023-12-06T07:47:21.843791Z","iopub.status.idle":"2023-12-06T07:47:22.070777Z","shell.execute_reply.started":"2023-12-06T07:47:21.843752Z","shell.execute_reply":"2023-12-06T07:47:22.069329Z"}}
plt.hist(reviews_df['Rating'])
plt.show()

# %% [code] {"execution":{"iopub.status.busy":"2023-12-06T07:47:22.072793Z","iopub.execute_input":"2023-12-06T07:47:22.073175Z","iopub.status.idle":"2023-12-06T07:47:22.30158Z","shell.execute_reply.started":"2023-12-06T07:47:22.073145Z","shell.execute_reply":"2023-12-06T07:47:22.300443Z"}}
plt.hist(reviews_df['Age'])
plt.show()

# %% [code] {"execution":{"iopub.status.busy":"2023-12-06T07:47:31.601738Z","iopub.execute_input":"2023-12-06T07:47:31.602172Z","iopub.status.idle":"2023-12-06T07:47:31.863476Z","shell.execute_reply.started":"2023-12-06T07:47:31.602135Z","shell.execute_reply":"2023-12-06T07:47:31.862331Z"}}
plt.hist(reviews_df['Positive Feedback Count'])
plt.show()

# %% [markdown]
# # Data Preprocess <a class="anchor" id="data-preprocess"></a>

# %% [code] {"execution":{"iopub.status.busy":"2023-12-06T05:26:06.363727Z","iopub.execute_input":"2023-12-06T05:26:06.364298Z","iopub.status.idle":"2023-12-06T05:26:06.436102Z","shell.execute_reply.started":"2023-12-06T05:26:06.36426Z","shell.execute_reply":"2023-12-06T05:26:06.434867Z"}}
reviews_df['Review Text'] = reviews_df['Review Text'].str.strip()
reviews_df['Title'] = reviews_df['Title'].str.strip()
reviews_df['Division Name'] = reviews_df['Division Name'].str.strip()
reviews_df['Department Name']= reviews_df['Department Name'].str.strip()
reviews_df['Class Name']= reviews_df['Class Name'].str.strip()

reviews_df['Review Text'] = reviews_df['Review Text'].fillna('')
reviews_df['Title'] = reviews_df['Title'].fillna('')
reviews_df['Division Name'] = reviews_df['Division Name'].fillna('')
reviews_df['Department Name']= reviews_df['Department Name'].fillna('')
reviews_df['Class Name']= reviews_df['Class Name'].fillna('')

# Replace typo
reviews_df['Division Name'] = reviews_df['Division Name'].replace({'Initmates': 'Intimates'})

# Replace for consistency between division, department, and class names
reviews_df['Department Name'] = reviews_df['Department Name'].replace({'Intimate': 'Intimates'})



# %% [code] {"execution":{"iopub.status.busy":"2023-12-06T05:26:06.439542Z","iopub.execute_input":"2023-12-06T05:26:06.440004Z","iopub.status.idle":"2023-12-06T05:26:12.325341Z","shell.execute_reply.started":"2023-12-06T05:26:06.439961Z","shell.execute_reply":"2023-12-06T05:26:12.324175Z"}}
from nltk.tokenize import word_tokenize
import nltk
from nltk.corpus import stopwords
nltk.download('stopwords')
words = set(nltk.corpus.brown.words())

# %% [markdown]
# We remove parentheses from our text and remove the last word from the text if it doesn't belong in the Brown corpus. This accounts for reviews that were cut off for exceeding the character limit.

# %% [code] {"execution":{"iopub.status.busy":"2023-12-06T05:26:12.327027Z","iopub.execute_input":"2023-12-06T05:26:12.327409Z","iopub.status.idle":"2023-12-06T05:29:34.013539Z","shell.execute_reply.started":"2023-12-06T05:26:12.327379Z","shell.execute_reply":"2023-12-06T05:29:34.012585Z"}}
reviews_text = reviews_df['Review Text'].copy()
for idx in range(len(reviews_text)):
    reviews_text[idx] = reviews_text[idx].lower().replace('(','').replace(')','') # Convert to lowercase and remove parentheses.
    reviews_text[idx] = word_tokenize(reviews_text[idx]) # Split into words.
    filtered_sentence = [w for w in reviews_text[idx] if not w.lower() in stopwords.words('english')]
    if filtered_sentence and filtered_sentence[-1] not in words:
        reviews_text[idx] = filtered_sentence[:-1]
    else:
        reviews_text[idx] = filtered_sentence
```

```
# %% [code] {"execution":{"iopub.status.busy":"2023-12-06T05:29:34.014757Z","iopub.execute_input":"2023-12-06T05:29:34.015526Z","iopub.status.idle":"2023-12-06T05:29:34.025678Z","shell.execute_reply.started":"2023-12-06T05:29:34.015494Z","shell.execute_reply":"2023-12-06T05:29:34.024677Z"}}
reviews_text.sample(5)

# %% [markdown]
# ### Lemmatize

# %% [markdown]
# **Lemmatizing vs Stemming**
#
# **Stemming** reduces a word to its stem form by removing the last few letters of the word. Requires shorter runtime.
# Ex: Caring -> Car
#
# **Lemmatizing** reduces a word to a meaningful base form by considering the word's context and shortening it into a *lemma*. More computationally expensive.
# Ex: Caring -> Care

# %% [code] {"execution":{"iopub.status.busy":"2023-12-06T05:29:34.028947Z","iopub.execute_input":"2023-12-06T05:29:34.029372Z","iopub.status.idle":"2023-12-06T05:29:35.487002Z","shell.execute_reply.started":"2023-12-06T05:29:34.029333Z","shell.execute_reply":"2023-12-06T05:29:35.485589Z"}}
import nltk
nltk.download('wordnet')
!unzip -o /usr/share/nltk_data/corpora/wordnet.zip -d /usr/share/nltk_data/corpora/

# %% [code] {"execution":{"iopub.status.busy":"2023-12-06T05:29:35.488818Z","iopub.execute_input":"2023-12-06T05:29:35.489654Z","iopub.status.idle":"2023-12-06T05:30:27.655877Z","shell.execute_reply.started":"2023-12-06T05:29:35.489618Z","shell.execute_reply":"2023-12-06T05:30:27.654754Z"}}
from nltk.stem.wordnet import WordNetLemmatizer
from nltk import pos_tag
lemmatizer = WordNetLemmatizer()
pos_reviews = [pos_tag(text) for text in reviews_text]
pos_reviews[0]

# %% [code] {"execution":{"iopub.status.busy":"2023-12-06T05:30:27.657392Z","iopub.execute_input":"2023-12-06T05:30:27.657748Z","iopub.status.idle":"2023-12-06T05:30:32.132744Z","shell.execute_reply.started":"2023-12-06T05:30:27.657718Z","shell.execute_reply":"2023-12-06T05:30:32.131345Z"}}
def lemmatize_pos_word(word, tag):
    wntag = tag[0].lower()
    wntag = wntag if wntag in ['a', 'r', 'n', 'v'] else None
    if not wntag:
        lemma = word
    else:
        lemma = lemmatizer.lemmatize(word, wntag)
    return lemma

text = [[lemmatize_pos_word(word, tag) for word, tag in review] for review in pos_reviews]

# %% [code] {"execution":{"iopub.status.busy":"2023-12-06T05:30:32.134136Z","iopub.execute_input":"2023-12-06T05:30:32.134464Z","iopub.status.idle":"2023-12-06T05:30:44.782327Z","shell.execute_reply.started":"2023-12-06T05:30:32.134435Z","shell.execute_reply":"2023-12-06T05:30:44.781029Z"}}
# Compute bigrams.
from gensim.models import Phrases

# Add bigrams and trigrams to docs (only ones that appear 20 times or more).
bigram = Phrases(text, min_count=20)
for idx in range(len(text)):
    for token in bigram[text[idx]]:
        if '_' in token:
            # Token is a bigram, add to document.
            text[idx].append(token)

# %% [code] {"execution":{"iopub.status.busy":"2023-12-06T05:30:44.794974Z","iopub.execute_input":"2023-12-06T05:30:44.79534Z","iopub.status.idle":"2023-12-06T05:30:45.93237Z","shell.execute_reply.started":"2023-12-06T05:30:44.795309Z","shell.execute_reply":"2023-12-06T05:30:45.930976Z"}}
from gensim.corpora import Dictionary

# Create a dictionary representation of the documents.
```

```
dictionary = Dictionary(text)

# Filter out words that occur less than 20 documents, or more than 50% of the documents.
dictionary.filter_extremes(no_below=20, no_above=0.5)

# %% [code] {"execution":{"iopub.status.busy":"2023-12-06T05:30:45.934155Z","iopub.execute_input":"2023-12-06T05:30:45.934499Z","iopub.status.idle":"2023-12-06T05:30:46.593369Z","shell.execute_reply.started":"2023-12-06T05:30:45.934468Z","shell.execute_reply":"2023-12-06T05:30:46.592204Z"}}
corpus = [dictionary.doc2bow(doc) for doc in text]

# %% [code] {"execution":{"iopub.status.busy":"2023-12-06T05:30:46.594574Z","iopub.execute_input":"2023-12-06T05:30:46.596896Z","iopub.status.idle":"2023-12-06T05:30:46.602938Z","shell.execute_reply.started":"2023-12-06T05:30:46.596865Z","shell.execute_reply":"2023-12-06T05:30:46.601817Z"}}
print('Number of unique tokens: %d' % len(dictionary))
print('Number of documents: %d' % len(corpus))

# %% [markdown]
# # LDA Model <a class="anchor" id="lda"></a>

# %% [code] {"execution":{"iopub.status.busy":"2023-12-05T22:37:45.636208Z","iopub.execute_input":"2023-12-05T22:37:45.636832Z","iopub.status.idle":"2023-12-05T22:37:45.652634Z","shell.execute_reply.started":"2023-12-05T22:37:45.6368Z","shell.execute_reply":"2023-12-05T22:37:45.651236Z"}}
# Train LDA model.
from gensim.models import LdaModel
from pprint import pprint
def create_lda_model(num_topics = 10):
    # Set training parameters.
    chunksize = 2000
    passes = 20
    iterations = 400
    eval_every = None  # Don't evaluate model perplexity, takes too much time.

    # Make an index to word dictionary.
    temp = dictionary[0]  # This is only to "load" the dictionary.
    id2word = dictionary.id2token

    model = LdaModel(
        corpus=corpus,
        id2word=id2word,
        chunksize=chunksize,
        alpha='auto',
        eta='auto',
        iterations=iterations,
        num_topics=num_topics,
        passes=passes,
        eval_every=eval_every,
        random_state  = 1
    )
    return model

def get_average_topic_coherence(model):
    top_topics = model.top_topics(corpus)
    # Average topic coherence is the sum of topic coherences of all topics, divided by the number of topics.
    avg_topic_coherence = sum([t[1] for t in top_topics]) / model.num_topics
    return avg_topic_coherence

def get_top_topics(model):
    top_topics = model.top_topics(corpus)
    # Average topic coherence is the sum of topic coherences of all topics, divided by the number of topics.
    return top_topics

# %% [code] {"execution":{"iopub.status.busy":"2023-12-05T22:37:45.654094Z","iopub.execute_input":"2023-12-05T22:37:45.654548Z","iopub.status.idle":"2023-12-05T23:16:21.051535Z","shell.execute_reply.started":"2023-12-05T22:37:45.654493Z","shell.execute_reply":"2023-12-05T23:16:21.049232Z"}}
from gensim.models import CoherenceModel
topic_coherences = []
models = []
for num_topics in range(2, 12):
```

```
    model = create_lda_model(num_topics = num_topics)
    average_topic_coherence = get_average_topic_coherence(model)
    print(f"Num topics: {num_topics}")
    coherence_model_lda = CoherenceModel(
        model=model, texts=text, dictionary=dictionary, coherence='c_v')
    coherence_lda = coherence_model_lda.get_coherence()
    print('Coherence Score: ', coherence_lda)
    topic_coherences.append(coherence_lda)
    models.append(model)

# %% [code] {"execution":{"iopub.status.busy":"2023-12-05T23:16:21.054263Z","iopub.execute_input":"2023-12-
05T23:16:21.055245Z","iopub.status.idle":"2023-12-05T23:16:21.367356Z","shell.execute_reply.started":"2023-12-
05T23:16:21.055192Z","shell.execute_reply":"2023-12-05T23:16:21.36605Z"}}
plt.plot(list(range(2, 12)), topic_coherences)

# %% [code] {"execution":{"iopub.status.busy":"2023-12-05T23:16:21.368884Z","iopub.execute_input":"2023-12-
05T23:16:21.369287Z","iopub.status.idle":"2023-12-05T23:16:21.926507Z","shell.execute_reply.started":"2023-12-
05T23:16:21.369254Z","shell.execute_reply":"2023-12-05T23:16:21.92519Z"}}
pprint(get_top_topics(models[5]))

# %% [markdown]
# #### Interpretation
#
# From the coherence scores, we see that performing LDA with 7 topics gives the best results. We can interpret
the topics as follows:
#
# Topic 1: love, wear, great, perfect
# Topic 2: good, quality, picture, look
# Topic 3: nice, top, shirt, fabric
# Topic 4: size, fit (talks about differents fits)
# Topic 5: dress, length, petite, waist
# Topic 6: not, one, get
# Topic 7: saw, store, can't, wait
#
# From this, we can see mentions of fit and material, particularly about the shirts and dresses.
# However, this can suggest a lack of emphasis on other catalog items like bottoms. With this, we can look in
depth into what is and isn't being mentioned more commonly with bottoms that we feel we can improve on.
#
# We also note that improving this dataset with sentiment annotations would allow us to extract opinions
regarding each of these topics, so we can see if there is more positive or negative sentiment regarding the
quality or size of products.

# %% [markdown]
# # BERTopic Model <a class="anchor" id="bertopic"></a>

# %% [code] {"execution":{"iopub.status.busy":"2023-12-06T05:30:46.604394Z","iopub.execute_input":"2023-
12-06T05:30:46.604801Z","iopub.status.idle":"2023-12-
06T05:32:10.738049Z","shell.execute_reply.started":"2023-12-
06T05:30:46.604774Z","shell.execute_reply":"2023-12-06T05:32:10.736786Z"}}
!pip install bertopic

# %% [code] {"execution":{"iopub.status.busy":"2023-12-06T05:32:10.74271Z","iopub.execute_input":"2023-12-
06T05:32:10.743548Z","iopub.status.idle":"2023-12-06T05:32:45.113007Z","shell.execute_reply.started":"2023-
12-06T05:32:10.743511Z","shell.execute_reply":"2023-12-06T05:32:45.111925Z"}}
from bertopic import BERTopic

# %% [code] {"execution":{"iopub.status.busy":"2023-12-06T05:32:45.114172Z","iopub.execute_input":"2023-12-
06T05:32:45.115252Z","iopub.status.idle":"2023-12-06T05:40:45.509896Z","shell.execute_reply.started":"2023-
12-06T05:32:45.115209Z","shell.execute_reply":"2023-12-06T05:40:45.508486Z"}}
topic_model = BERTopic()
topics, probs = topic_model.fit_transform(reviews_df['Review Text'])

# %% [code] {"execution":{"iopub.status.busy":"2023-12-06T05:40:45.511727Z","iopub.execute_input":"2023-12-
06T05:40:45.512583Z","iopub.status.idle":"2023-12-06T05:40:45.538332Z","shell.execute_reply.started":"2023-
12-06T05:40:45.512513Z","shell.execute_reply":"2023-12-06T05:40:45.537031Z"}}
topic_model.get_topic_info()[0:10]

# %% [code] {"execution":{"iopub.status.busy":"2023-12-06T05:40:45.540312Z","iopub.execute_input":"2023-12-
06T05:40:45.54079Z","iopub.status.idle":"2023-12-06T05:40:45.551984Z","shell.execute_reply.started":"2023-
12-06T05:40:45.540751Z","shell.execute_reply":"2023-12-06T05:40:45.55077Z"}}
```

```
topic_model.get_topic(0)

# %% [code] {"execution":{"iopub.status.busy":"2023-12-06T05:40:45.553984Z","iopub.execute_input":"2023-12-06T05:40:45.554448Z","iopub.status.idle":"2023-12-06T05:40:45.627225Z","shell.execute_reply.started":"2023-12-06T05:40:45.554405Z","shell.execute_reply":"2023-12-06T05:40:45.626123Z"}}
doc_info = topic_model.get_document_info(reviews_df['Review Text'])
doc_info[doc_info['Topic'] == 5]

# %% [markdown]
# We address this with LDA too, but one thing to note is that within each topic, there are different sentiments regarding the topic. Some reviews praise topics like the shirt's fit, while others will criticize the fit. As a prior step, we can gauge sentiment and separate positive sentiment into one dataset and negative sentiment into another, which would allow us to see which topics are received more positively or more negatively, given there will be different spectrums of thought among a given topic.

# %% [markdown]
# # Latent Semantic Analysis <a class="anchor" id="lsa"></a>
#
# For LSA, we apply SVD to achieve a decomposition of our data, where our number of singular values would equal the number of topics we predesignate.

# %% [code] {"execution":{"iopub.status.busy":"2023-11-29T05:06:32.885559Z","iopub.execute_input":"2023-11-29T05:06:32.886369Z","iopub.status.idle":"2023-11-29T05:06:44.71567Z","shell.execute_reply.started":"2023-11-29T05:06:32.886328Z","shell.execute_reply":"2023-11-29T05:06:44.714251Z"}}
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.decomposition import TruncatedSVD
from sklearn.pipeline import Pipeline
from sklearn.model_selection import train_test_split

X = reviews_df['Review Text'].copy().values
# for idx in range(len(X)):
#     X[idx] = X[idx].lower()
vectorizer = TfidfVectorizer(stop_words = stopwords.words('english'),
                             ngram_range = (1,2),
                             lowercase = False,
                             tokenizer = word_tokenize
                            )
train_X = train_test_split(X, train_size = 0.8)[0]
train_X_tfidf = vectorizer.fit_transform(train_X)

# %% [code] {"execution":{"iopub.status.busy":"2023-11-29T06:46:15.473316Z","iopub.execute_input":"2023-11-29T06:46:15.473715Z","iopub.status.idle":"2023-11-29T07:04:24.243377Z","shell.execute_reply.started":"2023-11-29T06:46:15.473684Z","shell.execute_reply":"2023-11-29T07:04:24.241187Z"}}
performance = []
compenents = range(2, 10, )
for n in compenents:
    svd = TruncatedSVD(n_components=n, n_iter=100, random_state=0)
    lsa = svd.fit(train_X_tfidf)
    performance.append(lsa.explained_variance_ratio_.sum())
plt.plot(list(compenents), performance, 'ro--')
plt.title('explained variance by n-components');

# %% [markdown]
# Even with SVD, the explained variance would be too small unless the number of components was very large. Since a large number of components would take too long to compute, we decide that LSA would not be reasonable to use here.
```

ДОДАТОК Б

РЕЗУЛЬТАТИ МОДЕЛЮВАННЯ BERTOPIC

Таблиця Б.1 - Результати моделювання BERTopic

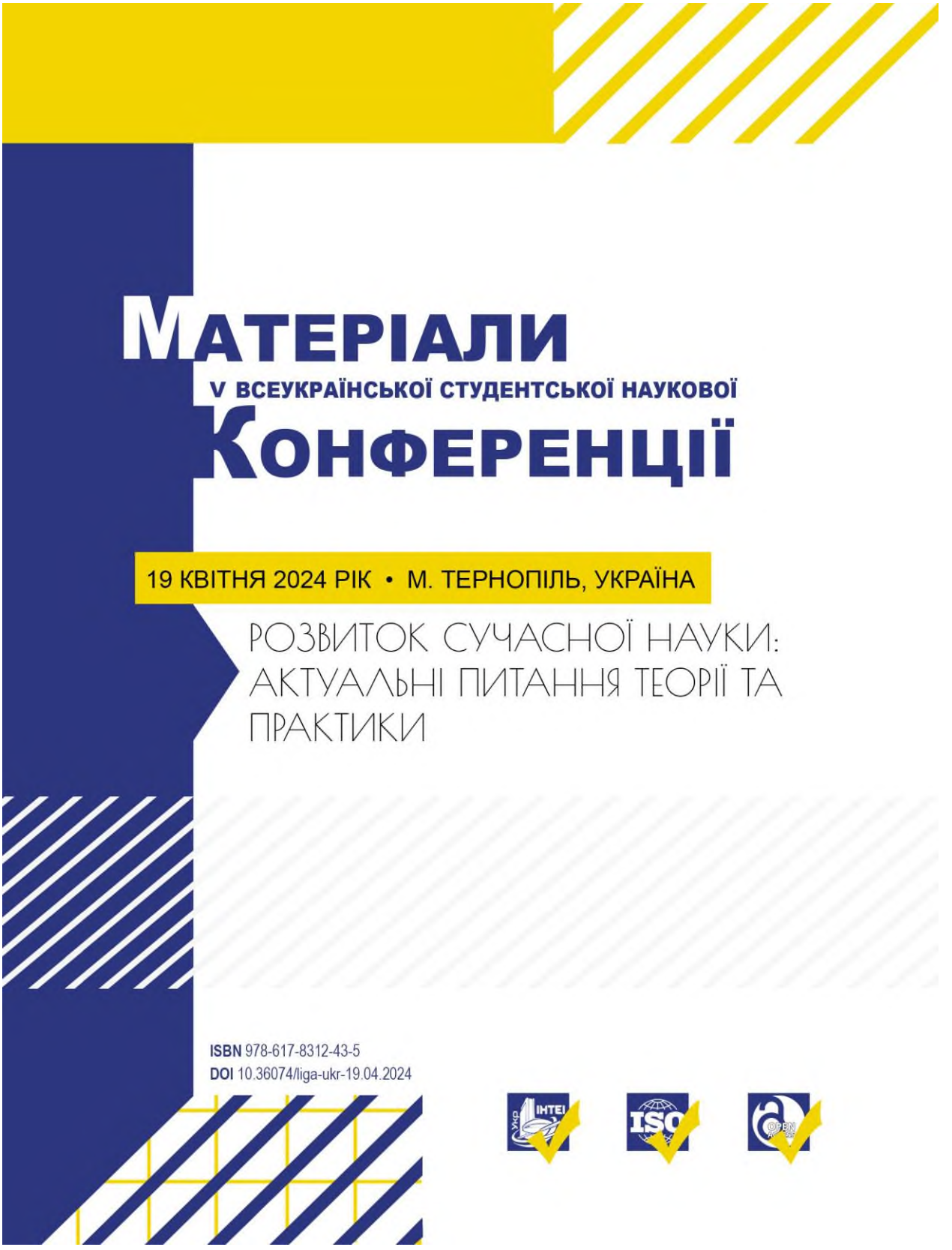
Document	Topic	Name	Representation	Representative_Docs	Top_n_words	Probability	Representative_document	
51	I absolutely love this bib tee! it's probably my favorite retailer purchase of all time. i'm 5'7", 140 pounds and the small was a perfect fit for me. i typically wear either a s or m tops.	5	5_tee_tees_basic_white	[tee, tees, basic, white, its, bought, this, the, is, and]	[Great soft tee. got the blue and really love the color., I love this tee. really comfortable and super cute too!, Love the way this tee looks and feels. so comfortable and fun! good length. perfect jeans tee.]	tee - tees - basic - white - its - bought - this - the - is - and	0.936653	False
129	This top is super comfy and casual. the slit/design in the front gives it more of a stylish look than your average white long sleeve tee. would definitely recommend.	5	5_tee_tees_basic_white	[tee, tees, basic, white, its, bought, this, the, is, and]	[Great soft tee. got the blue and really love the color., I love this tee. really comfortable and super cute too!, Love the way this tee looks and feels. so comfortable and fun! good length. perfect jeans tee.]	tee - tees - basic - white - its - bought - this - the - is - and	0.782965	False
332	I love this tee! the material is thick but not too thick. it's highly flattering- i love that the sleeves are a longer short sleeve to cover up the bingo wings. the shimmer on the front adds something special.	5	5_tee_tees_basic_white	[tee, tees, basic, white, its, bought, this, the, is, and]	[Great soft tee. got the blue and really love the color., I love this tee. really comfortable and super cute too!, Love the way this tee looks and feels. so comfortable and fun! good length. perfect jeans tee.]	tee - tees - basic - white - its - bought - this - the - is - and	0.801359	False
336	This is a great piece to help ease on in to the chillier days of fall and winter. it's versatile- i can see myself wearing it with plain short or long sleeved tees or even tops with more elaborate padderns or button downs! there's so many options i cannot wait to put this gorgeous piece to use! just as an fyi, the sleeve openings are slightly unfinished and have a frayed effect to them. this may not even be noticeable if you're not looking right up close. for me, this adds to the unique quality	5	5_tee_tees_basic_white	[tee, tees, basic, white, its, bought, this, the, is, and]	[Great soft tee. got the blue and really love the color., I love this tee. really comfortable and super cute too!, Love the way this tee looks and feels. so comfortable and fun! good length. perfect jeans tee.]	tee - tees - basic - white - its - bought - this - the - is - and	0.629009	False

Document	Topic	Name	Representation	Representative_Docs	Top_n_words	Probability	Representative_document	
376	I was looking for a basic tee, but this one was just ok...the quality is okay, but it us not as soft as i would have liked. unfortunately, i will be returning this item.	5	5_tee_tees_basic_white	[tee, tees, basic, white, its, bought, this, the, is, and]	[Great soft tee. got the blue and really love the color., I love this tee. really comfortable and super cute too!, Love the way this tee looks and feels. so comfortable and fun! good length. perfect jeans tee.]	tee - tees - basic - white - its - bought - this - the - is - and	1.000000	False
...	...	...	...	...	...	...	...	...
22466	I bought this tee for a light summer top as i love linen. this top does not disappoint. i bought my usual size xs and the length is good-slightly longer in back. i purchased the black and the grey. the grey will be great with jeans in any color and the black i can wear either dressed up or down. the quality seems good. great purchase!	5	5_tee_tees_basic_white	[tee, tees, basic, white, its, bought, this, the, is, and]	[Great soft tee. got the blue and really love the color., I love this tee. really comfortable and super cute too!, Love the way this tee looks and feels. so comfortable and fun! good length. perfect jeans tee.]	tee - tees - basic - white - its - bought - this - the - is - and	1.000000	False
22653	I had my eye on this tee for awhile, and finally ordered the turquoise/navy in the small. as soon as i received it, i ordered it in the taupe/peach and am contemplating one more! i plan to layer it now and can't wait to wear it solo next spring and summer. the colors are gorgeous (although the taupe is more dark gray which i love), the material is very soft, stretchy and comfortable, but it's the very flattering cut that sold me. it hangs beautifully and doesn't cling to my middle but skims it p	5	5_tee_tees_basic_white	[tee, tees, basic, white, its, bought, this, the, is, and]	[Great soft tee. got the blue and really love the color., I love this tee. really comfortable and super cute too!, Love the way this tee looks and feels. so comfortable and fun! good length. perfect jeans tee.]	tee - tees - basic - white - its - bought - this - the - is - and	0.681370	False
22793	I found this tee to run true to size with a comfortable relaxed fit. the neckline and elegant button gives this tee enough detail to dress it up if you wanted to take it in that direction.	5	5_tee_tees_basic_white	[tee, tees, basic, white, its, bought, this, the, is, and]	[Great soft tee. got the blue and really love the color., I love this tee. really comfortable and super cute too!, Love the way this tee looks and feels. so comfortable and fun! good length. perfect jeans tee.]	tee - tees - basic - white - its - bought - this - the - is - and	1.000000	False

Document	Topic	Name	Representation	Representative_Docs	Top_n_words	Probability	Representative_document	
23097	This tee is cute and unusual and i am sure i will wear it pretty often. i do wish i had ordered a size smaller than my normal large, but i am waiting for it to shrink in the wash a bit. the material is thin and it certainly doesn't feel like it is worth the price tag. i got it on sale and still think it is only worth about 20$.	5	5_tee_tees_basic_white	[tee, tees, basic, white, its, bought, this, the, is, and]	[Great soft tee. got the blue and really love the color., I love this tee. really comfortable and super cute too!, Love the way this tee looks and feels. so comfortable and fun! good length. perfect jeans tee.]	tee - tees - basic - white - its - bought - this - the - is - and	0.618761	False
23446	This tee is amazing. it's light weight and perfect for the summer heat. the detail of the cross and tie back is perfect to elevate the style. this is going to be a solid staple in my tee collection. i wish it came in other colors because i would buy them all.	5	5_tee_tees_basic_white	[tee, tees, basic, white, its, bought, this, the, is, and]	[Great soft tee. got the blue and really love the color., I love this tee. really comfortable and super cute too!, Love the way this tee looks and feels. so comfortable and fun! good length. perfect jeans tee.]	tee - tees - basic - white - its - bought - this - the - is - and	0.712580	False

ДОДАТОК В

АПРОБАЦІЯ ОТРИМАНИХ РЕЗУЛЬТАТІВ



УДК 082:001
Р 64

Голова оргкомітету: Коренюк І.О.

Верстка: Зрада С.І.
Дизайн: Бондаренко І.В.

Рекомендовано до видання Вченою Радою Інституту науково-технічної інтеграції та співпраці. Протокол № 32 від 18.04.2024 року.

Конференцію зареєстровано Державною науковою установою «УкрІНТЕІ» в базі даних науково-технічних заходів України та інформаційному бюлетені «План проведення наукових, науково-технічних заходів в Україні» (Посвідчення №25 від 05.01.2024).

Р 64 **Розвиток сучасної науки: актуальні питання теорії та практики**: матеріали V Всеукраїнської студентської наукової конференції, м. Тернопіль, 19 квітня, 2024 рік / ГО «Молодіжна наукова ліга». — Вінниця: ТОВ «УКРЛОГОС Груп», 2024. — 246 с.

ISBN 978-617-8312-43-5
DOI 10.62732/liga-ukr-19.04.2024

Викладено матеріали учасників V Всеукраїнської мультидисциплінарної студентської наукової конференції «Розвиток сучасної науки: актуальні питання теорії та практики», яка відбулася 19 квітня 2024 року у місті Тернопіль, Україна.

УДК 082:001

ISBN 978-617-8312-43-5

СЕКЦІЯ 12.
ТЕХНОЛОГІЇ ЛЕГКОЇ ТА ДЕРЕВООБРОБНОЇ ПРОМИСЛОВОСТІ

СЕКЦІЯ 13.
ЕНЕРГЕТИКА ТА ЕНЕРГЕТИЧНЕ МАШИНОБУДУВАННЯ

СЕКЦІЯ 13.
КОМП'ЮТЕРНА ТА ПРОГРАМНА ІНЖЕНЕРІЯ

СЕКЦІЯ 14.
ІНФОРМАЦІЙНІ ТЕХНОЛОГІЇ ТА СИСТЕМИ

Ковальчук Наталя Богданівна, здобувач вищої освіти
факультету комп'ютерних інформаційних технологій
Західноукраїнський національний університет, Україна

Науковий керівник: Ліп'яніна-Гончаренко Христина Володимирівна, канд. техн. наук, доцент, доцент кафедри інформаційно-обчислювальних систем і управління
Західноукраїнський національний університет, Україна

АЛГОРИТМ МОДЕЛЮВАННЯ ТЕМ ВІДГУКІВ В ЕЛЕКТРОННІЙ КОМЕРЦІЇ

Далі розглянемо по-кроковий алгоритм (рис.1), що описує процес впровадження NLP тематичного моделювання для аналізу відгуків клієнтів в електронній комерції, зокрема для бренду жіночого одягу. Метою є витяг тематичних груп, пов'язаних з перевагами та проблемами клієнтів, такими як якість продукції, час доставки, невідповідності розміру та кольору.

По-кроковий алгоритм:

1. Завантаження даних

1.1. Імпортуйте необхідні бібліотеки: numpy, pandas, matplotlib.pyplot.

1.2. Завантажте набір даних з відгуками про жіночий електронний комерційний одяг за допомогою pandas.

1.3. Відобразіть перші кілька рядків, щоб зрозуміти структуру даних.

2. Огляд даних

2.1. Використовуйте метод .info(), щоб отримати огляд набору даних, включаючи назви стовпців та типи даних.

2.2. Проаналізуйте розподіл відгуків за різними рейтингами, віком та кількістю позитивних відгуків, використовуючи гістограми.

3. Попередня обробка даних

3.1. Очистіть і попередньо обробіть текстові дані:

3.2. Видаліть зайві пробіли.

3.3. Заповніть пропущені значення порожніми рядками.

3.4. Виправте помилки друку та невідповідності в назвах категорій.

3.5. Токенізуйте текст відгуків і видаліть стоп-слова.

3.6. За бажанням, проведіть лематизацію для зведення слів до їх базових форм.

4. Моделювання тем

4.1. Модель LDA

4.1.1. Реалізуйте модель Latent Dirichlet Allocation (LDA) за допомогою бібліотек, таких як gensim.

4.1.2. Створіть словник і корпус, необхідні для LDA.

4.1.3. Навчіть модель LDA на корпусі.

4.1.4. Оцініть модель, використовуючи когерентність тем, щоб визначити оптимальну кількість тем.

4.2. Модель BERTopic

4.2.1. Встановіть та імпортуйте бібліотеку BERTopic.

4.2.2. Застосуйте модель BERTopic до очищених текстів відгуків.

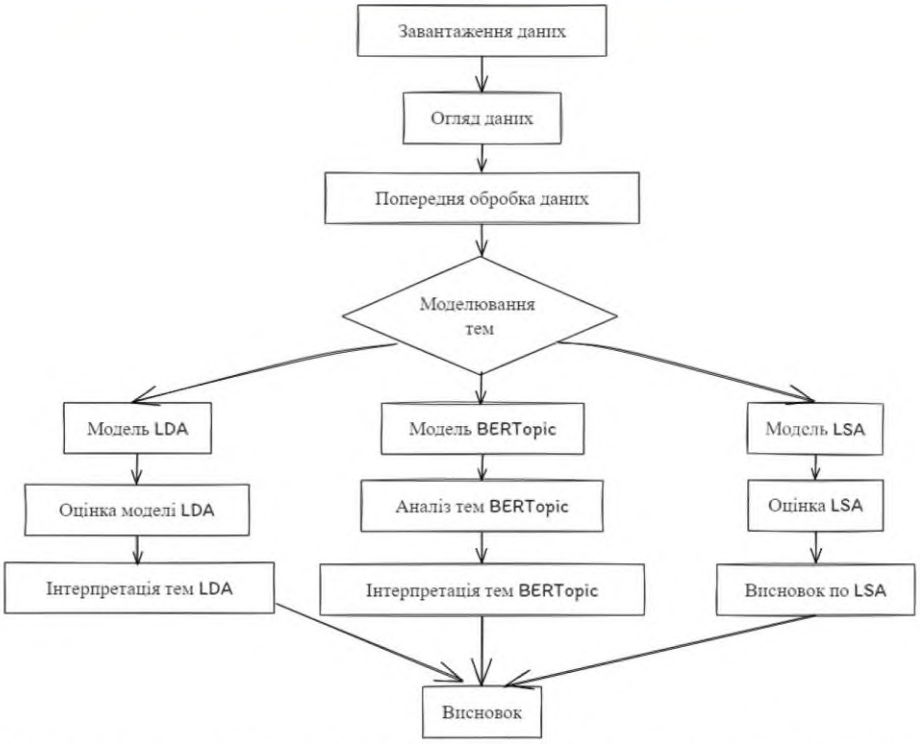


Рис. 1. **Алгоритм моделювання тем відгуків в електронній комерції**

4.2.3. Проаналізуйте теми для розуміння загальних тем у відгуках клієнтів.

4.3. Латентний семантичний аналіз

4.3.1. Використовуйте векторизацію TF-IDF для перетворення текстових даних на матрицю особливостей TF-IDF.

4.3.2. Застосуйте скорочений сингулярний розклад (Truncated SVD) для зниження розмірності.

4.3.3. Досліджуйте пояснювальну здатність, щоб вирішити кількість тем.

5. Висновок

Цей алгоритм надає всеосяжний підхід до аналізу відгуків клієнтів через тематичне моделювання, пропонуючи інсайти щодо особливостей продукції, які сприяють задоволенню клієнтів, і сфер для поліпшення.

Список використаних джерел:

1. Khanal, S. S., Prasad, P. W. C., Alsadoon, A., & Maag, A. (2020). A systematic review: machine learning based recommendation systems for e-learning. Education and Information Technologies, 25(4), 2635-2664.