Хмельницький національний університет
Факультет програмування та комп'ютерних і телекомунікаційних систем
Кафедра комп'ютерних наук та інформаційних технологій

ДИПЛОМНА РОБОТА

МАГІСТРА

на тему *Аналітична система рекомендацій закладів харчування на основі відгуків та рейтингу*

Галузь знань *12 – Інформаційні технології*
Шифр і назва галузі знань

Спеціальність *122 – Комп'ютерні науки*
Шифр і назва спеціальності

Виконав: студент *2* курсу, група *КНм-19-1* — Підпис — *І.Д. Тіторов* (Ініціали, прізвище)

Керівник: *к.т.н., доцент кафедри КНІТ* — Підпис — *Е.А. Манзюк* (Ініціали, прізвище)

Нормоконтроль: *к.т.н., доцент кафедри КНІТ* — Підпис — *Р.О. Багрій* (Ініціали, прізвище)

До захисту допускаю:
Зав. кафедри КНІТ, к.т.н., професор — Підпис — *О.В. Бармак* (Ініціали, прізвище)

7 *12* 2020 р.

Хмельницький 2020

ХМЕЛЬНИЦЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ

Факультет *програмування та комп'ютерних і телекомунікаційних систем*

Кафедра *комп'ютерних наук та інформаційних технологій*

Освітній ступінь *магістр*

Галузь знань *12 – Інформаційні технології*

Спеціальність *122 – Комп'ютерні науки*

ЗАТВЕРДЖУЮ

Завідувач кафедри *комп'ютерних наук та інформаційних технологій*

_____

(підпис)

д.т.н., доцент О.В. Бармак

«*7*» *09* *2020* року

ЗАВДАННЯ

НА ДИПЛОМНУ РОБОТУ МАГІСТРА

1. Тема дипломної роботи магістра: *«Аналітична система рекомендацій закладів харчування на основі відгуків та рейтингу»*

2. Завдання видано студентці *Тіторов Ігор Дмитрович*

(прізвище, ім'я, по батькові)

3. Керівник роботи *к.т.н., доцент Манзюк Едуард Андрійович*

(прізвище, ім'я, по батькові)

4. Затверджені наказом університету від «*9*» *09* *2020* р. № *22*

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

Мета роботи – розробка аналітичної системи рекомендації закладів харчування на основі відгуків та рейтингу. Об'єктом дослідження є методи отримування інформації з використанням текстової інформації та цифрового рейтингового оцінювання. Предметом дослідження є текстові дані у вигляді відгуків з соціальних мереж та додаткова інформація у вигляді відносних рейтингових оцінювань.

Реферат

Дипломна робота магістра присвячена розробці аналітичної системи рекомендацій закладів харчування на основі відгуків та рейтингу.

Актуальність теми. В магістерській роботі було розроблено та набуло практичної реалізації системного підходу щодо аналізу відгуків та рейтингу в системі закладів громадського харчування .

Метою дослідження є розробка аналітичної системи рекомендації закладів харчування на основі відгуків та рейтингу. Для досягнення зазначеної мети поставлені наступні **задачі**:

– провести вибір ознак та впливових факторів для використання в бізнес-аналітики;

– провести порівняння застосовності відомих методів дослідження щодо розробки та аналізу рекомендаційної системи.

При цьому передбачається розв'язок таких **підзадач**, як

– попередня даних та їх очищення;

– побудова списку ознак предметної області;

– дослідження методів визначення впливовості ознак;

– вибір моделей, виділення ознак і застосування методів машинного навчання;

– тестування методів на основі правил і з використанням машинного навчання;

– програмна реалізація система рекомендацій закладів харчування.

Об'єктом дослідження є методи отримування інформації з використанням текстової інформації та цифрового рейтингового оцінювання.

Предметом дослідження є текстові дані у вигляді відгуків з соціальних мереж та додаткова інформація у вигляді відносних рейтингових оцінювань.

Дослідження показало, що рейтинговий висновок є правдоподібним завданням, потенційні перешкоди існують у вивченні взаємозв'язку між рейтингами та відгуками. По-перше, є непослідовність при присвоєнні рейтингів

серед авторів, дивергенція крос-автора. Це легко пояснюється тим, що думки - суб'єктивна річ. Для того ж рейтинг по відгукам, це загально прийнято, щоб один відгук з'являється дуже позитивним, а інший менш позитивним або навіть трохи негативним. По-друге, оцінки не повністю підтримуються текстом. Коли людей просять призначити оцінку для продукту або послуги, вони зазвичай представляють їх в загальному вираженні, тоді як те, що вони пишуть в огляді, може бути просто більш емоційно виражено, хорошим або поганим відношенням. Насправді додавання даних персоналізації може бути корисним у покращенні релевантності рейтингової системи.

У минулому бізнес-аналітика була привілеєм великих компаній, які могли дозволити собі підтримувати команди IT-фахівців і вчених з обробки даних. Але в останнє десятиліття, оскільки технологія швидко розвивалася, програмне забезпечення стало не тільки більш легким і потужним, але і більш доступним. Малий бізнес може використовувати ті ж інструменти, що і основні гравці ринку, і змагатись зі своїми конкурентами. Нові інструменти самообслуговування доводять, що бізнес-аналітика скоріше корисний інструмент, який допоможе перетворити дані в обґрунтовані рішення. Тепер кожна компанія може використовувати силу сучасного програмного забезпечення, щоб підняти свою нижню лінію, оскільки бізнес-аналітика для малого бізнесу стала доступною.

Достовірність результатів забезпечується використанням сучасних методів та підходів та напрацювань в сфері досліджуваної області.

Велика кількість диспропорцій та операційних питань, доведених до уваги завдяки впровадженій системі бізнес-аналітики, яку потім можна було б вирішити належним чином. Наприклад, порівняння кількості об'єднаних людино-годин, необхідних для праці та доставки в середньому по країні, показало, що там виникла проблема, оскільки вони були вище середнього рівня. Таким чином переваги використання аналітики у бізнес процесах є очевидним та вимагає впровадження в практичних цілях.

У цій роботі запропонували інноваційний метод визначення різних особливостей для ресторанів різних кухонь. Метод базувався на моделі SVM, розрахунку балів слів і вимірюванні популярності.

Основні функції можуть не тільки допомогти клієнтам вибрати свою улюблену кухню, але і надати ресторанам свої переваги.

З іншого боку, подібні процедури можуть бути відтворені для відгуків і коментарів в інших областях, таких як огляди фільмів і публікації в соціальних мережах.

Практична значимість дослідження полягає в тому, що отримані практичні результати можуть бути застосовні для підвищення ефективності роботи системи.

Програмне забезпечення для аналітики ресторанів є хорошим інструментом для будь-якого сучасного, амбітного та далекоглядного ресторатора.

Аналітика даних ресторану дасть додаткову конкурентну перевагу, яка допоможе зрозуміти клієнтів набагато більш докладно. Також допоможете розкрити уявлення про бізнес,- саме через ці відкриття дають найбільше зростання.

На сьогоднішній час це не тільки важливо, але потенційно корисно - для ресторану потрібна кожна перевага, яку можете отримати, щоб виграти на харчовому полі бою, і аналітика ресторану допоможе вказати де потрібно бути.

Апробація дипломної роботи.

Основні положення і результати роботи опубліковані в збірнику наукових праць – Тіторов І. Д. Аналітична система рекомендацій закладівхарчування на основі відгуків та рейтингу / І. Д. Тіторов, Т. К. Скрипник // Збірник наукових праць за матеріалами Всеукраїнської науково-практичної конференції «Актуальні проблеми комп’ютерних наук - 2020» Хмельницький, 2020, – С.300-302.

Структура та обсяг роботи. Дипломна робота магістра складається з завдання, реферату, змісту, вступу, 4 розділів, висновків, переліку посилань із 32

найменувань та додатків. Загальний обсяг дипломної роботи магістра становить 74 сторінок, з них 70 сторінок основного тексту та 1 сторінки додатків. в роботі наведено 19 рисунків та 3 таблиць.

Ключові слова: Діріхте, класифікація, аналітична система.

Зміст

Вступ

У минулому бізнес-аналітика була привілеєм великих компаній, які могли дозволити собі підтримувати команди IT-фахівців і вчених з обробки даних. Але в останнє десятиліття, оскільки технологія швидко розвивалася, програмне забезпечення стало не тільки більш легким і потужним, але і більш доступним. Малий бізнес може використовувати ті ж інструменти, що і основні гравці ринку, і стикатися зі своїми конкурентами. Нові інструменти самообслуговування доводять, корисний інструмент, який допоможе перетворити дані в обґрунтовані рішення. Тепер кожна компанія може використовувати силу сучасного програмного забезпечення, щоб підняти свою нижню межу конкурентності, оскільки бізнес-аналітика для малого бізнесу стала доступною і це робить бізнес-аналітику незамінною.

Важливо знати, що бізнес-партнер може бути адаптований до будь-якої бізнес-моделі або галузі. Відомо, про те, як побудувати успішну стратегію бізнес-аналітики, або використовувати силу фінансової звітності та аналізу, як використовувати силу сучасної інформаційної панелі бізнесу і зробити більшу ефективність з аналітики даних малого бізнесу. Потреба у великому відділі збору та аналізу всіх зібраних даних, та проведення цих даних у різних відділах і показ їх багатьом зацікавленим сторонам – це модель минулого. У цифрову епоху доступ до даних майже в режимі реального часу має вирішальне значення, якщо необхідно залишатися на вершині ринку. Більшість звітів і аналізу, необхідних для прийняття швидких, обґрунтованих, детальних і надійних рішень, можуть бути доступні протягом декількох кліків, поділяються простим посиланням і аналізуються за допомогою простих діаграм, які можуть зробити процес прийняття бізнес-рішень і час дії набагато швидше.

Ефективне використання бізнес-аналітики та аналітики є вирішальною відмінністю між компаніями, які досягають успіху, та компаніями, які зазнають

невдачі в сучасному середовищі. Все змінюється і стає все більш конкурентоспроможним в кожному секторі бізнесу, а переваги бізнес-аналітики і правильне використання аналітики даних є ключовими для перевершеної конкуренції.

Наприклад, що стосується маркетингу, традиційні рекламні методи витрачання великих сум грошей на радіо та друковані оголошення без вимірювання рентабельності інвестицій не працюють так, як раніше. Споживачі все більше і більше застраховані від оголошень, які не націлені безпосередньо на них.

Компанії, які є найбільш успішними в маркетингу як в B2C, так і в B2B, використовують дані та дослідження для створення гіпер-конкретних кампаній, які простягають руку цільовим перспективам з куратором повідомлення. Все тестують, а потім кампанії, які досягають успіху, отримують більше грошей, вклав у них, а інші не повторюються.

Актуальність теми. В магістерській роботі було розроблено та набуло практичної реалізації системного підходу щодо аналізу відгуків та рейтингу в системі закладів громадського харчування .

Метою дослідження є розробка аналітичної системи рекомендації закладів харчування на основі відгуків та рейтингу. Для досягнення зазначеної мети поставлені наступні **задачі**:

– провести вибір ознак та впливових факторів для використання в бізнес-аналітики;

– провести порівняння застосовності відомих методів дослідження щодо розробки та аналізу рекомендаційної системи.

При цьому передбачається розв'язок таких **підзадач**, як

– попередня даних та їх очищення;

– побудова списку ознак предметної області;

– дослідження методів визначення впливовості ознак;

– вибір моделей, виділення ознак і застосування методів машинного навчання;

– тестування методів на основі правил і з використанням машинного навчання;

– програмна реалізація система рекомендацій закладів харчування.

Об'єктом дослідження є методи отримування інформації з використанням текстової інформації та цифрового рейтингового оцінювання.

Предметом дослідження є текстові дані у вигляді відгуків з соціальних мереж та додаткова інформація у вигляді відносних рейтингових оцінювань.

Дослідження показало, що рейтинговий висновок є правдоподібним завданням, потенційні перешкоди існують у вивченні взаємозв'язку між рейтингами та відгуками. По-перше, є непослідовність при присвоєнні рейтингів серед авторів, дивергенція крос-автора. Це легко пояснюється тим, що думки - суб'єктивна річ. Для того ж рейтинг по відгукам, це загально прийнято, щоб один відгук з'являється дуже позитивним, а інший менш позитивним або навіть трохи негативним. По-друге, оцінки не повністю підтримуються текстом. Коли людей просять призначити оцінку для продукту або послуги, вони зазвичай представляють їх в загальному вираженні, тоді як те, що вони пишуть в огляді, може бути просто більш емоційно виражено, хорошим або поганим відношенням. Насправді додавання даних персоналізації може бути корисним у покращенні релевантності рейтингової системи в використаннм аналітичного підходу.

У минулому бізнес-аналітика була привілеєм великих компаній, які могли дозволити собі підтримувати команди IT-фахівців і вчених з обробки даних. Але в останнє десятиліття, оскільки технологія швидко розвивалася, програмне забезпечення стало не тільки більш легким і потужним, але і більш доступним. Малий бізнес може використовувати ті ж інструменти, що і основні гравці ринку, і змагатись зі своїми конкурентами. Нові інструменти

самообслуговування доводять, що бізнес-аналітика скоріше корисний інструмент, який допоможе перетворити дані в обґрунтовані рішення. Тепер кожна компанія може використовувати силу сучасного програмного забезпечення, щоб підняти свою нижню лінію, оскільки бізнес-аналітика для малого бізнесу стала доступною.

Достовірність результатів забезпечується використанням сучасних методів та підходів та напрацювань в сфері досліджуваної області.

Велика кількість диспропорцій та операційних питань, доведених до уваги завдяки впровадженій системі бізнес-аналітики, яку потім можна було б вирішити належним чином. Наприклад, порівняння кількості об'єднаних людино-годин, необхідних для праці та доставки в середньому по країні, показало, що там виникла проблема, оскільки вони були вище середнього рівня. Таким чином переваги використання аналітики у бізнес процесах є очевидним та вимагає впровадження в практичних цілях.

У цій роботі запропонували інноваційний метод визначення різних особливостей для ресторанів різних кухонь. Метод базувався на моделі, розрахунку балів слів і вимірюванні популярності.

Основні функції можуть не тільки допомогти клієнтам вибрати свою улюблену кухню, але і надати ресторанам свої переваги.

З іншого боку, подібні процедури можуть бути відтворені для відгуків і коментарів в інших областях, таких як огляди фільмів і публікації в соціальних мережах.

Практична значимість дослідження полягає в тому, що отримані практичні результати можуть бути застосовні для підвищення ефективності роботи системи.

Програмне забезпечення для аналітики ресторанів є хорошим інструментом для будь-якого сучасного, амбітного та далекоглядного ресторатора.

Аналітика даних ресторану дасть додаткову конкурентну перевагу, яка допоможе зрозуміти клієнтів набагато більш докладно. Також допоможете розкрити уявлення про бізнес,- саме через ці відкриття дають найбільше зростання.

На сьогоднішній час це не тільки важливо, але потенційно корисно - для ресторану потрібна кожна перевага, яку можете отримати, щоб виграти на харчовому полі бою, і аналітика ресторану допоможе вказати де потрібно бути.

Розділ 1
Системи аналізу та генерування рекомендацій закладів харчування

1.1 Опис предметної області

Сучасне управління рестораном і ресторанна асоціація показали, що близько 60 000 нових ресторанів відкриваються щороку (примітка, дані для дослідження було використанно для США). Але 50 000 ресторанів щороку закривають свої двері.

Хоча немає швидкого рішення або остаточної відповіді на питання «розширити свій ресторанний бізнес», можна сказати, що інвестування в рішення на основі даних, інструменти звітності та використання потужності ресторанної аналітики допоможуть досягти успіху в цьому найбільш паретинайочому регіоні галузей.

Керуючи інформацією за допомогою інструментів аналізу даних, з метою щоб загострити конкурентну перевагу, підвищити рентабельність, підвищити прибуток і збільшити клієнтську базу. Дані пропонують можливість отримати об'єктивний, точний і всеосяжний погляд на щоденні функції ресторану.

Тут розглянемо аналітику даних ресторану, прогнозну аналітику ресторану, аналітичне програмне забезпечення для ресторанів, а також конкретні способи, якими великі дані можуть допомогти підвищити перспективи бізнесу по всій палітрі параметрів.

Почнемо з того, що подивимося на визначення. Що таке аналітика ресторанів.

За своєю суттю аналітика ресторанів – це концепція аналізу всіх даних, пов'язаних з ресторанним бізнесом, та перетворення їх у дієві ідеї за допомогою програмного забезпечення бізнес-аналітики, що в кінцевому підсумку призведе до значно підвищення ефективності.

У сучасному гіперзв'язаному цифровому ландшафті можна збирати, організовувати та презентувати кожен фрагмент інформації – від часу очікування до продуктивності персоналу та оптимізації меню – таким чином, щоб допомогти ресторану розвиватися та вдосконалюватися на постійній основі.

Чому аналітика ресторанів важлива. Бізнес-аналітика для ресторанів є невід'ємною частиною розуміння внутрішньої роботи бізнесу, але і усвідомлюючи, як можете поліпшити його, щоб сприяти сталому рівню успіху, який відірве від конкуренції.

Працюючи з відповідними ключовими показниками ефективності (KPI) і приладними дошками даних, зможете відстежувати, відстежувати та вимірювати найцінніші бізнес-аналітики таким чином, щоб очистити, стислі та засвоюваними, витягуючи з минулих, теперішніх і передбачуваних даних. Це дозволить забезпечити стійкі процеси управління KPI, які в кінцевому підсумку підвищать продуктивність і заощадять гроші [2].

Аналітика даних ресторану допоможе дістатися до серця питання і зрозуміти всю правду про бізнес. Поки знаходимося на цю тему, давайте подивимося, як бізнес-аналітика для ресторанів допоможе внести позитивні зміни, які отримають реальні результати.

У підсумку, аналітика даних на основі ресторанів має вирішальне значення для успіху ресторану, оскільки вони дозволяють:

Упорядкуйте свої дані та проімітуйте будь-які показники, які відносяться від цілей. Перетворіть найцінніші дані на дієві аналітичні огляди. Відстежуйте,вимірюйте та відстежуйте ефективність за допомогою інтерактивних KPI.

Знайдіть нові тенденції, які відійдуть від конкуренції.

Зробіть свій бізнес більш ефективним, більш розумним і прибутковим, ніж коли-небудь вважали можливим.

1.2 Роль прогнозної аналітики в ресторанах

Дослідження від The Perry Group і The Restaurant Brokers свідчать про те, що 90% ресторанів, які самостійно перебувають у власності, закриваються протягом одного року після відкриття. Крім того, 70% ресторанів, які примудряється вижити протягом 12 місяців, закривають свої двері протягом найближчих трьох-п'яти років. Отже, не перебільшенням є те, що переважна більшість ресторанів, які відкриваються в цьому світі, не досягають успіху. Але це не означає, що не може [4].

Дослідили, що таке аналітика ресторанів і як аналітика даних для ресторанів може допомогти зрозуміти бізнес на більш глибокому рівні. Але перш ніж деталізувати конкретні способи, в яких аналітика даних ресторану може підвищити ефективність, важливо зрозуміти загальну роль прогнозних даних у галузі послуг.

Хоча прогнозна аналітика не є тією чиєюсь формою магічної цифрової екстрасенси, ця галузь далекоглядних даних та розуміння може допомогти ресторану внести безцінні зміни на основі тенденцій, які свідчать про те, як конкретні елементи бізнесу, ймовірно, розгортатимуться.

1.3 Головна роль аналітики в роботі закладу

Ось основні ролі прогнозної аналітики в ресторанах:

1. Прогнозування тенденцій

Прогностична аналітика ресторану використовує історичні дані, а також дані в реальному часі для прогнозування майбутніх сильних сторін, слабких сторін і тенденцій. Отримуєте доступ до цієї інформації, як правило, за допомогою живої приладної дошки, зможете сформулювати стратегії та створити ініціативи, які допоможуть підвищити майбутній успіх бізнесу.

2. Панорамний зір.

Працюючи з прогнозною аналітикою, отримаєте можливість деталізувати минулі та теперішні тенденції, аналітичні огляди та візуалізації, а отже, створити розповідь з даними. При цьому будете насолоджуватися панорамним баченням підприємства, отримуєте перспективу, яку потрібно дійсно дізнатися про ресторан, який, у свою чергу, дасть натхнення, необхідне для розробки інноваційних стратегій підвищення бізнесу [2].

3. Операційна ефективність.

Від скорочення харчових відходів до сезонної оптимізації меню та майбутніх рівнів ефективності персоналу, прогнозна аналітика ресторану може допомогти в щоденних, щотижневих і довгострокових операціях бізнесу - переваги, які розглянемо в установленому порядку.

На даний момент, можете подумати: "Ну, дані все добре, але працював в ресторанної індустрії протягом тривалого часу. Давайте розслідувати це далі.

Припустимо, що були в ресторанній індустрії протягом десятиліть. Може бути, працювали свій шлях аж від посудомийної машини до власника або менеджера позиції. Або, може бути, сім'я була ресторан до тих пір, як пам'ятаєте, і були залучені з тих пір, як були молоді. У будь-якому випадку, розробили тонко налаштоване відчуття "що працює" у ресторані та географічному розташуванні, а що ні. спробували спеціальні пропозиції, спробували акції та переключили меню навколо.

Аналітика малого бізнесу відноситься до методів і практик вимірювання конкретної ефективності роботи невеликої компанії, на операційному або стратегічному рівні. Він використовується для оцінки невеликих наборів даних для отримання статистики щодо певного проекту або процесу компанії.

Невеликі дані є більш доступними, ніж великі дані, але це не означає, що ефективне його використання не вимагає будь-яких зусиль. Якщо хочете, щоб бізнес досяг кращих результатів, важливо набути правильного мислення і стати

організацією на основі даних. Це вимагає налаштувати спосіб управління щоденними операціями компанії від топ-керівників до рівня нижче. Це не складний процес, якщо встановити чіткі цілі, визначити інструменти звітування, з яких потрібно працювати, і почати створювати перший бізнес-проект для операцій малого бізнесу [4]. Щоб глибше впоратися з аналітикою для малого бізнесу, подивимося, як малі дані корелюють з більш широкомасштабним бізнесом- - великими даними.

Є багато визначень малих даних, що в тренді по всьому Інтернету - в більшості випадків побудований на його опозиції до великих даних. Інші визначення підкреслюють більш людську сторону малих даних, оскільки вони зазвичай генеруються і вводяться в систему людиною, а не машиною. Крім того, цей тип даних зазвичай міститься в операційній базі даних - CRM або ERP недостатньо велика, щоб називатися великими даними. Крім того, ним можна керувати в базі даних MySQL - і потужності буде достатньо. Аналітика малих даних базується на тому, що бізнес повинен ефективно використовувати ресурси, які він вже має, і уникати надмірного використання додаткових технологій або зовнішньої інфраструктури.

У своїй роботі ForbesМайк Кавіс має трохи інший погляд на невеликі дані, підкреслюючи той факт, що він включає в себе тільки дуже конкретні атрибути. Він використовується для визначення поточних станів і умов, які можуть бути згенеровані, наприклад, датчиками, розгорнутими на вітрових турбінах, невеликими пакетами або прикріпленими до дронів для надання дуже конкретної інформації - про місцезнаходження, температуру і т.д. Всі ці невеликі набори даних, зібрані в режимі реального часу, створюють більшу картину у вигляді великих наборів даних, які дають нам історичний, багатогранний вигляд.

Після всіх цих експериментів знаєте, що подобається клієнтам, не подобається, і що вони можуть бути зацікавлені в майбутньому. Ніхто не сперечається з цим. Дані не можуть запустити ресторан, і дані не можуть

замінити досвід. Дані також не можуть замінити творчість, стиль і пристрасть до бізнесу.

Дані не призначені для того, щоб "замінити" що-небудь. Натомість аналітика ресторанів є доповненням до вже спроможної бізнес-аналітики. Однак деякі з інтуїцій не є досконалими. Що знаєте:

- які види страв клієнтам подобаються найкраще;
- які сервери приносять найбільші замовлення послідовно;
- які нові акції, швидше за все, продаватимуть.

Знаєте ці речі, засновані на минулому досвіді, і тому зробили переконання для кожної з цих областей. Проблема в тому, що наш сучасний світ змінюється з прискоренням. Переконання та інтуїції можуть швидко стати неточними.

Дані можуть служити способом "перевірити себе" і дістатися до нижньої межі того, що дійсно робить бізнес успішним. Як писав про дані Пітер Чен, «аналітика не може придумати ідеї, але це може допомогти покращитися на хороших, уникнути спроб поганих і розкрити недоліки, які можна виправити».

Давайте проілюструвати деякі з цих принципів на роботі в прикладі. Dickey's Barbecue Pit - це мережа ресторанів у США з більш ніж 500 місцями. Одного разу генеральний директор Роланд Діккі поставив ідею своїй дружині Лорі: "Барбекю і великі дані - давайте зробимо цю роботу!"[8]. Більше, ніж просто великі дані, пара хотіла в режимі реального часу, дієві ідеї.

Після отримання системи аналітики ресторану вони почали збирати "безцінну" інформацію, таку як:

- Демографічні дані: Завдяки аналітикі, Діккі тепер знає, що їх середній гість 43-річний чоловік, який водить позашляховик на роботу. Вони навіть знають, що середній час поїздки цього клієнта становить 30 хвилин. В результаті, Dickey's тепер спеціально націлений на власників Ford, які живуть від 15 до 30 хвилин від місця розташування Діккі в своїй рекламі.

– Поведінкові дані: Діккі дізнався, що жінки з дітьми часто ходять в місце в середу, і насолоджуватися довгим обідом пізно вдень. Через ці дані Діккі тепер рекламує "Craft Wednesdays" на Pinterest як нічию для матерів та їхніх дітей [6].

– Спільні інтереси клієнтів: Діккі виявив, що їхні клієнти люблять - футбол і собак. В результаті вони почали рекламувати на футбольних сайтах і сайтах любителів собак, а також каналах Animal Planet. Вони навіть використовують собак у своїх фотографіях громадського харчування в якості фірмового ходу.

Нарешті, в результаті реального характеру своєї платформи аналітики ресторану, менеджери і власники франчайзингу можуть робити великі тактичні кроки, пов'язані з щоденними тенденціями продажів.

Поєднуючи свої знання з різними KPI продажів, можете оптимізувати свої операції.

Наприклад, можете виконувати місцеві ініціативи з продажу на основі елементів, які будуються в інвентарі, або робити місцеві продажі, якщо в певному місці менше бізнесу, ніж передбачалося спочатку.Ще один плюс для аналітики даних для ресторанів. Сучасна аналітика ресторанів відповідає на критичні питання бізнесу.

Yelp - американська транснаціональна корпорація, заснована в 2004 році, яка спрямована на те, щоб допомогти людям знайти місцевий бізнес на основі функціонально та оглядів соціальних мереж. Основна мета Yelp - надати платформу для клієнтів, щоб написати відгук разом з наданням зоряного рейтингу разом з відкритим коментарем. Дані Yelp є достовірними і має широке охоплення всіх видів бізнесу. Мільйони людей використовують yelp і емпіричні дані продемонстрували, що огляди ресторанів Yelp вплинули на прийняття рішень щодо вибору продуктів харчування споживачами; зростання однієї зірки призвело до зростання доходів незалежних ресторанів на 59% [5]. Завдяки

швидкому зростанню відвідувачів і користувачів, бачимо великий потенціал yelp ресторану відгуки набір даних як цінний репозиторій статистики.

Оскільки все більше клієнтів покладаються на Yelp для полювання на їжу. Тому огляд на Yelp став важливим показником для харчової промисловості. В останні роки зростає кількість досліджень, зосереджених на Yelp. Високі цитовані роботі включають огляд, репутацію та дослідження відносин з доходами [3], ефект груп [4] та дослідження того, чому люди використовують Yelp [5]. Оскільки огляди становить найбільший компонент для Yelp, дослідження їх за допомогою методів машинного навчання, як очікується, отримують цікаві відкриття. Наприклад, був розроблений фільтр огляду [7] і він перевірив ефективність нормального алгоритму спаму Yelp. Ця робота також відносить ідею обробки природної мови до даних Yelp, але вона зосереджена на області аналізу настроїв, які проводились за високоефективною моделлю векторної машини (SVM).

Аналіз настроїв, також відомий як видобуток думок, є процесом визначення того, чи є текстова одиниця позитивною або негативною. Він може мати широкий спектр додатків, таких як автоматичне виявлення зворотного зв'язку щодо продуктів, новин і персонажів або поліпшення моделі зв'язку клієнтів.

Для автоматизації видобутку або класифікації настроїв з відгуків настроїв, аналіз настроїв [8] використовує обробку природної мови, аналіз тексту та обчислювальні методики. Ставши однією з гарячих областей у прийнятті рішень, аналіз настроїв широко використовується в багатьох сферах, таких як споживча інформація, маркетинг, книги, програми, веб-сайти та соціальні медіа [5]. Цей аналіз поділяється на багато рівнів [9]: рівень документа [7], рівень речення [10], рівень слова / терміну [12] або рівень аспектів [11].

Різні підходи були використані для оцінки настроїв під словами і виразами або документами. Деякі з найбільш поширених алгоритмів машинного

навчання, які використовуються в полях, включають Naive Bayes (NB), максимальна ентропія (ME), векторної машини (SVM) і навчання без нагляду [12]. До стрімкого розвитку методів нейронної мережі [13] останнім часом лінійні SVM часто дають кращу продуктивність [14] в природних мовах.

Однією з фундаментальних переваг аналітики даних на основі ресторанів є той факт, що вона може допомогти розкрити критичні бізнес-питання.

Роблячи це, зробимо ресторан більш згуртованим, більш конкурентоспроможним, і звичайно ж – більш прибутковим.

Як власник ресторанного бізнесу, безсумнівно, є безліч питань, на які хоче знайти дієві відповіді на кожен день. Аналітика ресторану допоможе з процесом.

Для довідки за допомогою інтерактивних даних зможете знайти відчутні відповіді на такі бізнес-запитання на основі ресторану:

– які пункти меню або які пропозиції по меню найбільш популярні;

– який найкрищий період на щотижневій основі;

– який сервіс або офіціант виконує найкраще роботу;

– наскільки добре працює штатний розклад з точки зору прибутку та обслуговування;

– чи постійно зростає дохід з плином часу;

– давайте подивимося, як це працює на практиці.

6 способів аналітики ресторанів можуть допомогти бізнесу

Аналітика ресторану може допомогти бізнесу 6 наступними способами:

1. Збільшити розміри замовлень.

2. Отримати більше повторюваного бізнесу з аналітикою меню.

3. Краща продуктивність персоналу.

4. Можете побачити і передбачити тенденції.

5. Поліпшити фінансовий потік.

6. Скоротити харчові відходи.

Розширююсь на попередніх пунктах, настав час дослідити, як охоплення рішень на основі даних може допомогти відповісти на питання та підвищити ефективність у низці ключових сфер - у реальному контексті [8].

Отже шість ключових способів, якими аналітика на основі ресторану може допомогти бізнесу.

1) Збільшення розмірів замовлень з напоями.

Якщо розробляється нове меню напоїв, можете подивитися на аналітику ресторану, щоб побачити, які напої люди, як правило, замовляють з певними пунктами у меню.

Потім, можете збільшити ці продажі, пропонуючи ці вже популярні вина.

2) Отримання більш повторюваного бізнесу з аналітикою меню.

Ви можете використовувати аналітику ресторану, щоб визначити, які елементи у меню є шпильками, а які - дудами. Це найкраще працює в поєднанні з програмою лояльності клієнтів, так що можете відстежувати шаблони окремих клієнтів з плином часу.

Наприклад, скажімо, запускаєте дані за кілька місяців покупок. зможете розділити пункти меню на 4 категорії:

1. Ці пункти замовляються багато, і люди, як правило, переупорядкувати їх.

Безумовно, не возитися з ними - якщо що-небудь, розглянути питання робити більше реклами згадки цих страв або функції їх в якийсь інший спосіб. Крім того, якщо хочете додати нові елементи до свого меню, "все-таки великі" повинні бути першим місцем, де шукаєте натхнення. Якщо всі "великі" стейк страви, цілком можливо, що нова вегетаріанська страва не буде pan out так добре.

2. Ці предмети замовляються багато, але люди не схильні переупорядкувати їх.

Ви захочете дослідити їх далі. Або вони просто хотіли спробувати "інше" блюдо, щоб побачити, що це буде. Якщо це з першої причини, можете повторно працювати блюдо, щоб бути краще або позбутися від нього взагалі. Однак, якщо люди просто хотіли спробувати «іншу страву» – це теж корисна річ, яку потрібно знати. Перегляньте назву і опис меню елемента на предмет підказок, які можна застосувати до решти меню.

3. Люди не схильні замовляти ці страви, але як тільки хтось спробує один раз, вони підключені.

Це відмінний випадок для аналітики, тому що можете запускати акції та знижки, щоб змусити людей спробувати ці страви. Як тільки вони спробують їх, робота буде виконана. Ця категорія також може отримати вигоду з кращих імен меню та описів, щоб зробити їх більш привабливими для людей, які раніше не пробували їх.

Крім того, це чудові страви для очікування персоналу, щоб виділити, даючи їм можливість показати свої знання. Адже багато хто любить ідею «прихованого дорогоцінного каменю» але не добре відома.

4. Ці страви замовляють не дуже часто. Коли їх замовляють, люди не замовляють їх знову.

Повинні або переробити ці страви, або отримати їх з меню, тому що вони дають ресторану погане ім'я. Набагато краще мати "щільне" смачне меню з меншою кількістю виділень, ніж мати розлоге меню з деякими такими пунктами на ньому.

В якості останньої ноти тут, повинні навчити свій персонал, щоб дати рекомендації від категорії коли клієнт просить рекомендацію. Ці страви мають найкращі шанси справити хороше враження (і отримати повторний бізнес).

3) Можете (об'єктивно) побачити, хто зіркові виконавці.

Скажімо, покладаєтеся на думку менеджера, коли мова йде про найм і звільнення співробітників [7]. Це, ймовірно, працює нормально - але менеджер

може мати свої власні упередження, і спотворити їх сприйняття того, що чекати від персоналу.

Наприклад, якщо є аналітика даних ресторану, пов'язана з середнім розміром замовлення співробітників, можете мати набагато чіткіше уявлення про те, хто "приносить додому їжу".

4) Можете побачити (і передбачити) тенденції.

Після того, як використовуєте аналітику ресторану деякий час, зможете знати такі речі, як:

– якими є найжливіші часи доби;
– які найважливіші дні;
– які святкові дні.

І це коли інструменти візуалізації даних приєднуються до партії, щоб дати корисну інформацію для того, щоб організувати різні індикатори та заходи в переконливі бізнес-панелі. Завдяки цим аналітичним оглядам, можете планувати свої кадрові потреби краще і переконатися, що не закінчені або недоумововані для будь-яких заданих змін [10]. Для того, щоб зробити більшу частину цих даних, потрібно буде візуалізувати, щоб зрозуміти його краще.

5) Можете поліпшити свій фінансовий потік.

Однією з найважливіших складових будь-якого успішного ресторану є фінансова ефективність.

1.4 Роль програмного забезпечення

На додаток до кращого управління інвентарем за допомогою програмного забезпечення для прогнозної аналітики ресторану, також можна отримати глибше розуміння того, де можете коригувати прибуток, щоб збільшити дохід, зберігаючи додатковий рівень успіху. Використовуючи фінансову приладну програму, навіть можете отримувати сповіщення, якщо виникає бізнес-аномалія.

Крім того, дані, що обслуговуються програмним забезпеченням фінансової аналітики для ресторанів, пропонуватимуть всебічне уявлення про успішність угод, пропозицій та композицій. Якщо певні пропозиції працюють краще за інших, можете зосередитися на створенні рекламних стратегій, щоб охопити ширшу цільову аудиторію. Крім того, якщо виявите, що конкретна угода коштує більше грошей, ніж спочатку передбачалося, зможете видалити її або внести зміни, необхідні для забезпечення прибутковості.

Гарантовано збільшується рух грошових коштів та фінансова ефективність.

6) Можете скоротити харчові відходи.

Розглянули це коротко раніше, але як харчовий елемент, це плюс-точка програмного забезпечення ресторанної аналітики, безумовно, варто вивчити більш глибоко.

Як ресторан, їжа - це гроші, тому останнє, що хочете зробити, це витрачати його (втрата їжі у великих обсягах також неетична). Працюючи з правильним програмним забезпеченням для звітування про приладну дошку та KPI, можете збільшити елемент управління продуктами харчування в роботі ресторану.

Програмне забезпечення Analytics для ресторанів може допомогти зрозуміти, які продукти ймовірно, потребують найбільш або найменшої кількості відповідно до терміну їх придатності щодо попиту в певний час доби, тижня, місяця або сезону [4].

Крім того, аналітичні дані, що надаються аналітичним програмним забезпеченням для ресторанів, дадуть інформацію, необхідну для розробки графіків підвищення ефективності продуктів харчування, які забезпечать скорочення відходів, максимізацію прибутку та покращення загальної організаційної інфраструктури ресторану.

"Як ресторатор, моя робота полягає в тому, щоб в основному контролювати хаос і драму. Там завжди буде хаос в ресторанному бізнесі." - Рокко ДіСпіріто

1.5 Постановка задачі

Метою дослідження є розробка аналітичної системи рекомендації закладів харчування на основі відгуків та рейтингу. Для досягнення зазначеної мети поставлені наступні задачі:

– провести вибір ознак та впливових факторів для використання в бізнес-аналітики;

– провести порівняння застосовності відомих методів дослідження щодо розробки та аналізу рекомендаційної системи.

При цьому передбачається розв'язок таких підзадач, як

– попередня даних та їх очищення;

– побудова списку ознак предметної області;

– дослідження методів визначення впливовості ознак;

– вибір моделей, виділення ознак і застосування методів машинного навчання;

– тестування методів на основі правил і з використанням машинного навчання;

– програмна реалізація система рекомендацій закладів харчування.

Висновок до розділу 1

Програмне забезпечення для аналітики ресторанів є безцінним інструментом для будь-якого сучасного, амбітного та далекоглядного ресторатора.

Аналітика даних ресторану дасть додаткову конкурентну перевагу, яка не допоможе зрозуміти клієнтів набагато більш докладно. також зможете розкрити уявлення про свій бізнес, який ніколи навіть не знали, існували - саме через ці відкриття відбувається найбільше зростання.

На сьогоднішній час це не тільки важливо, але потенційно корисно - для ресторану потрібна кожна перевага, яку можете отримати, щоб виграти на харчовому полі бою, і аналітика ресторану допоможе де потрібно бути.

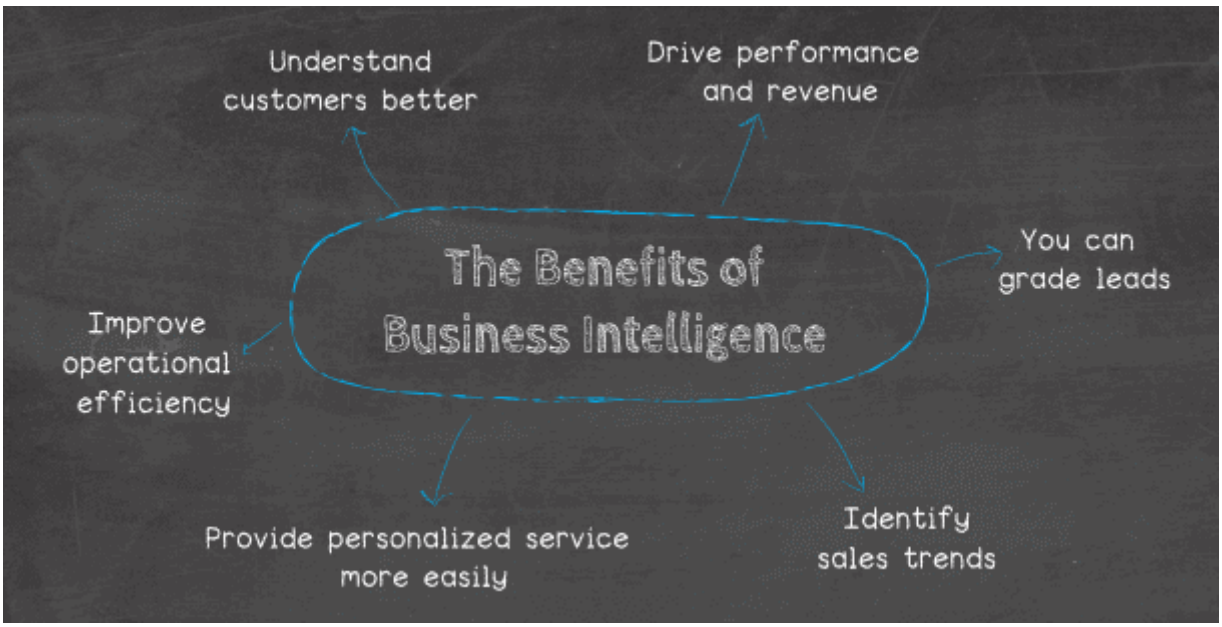
Розділ 2
Використання бізнес-аналітики
2.1 Використання бізнес – аналітики

Чому бізнес-аналітика настільки важлива. Основне використання бізнес-аналітики полягає в тому, щоб допомогти бізнес-підрозділам, керівникам, топ-керівникам та іншим оперативним працівникам приймати більш обґрунтовані рішення, підкріплені точними даними. Це в кінцевому рахунку допоможе їм виявити нові можливості для бізнесу, скоротити витрати або визначити неефективні процеси, які потребують реінжинірингу.

Бізнес-аналітика використовує програмне забезпечення та алгоритми для вилучення дієвих аналітичних даних з даних компанії та керівництва їх стратегічними рішеннями. Користувачі бізнес-аналітики аналізують і представляють дані у вигляді данихі звітів бізнес-аналітики, візуалізуючи складну інформацію простішим, більш підходящим і зрозумілим способом. Бізнес-аналітику також можна назвати "описову аналітику", оскільки вона показує лише минулий та поточний стан: вона не говорить, що робити, а те, що є чи було. Відповідальність вжити заходів досі лежить на руках керівників.

Ця методологія «подивіться на дані, налаштуйте» є в основі бізнес-аналітики. Це все про використання даних, щоб отримати чіткіше розуміння реальності, щоб компанія могли приймати більш стратегічно обґрунтовані рішення (замість того, щоб покладатися тільки на інстинкт передбачення або корпоративну інерцію).

В кінцевому рахунку, бізнес-аналітика та аналітика - це набагато більше, ніж технологія, яка використовується для збору та аналізу даних. Вона про те, щоб мати мислення експериментатора, і дозволити даним керувати процесом прийняття рішень компанії.



Рисунки 2.1 – Переваги використання бізнес-аналітики [11]

Використання даних, щоб отримати чіткіше розуміння реальності в даних, щоб аналітики могли приймати більш стратегічно обґрунтовані рішення (замість того, щоб покладатися тільки корпоративну інерцію в роботі).

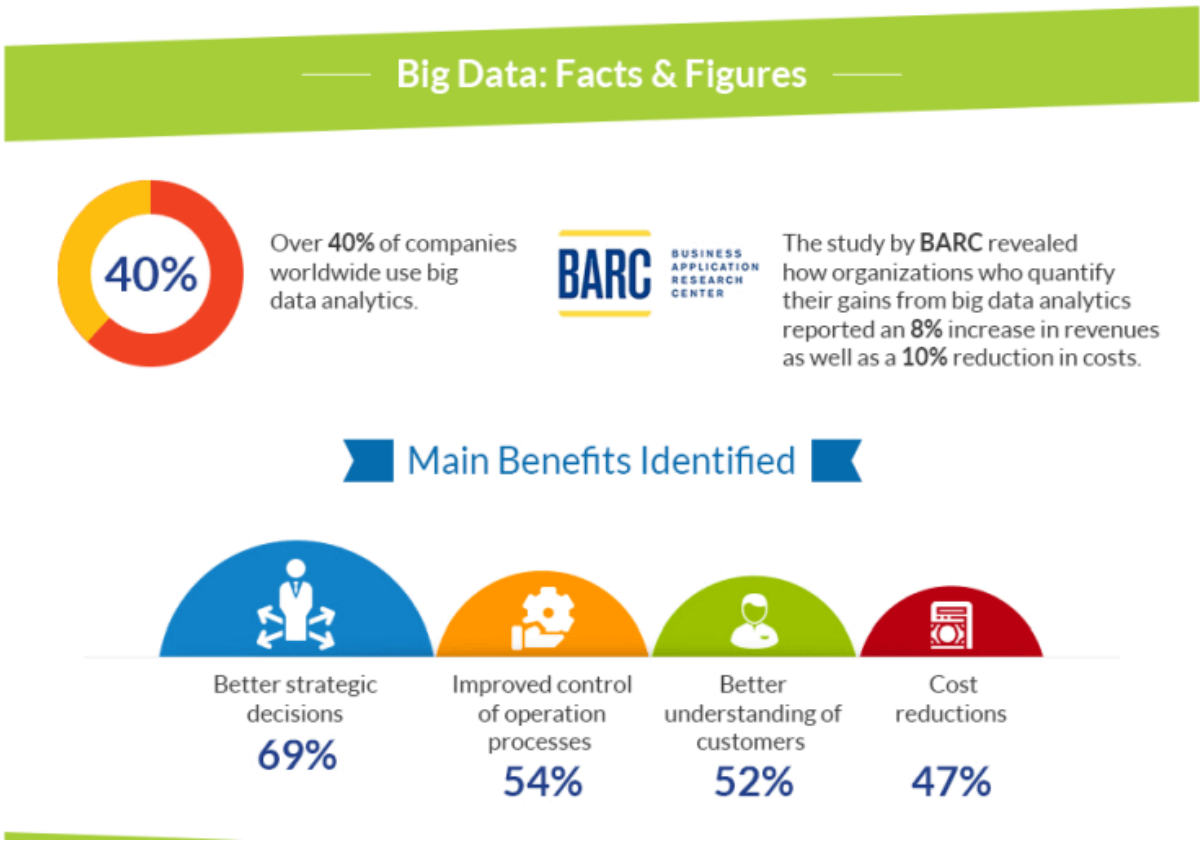
бізнес-аналітика - це набагато більше, ніж технологія отримання інформації, яка використовується для збору та аналізу складних даних. Вона про те, щоб мати мислення експериментатора-дослідника, і готова дозволити складним даним керувати процесом прийняття рішень компанії.

Які переваги бізнес-аналітики. Переваги бізнес-аналітики та аналітики різні і різноманітні, але всі вони мають одну спільну рису: вони приносять владу і силу знань. На який би підрозділ вони не вплинули, вони можуть перетворити організацію і спосіб ведення бізнесу глибоко. Ось огляд основних переваг бізнес-аналітики:

1. Ефективніше зрозумійте своїх клієнтів.
2. Підвищення продуктивності та доходу.
3. Можете сортувати потенційних клієнтів.
4. Визначення тенденцій продажів.

5. Легше надавати персоналізовані послуги.
6. Підвищення операційної ефективності.

У цьому розділі розкриємо переваги бізнес-аналітики, підкріплені деякими реальними прикладами на цьому шляху. Якщо відчується необхідність подвоїти створення культури на основі даних у компанії, і необхідно мати деякі жорсткі докази, які можна використовувати, щоб переконати скептичних товаришів по команді.



Рисунки 2.2 – Переваги використання великих даних [12]

Переваги бізнес-аналітики. Ось випадки використання, які ілюструють різні переваги бізнес-аналітики.

1) Можете зрозуміти своїх клієнтів більш ефективно.

Перша перевага бізнес-аналітики, яку будемо вирішувати тут, це відносини з клієнтами. Відомий вчений Андре Чаперон каже, що "бізнес, який

досягає успіху найбільше, це той, який розуміє своїх клієнтів найкраще". Це приклад про німецького телекомунікаційного провайдера, показує правду цього твердження і про те, як вони отримали великі переваги бізнес-аналітики.

Рисунок 2.3 – Аналітична індентифікація особливостей з бажаними результатами [6]

Система добре справлялася на своєму ринку, але зіткнулася зі зростаючою конкуренцією і ціновим тиском, що змусило вище керівництво шукати нові способи зниження ставок клієнтів. Зрештою, набагато простіше продовжувати продавати свої послуги існуючому клієнту, ніж отримати цілий новий.

Після копання в цьому питанні, виявили, що вони повинні розуміти потреби та переваги своїх клієнтів більш ретельно, щоб отримати більше оновлення. Вони вже використовували ручні методи для цього, але їм потрібно було оновити їх виконання. Саме туди зайшла бізнес-аналітика.

Як говориться в дослідженні Research, "розгортання [бізнес-аналітики] дозволило глибше зрозуміти переваги та поведінку клієнтів, щоб могли підвищити ефективність свого маркетингу".

Одна велика знахідка була в тому, що їхнім клієнтам дійсно не подобалося мати справу з аутсорсинговим колл-центром підтримки - вони просто хотіли поговорити безпосередньо з клієнтом, коли справи йшли не так добре. Усунувши свій аутсорсинговий колл-центр і поклавши речі назад в будинок, змогли збільшити свої ставки оновлення клієнтів. Іншими словами, краще розуміючи потреби своїх клієнтів за допомогою аналітики ринкових досліджень, змогли зберегти свої показники найнижчими у своїй галузі в Німеччині [14].

Завдяки цьому покращеній підтримці клієнтів, також різко скоротили кількість викликів скарг, які вони отримали в цілому. Це призвело до скорочення часу очікування дзвінків для підтримки клієнтів, що, в свою чергу, призвело до підвищення задоволеності клієнтів. Все це важливі KPI клієнтів, які повинні регулярно вимірюватися та відстежуватися для покращення обслуговування та утримання клієнтів.

Щоб все це вимкнути, скоротили зовнішні витрати на IT в декількох областях за допомогою платформи бізнес-бізнесу. Оскільки вони мали інструмент самообслуговування BI у своєму розпорядженні, не довелося витрачати стільки грошей, скільки платять за межами фірм, щоб зробити звіти для них.

Повернення своїх інвестиційних грошей, а потім 62% більше, отримання своїх грошей, виплачених назад в 1,9 років, а потім пожинаючи середньорічну вигоду в розмірі € 454,075.

Підвищення продуктивності та доходу. McKinsey реалізувала кейс-дослідження про ресторанну компанію мережі фаст-фудів з тисячами торгових точок по всьому світу. Ця компанія хотіла зосередитися на своєму персоналі і глибше проаналізувати будь-які дані, що стосуються їх персоналу, зрозуміти, що рухає ними і що вони можуть зробити, щоб підвищити ефективність бізнесу.

Після вичерпання більшості своїх традиційних методів, компанія шукала інші способи поліпшити клієнтський досвід, в той же час вирішуючи їх високу річну плинність кадрів, чий показник був вище середнього показника своїх конкурентів. Вище керівництво вважало, що вирішення цього обороту буде ключовим у поліпшенні клієнтського досвіду, і що це призведе до збільшення доходів.

Для цього компанія почала з визначення цілей, і знайти спосіб перевести поведінку співробітників і досвід в дані, щоб моделювати проти фактичних результатів. Цілі були множинними: зростання доходів, задоволеність клієнтів і швидкість обслуговування. Потім вони приступили до аналізу трьох напрямків: підбір і відбір співробітників, щоденне управління персоналом, і, нарешті, поведінка співробітників і взаємодія в ресторанах.

2.2 Вимірювання функцій загального ранжирування

Tf-idf є найбільш класичним методом для розподілу термінів моделі. tf означає частоту термінів, a idf означає частоту зворотного документа. Інтуїціям говорить про те що є документ, який містить більше термінів запиту (тобто з вищим tf), швидше за все, буде актуальним; і термін запиту, який виникає в декількох документах (тобто з нижнім idf) є більш важливим.

Функція ранжирування тексту, яка збирає частоту термінів, частоту зворотного документа та тривалість документа разом для вимірювання актуальності.

$$\sum_{t \in Q} w^{(1)} \frac{(k_1 + 1) \cdot tf}{K + tf} \tag{2.1}$$

K визначається як

$$K = k_1 \cdot [(1 - b) + b \cdot \frac{l}{avdl}] \tag{2.2}$$

де l - довжину документа;

$avdl$ - середня довжиною документа в корпусі;

b, k_1 є константами,

$w^{(1)}$ – вага, яка визначається як

$$\log \frac{N - n + 0.5}{n + 0.5} \tag{2.3}$$

де: N – це сума документів у всіх колекціях,

n - це кількість документів, що містять термін t у всіх колекціях.

Очевидно, w(1) є варіантом idf. Таким чином, - це комбінація з 3 вимірювань: tf, idf і довжини документа.

У вищевказаних функціях ранжирування, для запиту, що містить кілька термінів, терміни розглядаються як взаємні незалежні і, таким чином, відстань

термінів входжень в документах не розглядається для вимірювання розподілу термінів взагалі. Однак, термін близькість, вимірювання відстані термінів виникнення, може бути важливим фактором, щоб вплинути на релевантність.

Нарешті, актуальність документа до теми - це сума балів усіх прольотів. Існує невелика різниця в тому, щоб забити проміжок між двома функціями рейтингу. У найкоротшій функції ранжирування оцінки пропорційне взаємній довжині проміжку часу, тоді як оцінка пропорційна оберненому квадратному кореню довжини.

Але в реальних додатках, особливо в веб-пошуку, поширені короткі і неструктуровані запити і немає поняття проміжку. Таким чином, концепція поширюється як термін запиту, і подальший проміжок більше не потрібно включати всі терміни запиту.

2.3 Включення наближення термінів до функції ранжування

Припущення узгоджується з досвідом користувачів, які часто очікують, що документ, який містить більшість або всі терміни запиту, буде ранжований перед документом, що містить менше термінів.

Ці їхні підходи досягають дуже хорошого точного оцінювання в експериментах. Основним недоліком цих підходів є те, що кілька важливих вимірювань у традиційних функціях ранжирування, таких як tf, idf і довжина документа.

Поєднуючи своєрідну просту близькість термінів на основі пар слів у традиційну функцію ранжирування, можна підвищити ефективність отирмання інформації . Для запиту $q=(t_i,t_j,t_k)$ виходить такий набір S термінів пар $\{(t_i,t_j),(t_i,t_k),(t_j,t_k)\}$.

Потім функція оцінки розширюється додаванням аналогічної функції для термінальних пар (розглядаються лише термінові пари на максимальній відстані до п'яти).

$$\sum_{(t_i,t_j)\in S} \min(w_i^{(1)}, w_j^{(1)}) \cdot \frac{(k_1+1)\cdot \sum_{occ(t_i,t_j)} tpi(t_i,t_j)}{K + \sum_{occ(t_i,t_j)} tpi(t_i,t_j)} \quad (2.4)$$

І вони обчислюють вагу екземпляра пари термінів (tpi) наступним чином:

$$tpi(t_i,t_j) = \frac{1.0}{d(t_i,t_j)^2} \quad (2.5)$$

Де відстань, виражена кількістю слів. У формулі (2.5) найвище значення – 1,0, що відповідає відстані одного (терміни суміжні), а найменше значення – 0,04, що відповідає відстані 5.

Термін пара тут можна розглядати як свого роду вільну фразу, і підхід схожий на фразу пошуку та індексації. Остаточна оцінка релевантності - це лінійне поєднання показників релевантності одиночних термінів і тих, що мають вільні фрази. Але є дві проблеми, що лежать в основі такого роду підходів:

1. Оскільки фрази не є незалежними від уніграм (єдині терміни) і загальні функції ранжирування, вважаються уніграмами. Таким чином, важко оцінити важливість фраз і їх додатковий внесок у оцінку релевантності. Наприклад, якщо використовуємо n_i для представлення кількості документів, що містять термін t_i у всіх колекціях, і n_{ij} кількості документів, що містять терміни t_i та t_j . Якщо вважаємо, що перекриття між вільними фразами і одним запитом терміни,

повинно бути менше, ніж обидва і тому, що $n_{ij} < \min(n_i, n_j)$. Оцінити двограми набагато складніше при розгляді перекриття.

2. Лінійна комбінація балів уніграмів і неповних фраз може порушити властивість термінної частоти. У більшості сучасних функцій ранжирування нелінійна частота термінів бажана через статистичну залежність строкового виникнення: інформація, отримана при дотриманні терміну, вперше важливіша, ніж інформація, отримана згодом. Наприклад, якщо $l = avdl$, і $k_1 = 1.0$, при отриманні терміну 4 рази $\frac{(k_1+1)\cdot tf}{K+tf}$, ($\approx 1.6$) це лише трохи більше, ніж при отриманні терміну 3 рази (≈ 1.5). Припустимо $tpi(t_i, t_j) = 1.0$, вільна фраза $d(t_i, t_j) = 5$ збільшить значення $\frac{(k_1+1)\cdot tf}{K+tf}$ на 1, а інша пара $d(t_i, t_j) = 6$ нічого не важить. Однак різниця між цими двома випадками не така велика, як показують оцінки. Формула (2.5) здається більш важливо, тому що tpi швидко розпадеться зі збільшенням відстані, що частково зберігає нелінійну властивість термінної частоти.

У цьому дослідженні застосовуємо стратегію інтеграції термінів близькості до функцій ранжирування, уникаючи при цьому вищезазначених труднощів одночасно.

Висновки до розділу 2

Велика кількість диспропорцій та операційних питань, доведених до уваги завдяки впровадженій системі бізнес-аналітики, яку потім можна було б вирішити належним чином. Наприклад, порівняння кількості даних, показало, що виникла проблема в аналізі даних по відгукам. Таким чинов переваги

використання аналітики у бізнес процесах є очевидним та вимагає впровадження в практичних цілях

Використання в аналітиці текстових даних дозволяє отримати додаткову інформацію про дані. Замість того, щоб безпосередньо інтегрувати оцінки вільних фраз, ймовірність того, що термін вказує на релевантність, залежить від властивостей контексту терміна. Як наслідок, коли введено термін близькість, додаткові елементи не буде додано до функції ранжирування, а уточнена частина частоти оберненого документа все ще ефективна.

Розділ 3
Розробка системи аналізу на основі відгуків та рейтингу
3.1 Структурні елементи системи аналітики

Що робить хороший ресторан. Які основні проблеми клієнтів для великої їжі. Загальні знання можуть дати загальні відповіді, такі як смачна їжа, чудові послуги або приємні умови, але вони не можуть бути вірними для різних типів ресторанів. У цій роботі збираємося розкрити ці основні функції, що стоять за ресторанами, за допомогою аналізу настроїв на даних Yelp.

У цій роботі провели аналіз даних текстових і виявили кілька цікавих наслідків з відгуків Yelp. Загальна полярність настроїв показала перевагу в сервісі і відгуках, що може вказувати на те, що клієнти "самостійно відбирати" їжу, яка їм подобається.

З іншого боку, могли б повідомити багато цінних ідей, які не можуть бути безпосередньо виявлені на інформаційній панелі веб-сайтів, таких як Yelp. Інформаційна панель Yelp просто показує загальний рейтинг по відношенню до бізнесу, а не кілька рейтингів для різних аспектів бізнесу, в той час як розглядали його на більш дктальному рівні слів.

3.2. Дані та методи

Набір даних походить від онлайнового набору даних Yelp. Задача, що складається з п'яти частин, які надають нам 566 000 основних ділових відомостей (наприклад, годин, адрес, атмосфери), 2,2 мільйона відгуків клієнтів, а також 519 000 порад 552 000 користувачів. Загальний розмір даних становить близько 2.39GB.

Для цього аналізу зосередилися на оглядах ресторанів і використали відгуки клієнтів і дані бізнес-атрибутів. Ці два набори даних обидва у форматі

json. Після фільтрації ресторанів з усіх видів бізнесу, було 1 363 242 відгуки клієнтів, зібрані з 77 445 різних ресторанів. Більшість з цих ресторанів знаходяться в Арізоні, Неваді та Північній Кароліні, а набір даних включає в себе величезну різноманітність типів кухні.

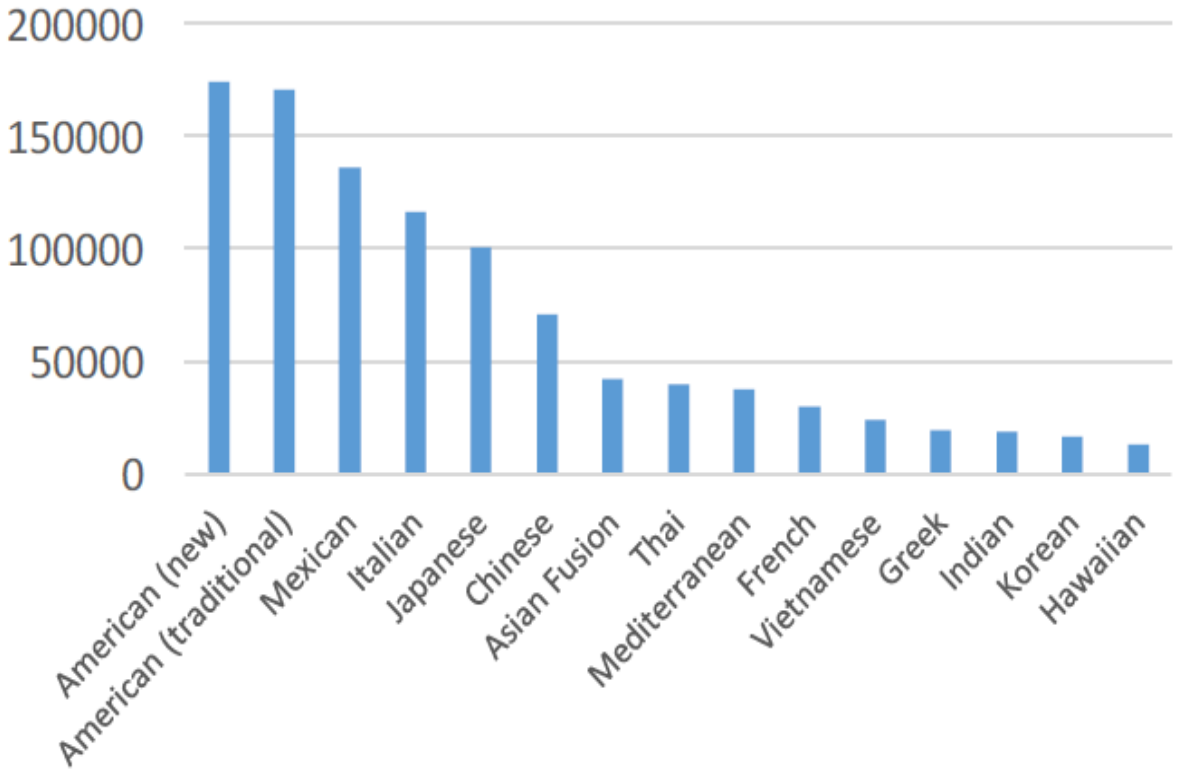


Рисунок 3.1 – Кількість ресторанів за типами [6]

Атрибути в даних рецензування включають ідентифікатор компанії, повну адресу, ціновий діапазон, категорії бізнесу тощо. Атрибути в даних рецензування включають вміст рецензування, рейтинг, ідентифікатор компанії тощо.

\

Таблиця 3.1 - Розподіл ресторанів по штатам

Штат	AZ	NV	NC	Інші
Кількість ресторанів	32615	21233	6162	17435

Атрибути, які використовували, були ідентифікатором компанії, категоріями бізнесу, переглядом вмісту та оцінки. Зокрема, перегляд контенту був корпусом для аналізу; був ідентифікатором дискримінації позитивних або негативних настроїв; бізнес-ідентифікатор служив ключем для очищення даних, а бізнес-категорії служили ключем для групування.

3.3. Очищення даних

Набір бізнес-даних було об'єднано з набором даних відгуків за атрибутом "business id". Після цього слова в кожному огляді були розділені, а пунктуації були видалені так, що для кожного огляду генерувалися «мішок слів». Нарешті, очищали, лемматизували і відфільтрували стоп-слова в кожному мішку слів, використовуючи як вбудований список в пакеті NLTK Python.

Об'єднані дані рецензування були випадково розділені на навчання, перевірку та тестування, встановлені відповідно до співвідношення 3:2:5.

Зокрема, припускали і позначали відгуки з рейтингами, більшими або рівними 4 як позитивні, а решта як "негативні". Таке рішення було прийнято, виходячи з нашого спостереження за розподілом рейтингів.

Методи вилучення характеристик різних типів ресторанів складаються з двох частин. По-перше, модель Support Vector Machine (SVM) була застосована для диференціації позитивних і негативних слів у відгуках, а далі, щоб отримати оцінку слова, щоб зрозуміти, наскільки позитивними або наскільки негативними

були слова. Потім було проаналізовано вплив балів, негативних або позитивних від різних слів в рамках відгуків різної категорії ресторанів.

Застосували два різних методи вибору функцій для SVM: "мішок слів": частоти різних слів з'являлися в кожному огляді і "tf-idf": частота термінів – зворотна статистика частоти документів. Мітки були "позитивними" або "негативними", що відрізнялися залежно від значення рейтингу

Статистика TF-idf, яку будемо використовувати,

$$\text{tfidf}(t,d) = \text{tf}(t,d)\log\frac{N}{|\{d \in D: t \in d\}|}, \tag{3.1}$$

де tfidf t, d є значенням для терміну t в документі d;

tf t, d - це частота терміну t в документі d;

N – це загальна кількість документів;

d ∈ D: t ∈ d – це кількість документів, що містять.

Алгоритм Pegasos використовувався для вирішення SVM, оскільки було доведено високу обчислювальну ефективність. Щоб узгодити позначення, що використовується в документі Pegasos, розглядаємо наступне формулювання об'єктивної функції SVM.

$$\min_{w \in R^n} \frac{\lambda}{2}\|\omega\|^2 + \frac{1}{m}\sum_{i=1}^{m} \max\{0, 1 - y_i \omega^T x_i\} \tag{3.2}$$

де λ є терміном узагальнення;

ω - це вектор, який прагнемо оцінити, вказуючи оцінку для кожного слова;

m - це кількість зразків (оглядів);

yD є 1 або -1, що представляє, якщо відгук позитивний або негативний;

xD є вектором функції кожного огляду.

Pegasos є стохастичним методом субградієнтного спуску з різним розміром кроку. Псевдо код наведено нижче.

Input: $\lambda > 0$. Choose ωI = 0, $t = 0$

While epoch ≤ max_epochs

For $j = 1,\ldots,m$ (рандомно перемішані)

$t = t + 1$

ηP = $1/(t\lambda)$;

If $y\omega_t^T x_j < 1$

$\omega_{t+1} = (1 - \eta_t\lambda)\omega_t + \eta_t y_j x_j$

Else

$\omega_{t+1} = (1 - \eta_t\lambda)\omega_t$

Оскільки кожне слово розглядається як індивідуальна особливість, в моделі буде згенерована розріджена матриця функцій з дуже високими вимірами. Щоб вирішити цю проблему, для кожного рецензування буде налаштовано словник Python замість списку (вектора), який надає інформацію лише для слів, які з'являються в огляді. Цього вдалося уникнути за участю численних нулів у списку. Масивна обчислювальна потужність була збережена, хоча це оптимізація.

Нарешті, були застосовані оцінки слів до набору тестових даних і оцінювалась точність нашого класифікатора.

Обидві моделі з “мішком слів” і функцією tf-idf будуть експериментувати з різними параметрами регулярності. Різні параметри узагальнення були використані для функцій, витягнутих за допомогою мішка слів і методу tf-idf. Проводимо експерименти для двох різних функцій слова, мішок слів, яка є

простим підрахунком кожного слова, і функція tf-idf, яка включає в себе розгляд частоти документів. Відповідно, для найкращої точності перевірки присвоюється параметр регулярності, що призведе до оптимізації моделі. Порівняння моделей буде зроблено, і модель з найкращою продуктивністю тесту буде обрана для реалізації наступного кроку.

Для того, щоб знайти конкретні слова, які були використані для позначення відгуків клієнтів для ресторану, або шляхом просування вперед вивчення унікальної характеристики кожної категорії ресторану, прикметниками, які просто описують полярність настроїв (тобто "добре", "дивно", "жахливо" і т.д.) були знехтувані.

Забезпечуємо для оновленням опису цього відгука повторюваність даних. Щоб отримати "загальну оцінку полярності" для кожного слова, оцінка настроїв кожного слова була помножена на його середню частоту серед усіх відгуків. І так само, щоб отримати «оцінку полярності» (значення, яке відображає полярність настроїв) по відношенню до кожної категорії ресторану, оцінка настроїв кожного слова спочатку була помножена на його частоту, а потім нормалізована на загальну кількість відгуків для конкретної категорії ресторанів.

$$overall\_polarity\_score\ t = \\ = score(t) \times total\_frequency(t)/total\_number\_of\_reviews \qquad (3.3)$$

$$polarity\_score\ t, c = \\ = score(t) \times total\_frequency(t, c)/number\_of\_reviews(c) \qquad (3.4)$$

де overall_polarity_score t є індексом для вимірювання того, наскільки важливим є слово t серед усіх відгуків;

score(t) – це оцінка слова, розрахована за моделлю SVM;

total_frequency (t, c) - це загальна частота слова t у всіх відгуках;

polarity_score (т, с) - є індексом для вимірювання того, наскільки важливим є слово t серед ресторанів типу с;

total_frequency (t, c) - це загальна частота слова t у всіх оглядах ресторанів типу с;

number_of_ відгуками(с) є загальна кількість відгуків про ресторани типу с.

Інтуїтивно, оскільки модель SVM фактично обчислює загальну оцінку для кожного огляду, і ця оцінка певною мірою вказує на те, наскільки клієнт задоволений або незадоволений. Бал полярності, який розрахували, показує, наскільки слово сприяє рахунку всіх ресторанів певного типу. Наприклад, оцінка французьких ресторанів знижується на 0,15 в середньому через "завищені", в той час як знижується лише на 0,02 через "заниження". Тоді можемо стверджувати, що "завищені" незадоволені клієнти набагато більше, ніж "занижені", і, таким чином, "завищені" є більш важливою (негативною) характеристикою французьких ресторанів.

Потім для кожної категорії ресторанів витягуються топові позитивні і негативні слова. Можемо виявити, які особливості для кожного типу і невідповідність цих ресторанів, що забезпечують велику оцінку по всіх даних.

Використовуючи функцію “мішка слів”, а також оцінюючи продуктивність різних лямбда на наборі перевірки, виявили, що результат з найкращою точністю був досягнутий, коли лямбда була встановлена на 0,0003. Відповідна помилка перевірки становить 11,035%.

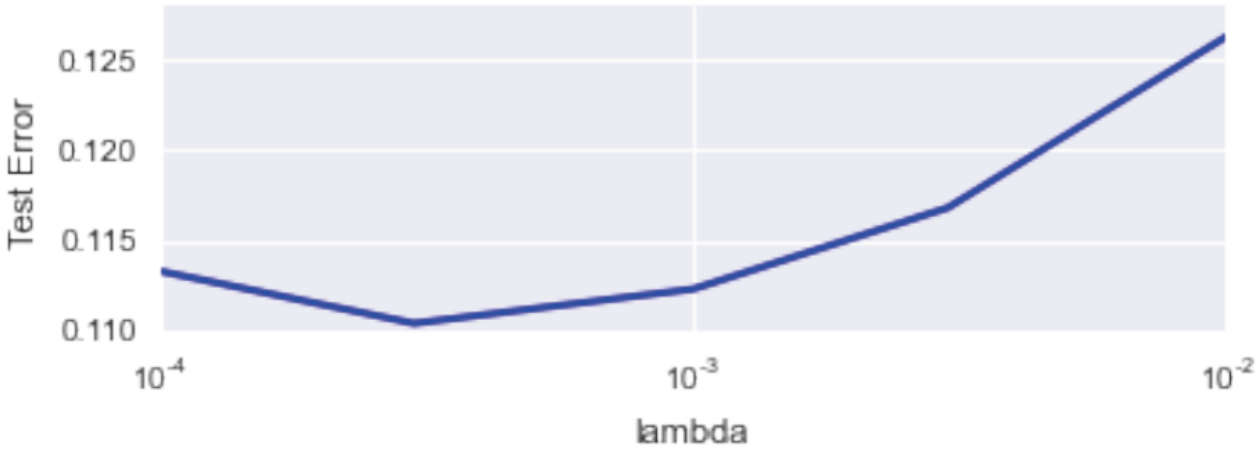


Рисунок 3.2 – Помилка тесту в різних термінах регулярності [5]

Точність класифікатора SVM на тестовому наборі даних становить 88,906% з налаштуванням лямбда до 0,0003. Для іншого методу вибору функцій 'tf-idf', жоден з параметрів узагальнення не призводить до точності понад 88%.

Також бачимо, що після очищення, лемматизації та видалення стоп-слів точність нашої моделі нижча. Наше припущення, що деякі слова зупинки все ще можуть мати тенденцію до настроїв. Ігноруючи ці слова, потенційна інформація була втрачена і призвела до зниження точності. Ця точка зору підтримується в робочі версії.

Несподівано смак страв не займає перше місце серед всіх позитивних відгуків. Натомість обслуговування ресторанів, здається, є пріоритетним для більшості клієнтів, оскільки слово friendly займає перше місце і слово attentive також входить в топ-5. Можна також спостерігати, що коли справа доходить до аромату їжі, клієнти цінують свіжість більше, ніж смачність, а гостра їжа, здається, сподобається багатьом клієнтам.

Таблиця 3.2 - Найвища полярність за типом ресторану (позитивна)

Category	Top 5 Positive Words				
American (New)	friendly	fresh	tasty	attentive	huge
American (Traditional)	friendly	fresh	tasty	attentive	huge
Asian Fusion	friendly	fresh	spicy	tasty	attentive
Chinese	friendly	fresh	tasty	spicy	authentic
French	friendly	fresh	wine	tasty	dessert
Greek	friendly	fresh	gyro	tasty	pita
Hawaiian	friendly	fresh	tasty	spicy	chicken
Indian	friendly	fresh	tasty	spicy	lamb
Italian	friendly	pizza	fresh	tasty	wine
Japanese	fresh	friendly	spicy	tasty	attentive
Korean	friendly	spicy	fresh	kimchi	tasty
Mediterranean	friendly	fresh	tasty	pita	lamb
Mexican	friendly	fresh	tacos	authentic	tasty
Southern	friendly	tasty	fresh	chicken	fried
Thai	spicy	friendly	curry	fresh	tasty
Vietnamese	fresh	friendly	pho	tasty	clean

Таблиця 3.3 - Найвища полярність за типом ресторану (негативна)

Category	Top 5 Negative Words				
American (New)	dry	overpriced	slow	cold	rude
American (Traditional)	dry	slow	cold	rude	overpriced
Asian Fusion	overpriced	dry	slow	salty	rude
Chinese	dry	rude	salty	overpriced	cold
French	dry	overpriced	cold	slow	salty
Greek	dry	slow	rude	cold	overpriced
Hawaiian	dry	salty	rude	cold	slow
Indian	dry	slow	overpriced	rude	cold
Italian	overpriced	dry	rude	slow	cold
Japanese	slow	overpriced	cold	salty	rude
Korean	slow	overpriced	cold	rude	dry
Mediterranean	dry	slow	cold	overpriced	rude
Mexican	dry	slow	overpriced	rude	cold
Southern	dry	slow	cold	overpriced	salty
Thai	dry	slow	overpriced	salty	rude
Vietnamese	rude	dry	slow	cold	salty

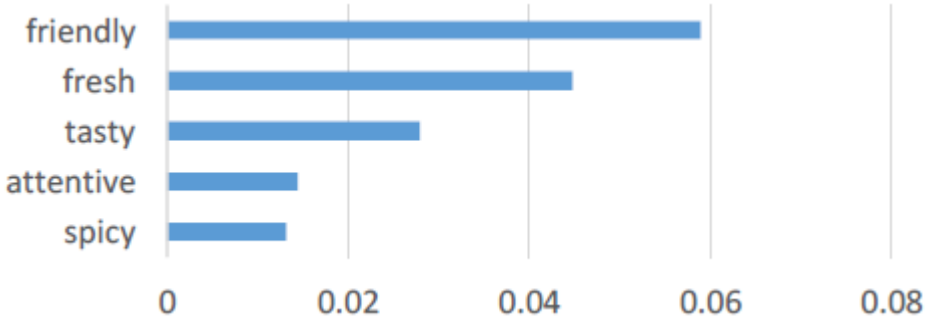


Рисунок 3.3 – Топ-5 позитивних слів всіх типів ресторанів

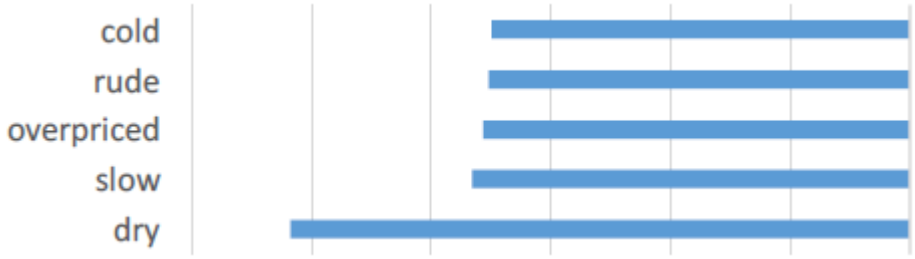


Рисунок 3.4 – Топ-5 негативних слів всіх типів ресторанів

З топ-5 список негативних слів, можна легко спостерігати, що суха їжа дійсно табу, коли справа доходить до якості їжі, так як вона перевершує інших на величезну кількість. Крім того, можна також зробити висновок, що власники ресторанів повинні уникати повільного або грубого обслуговування. Що стосується появи слова холодно, то його значення залишається неоднозначним, оскільки не впевнені, чи відноситься воно до їжі або навколишнього середовища ресторану.

Також були виявлені слова з найвищим рейтингом для різних типів ресторанів що забезпечує базове розуміння характеристик для кожної категорії ресторанів.

З негативного списку слів можна було спостерігати, що завищеня є однією з головних проблем для італійських, французьких і східноазійських ресторанів. Для китайських і в'єтнамських ресторанів грубе ставлення персоналу, швидше за все, буде причиною низької оцінки. З іншого боку, помічаємо, що свіжість перша серед позитивних слів для японської та в'єтнамської їжі. Є також деякі назви страв, присутні в позитивному списку слів, які можуть свідчити про те, що люди віддають перевагу певним ресторанам для своїх конкретних страв, таких як pho у в'єтнамській їжі та піца в італійській їжі.

У цьому розділі розробили ефективну модель SVM для дискримінації позитивних або негативних настроїв на відгуках Yelp з точністю 88,906% на тестовому наборі. Крім збору ключових слів з різних кухонь, модель також може бути використана для автоматичного генерації оцінок для порад (короткі відгуки, які не супроводжуються рейтингами) на Yelp, призначаючи ваги порадам, використовуючи оцінку настроїв слів і, таким чином, даючи більш розумні загальні оцінки для ресторанів.

Виходячи з нашого аналізу, з'ясували, що для більшості типів ресторанів дружні оцінки в першу чергу перед усіма іншими позитивними коментарями, що вказує на те, що сервіс може важити більше, ніж смак, коли люди говорять про ресторан. Також для всіх категорій кухонь, крім японської та в'єтнамської, смак займає перше місце. Одним з можливих пояснень цього збігу є те, що коли клієнти вирішують, де поїсти, вони зазвичай вибирають конкретні види кухні, яким вони віддають перевагу. Такий "самовибір", можливо, призведе до того, що клієнти більше зосередяться на сервісі або середовищі.

Крім того, для різних категорій ресторанів показані різні характеристики. Японська і в'єтнамська їжа отримали позитивні відгуки через свіжість, в той час як корейські та тайські ресторани отримали позитивні відгуки про свою гостру їжу.

Можна також помітити, що більшість азіатських ресторанів вважаються солоними, включаючи китайські, японські, південні, тайські та в'єтнамські. Хоча французька, італійська, японська та корейська їжа вважається високою кухнею, що може мати щось спільного з їх відносно кращим середовищем і сервісом. Навпаки, офіціанти в китайських і в'єтнамських ресторанах, як згадується, грубі, що збігається зі спільною причиною клієнтів, або, можливо, з деякими упередженнями, пов'язаними зі скаргами на китайський ресторан.

Оскільки аналіз може допомогти витягти конкретні функції з будь-якого набору відгуків, власники ресторанів можуть добре використовувати його для важливої інформації, як тільки вони отримали певну кількість відгуків Yelp.

З цих відгуків вони можуть зрозуміти, чому клієнти люблять або не люблять свої ресторани, можуть бути, чудові відгуки в першу чергу через свіжу їжу, або, можливо, незадоволені відгуки, викликані занадто високою ціною. Тим часом вони також можуть порівняти ресторан з аналогічними ресторанами в межах одного типу.

З точки зору клієнтів, більш задовольняючою рекомендацією користувачів і більш важливою оцінкою ресторанів можна очікувати, якщо Yelp включає більше функцій у загальній оцінці кожного ресторану за допомогою техніки, яку вони розробили.

Хоча продуктивність моделі прийнятна, просторів для вдосконалення ще багато. Одна з пропозицій для майбутньої роботи полягає в тому, щоб спробувати інші класифікатори, такі як випадкові лісові або нейронні мережі, перевірити, чи можуть вони перевершити модель SVM.

Однак, оскільки наступним кроком є диференціація впливу різних слів у різноманітних типах ресторанів, лінійний класифікатор, такий як SVM або логістична регресія, буде зручним.

Якщо для класифікації застосовуються методи ансамблю, такі як випадкові лісові або нейронні мережі, вирішення цього питання буде важливим. Для частини вибору функцій можуть бути розглянуті варіанти вимірювання tf-idf або гібридна модель, що має більше притаманних слову значень.

Крім того, враховуючи великий розмір даних, є сенс виконати роботу над рамками великих даних, таких як Spark або повторно зробити навчальний процес більш "оригінальними" мовами, такими як C або Java, щоб підвищити ефективність обчислень.

Висновки до розділу 3

У цій роботі запропонували інноваційний метод визначення різних особливостей для ресторанів різних кухонь. Метод базувався на високотехнійності моделі, розрахунку балів слів і вимірюванні полярності.

Основні функції, які виявили, можуть не тільки допомогти клієнтам вибрати свою улюблену кухню, але і надати ресторанам свої переваги в покрещені роботи.

З іншого боку, подібні процедури можуть бути відтворені для відгуків і коментарів в інших областях, таких як огляди фільмів і публікації в соціальних мережах. Припустимо, хтось хотів би знайти фільм шляхом аналізу операційних настроїв і виявлення полярності на IMDB.

Розділ 4

Дослідження ефективности методів прогнозування

4.1 Чисельні результати проведених досліджень

Дані були сильно перекошені з більшістю ресторанів, що мають менше 1000 відгуків, і деякі з них мають до 10000 відгуків. Для поліпшення інтерпретованості набір даних був скорочений, щоб включати тільки ресторани від 20 до 800 відгуків. Це зменшило розмір до 82% від початкових зібраних даних для остаточного примірки моделі.

Загальний навчальний процес — OLS, LassoCV, RidgeCV або ElasticNetCV.

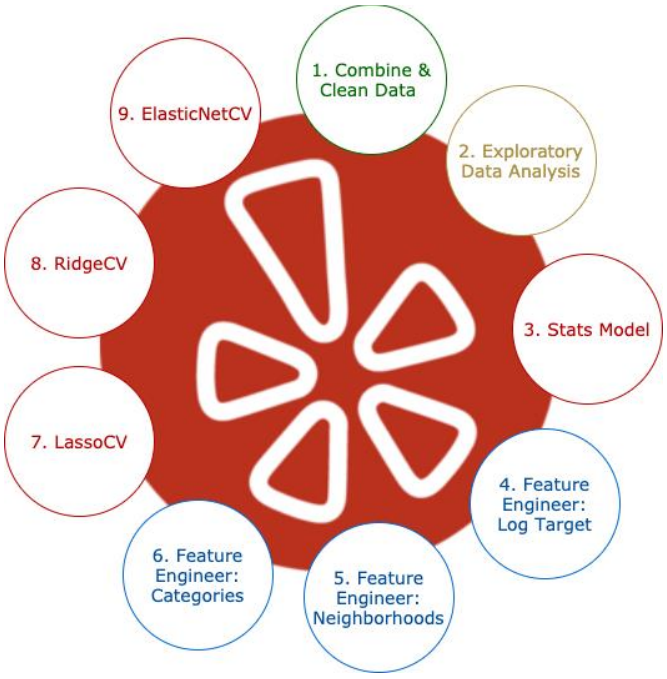


Рисунок 4.1 - Модель навчального процесу [10]

Зображення вище приблизно моделює кроки, зроблені для прогнозування кількості огляду моделі, хоча це було більш ітераційним, ніж лінійним. У двох словах, використано просту лінійна регресія, щоб надати інформацію про

предикторів при виконанні інженерних функцій. Методи регулярності допомогли в управлінні незначучими та коллінарними функціями. Код докладно описаний нижче.

Проста лінійна регресія з стандартизацією функцій дала уявлення під час інженерії процесу функції. Оскільки будемо використовувати це часто, було сенс створити функцію:

```
def split_and_validate(X, y):
    """
    Split data to train, val and test set and perform linear   regression
    """
    columns = X.columns
    X = X.values
    std = StandardScaler()

    # perform a train, val and test set
    X_train_val, X_test, y_train_val, y_test = train_test_split(X, y,
test_size=0.2,random_state=42)
    X_train, X_val, y_train, y_val = train_test_split(X_train_val, y_train_val,
test_size=.25, random_state=42)

    # fit linear regression to training data
    lr_model = LinearRegression()
    lr_model.fit(std.fit_transform(X_train), y_train)    # score fit model on
validation data
    val_score = lr_model.score(std.transform(X_val), y_val)
    adjusted_r_squared = 1 - (1-val_score)*(len(y)-1)/(len(y)-X.shape[1]-1)

    # report results
    print('\nValidation R^2 score was:', val_score)
    print('Validation R^2 adj score was:', adjusted_r_squared)
```

LassoCV (Lasso з K-fold перевіркою) з стандартизацією функцій як метод вибору функцій. Лассо обнуляє коефіцієнти, які не мають сильної прогностичної сили, отже, зосереджуючи інтерпретацію на кількох ключових особливостях.

```
std = StandardScaler()# Splitting data to train and val
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2,
random_state=42)

# Setting up alpha list and letting the model select the best alpha
alphalist = 10**(np.linspace(-2,2,200))
lasso_model = LassoCV(alphas = alphalist, cv=5) # setting K-fold
lasso_model.fit(std.fit_transform(X_train), y_train)# printing coefficients
list(zip(X_train.columns, lasso_model.coef_ / std.scale_))
```

Ridge (Ridge з K-fold перевіркою) допомагає згладити коефіцієнти, наближаючи коефіцієнти до нуля, не видаляючи їх. Але що ще важливіше, вони надають приблизно однакову вагу двом високо коллінарним особливостям.

```
std = StandardScaler()# Split data to train and test
X_train, X_test, y_train, y_test = train_test_split(X7, y7, test_size=0.2,
random_state=42)# Fit and train model
alphalist = 10**(np.linspace(-1,2,200))
ridge_model = RidgeCV(alphas = alphalist, cv=5)
std.fit(X_train)
ridge_model.fit(std.transform(X_train), y_train)# print coefficients
list(zip(X_train.columns, ridge_model.coef_))
```

ElasticNetCV (Ridge з перевіркою K-fold) лінійно поєднує в собі штрафи L1 і L2 методів Лассо і Рідж, як зазначено вище. Він вирішує деякі обмеження обох методів. Ця модель була остаточно обрана, з більш високою вагою по відношенню до моделі Ridge.

```
std = StandardScaler()l1_ratios = [.1, .5, .7, .9, .95, .99, 1]
alphas = 10**np.linspace(-2,2, 200)model = ElasticNetCV(cv=5, l1_ratio = l1_ratios,
alphas = alphas, random_state=42)
std.fit(X_train)
model.fit(std.transform(X_train), y_train)
pred = model.predict(std.transform(X_test))
```

```
score = model.score(std.transform(X_test), y_test)
model.alpha_mse = mean_squared_error(np.exp(y_test), np.exp(pred))adj_r_squared = 1 -
(1-score)*(len(y_test)-1)/(len(y7)-X7.shape[1]-1)print("Alpha:{0:.4f},
R2:{1:.4f},ajd_R2{2: 4f}, MSE:{2:.2f}, RMSE:{3:.2f}"
   .format(model.alpha_, score,adj_r_squared, mse, np.sqrt(mse)))print(f'l1 ratio is
{model.l1_ratio_}')
list(zip(X_train.columns, model.coef_ / std.scale_))
```

У цьому розділі аналізуються лише елементи відгуків, які можемо отримати від моделі, однак опишемо структурування блоків споріжнених страв в ресторанах.

Які страви повинні служити в ресторані і ресторан може мати більше однієї категорії продуктів харчування, для маркування Yelp. В рамках 1165 ресторанів, які були проаналізовані, було 146 різних кухонь.

Використовуючи хмару Word Python, можете побачити, що кілька категорій з'являються багато разів (мексиканські ресторани з'являються принаймні 135 разів), тоді як багато кухонь здаються настільки екзотичними, наприклад, Live / Raw або малайзійська їжа з'являються лише один раз.

Рисунок 4.2 - Word Cloud (хмара слів)

Word Cloud візуалізує частоту категорій продуктів харчування в наборі даних.

Було два способи включити кухні в якості особливостей. Одна особливість: Top N Кухні (наприклад, Топ 5, 10 кухонь або топ-20 кухонь) як фіктивна змінна

Багато особливостей: Кожна їжа є фіктивною змінною сама по собі з до N фіктивних змінних.

4.2 Кориговані оцінки розподілів

Як визначити N: запустити модель перехресної перевірки для кожного N. Ліве зображення під графіками кожного N для одного методу функції та правого зображення графіків кожного N для методу багатьох функцій. Використовуючи цю візуалізацію, вибрано функцій методу створення 21 маркерної змінної для кожної з популярних кухонь. Це все ще багато функцій, але оскільки був під керуванням моделі Лассо, очікували видалити деякі з цих 21 кухні.

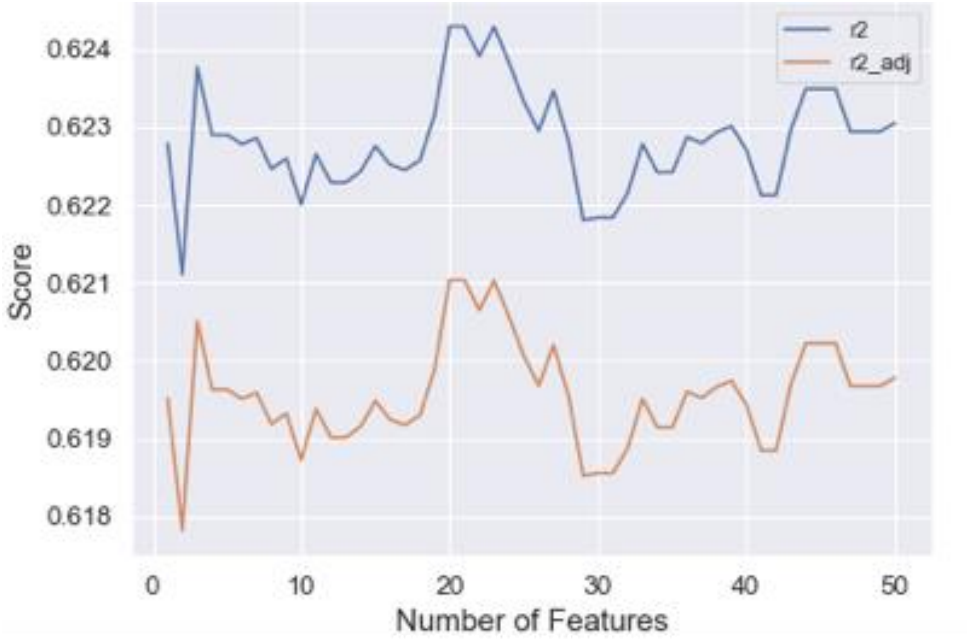


Рисунок 4.3 – r2 та r2 скориговані оцінки за x найпопулярнішими кухнями

У фінальній моделі було відібрано 10 кухонь. Зверніть увагу на значення коефіцієнта: оскільки цільова змінна була зареєстрована, можна прочитати значення як відсоткову зміну в кількості відгуків, якщо ресторан включив кухню як категорію на Yelp.

Мексиканські, американські, бутерброди і фаст-фуд були негативними предикторами, а решта були позитивними предикторами.

Не дивно, що визначення ресторану як фаст-фуд призведе до зниження прогнозованих показників огляду на 11,9%. Однак було виявлено, що мексиканська їжа призведе до зниження кількості відгуків на 9,7% - в місті ніколи не може бути достатньо мексиканських ресторанів.

Рисунок 4.4 – r2 та r2 скориговані оцінки за особливостями їжі

Рішення про те, як включити N кухні в якості предикторів в моделі. Використовуючи ту ж техніку, що і вище в Кухнях, щоб визначити, чи мають значення околиці значення елемента даних, було встановлено, що це зовсім не поліпшить R-квадрат моделей. Хоча нульова гіпотеза не була відкинута,

Чи має значення їжа або атмосфера - тест Instagram набір фотоданих Yelp містить інформацію про фото етикетку з наступними мітками:

- продовольство;
- пити;
- всередині;
- за межами;
- меню;

Cuisine	Coefficient Values
Mexican	-9.7%
Breakfast and Brunch	1.9%
American (New)	-1%
Sandwiches	-5.7%
Sushi Bars	2.9%
Fast Food	-11.9%
Burgers	1.4%
Thai	5.3%
Barbeque	4%
Buffets	6%

Рисунок 4.5 – Частка виду кухні

Ці етикетки корисні як індикатор, щоб дізнатися, чи естетика їжі / напоїв або навколишнього середовища (всередині / зовні етикетки) має значення більше для успіху ресторану. Встановили цей показник для ресторанів з більш ніж 10 фотографіями заради актуальності (ресторан з 20 фотографіями навколишнього середовища і 10 фото їжі передає більш сильну інформацію про естетику, ніж ресторан з 2 фотографіями навколишнього середовища і 1 фото їжі).

Загалом у ресторанах розміщено більше фотографій їжі, ніж фотографій навколишнього середовища:

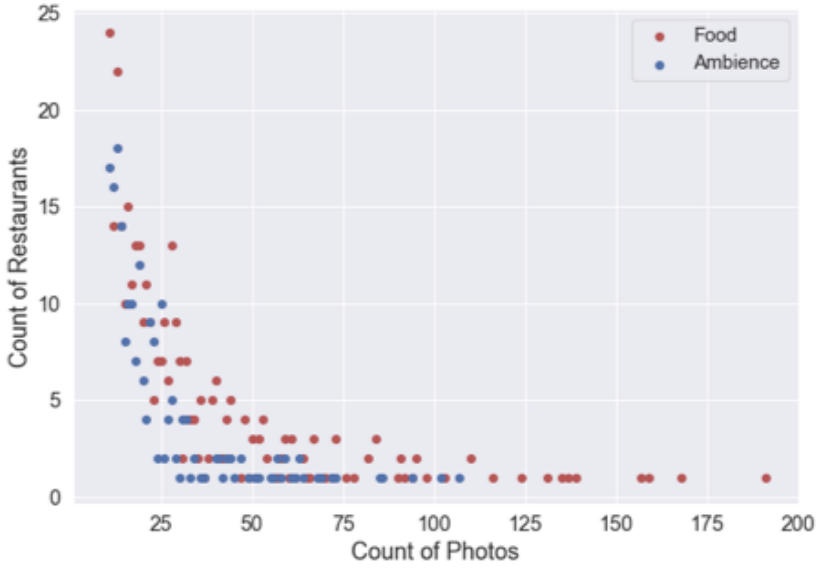


Рисунок 4.6 – Розподіл фотографій по ресторанам

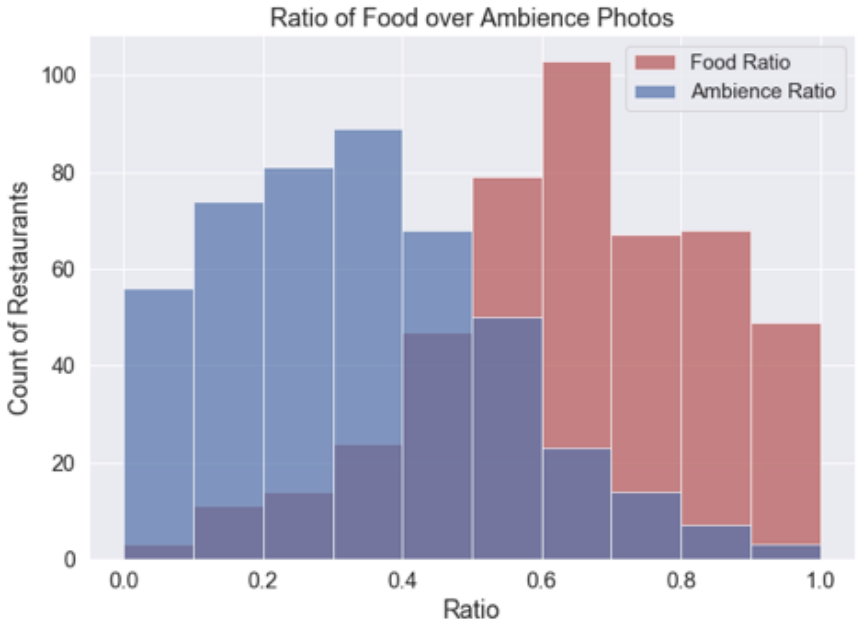


Рисунок 4.7 – Співводношення частки їжі відносно оточення

Отже, створюю фіктивну змінну під назвою food 70, щоб вказати, чи є в ресторані співвідношення їжі / напоїв принаймні 70:100.

Наступною особливістю є кількість фотографій. Хоча може бути зворотній зв'язок між кількістю фотографій і кількість оглядів (більш популярні ресторани мають більше фотографій), вважав, що буде більш слабкий зворотний зв'язок цикл для ресторанів з меншою кількістю фотографій (менше 10).

Отже, було створено три фіктивні змінні: ресторани з нульовими фотографіями, 1–5 фотографій та 6–10 фотографій.

Перш ніж кидати функції в модель, деякі прості Exploratory Analysis показує, що ресторани з більшою кількістю відгуків мають, в середньому, більш збалансовану їжу: співвідношення атмосфери (близько до 1.5:1, ніж 3:0 для ресторанів з менш ніж 50 відгуками).

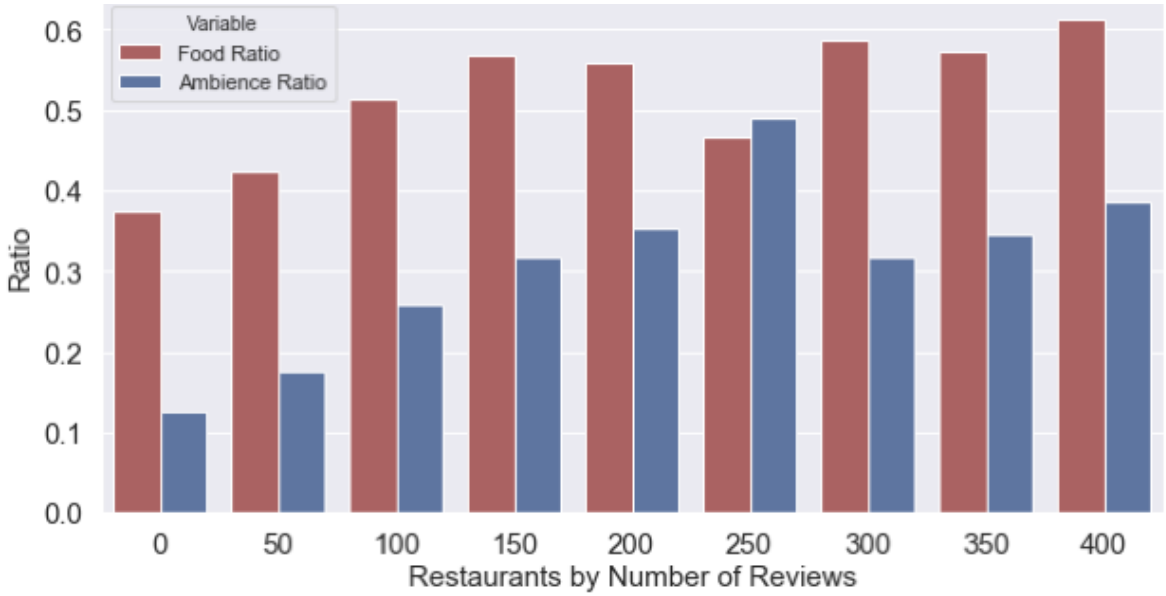


Рисунок 4.8 – Співвідношення їжі та оточення у відгуках ресторанів

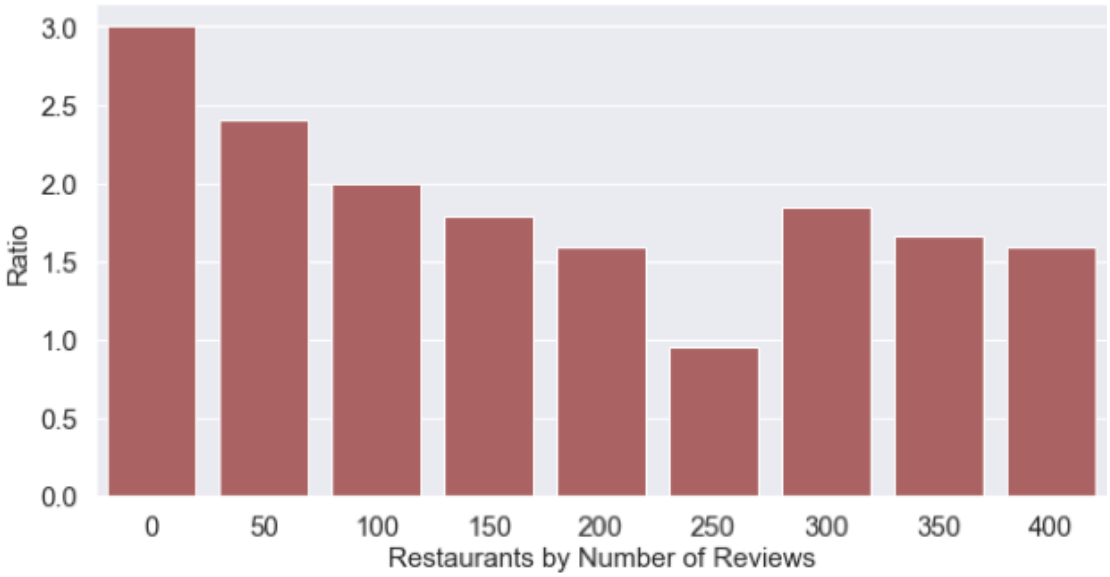


Рисунок 4.9 – Співводношення оточення і відгуках ресторанів

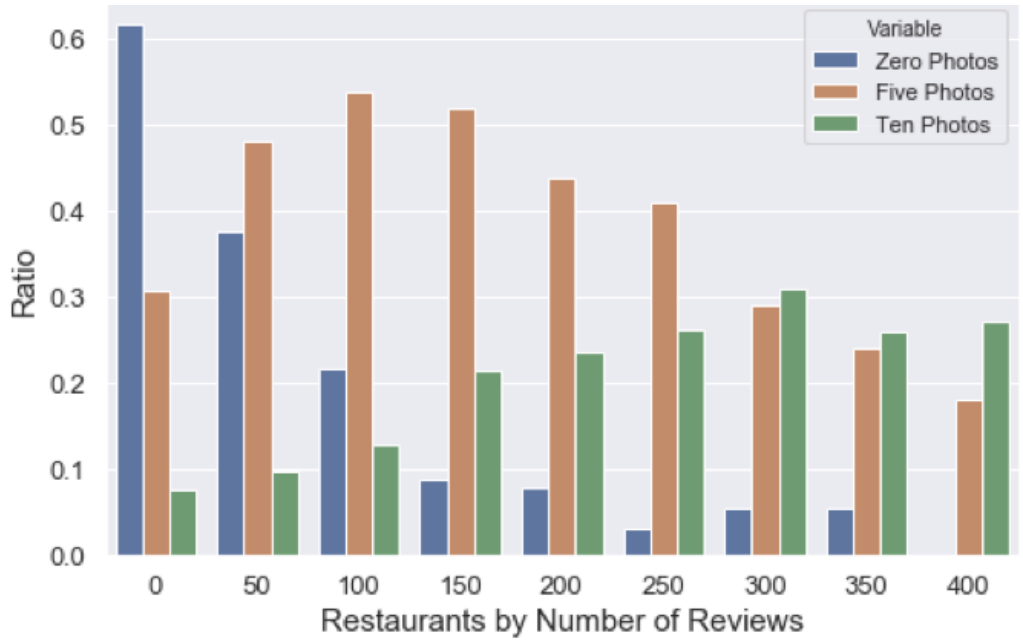


Рисунок 4.10 – Співводношення кількості фото

Якщо у ресторані немає фотографій, шанси мати менше відгуків значно збільшуються.

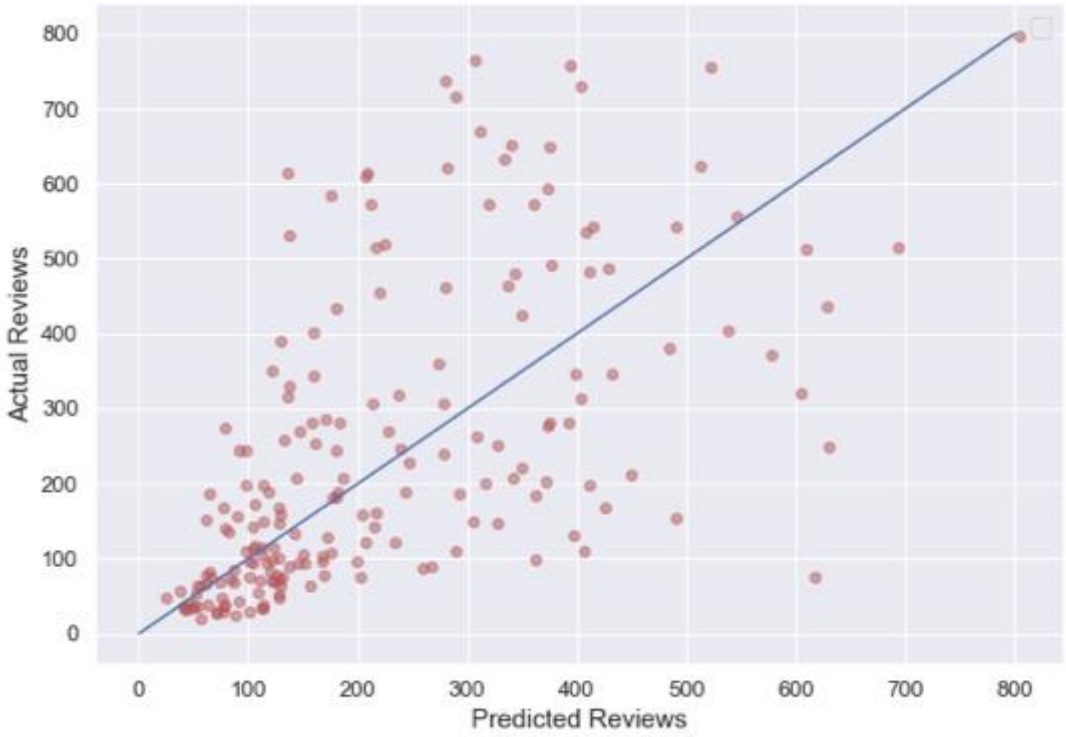


Рисунок 4.11 – Дійсні та прогнозовані відгуки

Coefficients	Beta Values (% impact to Reviews)
Price Range	8.5%
Ratings	27.9%
Zero Photos	-133.9%
One to Five Photos	-96.6%
Six to 10 Photos	-54.0%
More than 10 photos and Food	-21.6%
Useful ratio	-16.1%
Mexican Restaurant	-26.3%
Breakfast & Brunch	6.9%
American (New)	-3.4%
Sandwiches	-20.1%
Sushi Bars	13.4%
Fast Food	-37.2%
Burgers	4.5%
Thai	27.9%
Barbeque	20.6%
Buffets	34.4%
Mean Text * Std Dev	113.2%

Рисунок 4.12 – Розподіл то їжі

Не дивно, що наявність нульових фотографій не допоможе в отриманні більше, ніж навіть 50 відгуків.

Прогнозовані та фактичні відгуки. Використано R Square, щоб вибрати свою остаточну модель і досяг 0,50 балів. При побудові прогнозованих відгуків проти фактичних відгуків, модель все ще робить погану роботу в прогнозуванні ресторанів з оглядом розраховувати на більш високий діапазон.

Огляд відрізняється від проблеми аналізу настроїв прогнозування того, позитивний чи негативний відгук, в якому дві полярності (позитивні та негативні) є незалежними, рейтинги є реальними цифрами, і вони пов'язані один з одним. Наприклад, 5-зірок більш позитивні, ніж 4-зірки, а також 5-зірок ближче до 4-зірок, ніж 3-зірок. Тому відношення до цієї проблеми як до проблеми регресії і використано лінійний алгоритм регресії для прогнозування оцінок. Мета полягає в тому, щоб вивчити функцію *f(x)*, яка співставить вхідні функції *x* з числовим рейтингом *y* і мінімізувати різницю між прогнозованим рейтингом та істинним рейтингом *y*. Продуктивність оцінюється як середня абсолютна помилка (MAE) і середня квадратична помилка (MSE).

Є чотири моделі: модель сентименту, тематична модель, сентимент і тема комбінованої моделі і остання, модель, яка включає в себе всі особливості. Як ілюстровано раніше, огляд настроїв зазвичай погоджується з рейтингами. Тому базова модель настроїв, заснована на лексиконі думки. Тематична модель реалізована в LDA. Щоб вибрати номери тем, моделі LDA з різними номерами тем навчаються та перевіряються на наборі даних перевірки для налаштування параметрів. Взаємодія між моделлю настроїв та моделлю теми в комбінованій моделі досягається елементарно множення між оцінками настроїв і розподілами тем. Остання модель - це повноправна модель, яка включає в себе всі функції в попередніх трьох моделях.

4.3 Базовий план: модель настроїв (лексикон думки)

Рейтинги є приблизною функцією прикметників позитивних і негативних конотацій: огляд з переважно позитивними словами, швидше за все, буде рейтингом 4- або 5 зірок, а огляд з рівною кількістю позитивних і негативних прикметників, швидше за все, буде рейтингом 3 зірок. Базова модель настроїв слідує аналогічній ідеї.

Реалізовано список англійських позитивних і негативних слів або сентиментальних слів, загальна кількість яких становить близько 6800 слів. Всі слова на думку lexicon файл є нижнім регістром, отже, всі слова огляду перетворюються в нижній регістр в попередній обробці. Після цього кількість позитивних і негативних слів лексикону враховується для кожного огляду, і ці два числа, які розглядаються як позитивні та негативні оцінки для текстового огляду, використовуються як функції для прогнозування оцінок рецензування.

4.4 Модель теми для використаннм латентного розміщення Діріхте

LDA (*Latent Dirichlet allocation*) є однією з найпопулярніших тематичних моделей, заснованих на припущенні, що документи є сумішшю тем, де тема є ймовірністю розподілу слів. LDA має кращу статистичну основу, визначаючи розподіл тематичних документів, що дозволяє випускати новий документ на основі раніше оціненої моделі та уникає проблеми переоснащення. LDA вибирається як алгоритм моделювання тем завдяки своїй популярності, а також перспективній продуктивності.

LDA дуже залежить від текстів, тому відгуки попередньо обробляються перед тим, як вписуватися в LDA, щоб отримати тематичні моделі. Стоп слова, пунктуації та всі інші символи, крім алфавітів і чисел, видаляються. Крім того, всі слова перетворюються в нижній регістр і застосовується також.

Після встановлення оброблених відгуків у навчальних даних в LDA, тематична модель навчена. Ця модель теми потім застосовується назад до відгуків у навчальних даних і отримує їх розподіли тем. Ці теми ймовірності використовуються як функції для прогнозування рейтингу та внасяться в лінійний алгоритм регресії. Навчена модель лінійної регресії потім застосовується до невидимих даних для оцінки ефективності.

Щоб вибрати перспективний номер теми цього набору даних yelp review, 5 моделей LDA навчаються з 5, 10, 20, 30 та 40 темами відповідно. Їх виступи перевіряються на наборі даних перевірки, і вибрано найкращу модель продуктивності. Модель LDA реалізована в пакеті Python genism7, який є широко використовуваним пакетом моделювання тем.

4.5 Комбіновані настрої та модель теми LDA

В останніх двох моделях спостерігається, як в цілому настрої та теми огляду є впливовими для прогнозування рейтингу. Однак, може бути кілька тем, які їхні настрої важливіші за інших, коли клієнти призначають оцінку. Наприклад, в огляді клієнти можуть написати багато позитивних слів, що описують хороший сервіс і красиву прикрасу, що робить цей огляд схожим на 5-зірковий огляд. Тим не менш, це може бути в кінцевому підсумку з 3-зірковим оглядом, оскільки "їжа", про яку клієнти більше пише відгук, негативна. Тому в цій моделі спостерігаю, як тематичні настрої впливають на прогнозування рейтингу.

Отримуємо тематичні настрої, помножуючи оцінки настроїв з тематичними розподілами. Зокрема, робимо елементарне множення між оцінками настроїв і розподілами тем і використовуємо ці цифри як функції. Наприклад, якщо відгук має (5, 3) як позитивні та негативні оцінки і має (0.2, 0.4,

0.4) як свої тематичні розподіли, функції для цієї взаємодії моделі будуть 5*0.2, 5*0.4, 5*0.4, 3*0.2, 3*0.4, 3*0.4, які в цілому 6 числа.

Для цієї моделі використовуються всі функції з останніх трьох моделей, які включають оцінки настроїв (позитивні та негативні), розподіл тем, множення показників настроїв та розподілів тем.

Продуктивність оцінюється в середній абсолютній помилці, яка є середнім абсолютної похибки, і середній квадрат помилки, яка є середнім квадратів помилок, "відстань" між оцінювачем і його істинним значенням.

Крім того, використовую R2, коефіцієнт визначення, щоб виміряти, наскільки добре модель відповідає даним. Нижче MAE і MSE вказують на меншу похибку, а вищий R2 вказує на кращу придатність.

Модель базових настроїв MAE дорівнює 0,99, що означає перебір, абсолютне значення різниці між прогнозованим рейтингом і реальним рейтингом становить 0,99. MSE дорівнює 1,38, що означає, що в середньому квадратна різниця між прогнозованим рейтингом і реальним рейтингом дорівнює 1,38. R2 дорівнює 0,31, що означає, що 31% очок підпадають під лінію регресії.

Для всіх п'яти тематичних моделей, з різними номерами тем, отримуємо топ-10 тем (топ-5 тем для моделі 5 тем) і для кожної теми отримуємо топ-5 тем термінів і їх ймовірності. Терміни відображаються через дані. Багато продуктів з'являються в навчених темах, таких як "піца", "суші" і "курка". Крім того, деякі прикметники з'ються, як "добре" і "щасливий". Деякі теми мають сенс, наприклад п'ятої теми в 10 темах моделі: "сендвіч" + "сніданок" + "кава" + "яйце" + "французька", що, ймовірно, є темою для сніданку.

4.6 LDA для прогнозування рейтингу

Три моделі, які включають LDA (тематична модель, тема і настрої комбінованої моделі і всі функції моделі) навчаються п'ять разів з різними номерами тем. Продуктивність усіх моделей на наборі даних перевірки відображається в даних. Усі моделі досягають найнижчого рівня MAE та MSE (найкраща продуктивність) за допомогою 40 тем. Для всіх номерів тем модель "Усі функції" має найкращу продуктивність. Крім того, хоча абсолютні помилки моделі LDA і Lexicon * LDA в кожному номері теми близькі один до одного, різниця MAE між моделлю "Всі функції" та двома іншими моделями є більш значущою. Модель LDA має найбільше покращення продуктивності від моделі 5 тем до моделі 40 тем, виміряної як в MAE, так і в MSE. Як правило, від 5 тем до 40 тем, MAE і MSE зменшення (за винятком MAE з 20 тем LDA модель) і R^2 збільшується в міру збільшення числа теми. Сорок тем модель «Всі особливості» має найкращу продуктивність у всій таблиці, яка становить 0,80 MAE, 1.01 MSE і 0.50 R^2.

Найкращий номер теми для кожної моделі вибирається для всіх моделей, це трапляється однакове число, 40. Після вибору значення номера теми всі моделі перевіряються на тестовому наборі даних для оцінки їх загальної ефективності та результат відображається. Всі моделі, які беруть участь LDA мають кращу продуктивність моделі лексикону базових настроїв та моделі «Усі функції» мають найкращі продуктивність з 0.80 MAE і 0.98 MSE, що набагато краще, ніж базовий. R^2 0,51 вказує на те, що більше половини всіх пунктів підпадають під регресійну лінію.

Найкращий виконаний номер теми для кожної моделі вибирається з таблиці 7 і для всіх моделей це трапляється однакове число, 40. Після вибору значення номера теми всі моделі перевіряються на тестовому наборі даних, щоб оцінити їх загальну продуктивність. Всі моделі, які використовували LDA,

мають кращу продуктивність, ніж базова модель лексикону настроїв і модель "усі особливості", мають найкращу продуктивність з 0.80 MAE і 0.98 MSE, що набагато краще, ніж базовий. R^2 значення 0,51 вказує на те, що більше половини всіх точок підпадають під лінію навченої регресії.

Питання дослідження в полягає в тому, чи працюють тематичні моделі, на додаток до настроїв, прогнозування рейтингу вигоди і чотири моделі, щоб відповісти на це питання. Як бачимо, всі моделі, які залучали LDA, навіть саму LDA, мають кращу продуктивність, ніж використання лише лексикону настроїв. Можна сказати, що тематичні моделі дійсно сприяють прогнозу рейтингу. Модель LDA перевершує як індивідуальну модель Lexicon, так і LDA, тому тематичні настрої, як очікувано, є кращим показником для рейтингів. Щоб додатково інтерпретувати результат, випадковим чином вибираю п'ять прикладів з тестового набору даних і отримуємо прогнозовані оцінки з чотирьох моделей.

Висновки до розділу 4

Ціновий діапазон має значення для більш дорогостоячих продуктів. Наявність фото, маючи нуль або менше 6 фотографій були основними негативними предикторами популярності.

Співвідношення "їжа:атмосфера": ресторани,як правило, мають більше їжі, ніж фотографії навколишнього середовища. Але щоб ресторан мав принаймні 30% ваги в атмосфері (хороший декор має значення.)

З огляду на оцінку R^2 0.5, лінійна регресія може бути не найкращою моделлю для прогнозування кількості відгуків. Однак модель дійсно дає сильні пропозиції про те, що ресторани повинні враховувати при відкритті в новому місті.

Загальні висновки

Дослідження показало, що рейтинговий висновок є правдоподібним завданням, потенційні перешкоди існують у вивченні взаємозв'язку між рейтингами та відгуками. По-перше, є непослідовність при присвоєнні рейтингів серед авторів, дивергенція крос-автора. Це може легко уявити, оскільки, думки суб'єктивна річ. Для того ж рейтингу, це загальноприйнято, щоб один відгук з'являється дуже позитивним, а інший менш позитивним або навіть трохи негативним. По-друге, оцінки не повністю підтримуються текстом. Коли людей просять призначити оцінку для продукту або послуги, вона зазвичай представляє їх загальне враження, тоді як те, що вони пишуть в огляді, може мати просто найбільш позитивні частини, хороші або погані. Насправді додавання даних персоналізації може бути корисним у покращенні продуктивності.

Навчені тематичні моделі, відрізняються від того, що очікували. Спочатку очікувані теми схожі на "сервіс", "якість їжі" тощо. Тим не менш, теми, навчені більше схожі на різні види їжі, такі як "барбекю", "шашлик" і так далі. Однією з потенційних причин є набір даних. Незважаючи на те, що звузили набір даних до категорії "ресторан", він все ще занадто широкий і грубий, оскільки є ще 116 категорій в категорії "ресторан". Різні категорії, ймовірно, використовувати різні терміни, навіть якщо всі вони ресторани. Подальше звуження категорії, моделі тем LDA, моделюють різні аспекти огляду ресторану (наприклад, якість їжі, сервіс і так далі), а не різні типи кухонь.

Перелік посилань

1. Dos Santos, C. N., & Gatti, M. (2014). Deep Convolutional Neural Networks for Sentiment Analysis of Short Texts. In COLING (pp. 69-78).

2. Mullen, T., & Collier, N. (2004, July). Sentiment Analysis using Support Vector Machines with Diverse Information Sources. In EMNLP (Vol. 4, pp. 412-418).

3. Kiritchenko, S., Zhu, X., Cherry, C., & Mohammad, S. (2014, August). NRC-Canada-2014: Detecting aspects and sentiment in customer reviews. In Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014) (pp. 437-442). Dublin, Ireland: Association for Computational Linguistics and Dublin City University.

4. Huang, J., Rogers, S., & Joo, E. (2014). Improving restaurants by extracting subtopics from yelp reviews. iConference 2014 (Social Media Expo).

5. Shalev-Shwartz, S., Singer, Y., Srebro, N., & Cotter, A. (2011). Pegasos: Primal estimated sub-gradient solver for svm. Mathematical programming, 127(1), 3-30.

6. Byers, J. W., Mitzenmacher, M., & Zervas, G. (2012, June). The groupon effect on yelp ratings: a root cause analysis. In Proceedings of the 13th ACM conference on electronic commerce (pp. 248-265). ACM.

7. Hicks, A., Comp, S., Horovitz, J., Hovarter, M., Miki, M., & Bevan, J. L. (2012). Why people use Yelp. com: An exploration of uses and gratifications. Computers in Human Behavior, 28(6), 2274-2279.

8. Lin, C., & He, Y. (2009, November). Joint sentiment/topic model for sentiment analysis. In Proceedings of the 18th ACM conference on Information and knowledge management (pp. 375-384). ACM.

9. Linshi, J. (2014). Personalizing Yelp star ratings: A semantic topic modeling approach. Yale University.

10. Manke, S. N., & Shivale, N. (2015). A Review on: Opinion Mining and Sentiment Analysis based on Natural Language Processing. International Jour

11. Mei, Q., Ling, X., Wondra, M., Su, H., & Zhai, C. (2007, May). Topic sentiment mixture: modeling facets and opinions in weblogs. In Proceedings of the 16th international conference on World Wide Web (pp. 171-180). ACM.

12. Hoffman, M., Bach, F. R., & Blei, D. M. (2010). Online learning for latent dirichlet allocation. In advances in neural information processing systems (pp. 856-864).

13. Hu, L., Sun, A., & Liu, Y. (2014, July). Your neighbors affect your ratings: on geographical neighborhood influence to rating prediction. In Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval (pp. 345-354). ACM.

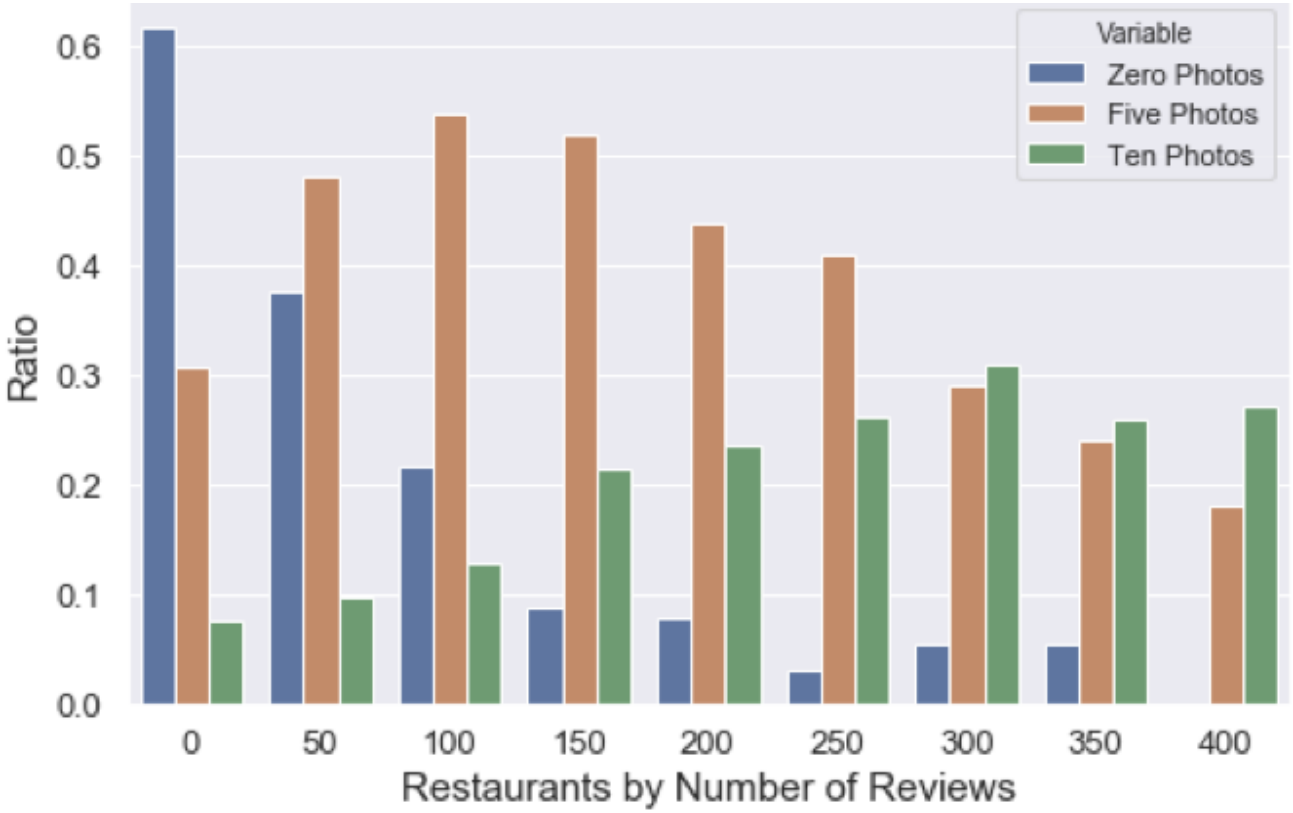
14. Hu, M., & Liu, B. (2004, August). Mining and summarizing customer reviews. In Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 168-177). ACM.

15. Huang, J., Rogers, S., & Joo, E. (2014). Improving restaurants by extracting subtopics from yelp reviews. iConference 2014 (Social Media Expo).

16. Kummer, O., Savoy, J., & Argand, R. E. (2012). Feature selection in sentiment analysis.

17. Li, W., & McCallum, A. (2006, June). Pachinko allocation: DAG-structured mixture models of topic correlations. In Proceedings of the 23rd international conference on Machine learning (pp. 577-584). ACM.

Додатки


Variable
Zero Photos
Five Photos
Ten Photos
Ratio
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0
50
100
150
200
250
300
350
400
Restaurants by Number of Reviews

УДК 004.4

Тіторов І. Д., Скрипник Т. К.

Хмельницький національний університет

АНАЛІТИЧНА СИСТЕМА РЕКОМЕНДАЦІЙ ЗАКЛАДІВ ХАРЧУВАННЯ НА ОСНОВІ ВІДГУКІВ ТА РЕЙТИНГУ

Розроблено та набуло практичної реалізації системного підходу щодо аналізу відгуків та рейтингу в системі закладів громадського харчування. Дослідження показало, що рейтинговий висновок є правдоподібним завданням, потенційні перешкоди існують у вивченні взаємозв'язку між рейтингами та відгуками. Є непослідовність при присвоєнні рейтингів серед авторів, дивергенція крос-автора. Це може легко уявити, як всі знаємо, думки така суб'єктивна річ. Для того ж рейтингу, це загальноприйнято, щоб один відгук з'являється дуже позитивним, а інший менш позитивним або навіть трохи негативним.

Developed and acquired a practical implementation of a systematic approach to the analysis of feedback and rating in the system of catering establishments. The study showed that the rating conclusion is a plausible task, there are potential obstacles in studying the relationship between ratings and reviews. There is inconsistency in assigning ratings among authors, cross-author divergence. It can easily imagine, as we all know, opinions such a subjective thing. For the same rating, it is common for one review to appear very positive and the other less positive or even slightly negative.

У минулому бізнес-аналітика була привілеєм великих компаній, які могли дозволити собі підтримувати команди IT-фахівців і вчених з обробки даних. Але в останнє десятиліття, оскільки технологія швидко розвивалася, програмне забезпечення стало не тільки більш легким і потужним, але і більш доступним. Малий бізнес може використовувати ті ж інструменти, що і основні гравці ринку, і стикатися зі своїми конкурентами. Нові інструменти самообслуговування доводять, що бізнес-аналітика не ракетобудування, а скоріше корисний інструмент, який допоможе перетворити дані в обґрунтовані рішення. Тепер кожна компанія може використовувати силу сучасного програмного забезпечення , щоб підняти свою нижню лінію, оскільки бізнес-аналітика для малого бізнесу стала доступною і доступною.

LDA (Latent Dirichlet allocation) є однією з найпопулярніших тематичних моделей, заснованих на припущенні, що документи є сумішшю тем, де тема є ймовірністю розподілу слів. LDA має кращу статистичну основу, визначаючи розподіл тематичних документів, що дозволяє випускати новий документ на основі

раніше оціненої моделі та уникає проблеми переоснащення [15]. У цій роботі LDA вибирається як алгоритм моделювання тем завдяки своїй популярності, а також перспективній продуктивності.

Навчені тематичні моделі, відрізняються від того, що очікував. Спочатку очікувані теми схожі на "сервіс", "якість їжі" тощо. Тим не менш, теми, навчені більше схожі на різні види їжі, такі як "барбекю", "тайський" і так далі. Однією з потенційних причин є набір даних. Незважаючи на те, що звузив набір даних до категорії "ресторан", він все ще занадто широкий і грубий, оскільки є ще 116 категорій в категорії "ресторан". Різні категорії, ймовірно, використовувати різні терміни, навіть якщо всі вони ресторани. Подальше звуження категорії, моделі тем LDA, як очікується, моделюють різні аспекти огляду ресторану (наприклад, якість їжі, сервіс і так далі), а не різні типи кухонь.

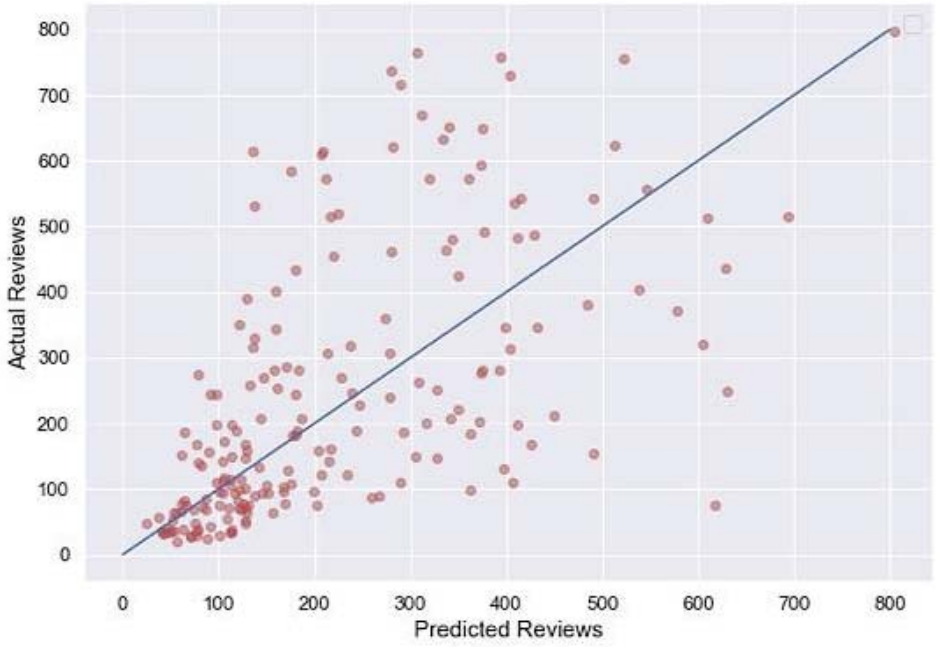


Рисунок 1 – Дійсні та прогнозовані відгуки

З негативного списку слів можна було спостерігати, що завищені є однією з головних проблем для італійських, французьких і східноазійських ресторанів. Для китайських і в'єтнамських ресторанів грубе ставлення серверів, швидше за все, буде

причиною низької оцінки. З іншого боку, помічаємо, що свіжі ряди першими серед позитивних слів для японської та в'єтнамської їжі. Є також деякі назви страв, присутні в позитивному списку слів, які можуть свідчити про те, що люди віддають перевагу певним ресторанам для своїх конкретних страв, таких як pho у в'єтнамській їжі та піці в італійській їжі або, можливо, такі страви просто легше задовольнити Yelpers, ніж інші.

Оцінки не повністю підтримуються текстом. Коли людей просять призначити оцінку для продукту або послуги, він зазвичай представляє їх загальне або загальне враження, тоді як те, що вони пишуть в огляді, може бути просто найбільш вражаючими частинами, хорошими або поганими. Насправді додавання даних персоналізації може бути корисним у покращенні продуктивності.

Перелік посилань

1. Mei, Q., Ling, X., Wondra, M., Su, H., & Zhai, C. (2007, May). Topic sentiment mixture: modeling facets and opinions in weblogs. In Proceedings of the 16th international conference on World Wide Web (pp. 171-180). ACM.
2. Hoffman, M., Bach, F. R., & Blei, D. M. (2010). Online learning for latent dirichlet allocation. In advances in neural information processing systems (pp. 856-864).
3. Hu, L., Sun, A., & Liu, Y. (2014, July). Your neighbors affect your ratings: on geographical neighborhood influence to rating prediction. In Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval (pp. 345-354). ACM.

МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
ХМЕЛЬНИЦЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ

МАГІСТЕРСЬКА РОБОТА

Аналітична система рекомендацій закладів харчування на основі відгуків та рейтингу

Розробив ст. гр. КНм-19-1:
Тіторов І.Д.

Хмельницький - 2020

В магістерській роботі було розроблено та набуло практичної реалізації системного підходу щодо аналізу відгуків та рейтингу в системі закладів громадського харчування.

Метою дослідження є розробка аналітичної системи рекомендації закладів харчування на основі відгуків та рейтингу.

Для досягнення зазначеної мети поставлені наступні **задачі**:
- провести вибір ознак та впливових факторів для використання в бізнес-аналітики;
- провести порівняння застосовності відомих методів дослідження щодо розробки та аналізу рекомендаційної системи.

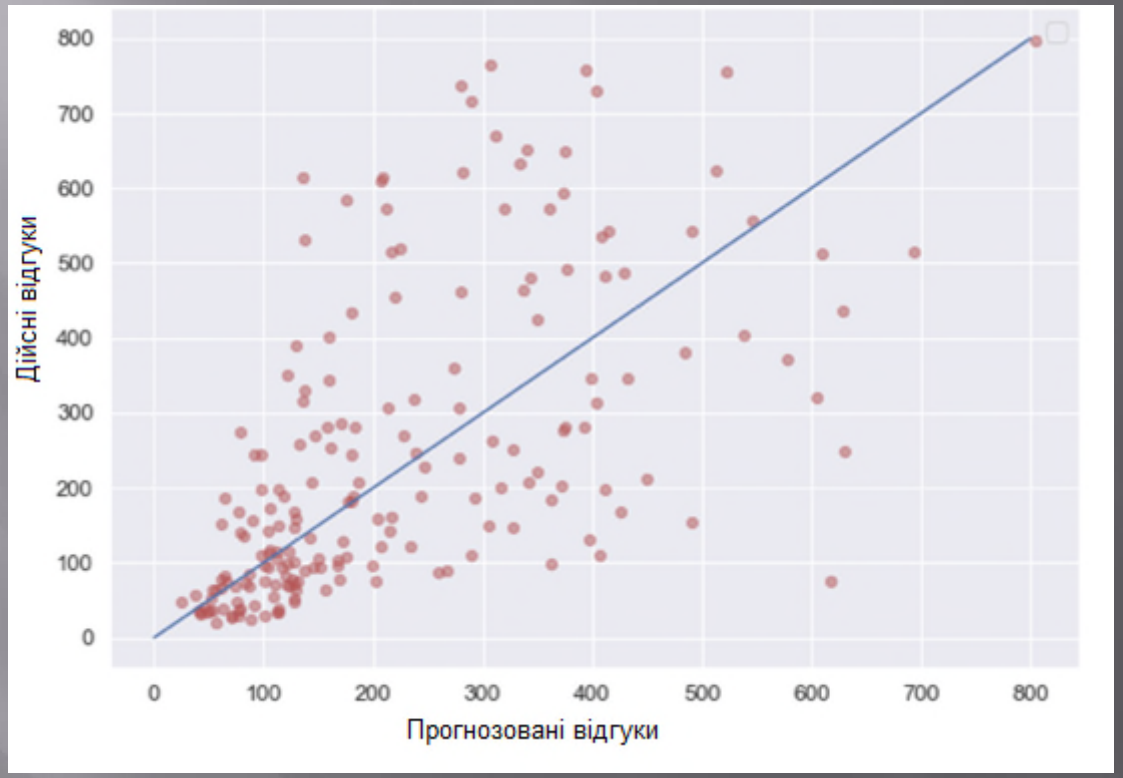
При цьому передбачається розв'язок таких підзадач, як
- попередня обробка даних та їх очищення;
- побудова списку ознак предметної області;
- дослідження методів визначення впливовості ознак;
- вибір моделей, виділення ознак і застосування методів машинного навчання;
- тестування методів на основі правил і з використанням машинного навчання;
- програмна реалізація система рекомендацій закладів харчування.

Об'єктом дослідження є методи отримування інформації з використанням текстової інформації та цифрового рейтингового оцінювання.

Предметом дослідження є текстові дані у вигляді відгуків з соціальних мереж та додаткова інформація у вигляді відносних рейтингових оцінювань.

Практична значимість дослідження полягає в тому, що отримані практичні результати можуть бути застосовні для підвищення ефективності роботи системи.
Програмне забезпечення для аналітики ресторанів є хорошим інструментом для будь-якого сучасного, амбітного та далекоглядного ресторатора.
Аналітика даних ресторану дасть додаткову конкурентну перевагу, яка допоможе зрозуміти клієнтів набагато більш докладно. Також допоможете розкрити уявлення про бізнес,- саме через ці рекомендації дають найбільше зростання.

Дослідження показало, що рейтинговий висновок є правдоподібним завданням, потенційні перешкоди існують у вивченні взаємозв'язку між рейтингами та відгуками. По-перше, є непослідовність при присвоєнні рейтингів серед авторів, дивергенція крос-автора. Це легко пояснюється тим, що думки - суб'єктивна річ. Для того ж рейтинг по відгукам, це загально прийнято, щоб один відгук з'являється дуже позитивним, а інший менш позитивним або навіть трохи негативним. По-друге, оцінки не повністю підтримуються текстом. Коли людей просять призначити оцінку для продукту або послуги, вони зазвичай представляють їх в загальному вираженні, тоді як те, що вони пишуть в огляді, може бути просто більш емоційно виражено, хорошим або поганим відношенням. Насправді додавання даних персоналізації може бути корисним у покращенні релевантності рейтингової системи.



Дійсні та прогнозовані відгуки

Висновки

Дослідження показало, що рейтинговий висновок є правдоподібним завданням, потенційні перешкоди існують у вивченні взаємозв'язку між рейтингами та відгуками. По-перше, є непослідовність при присвоєнні рейтингів серед авторів, дивергенція крос-автора. Це може легко уявити, оскільки, думки суб'єктивна річ. Для того ж рейтингу, це загальноприйнято, щоб один відгук з'являється дуже позитивним, а інший менш позитивним або навіть трохи негативним. По-друге, оцінки не повністю підтримуються текстом. Коли людей просять призначити оцінку для продукту або послуги, вона зазвичай представляє їх загальне враження, тоді як те, що вони пишуть в огляді, може мати просто найбільш позитивні частини, хороші або погані. Насправді додавання даних персоналізації може бути корисним у покращенні продуктивності.

Дякую за увагу

Sun Nov 29 15:26:17 EET 2020, Петровський Сергій Степанович, Хмельницький національний університет, ХНУ

Anti-Plagiarism v-15.257

Максимальне співпадіння з одним документом 0.0%

Словники перевірки: en_US, ru_RU, ua_UA. **Помилок в документах: 5%**

	Документ		Сумарний збіг по Базі Даних	
ID: 81552 Назва: Аналітична система рекомендацій закладів харчування на основі відгуків та рейтингу Додано в БД: 2020-11-29 Автора: Тіторов Ігор Дмитрович Керівники: Манзюк Е.А. Консультанти: Опоненти:	Символи	Лексеми	Символи	Лексеми
	74144	606	339 (0%)	2 (0%)

Джерело плагіату

ID	Опис	Наявність плагіату в документі	
		Символи	Лексеми

РІШЕННЯ КАФЕДРИ **КОМП'ЮТЕРНИХ НАУК ТА ІНФОРМАЦІЙНИХ ТЕХНОЛОГІЙ** ПРО ДОПУСК КВАЛІФІКАЦІЙНОЇ РОБОТИ ДО ЗАХИСТУ

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

Назва: **Аналітична система рекомендацій закладів харчування на основі відгуків та рейтингу**
Автор: **Тіторов І.Д.**
Спеціальність: **122 Комп'ютерні науки**
Науковий керівник: **к.т.н. доцент Манзюк Е.А.**

Після аналізу звіту подібності зроблено такий висновок:

№	Висновок	Позначка про відповідність
1	Запозичення, виявлені в роботі, є законними і не є плагіатом (далі – зазначаються підстави віднесення запозичень до правомірних). Робота приймається до захисту.	**відповідає**
2	Виявлені запозичення не є плагіатом, розміщені в розділах, які не описують безпосередньо авторське дослідження, але кількість цитат перевищує обсяг, виправданий поставленою метою роботи (далі – зазначаються детальні та аргументовані підстави віднесення запозичень до правомірних). Робота приймається до захисту, але має бути відкоригована. Відкоригований варіант має бути поданий на кафедру за 2 дні до захисту, разом із заявою щодо самостійності виконання письмової роботи та ідентичності друкованої та електронної версії роботи	-
3	Робота містить навмисні текстові спотворення, передбачувані спроби укриття запозичень або інші прояви академічного плагіату. Робота містить фабрикацію або фальсифікацію даних. Робота не допускається до захисту.	-
4	Інше:	-

Підтвердження: Виявленні запозичення не є плагіатом так як відносяться до загальновживаних типових фраз і складають 0.57%.

01.11.2020
Дата

Підпис керівника

Підпис завідувача кафедри

Хмельницький національний університет

Кафедра комп'ютерних наук та інформаційних технологій

ВІДГУК ОПОНЕНТА

на дипломну роботу магістра

Магістра *гр. КНм-19-1 Тіторова Ігоря Дмитровича*

На тему: *Аналітична система рекомендацій закладів харчування на основі відгуків та рейтингу.*

1. Актуальність і значення теми

Актуальність полягає в необхідності розробки та застосування аналітичної системи регулювання факторами впливу на ефективність системи.

У минулому бізнес-аналітика була привілеєм великих компаній, які могли дозволити собі підтримувати команди IT-фахівців і вчених з обробки даних. Але в останнє десятиліття, оскільки технологія швидко розвивалася, програмне забезпечення стало не тільки більш легким і потужним, але і більш доступним. Малий бізнес може використовувати ті ж інструменти, що і основні гравці ринку, й стикатися зі своїми конкурентами. Нові інструменти самообслуговування доводять, що бізнес-аналітика не ракетобудування, а скоріше корисний інструмент, який допоможе перетворити дані в обґрунтовані рішення.

2.Оцінка якості та достовірності проведених досліджень.

Отримані результати добре співвідносяться з результатами, наведеними в наукових роботах і довідниках.

3. Оцінка запропонованих заходів та пропозицій, практичної цінності та ефективності.

Проведені дослідження представляють науково-технічну цінність, є ефективним дослідженням в галузі машинобудування, їх можна використати з метою підвищення стійкості ріжучого інструменту.

4. Загальний висновок та оцінка

Робота виконана в повному обсязі. Досліджені та проаналізовані дані за допомогою комплексу входять в рамки допустимих відхилень. Пояснювальна записка оформлена в відповідності з нормами. Відмічені недоліки не знижують цінності дипломної роботи. За своєю структурою, практичними цінностями, поставленій меті та вирішеними задачами робота відповідає вимогам вищої школи і вимогам, що пред'являються до освітньо-кваліфікаційного рівня «магістр», а її автор Тіторов І.Д. заслуговує присвоєння кваліфікації магістра з комп'ютерних наук та інформаційних технологій.

Робота заслуговує на оцінку « Задовільно ».

Опонент Дихтов В., д.т.н., проф., зав.каф. ТАМ ХНУ