**КИЇВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ ІМЕНІ ТАРАСА ШЕВЧЕНКА**

**Факультет інформаційних технологій**

Кафедра технологій управління

Спеціальність 122 – Комп'ютерні науки,
освітня програма «Інформаційна аналітика та впливи»

# КВАЛІФІКАЦІЙНА РОБОТА МАГІСТРА

на тему:

**"**Аналіз емоційного забарвлення текстів з використанням засобів машинного навчання**"**

**Студента 2-го курсу групи ІАВ-21**

*Гліб ЄФРЕМОВ*
(прізвище, ім'я, по батькові)

(підпис студента)

**Науковий керівник:**

*д.т.н., доцент*
(науковий ступінь, вчене звання)

*Юлія ХЛЕВНА*
(прізвище, ім'я, по батькові)

(дата) (підпис)

**Попередній захист:**

(Висновок: «До захисту в Екзаменаційній комісії»)

Завідувач кафедри технологій управління *Віктор МОРОЗОВ*

(підпис) (прізвище, ініціали) (дата)

**Київ – 2023**

**КИЇВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ**
**ІМЕНІ ТАРАСА ШЕВЧЕНКА**
**Факультет інформаційних технологій**

Кафедра технологій управління
Освітньо-кваліфікаційний рівень Магістр
Спеціальність 122 - Комп'ютерні науки
Освітня програма Інформаційна аналітика та впливи

**ЗАТВЕРДЖУЮ**
Завідувач кафедри
професор Морозов В.В.

_____
«____» _____20___року

**З А В Д А Н Н Я**
**НА ВИКОНАННЯ КВАЛІФІКАЦІЙНОЇ РОБОТИ**

Студент *Гліб ЄФРЕМОВ*

Група ІАВ-21

**1. Тема кваліфікаційної роботи** "Аналіз емоційного забарвлення текстів з використанням засобів машинного навчання"

Затверджена наказом по від «17» листопада 2021 р. № 4.

**2. Строк подання студентом готової роботи –** "26" травня 2023 р.

**3. Цільова установка та вихідні дані до роботи**

Дослідження методів датамайнингу для коротких текстів та розробка системи для автоматизації аналізу таких текстів.

**4. Зміст роботи**

Дослідження підходів структуризації, обробки даних та валідації отриманих результатів, проєктування та імплементація програмного засобу для порівняння обраного набору підходів.

**5. Перелік графічного матеріалу (слайдів)**

9 рисунків, 11 слайдів презентації доповіді.

**6. Календарний план виконання роботи:**

| № з/п | Назва частин роботи | % | Виконання роботи | |
|---|---|---|---|---|
| | | | За планом | Фактично |
| 1 | Вибір теми дипломної роботи | 3 | 01.10.23 | 01.10.23 |

| | | | | |
|---|---|---|---|---|
| 2 | Протокол кафедри ТУ про затвердження тем дипломних робіт та призначення наукових керівників | 2 | 24.12.23 | 24.12.23 |
| 3 | Формування переліку нормативних матеріалів, літератури з проблематики дипломної роботи | 10 | 07.01.23 | 07.01.23 |
| 4 | Складання розгорнутого плану кваліфікаційної роботи | 5 | 18.01.23 | 18.01.23 |
| 5 | Ознайомлення наукового керівника з розгорнутим планом кваліфікаційної роботи. Внесення змін. | 5 | 19.01.23 - 20.01.23 | 19.01.23 - 20.01.23 |
| 6 | Написання розділу 1 дипломної роботи | 11 | 14.02.23 | 14.02.23 |
| 7 | Написання розділу 2 дипломної роботи | 15 | 08.03.23 | 08.03.23 |
| 8 | Написання розділу 3 дипломної роботи | 14 | 01.04.23 | 01.04.23 |
| 9 | Написання розділу 4 дипломної роботи | 12 | 01.05.23 | 01.05.23 |
| 10 | Оформлення кваліфікаційної роботи. Підготовка висновків і пропозицій | 15 | 03.05.23 | 03.05.23 |
| 11 | Передача кваліфікаційної роботи науковому керівникові | 1 | 04.05.23 | 04.05.23 |
| 12 | Передача кваліфікаційної роботи рецензенту для рецензування | 2 | 11.05.23 | 11.05.23 |
| 13 | Попередній захист кваліфікаційної роботи | 5 | 17.05.23 | 17.05.23 |

Дата видачі завдання «17» листопада 2023 р.

Керівник роботи д.т.н., доцент Юлія ХЛЕВНА

_____
(підпис)

Завдання прийняв до виконання:

Здобувач освіти групи ІАВ-21 Гліб ЄФРЕМОВ

_____
(підпис)

**ЗМІСТ**

# АНОТАЦІЯ

**КИЇВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ**
**ІМЕНІ ТАРАСА ШЕВЧЕНКА**
**Факультет інформаційних технологій**
Кафедра технологій управління
Спеціальність 122 - Комп'ютерні науки,
освітня програма "Інформаційна аналітика та впливи"

Дипломна робота магістра Єфремова Гліба Денисовича.

Тема роботи – «Аналіз емоційного забарвлення текстів з використанням засобів машинного навчання».

Мета дипломної роботи магістра – підвищення якості визначення емоційного забарвлення коротких текстів методами аналізу даних та засобами машинного навчання.

Об'єктом дослідження є процеси емоційного забарвлення текстів - Opinion Mining.

Предмет дослідження – визначення емоційного навантаження, властивостей та/або ключових особливостей, описаних в тексті, з великої кількості дописів з мережі Інтернет.

Наукова новизна роботи полягає у створенні нового програмного продукту, який дозволяє розгорнути систему для вилучення ключової інформації з великих баз із текстом, із вибором бажаного підходу. В роботі досліджені та перевірені на практиці сучасні методи обробки даних та методи інтерпретації отриманої інформації.

Були отримані наступні результати: виконано загальний огляд методів датамайнингу для коротких текстів та розробка системи для автоматизації аналізу таких текстів, розроблено програмний засіб для тестування досліджених підходів. Проаналізовано велику кількість текстової інформації у вільному доступі, порівняно результати аналізу різними методами та їх загальну ефективність в різних умовах. Скомпоновано кілька найбільш ефективних підходів для наявних даних та побудовано модель для розпізнавання емоційного навантаження текстів та ключових особливостей продуктів, подій чи політиків, на які звертають увагу автори текстів.

Дипломна робота складається зі вступу, основної частини, яка налічує чотири розділи, висновки та список використаних джерел. Всього налічує 75 сторінок, 9 ілюстрацій, 6 таблиць, 1 додаток та перелік посилань з 30 джерел на 3 сторінках.

Ключові слова: аналіз текстів, сентиментний аналіз, датамайнинг.

# СПИСОК СКОРОЧЕНЬ ТА ПОЗНАЧЕНЬ

Opinion – судження

Opinion mining – визначення судження, аналіз думок

Sentiment analysis – аналіз сентиментів

# ВСТУП

Кожний бізнес сьогодні, а особливо онлайн-бізнес, отримує клієнтів, переважно, за допомогою інтернет-маркетингу. Компанії, будь-то рітейл, маркетплейси, ігрові компанії, або B2B вкладають дуже великі гроші в інтернет-рекламу для залучення клієнтів своєї цільової аудиторії, і це, мабуть, один з найвагоміших пунктів зі списку витрат. Тим більше, кожного року ціна за приваблення одного клієнта росте невпинно протягом багатьох років і, скоріше за все, продовжуватиме рости. З вищесказаного у маркетологів, генеральних директорів та засновників компаній виникає раціональне і актуальне бажання оптимізувати великі витрати на маркетинг так, щоб вкладені гроші не були витрачені попусту. З розвитком комп'ютерних технологій та зі зростанням доступності потужних комп'ютерів все більше компаній та організацій (комерційних, благодійних, політичних, дослідницьких, тощо) можуть дозволити собі розробку, налаштування, та підтримку систем автоматичного збору та обробки інформації, яка могла б, певною мірою, замінити цілий відділ маркетологів чи політологів, чи навіть зовнішнього партнера (наприклад, агенцію соціологічних досліджень). Хоча перші важливі кроки в розвитку цього напряму були зроблені ще в далекому 1950 році (формування принципів кластерного аналізу, поява "генетичних" алгоритмів, теоретично описані нейронні мережі), дослідження все ще продовжуються, адже для кожної окремої задачі часто можна знайти кілька різних підходів, і кожна така дилема вибору підходів потребує окремого дослідження та оцінки на практиці, а маркетологія та політологія, в свою чергу, теж не стоять на місці

— з'являються нові "філософії" досліджень (тобто методології), вносяться зміни в старі, а системам аналізу інформації потрібно під ці зміни підлаштовуватись.

Станом на 2023 рік маркетологія та, певною мірою, політологія стали помітно більш впливовими сферами досліджень — спочатку через розгул коронавірусу, а потім через війну впав попит на велику кількість товарів, що не входять в перелік продуктів базової необхідності (ліки, продукти харчування,

засоби гігієни). Несучи втрати, компанії намагаються зрозуміти, як пандемія та бойові дії повпливали на “споживацькі” звички та отримати підтримку якомога більшої кількості нових клієнтів. Обидві потреби може задовольнити машинний opinion mining(з англійської – визначення судження), тобто програмні системи, що автоматично та автономно обробляють необроблені дані, які є в доступі компанії, наприклад: статті про новини у виданнях, дописи в соціальних мережах, відео-відгуки, тощо.

**Мета й завдання роботи.**

Метою роботи є підвищення якості визначення емоційного забарвлення коротких текстів методами аналізу даних та засобами машинного навчання.

**Об’єктом дослідження** є процеси емоційного забарвлення текстів - Opinion Mining.

**Предметом дослідження** є визначення емоційного навантаження, властивостей та/або ключових особливостей, описаних в тексті, з великої кількості дописів з мережі Інтернет.

**Наукова новизна одержаних результатів**

В ході виконання кваліфікаційної роботи було створено новий програмний продукт, який дозволяє розгорнути систему для вилучення ключової інформації з великих баз із текстом, із вибором бажаного підходу. В роботі досліджені та перевірені на практиці сучасні методи обробки даних та методи інтерпретації отриманої інформації.

**Практична значимість**

Практичне значення одержаних результатів полягає у тому, що отриманий програмний продукт можна використовувати також і у комерційних цілях, наприклад для аналізу відгуків на продукт організації в мережі Інтернет, або на політичну діяльність тощо. Програма дозволяє швидко обрати бажаний метод обробки.

# РОЗДІЛ 1. ДОСЛІДЖЕННЯ ТЕОРЕТИЧНИХ ЗАСАД ВИКОРИСТАННЯ ІНФОРМАЦІЙНОЇ АНАЛІТИКИ ДЛЯ ВИЗНАЧЕННЯ ЕМОЦІЙНОГО ЗАБАРВЛЕННЯ ТЕКСТУ

## 1.1 Стан та перспективи емоційного аналізу тексту в мережі Інтернет

Думки (opinions, інакше кажучи точки зору, погляди) є центральними для майже всіх людських дій і є ключовими факторами, що впливають на нашу поведінку. Наші переконання та сприйняття реальності, а також вибір, який ми робимо, значною мірою залежать від того, як інші бачать і оцінюють світ. З цієї причини, коли нам потрібно прийняти рішення, ми часто запитуємо думку інших. Це стосується не лише окремих осіб, а й організацій.

Думки є центральними для майже всіх людських дій, оскільки вони є ключовими факторами, що впливають на нашу поведінку. Щоразу, коли нам потрібно прийняти рішення, ми хочемо знати думку інших. У реальному світі компанії та організації завжди хочуть дізнатися думку споживачів або громадськості про їхні продукти та послуги. Індивідуальні споживачі також хочуть знати думку існуючих користувачів продукту перед його придбанням, а також думки інших про політичних кандидатів перед тим, як прийняти рішення про голосування на політичних виборах. У минулому, коли людині потрібна була думка, вона питала друзів і родину. Коли організації чи бізнесу потрібна була думка громадськості чи споживачів, вони проводили опитування, опитування громадської думки та фокус-групи. Отримання думок громадськості та споживачів вже давно стало величезним бізнесом для маркетингових компаній, компаній зі зв’язків з громадськістю та проведення політичних кампаній.

Думки та пов’язані з ними поняття, такі як настрої, оцінки, ставлення та емоції, є об’єктами дослідження аналізу настроїв та аналізу думок. Початок і швидке зростання галузі збігаються з соціальними мережами в Інтернеті, наприклад, оглядами, обговореннями на форумах, блогами, мікроблогами, Twitter і соціальними мережами, тому що вперше в історії людства ми маємо величезний обсяг думок, записаних у цифрових формах.

Сьогодні мікроблоги стали дуже популярним інструментом спілкування серед користувачів Інтернету. Мільйони повідомлень щодня з'являються на популярних веб-сайтах, які надають послуги мікроблогів, таких як Twitter, Tumblr, Facebook. Автори цих повідомлень пишуть про своє життя, діляться думками на різні теми та обговорюють актуальні проблеми. Через вільний формат повідомлень і легкий доступ до платформ мікроблогів користувачі Інтернету, як правило, переходять від традиційних засобів спілкування (таких як традиційні блоги чи списки розсилки) до послуг мікроблогів. Оскільки все більше користувачів розміщують повідомлення про продукти та послуги, якими вони користуються, або висловлюють свої політичні та релігійні погляди, веб-сайти мікроблогів стають цінним джерелом думок і настроїв людей. Такі дані можна ефективно використовувати для маркетингу чи соціальних досліджень.

З початку 2000 року аналіз настроїв, або аналіз сентиментів став одним із найактивніших напрямів дослідження обробки природної мови. Його також широко вивчають у інтелектуальному аналізі даних, веб-майнінгу та текстовому аналізі. Фактично, ця методологія поширилася з інформатики на науки про управління та соціальні науки через свою важливість для бізнесу та суспільства в цілому. В останні роки промислова діяльність, пов'язана з аналізом настроїв, також процвітала. Виникло багато стартапів. Багато великих корпорацій створили власні власні можливості. Системи аналізу настроїв знаходять своє застосування майже в кожному бізнесі та соціальній сфері.

## 1.2 Огляд задач датамайнингу для визначення емоційного забарвлення тексту

Датамайнингом (data mining) називають процес обробки великих наборів даних для знаходження спільних характеристик та закономірностей та їх використання для вирішення задач, що інакше потребували б великої кількості людської роботи. Зі збільшенням кількості гаджетів та цифрових сервісів, якими користуються звичайні люди, кількість даних, які вони залишають у кіберпросторі, зростає майже в геометричній прогресії, залишаючись, по суті, мертвим вантажем, невикористаними. Аналіз цих даних, безумовно, має

вирішальне значення для отримання та/або підтримання хорошого розуміння вподобань і думок клієнтів будь-якої організації. В процесі обробки інформації знайдені закономірності використовуються для оцінки найбільш важливих в конкретному заданому контексті характеристик тексту чи іншої інформації, далі за цими даними можна робити різноманітні прогнози чи оцінки актуальної ситуації. Термін Data mining почали вживати в 1990х роках, хоча напрям має довшу історію, але називався іншими термінами або існував як частина ширших процесів, наприклад датамайнингом можна назвати етап аналізу в KDD (knowledge discovery in databases, англ. виявлення знань в базах даних), в деяких старіших джерелах datamining називають як knowledge extraction (англ. витягнення знань) або information discovery (англ. виявлення інформації) [1]. На абстрактному рівні область KDD пов’язана з розробкою методів і технік для осмислення даних. Основна проблема, яку вирішує процес KDD, полягає у відображенні низькорівневих даних (які, як правило, надто об’ємні, щоб їх легко зрозуміти та засвоїти) в інші форми, які можуть бути більш компактними (наприклад, короткий звіт), більш абстрактними (наприклад, , описове наближення або модель процесу, який створив дані), або більш корисних (наприклад, прогностична модель для оцінки цінності майбутніх випадків). В основі процесу лежить застосування спеціальних методів аналізу даних для виявлення та вилучення патернів.

Традиційний метод перетворення даних у знання базується на ручному аналізі та інтерпретації. Наприклад, у галузі охорони здоров’я спеціалісти часто аналізують поточні тенденції та зміни в даних про охорону здоров’я, скажімо, щокварталу. Потім спеціалісти надають звіт із детальним описом аналізу спонсорській медичній організації; цей звіт стає основою для майбутнього прийняття рішень і планування управління охороною здоров'я. Як зовсім інший приклад, планетарні геологи переглядають зображення планет і астероїдів, отримані дистанційним зондуванням, ретельно виявляючи та каталогізуючи такі цікаві геологічні об’єкти, як ударні кратери. Будь то наука, маркетинг, фінанси, охорона здоров’я, роздрібна торгівля чи будь-яка інша галузь, класичний підхід

до аналізу даних базується на тому, що один або більше аналітиків добре знайомляться з даними та служать інтерфейсом між даними, користувачами та продуктами. У бізнесі основні сфери застосування інтелектуального аналізу даних включають маркетинг, фінанси (особливо інвестиції), виявлення шахрайства, виробництво, телекомунікації та Інтернет-агенції.

Для цих (та багатьох інших) застосувань ця форма ручного дослідження набору даних є повільною, дорогою та дуже суб'єктивною. Фактично, оскільки обсяги даних різко зростають, цей тип ручного аналізу даних стає абсолютно непрактичним у багатьох областях. Розмір баз даних збільшується двома способами: кількість N записів або об'єктів у базі даних і кількість d полів або атрибутів об'єкта. Бази даних, що містять близько $N = 10^9$ об'єктів, стають все більш поширеними, наприклад, в астрономічних науках[1]. Хто міг би переварити мільйони записів, кожен з яких має десятки чи сотні полів? Я вважаю, що аналітична робота має бути автоматизована, принаймні частково.

Необхідність розширити можливості людського аналізу для роботи з великою кількістю байтів, які ми можемо зібрати, є як економічною, так і науковою. Компанії використовують дані, щоб отримати конкурентну перевагу, підвищити ефективність і надавати більш цінні послуги клієнтам. Дані, які ми збираємо про навколишнє середовище, є основними доказами, які ми використовуємо для побудови теорій і моделей всесвіту, у якому ми живемо. Оскільки комп'ютери дозволили людям збирати більше даних, ніж ми можемо переварити, цілком природно звернутися до обчислювальних методів, щоб допомогти нам викопувати значущі моделі та структури з величезних обсягів даних. Таким чином, датамайнинг – це спроба вирішити проблему, яку ера цифрової інформації зробила фактом життя для всіх нас: перевантаження даними. Історично поняття пошуку корисних шаблонів у даних називалося різними назвами, включаючи інтелектуальний аналіз даних, витяг знань, відкриття інформації, збір інформації, археологію даних і обробку шаблонів даних.

Інтелектуальний аналіз даних отримав свою назву через схожість між пошуком цінної бізнес-інформації у великій базі даних (наприклад, пошук пов’язаних продуктів у гігабайтах даних сканера магазину) та видобутком гори в пошуках цінної руди. Обидва процеси вимагають або просіювання величезної кількості матеріалу, або розумного його дослідження, щоб точно знайти цінність. За наявності баз даних достатнього розміру та якості технологія інтелектуального аналізу даних може створювати нові можливості для бізнесу, надаючи такі можливості:

Автоматичне передбачення тенденцій і поведінки. Інтелектуальний аналіз даних автоматизує процес пошуку прогнозної інформації у великих базах даних. На питання, які традиційно вимагали ретельного аналізу, тепер можна швидко відповісти безпосередньо з даних. Типовим прикладом проблеми прогнозування є цільовий маркетинг. Інтелектуальний аналіз даних використовує дані про минулі рекламні розсилки, щоб визначити цілі, які, швидше за все, максимізують повернення інвестицій у майбутні розсилки. Інші проблеми прогнозування включають прогнозування банкрутства та інших форм дефолту, а також визначення сегментів населення, які, ймовірно, подібно відреагують на дані події.

Автоматизоване виявлення раніше невідомих шаблонів. Інструменти інтелектуального аналізу даних проходять через бази даних і визначають раніше приховані шаблони за один крок. Прикладом виявлення закономірностей є аналіз даних про роздрібні продажі для виявлення, здавалося б, не пов’язаних між собою продуктів, які часто купують разом. Інші проблеми виявлення шаблонів включають виявлення шахрайських транзакцій кредитних карток і виявлення аномальних даних, які можуть представляти помилки введення даних.

Фактично датамайнинг стоїть на перетині трьох дисциплін: статистики, машинного навчання та, в деяких випадках, штучного інтелекту, тому в розпорядженні розробників завжди є широкий вибір технологій та підходів, конкретні алгоритми реалізацій та модифікацій яких відрізняються від задачі до задачі та залежать від самого набору даних.


Рисунок 1.1 Головні кроки датамайнингу.

На початку роботи над частиною проекту, що потребує датамайнингових підходів, потрібно розглянути специфіку контексту розробки, наприклад компанії чи команди, якій потрібні висновки з датамайнингу. Для цього розглядають не тільки конкретні вимоги до розробки, а і бізнес в цілому, щоб зрозуміти на яку перспективу потрібно орієнтувати проект: чи потрібно продумувати стратегії доробки моделі аналізу, чи потрібно орієнтуватись на розширення сфер використання, чи знайде використання та чи інша додаткова властивість та на які обчислювальні потужності потрібно орієнтуватись. Більшість досліджень моделювання на основі даних виконуються в певній області застосування. Отже, для того, щоб придумати значущу постановку проблеми, зазвичай необхідні знання та досвід, що стосуються певної галузі. На жаль, багато прикладних досліджень, як правило, зосереджені на техніці інтелектуального аналізу даних за рахунок чіткого визначення проблеми. На цьому етапі модельєр зазвичай визначає набір змінних для невідомої залежності та, якщо можливо, загальну форму цієї залежності як початкову гіпотезу. На цій стадії може бути сформульовано кілька гіпотез щодо однієї проблеми. Перший крок вимагає комбінованого досвіду прикладної області та моделі інтелектуального аналізу даних. На практиці це зазвичай означає тісну взаємодію між експертом з аналізу даних і експертом із застосування. У успішних програмах інтелектуального аналізу даних ця співпраця не припиняється на початковому етапі; це триває протягом усього процесу інтелектуального аналізу даних.

Наступний крок, збір даних, стосується того, як дані генеруються та збираються. Загалом, є дві різні можливості. Перший – це коли процес генерації

даних знаходиться під контролем експерта (модельєра): цей підхід відомий як спланований експеримент. Друга можливість полягає в тому, що експерт не може вплинути на процес генерування даних: це відоме як підхід спостереження. Налаштування спостереження, а саме генерація випадкових даних, передбачається у більшості програм інтелектуального аналізу даних. Як правило, розподіл вибірки повністю невідомий після збору даних, або він частково та неявно надається в процедурі збору даних. Однак дуже важливо зрозуміти, як збір даних впливає на їх теоретичний розподіл, оскільки такі апріорні знання можуть бути дуже корисними для моделювання, а згодом і для остаточної інтерпретації результатів. Крім того, важливо переконатися, що дані, які використовуються для оцінки моделі, і дані, які пізніше використовуються для тестування та застосування моделі, надходять з того самого, невідомого розподілу вибірки. Якщо це не так, оцінена модель не може бути успішно використана для остаточного застосування результатів.

Наступним кроком виступає аналіз (розуміння) даних, тобто огляд наявних та можливих наборів інформації, доступних команді — скільки є даних, наскільки повні ці дані, наскільки вони точні та які в них є тенденції, що видно неозброєним оком або на простих графіках візуалізації даних. На цьому етапі можна провести простий статистичний аналіз за певним параметром, щоб зрозуміти, які дані є аномальними (та прибрати їх, якщо це необхідно). Дані, які стосуються аналітичної програми, визначаються та збираються. Дані можуть знаходитися в багатьох вихідних системах, сховищі даних або озері даних, яке стає все більш популярним у середовищах великих даних із поєднанням структурованих і неструктурованих даних. У обстановці спостереження дані зазвичай "збираються" з існуючих баз даних, сховищ даних і вітрин даних. Попередня обробка даних зазвичай включає принаймні два загальних завдання:

Виявлення (і видалення) викидів – це незвичайні значення даних, які не узгоджуються з більшістю спостережень. Зазвичай викиди є результатом помилок вимірювання, помилок кодування та запису, а іноді є природними

ненормальними значеннями. Такі нерепрезентативні зразки можуть серйозно вплинути на модель, виготовлену пізніше. Існує дві стратегії роботи з викидами:

-виявляти та зрештою видаляти викиди як частину етапу попередньої обробки, або

-розробити надійні методи моделювання, нечутливі до викидів.

Функції масштабування, кодування та вибору – попередня обробка даних включає кілька етапів, наприклад масштабування змінних і різні типи кодування. Наприклад, одна ознака з діапазоном [0, 1] та інша з діапазоном [−100, 1000] не матимуть однакові ваги в застосовуваній методиці; вони також по-різному впливатимуть на кінцеві результати аналізу даних. Тому рекомендується масштабувати їх і привести обидві функції до однакової ваги для подальшого аналізу. Крім того, спеціальні методи кодування програми зазвичай досягають зменшення розмірності шляхом надання меншої кількості інформаційних функцій для подальшого моделювання даних.

Ці два класи завдань попередньої обробки є лише ілюстративними прикладами широкого спектру дій попередньої обробки в процесі інтелектуального аналізу даних. Етапи попередньої обробки даних не слід вважати абсолютно незалежними від інших етапів інтелектуального аналізу даних. У кожній ітерації процесу інтелектуального аналізу даних усі дії разом можуть визначати нові та вдосконалені набори даних для наступних ітерацій. Як правило, хороший метод попередньої обробки забезпечує оптимальне представлення методу інтелектуального аналізу даних шляхом включення апріорних знань у формі програмного масштабування та кодування.

Наступним етапом є підготовка даних. Сьогодні доступно багато інформації про сховища даних, інтелектуальний аналіз даних, KDD, OLTP, OLAP та цілу абетку інших абревіатур, які описують техніки та методи зберігання, доступу, візуалізації та використання даних. Існують книги та журнали про створення моделей для прогнозування всіх типів — шахрайства, маркетингу, нових клієнтів, споживчого попиту, економічної статистики, руху акцій, цін на опціони, погоди, соціологічної поведінки, попиту на трафік,

потреби в ресурсах і багато іншого. Щоб використовувати ці методи чи робити прогнози, професіонали галузі майже повсюдно погоджуються, що підготовка даних є однією з найважливіших частин будь-якого такого проекту, а також однією з найбільш трудомістких і складних. На жаль, підготовка даних була дуже схожа на погоду, як каже старий афоризм: «Усі про це говорять, але ніхто нічого з цим не робить».

Наскільки важлива адекватна підготовка даних? Після знаходження правильної проблеми для вирішення підготовка даних часто є ключем до вирішення проблеми. Це може легко бути різницею між успіхом і невдачею, між корисними розумінням та незрозумілою тьмяною, між вартими уваги передбаченнями та марними здогадками.

Останні розроблені інструменти для дослідження даних, сьогодні відомі як інструменти інтелектуального аналізу даних, лише починають процес автоматизації пошуку. На сьогоднішній день більшість сучасних інструментів інтелектуального аналізу даних зосереджено майже виключно на створенні моделей — ідентифікації «риби». Проте застосування інструментів моделювання до правильно підготовлених даних приносить величезні дивіденди. Але підготовка даних для моделювання була надзвичайно трудомістким процесом, традиційно виконуваним вручну та дуже важко автоматизованим. Сьогодні дуже часто замість адекватної підготовки даних і точного дослідження даних, для розуміння даних будуються та перебудовуються моделі, що потребують багато часу. Моделювання та реконструкція не є найвигіднішим чи найефективнішим способом виявлення того, що міститься в наборі даних. Якщо потрібна модель, опитування даних показує, яка саме модель (або моделі, якщо декілька найкраще відповідають потребам) підходить, як її побудувати, наскільки добре вона працюватиме, де її можна застосувати, і наскільки вона буде надійною та його межі продуктивності. Усе це можна зробити до того, як буде створено будь-яку модель, і за невелику частину часу, необхідного для дослідження даних шляхом моделювання.

Підготовка даних була поміщена в контекст дослідження даних, у якому першочерговою є проблема, яку потрібно вирішити, а не технологія. Не визначивши проблему, яку потрібно вирішити, важко визначити, як витягнути цінність із подальших дій з аналізу даних. Не менш важливим є визначення форми рішення. Не маючи чіткого уявлення про те, як виглядає успіх, важко визначити, чи дійсно досягнутий результат і форма, в якій він представлений, справді досягли успіху. Визначивши, як виглядає відповідне рішення, і зібравши або виявивши відповідні дані, ви можете почати процес інтелектуального аналізу даних. Дані повинні бути підготовлені таким чином, щоб інструменти видобутку могли легко отримати доступ до інформації, що міститься в них. Відсутня частина, міст до розуміння, це пояснення того, як виглядає загальний процес. Огляд процесу в цілому забезпечує структуру та посилання, щоб зрозуміти, як кожен компонент вписується в загальний дизайн.

Дані очищаються від зайвих відомостей, їх підлаштовують до обраного формату для спрощення чи пришвидшення подальшої обробки і, власне, обробляють за допомогою обраних алгоритмів. Цей етап може займати дуже великий проміжок часу, величина якого залежить від складності алгоритмів обробки, розміру набору даних, кількості наборів та ефективності системи збереження результатів. Станом на 2023 рік для збереження та пошуку серед таких даних, в основному, використовують системи менеджменту розподілених баз даних (DDBMS), що дозволяє збільшити швидкість процесу та зняти обмеження на оперативну пам’ять пристрою обробки. Використання розподіленої системи керування базами даних (DDBMS) може бути корисним для аналізу думок і настроїв. DDBMS може забезпечити масштабованість, дозволяючи розподілено обробляти дані на кількох вузлах або машинах. Це означає, що більші набори даних і набори даних, розповсюджені в кількох місцях, можуть надійно оброблятися DDBMS за розумний проміжок часу. Крім того, DDBMS може зменшити затримку, дозволяючи обслуговувати інформацію вузлом, найближчим до місця призначення. Нарешті, DDBMS підвищує надійність, оскільки зберігає кілька копій бази даних на вузлах, зменшуючи

ризик втрати даних. Це може бути особливо корисним під час виконання аналізу настроїв, оскільки будь-які відсутні дані можуть призвести до неправильних висновків. Це важливо, бо оперативної (швидкої) пам’яті зазвичай менше, ніж звичайної (повільної), через те, що вона дорожча, а також максимальна кількість оперативної пам’яті обмежується архітектурою системи.

Виявлені закономірності в даних потім переглядаються людьми для оцінки знайденої інформації. Деякі явища можуть бути пояснені вже наявною інформа цією. Деякі можуть бути ігноровані як артефакти (аномалії) процесу або як побічні дані. А закономірності, які залишаться, можна інтерпретувати в певну модель, за допомогою якої можна отримати якесь припущення з високим рівнем достовірності, що можна використати в практичних цілях. Такі моделі можна поділити на три широкі категорії:

- Предиктивні моделі
- Дескриптивні (кластеризуючі) моделі ( вони ж – моделі для виявлення аномалій)
- Моделі для пошуку закономірностей.

Предиктивні моделі використовують, коли потрібно отримати приблизну оцінку чи припущення щодо значення якогось атрибуту об’єкту або явища за неповною інформацією, якщо в розпорядженні апроксимуючої системи є велика кількість схожих даних, де значення даного атрибуту вже відоме. Основна мета предиктивного майнингу — сказати щось про майбутні результати, а не про поточну поведінку. Він використовує функції навчання під наглядом, які використовуються для прогнозування цільового значення. Методи, які належать до цієї категорії видобутку, називаються класифікацією, аналізом часових рядів і регресією. Моделювання даних є необхідністю прогностичного аналізу, і воно працює, використовуючи кілька поточних змінних для прогнозування майбутніх невідомих значень даних для інших змінних.

Приклади інтелектуального аналізу даних включають регресійний аналіз, дерева рішень і нейронні мережі. Регресійний аналіз передбачає прогнозування безперервної змінної результату на основі однієї або кількох змінних

предиктора. Дерева рішень включають побудову деревоподібної моделі для прогнозування на основі набору правил. Нейронні мережі передбачають створення моделі на основі структури людського мозку для прогнозування. Відомий приклад техніки для формування схожої моделі — регресійний аналіз (побудова регресійної моделі для передбачення чисельних значень). Регресія — це вивчення функції, яка відображає елемент даних у змінну передбачення з дійсним значенням. Застосувань регресії багато, наприклад, передбачення кількості біомаси, присутньої в лісі, за допомогою мікрохвильових вимірювань дистанційного зондування, оцінка ймовірності того, що пацієнт виживе за результатами набору діагностичних тестів, прогнозування споживчого попиту на новий продукт як функція витрат на рекламу та прогнозування часових рядів, де вхідні змінні можуть бути версіями змінної прогнозу із затримкою в часі. Предиктивні моделі передбачають використання деяких змінних або полів у базі даних для прогнозування невідомих або майбутніх значень інших змінних, що цікавлять, а дескриптивні зосереджуються на пошуку інтерпретованих людиною шаблонів, що описують дані.

Дескриптивні (або кластеризуючі) моделі використовують якщо є потреба в поділі даних за певною ознакою або правилом. Цей термін в основному використовується для створення кореляції, перехресної таблиці, частоти тощо. Ці технології використовуються для визначення подібності в даних і пошуку існуючих закономірностей. Ще одним застосуванням описового аналізу є розробка захоплюючих підгруп у більшій частині наявних даних. Ця аналітика наголошує на узагальненні та перетворенні даних у значущу інформацію для звітності та моніторингу.

Приклади описового інтелектуального аналізу даних включають кластеризацію, аналіз правил асоціації та виявлення аномалій. Кластеризація передбачає групування подібних об’єктів разом, тоді як аналіз правил асоціації включає визначення зв’язків між різними елементами в наборі даних. Виявлення аномалій включає виявлення незвичайних шаблонів або викидів у даних. Кластеризація — це найбільш поширена дескриптивна задача, в якій

намагаються визначити кінцевий набір категорій або кластерів для опису даних. Категорії можуть бути взаємовиключними та вичерпними або складатися з більш детального представлення, наприклад, ієрархічних категорій або категорій, що перекриваються. Приклади кластеризації додатків у контексті виявлення знань включають виявлення однорідних субпопуляцій для споживачів у базах даних маркетингу та ідентифікацію підкатегорій спектрів із вимірювань інфрачервоного неба. З кластеризацією тісно пов'язана задача оцінки щільності ймовірності, яка складається з методів оцінки на основі даних спільної багатовимірної функції щільності ймовірності всіх змінних або полів у базі даних. Крім звичайних методів кластеризації для поділу об'єктів або явищ на групи також використовуються методи пошуку закономірностей, що не відображені в одиничних входженнях інформації, а проглядаються лише в контексті якогось набору інформації. Прикладом таких методологій є системи створення рекомендацій, що відносять користувачів за неповною інформацією про їх вподобання до певної групи користувачів і надають рекомендації, базуючись на інформації про всю групу. Інше застосування таких моделей — пошук аномалій, в тому числі і як побічний ефект кластеризації.

Хоча межі між дескриптивними і предиктивними моделями не є однозначно різкими (деякі прогнозні моделі можуть бути описовими, і навпаки), відмінність корисна для розуміння загальної мети відкриття. Відносна важливість передбачення та опису для окремих програм інтелектуального аналізу даних може значно відрізнятися. Цілі передбачення та опису можуть бути досягнуті за допомогою різних конкретних методів аналізу даних.

Основні відмінності між описовим і прогнозним аналізом даних:

Мета: Описовий інтелектуальний аналіз даних використовується для опису даних і виявлення закономірностей і зв'язків. Прогнозний аналіз даних використовується для прогнозування майбутніх подій.

Підхід. Описовий інтелектуальний аналіз даних передбачає аналіз історичних даних для виявлення закономірностей і зв'язків. Інтелектуальний аналіз даних передбачає використання статистичних моделей і алгоритмів

машинного навчання для визначення закономірностей і зв’язків, які можна використовувати для прогнозування.

Результат: Описовий інтелектуальний аналіз даних створює зведення та візуалізацію даних. Прогнозний інтелектуальний аналіз даних створює моделі, які можна використовувати для прогнозування.

Часові рамки: описовий інтелектуальний аналіз даних зосереджений на аналізі історичних даних. Прогнозний аналіз даних зосереджений на прогнозуванні майбутніх подій.

Застосування: Описовий інтелектуальний аналіз даних використовується в таких програмах, як сегментація ринку, профілювання клієнтів і рекомендації продуктів. Прогнозний аналіз даних використовується в таких програмах, як виявлення шахрайства, оцінка ризиків і прогнозування попиту.

Підсумовуючи, описовий і прогнозний інтелектуальний аналіз даних є двома важливими методами виявлення закономірностей і тенденцій у великих наборах даних. Описовий інтелектуальний аналіз даних використовується для узагальнення та опису даних, а прогнозний інтелектуальний аналіз даних використовується для прогнозування майбутніх подій. Обидві методики мають свої переваги та застосування, а вибір техніки залежить від конкретної проблеми та характеру даних.

Моделі для пошуку закономірностей - це високо спеціалізовані моделі для пошуку неявних або неочевидних взаємозв’язків між подіями, тобто,
зазвичай, для кожної окремої задачі проектується своя система побудови моделі через те, що інформація в наборах даних з різних сфер роботи можуть мати кардинальні відмінності, коли мова йде про неявні взаємозв’язки.

Програмний засіб, створений під час виконання цієї роботи, демонструє використання дескриптивних моделей (бінарна класифікація відгуків) та моделей для пошуку закономірностей (пошук важливих аспектів в відгуках). Використання предиктивних моделей, на думку автора, краще підходить для робіт, тісніше пов’язаних або з статистичним аналізом, або зі штучним

інтелектом (апроксимаціями за допомогою нейронних мереж), тому в цій роботі не розглядається.

Найбільш часті виклики, які виникають під час індивідуального аналізу даних:

1. Видобуток різних видів знань у базах даних. – Потреби різних користувачів неоднакові. І різні користувачі можуть бути зацікавлені в різних видах знань. Тому необхідно, щоб інтелектуальний аналіз даних охоплював широкий спектр завдань виявлення знань.

2. Інтерактивний видобуток знань на кількох рівнях абстракції. – Процес інтелектуального аналізу даних має бути інтерактивним, оскільки він дозволяє користувачам зосередитися на пошуку шаблонів, надаючи та уточнюючи запити інтелектуального аналізу даних на основі отриманих результатів.

3. Включення базових знань. – Щоб керувати процесом відкриття та виражати виявлені закономірності, можна використовувати базові знання. Базові знання можуть бути використані для вираження виявлених закономірностей не тільки в стислих термінах, але й на багатьох рівнях абстракції.

4. Мови запитів інтелектуального аналізу даних і спеціальний інтелектуальний аналіз даних. – Мова запитів інтелектуального аналізу даних, яка дозволяє користувачеві описувати спеціальні завдання інтелектуального аналізу даних, повинна бути інтегрована з мовою запитів до сховища даних і оптимізована для ефективного та гнучкого інтелектуального аналізу даних.

5. Презентація та візуалізація результатів аналізу даних. – Після виявлення шаблонів їх потрібно виразити мовами високого рівня, візуальними представленнями. Ці представлення мають бути легко зрозумілими користувачам.

6. Обробка шумних або неповних даних. – Потрібні методи очищення даних, які можуть обробляти шум, незавершені об’єкти під час аналізу закономірностей даних. Якщо методів очищення даних немає, то точність виявлених патернів буде низькою.

7. Оцінка патернів – Це стосується цікавості проблеми. Виявлені закономірності мають бути цікавими, тому що вони можуть або бути загальновідомими, або не мати новизни.

8. Ефективність і масштабованість алгоритмів інтелектуального аналізу даних. – Щоб ефективно витягувати інформацію з величезної кількості даних у базах даних, алгоритм інтелектуального аналізу даних повинен бути ефективним і масштабованим.

9. Алгоритми паралельного, розподіленого та інкрементального майнінгу. – Такі фактори, як величезний розмір баз даних, широке поширення даних і складність методів інтелектуального аналізу даних спонукають до розробки паралельних і розподілених алгоритмів інтелектуального аналізу даних. Цей алгоритм розділяє дані на розділи, які далі обробляються паралельно. Потім результати з розділів об'єднуються. Інкрементні алгоритми оновлюють бази даних без повторного видобутку даних з нуля.

## 1.3 Аналіз функціонування програмних засобів визначення емоційного забарвлення тексту на прикладі аналізу даних мережі

Оскільки аналіз суджень є доволі молодим напрямом в інтелектуальному аналізі даних, приклади використання його не досить численні (принаймні ті, які знаходяться у відкритому доступі). Одним з прикладів успішного використання методологій можна назвати Університет Париж-Південь XI, чиї академіки в праці [9] розглянули проблему автоматизації збору корпусу даних з цілями аналізу настроїв і аналізу думок.

Науковці використали TreeTagger для POS-тегування та спостерігали різницю в розподілі між позитивними, негативними та нейтральними наборами. TreeTagger — це інструмент для анотування тексту частинами мови та інформацією про леми. Він був розроблений Гельмутом Шмідтом в Інституті комп'ютерної лінгвістики Університету Штутгарта. Зі спостережень вони дійшли висновку, що автори твітів у мережі Twitter використовують синтаксичні структури для опису емоцій або констатації фактів. Деякі POS-теги можуть бути сильними індикаторами емоційного тексту.

Науковці провели лінгвістичний аналіз зібраного корпусу даних. Використовуючи корпус, вони створили класифікатор настроїв, який здатний визначати позитивні, негативні та нейтральні настрої у документі. Класифікатор базувався на мультиноміальному класифікаторі наївного Байєса, який використовує N-грами та POS-теги як ознаки емоційного забарвлення. У дослідженні йшла робота з англійською мовою, однак автори наукової праці стверджують, що запропоновану методику можна буде використовувати з будь-якою іншою мовою. Детальніше про n-грами та теги, які було використано, буде розкрито в розділі 3.

### 1.4 Постановка задачі дослідження

Для досягнення мети кваліфікаційної роботи магістра потрібно виконати такі завдання:

1. Дослідити особливості та характеристики різних підходів до
   - структуризації та збереження даних
   - обробки даних та збір інформації
   - валідації результатів обробки.
2. Спроектувати та імплементувати програмний засіб для перевірки конкурентоспроможності обраного набору підходів.
3. Легально отримати дані для некомерційного тестування.
4. Підібрати параметри для оптимальних результатів обробки та отримати оброблену інформацію.

# РОЗДІЛ 2. АНАЛІЗ МЕТОДІВ ТА МЕТОДОЛОГІЙ АНАЛІЗУ ДАНИХ ДЛЯ ВИЗНАЧЕННЯ ЕМОЦІЙНОГО ЗАБАРВЛЕННЯ ТЕКСТІВ

## 2.1Огляд методологій Opinion Mining та Sentiment Analysis

Ці терміни часто використовують як синоніми, проте, з огляду на історію цих напрямків, все ж можна виділити різницю.

Є два важливі поняття, які тісно пов'язані з сентиментами та думками, я маю на увазі суб'єктивність та емоції. Об'єктивне речення представляє деяку фактичну інформацію про світ, тоді як суб'єктивне речення виражає деякі особисті почуття, погляди чи переконання. Приклад об'єктивного речення: «iPhone — продукт Apple». Прикладом суб'єктивного речення є «Мені подобається iPhone». Суб'єктивні прояви мають багато форм, наприклад, думки, твердження, бажання, переконання, підозри та припущення. Існує певна плутанина серед дослідників, які ототожнюють суб'єктивність із думкою. Під упевненістю ми маємо на увазі, що документ або речення виражає або має на увазі позитивне чи негативне почуття. Ці два поняття не є еквівалентними, хоча й мають великий перетин. Завдання визначити суб'єктивне чи об'єктивне речення називається класифікацією суб'єктивності. Тут слід зазначити наступне:

-Суб'єктивне речення може не виражати жодного почуття. Наприклад, «Я думаю, що він пішов додому» є суб'єктивним реченням, але не виражає жодних почуттів.

-Об'єктивні речення можуть передбачати думки чи почуття через бажані та небажані факти. Наприклад, наступні два речення, у яких наводяться деякі факти, чітко передбачають негативні настрої (які є прихованими думками) щодо відповідних продуктів, оскільки факти є небажаними: «Навушник зламався за два дні». «Я приніс матрац тиждень тому, і утворилася долина».

Окрім явних думок із суб'єктивним виразом, також досліджувалися багато інших типів суб'єктивності, хоча й не настільки широко, наприклад, афект, судження, оцінка, спекуляція, перспектива, суперечка, згода та незгода, політична позиція абощо. Багато з них також можуть означати почуття.

Емоції – це наші суб'єктивні відчуття і думки. Емоції вивчали в багатьох областях, наприклад, психологія, філософія та соціологія. Дослідження дуже широкі: від емоційних реакцій фізіологічних реакцій (наприклад, зміни частоти серцевих скорочень, артеріального тиску, потовиділення тощо), міміки, жестів і поз до різних типів суб’єктивних переживань душевного стану індивіда. Вчені класифікували емоції людей на кілька категорій. Однак серед дослідників досі немає узгодженого набору основних емоцій. Згідно з [16], люди мають шість основних емоцій, тобто любов, радість, подив, гнів, смуток і страх, які можна розділити на багато вторинних і третинних емоцій. Кожна емоція також може мати різну інтенсивність. Емоції тісно пов'язані з почуттями. Сила почуттів або думок зазвичай пов’язана з інтенсивністю певних емоцій, наприклад радості чи гніву. Думки, які ми вивчаємо в аналізі настроїв, здебільшого є оцінками (хоча не завжди). Згідно з дослідженнями поведінки споживачів, оцінки можна розділити на два типи: раціональні оцінки та емоційні оцінки.

Раціональна оцінка: такі оцінки базуються на раціональних міркуваннях, реальних переконаннях і утилітарних установках. Наприклад, такі речення виражають раціональні оцінки: «Голос цього телефону чистий», «Ця машина варта своєї ціни» і «Я задоволений цією машиною».

Емоційна оцінка: такі оцінки базуються на нематеріальних та емоційних реакціях на сутності, які проникають глибоко в душевний стан людей. Наприклад, наступні речення виражають емоційну оцінку: «Я люблю iPhone», «Я так злий на їхніх обслуговуючих людей» і «Це найкраща машина, коли-небудь створена».

Щоб використати ці два типи оцінок на практиці, ми можемо розробити 5 рейтингів настроїв: емоційно негативний (-2), раціонально негативний (-1), нейтральний (0), раціонально позитивний (+1) і емоційно позитивний (+ 2). На практиці нейтральний часто означає відсутність висловленої думки чи почуття. Нарешті, ми повинні зазначити, що поняття емоції та думки явно не еквівалентні. Раціональні думки не виражають емоцій, наприклад, «Голос цього телефону чистий», а багато емоційних речень не виражають жодної думки/почуття щодо

будь-чого, наприклад, «Я так здивований, що вас тут бачу». Що ще важливіше, емоції можуть не бути спрямованими на якийсь об’єкт, а лише відображати внутрішні почуття людей, наприклад, «Мені сьогодні так сумно».

Однак слід розуміти, що всі ці поняття та їх визначення досить розмиті та суб'єктивні. Наприклад, досі не існує певного набору емоцій, з яким погоджуються всі дослідники. Сама думка також є широким поняттям. Аналіз настроїв в основному стосується оцінки думок або думок, які передбачають позитивні чи негативні настрої. Мета цього розділу — дати достатньо точне визначення сентиментного аналізу і пов’язаних з ним питань.

Отож, sentiment analysis (з англ. сентиментний аналіз) - це процес (або методологія) інтерпретації та класифікації емоцій в текстовій інформації з використанням технологій текстового аналізу. Такі підходи використовуються для визначення відношення клієнтів до продуктів, брендів, сервісів без необхідності в опитуваннях чи, популярного в минулому десятилітті, випрошування оцінок (прохання типу “будь ласка, залиште відгук про наш продукт”). Також зараз стало популярним інше використання таких методів — політичні та соціологічні дослідження. Тепер замість дорогих опитувань є варіант зчитати велику кількість дописів з соціальних мереж і дістати інформацію звідти. Аналіз настроїв – це сфера дослідження, яка аналізує думки, настрої, оцінки, схвалення/засудження, ставлення та емоції людей щодо таких об’єктів, як продукти, послуги, організації, особи, питання, події, теми та їхні атрибути. Це дуже великий проблемний простір.

Opinion mining же має за свою мету: дослідити причини, що стоять за тими чи іншими емоціями в тексті, щоб ідентифікувати певні аспекти продукту чи події, які потрібно переглянути, змінити, чи на яких треба наголосити в маркетинговій кампанії. Текстову інформацію у світі можна умовно поділити на дві основні категорії: факти та думки. Факти – це об'єктивні твердження про суб'єктів та події у світі. Точки зору – це суб'єктивні твердження, які відображають почуття або сприйняття людей щодо об'єктів та подій. Саме поняття «думки» все ще дуже широке. Аналіз настроїв і аналіз думок переважно

зосереджені на думках, які виражають або передбачають позитивні чи негативні настрої.

Хоча лінгвістика та обробка природної мови (NLP) мають довгу історію, до 2000 року було проведено мало досліджень щодо думок і настроїв людей. Відтоді ця сфера стала дуже активною дослідницькою областю. На це є декілька причин. По-перше, він має широкий спектр застосувань, майже в кожному домені. Галузь аналізу настроїв також процвітала завдяки поширенню комерційних програм. Це створює сильну мотивацію для дослідження. По-друге, ця галузь пропонує багато складних дослідницьких проблем, які ніколи раніше не вивчалися. У цій книзі систематично визначено та обговорено ці проблеми, а також описано сучасні методи їх вирішення. По-третє, вперше в історії людства ми тепер маємо величезний обсяг думок у соціальних мережах Інтернету. Без цих даних багато досліджень були б неможливими. Не дивно, що зародження та стрімке зростання аналізу настроїв збігаються з соціальними медіа. Насправді аналіз настроїв зараз знаходиться в центрі досліджень соціальних мереж. Таким чином, дослідження в області аналізу настроїв не тільки мають важливий вплив на НЛП, але також можуть мати глибокий вплив на науки про управління, політологію, економіку та соціальні науки, оскільки всі вони залежать від думок людей. Тож різницю між цими напрямками можна коротко підсумувати так: сентиментний аналіз - це попередник opinion mining, що розглядає емоції в людей за наданою інформацією, а сам напрям opinion mining шукає причини тих чи інших емоцій для визначення аспектів досліджуваних подій чи об’єктів, що є важливими з точки зору певної групи людей чи широкого загалу.


Рисунок 2.1 Загальна схема роботи сентиментного аналізу

Обидва напрями звертають увагу на велику кількість явних характеристик тексту, такі як довжина слів, кількість стоп-слів, розмір використаного словникового запасу, повнота пунктуації, орфографічні помилки, ключові слова. Потужні системи також можуть, крім самих слів, при аналізі знаходити також їх синоніми, можливі неправильні написання тощо.

Сфери використання обох методологій широкі та добре розвинені як в промисловості, так і в політичному секторі.

Наприклад перед виборчою кампанією 2012 року в США адміністрація Барака Обами використала сентиментний аналіз, щоб швидко зрозуміти ставлення виборців до свого кандидата та оцінити реакцію на окремі дії в межах кампанії [2]. Ще одним відомим прикладом є порятунок провальної рекламної кампанії Expedia Canada. Причину її негативного впливу на продажі не було зрозуміло аж поки не застосували opinion mining з дописами з кількох популярних соціальних мереж в якості вхідних даних. Виявилось, що багатьох людей дратувала музика в рекламному відеоролику. З розумінням цієї, раніше прихованої, проблеми маркетологи розіграли все на користь компанії, висміявши свою невдачу в наступному рекламному ролику [2]. Також відомо що спеціальні служби різних країн (в тому числі програма PRISM уряду США [3]) автоматизують знаходження потенційно небезпечних людей через соціальні мережі. Ще один відомий спосіб використання — аналіз транскриптів розмов з

кол-центрів з метою виявити часті проблеми в користувачів або ситуації, що викликають найбільше проблем.

Підходи датамайнингу в мережі Інтернеті покладаються на той факт, що в мережі люди напряму висловлюють свою думку з приводу речей, що їм подобаються чи не подобаються. Для аналітиків великих компаній чи політиків, відгуки про продукти чи дії яких рахуються десятками чи сотнями тисяч, переглядати всі ці відгуки зовсім недоцільно. Для аналізу цих відгуків використовують автоматизовані системи збору інформації, що спираються на методи датамайнингу. Під кожен окремий випадок налаштовують різні методи, які потребують низки правильно заданих параметрів для ефективної роботи.

Будь яка зібрана інформація може бути корисною. Наприклад, книжкове видавництво могло б подивитись на статистику використання слів “обкладинка”, “персонажі”, “історія” в позитивних та негативних контекстах серед відгуків та дізнатись, як саме ці аспекти впливають на сприйняття продукту. Зібравши певну кількість “сирої” інформації, можна побудувати більш специфічну інформацію та зібрати менш очевидні відомості.

Загалом сентиментний аналіз досліджувався в основному на трьох рівнях поглиблення в дані.

На рівні документів: Завдання на цьому рівні полягає в тому, щоб класифікувати, чи виражає весь документ думки позитивні чи негативні почуття. Наприклад, з огляду на продукт система визначає, чи відгук виражає загальну позитивну чи негативну думку про продукт. Це завдання широко відоме як класифікація настроїв на рівні документа. Цей рівень аналізу припускає, що кожен документ висловлює думку про одну сутність (наприклад, окремий продукт). Таким чином, він не застосовується до документів, які оцінюють або порівнюють кілька об’єктів.

На рівні окремих речень: завдання на цьому рівні стосується речень і визначає, чи кожне речення виражає позитивну, негативну чи нейтральну думку. Зазвичай нейтральність означає відсутність думки. Цей рівень аналізу тісно пов’язаний із класифікацією суб’єктивності, яка розрізняє речення (так звані

об’єктивні речення), які виражають фактичну інформацію, від речень (так звані суб’єктивні речення), які виражають суб’єктивні погляди та думки. Однак слід зазначити, що суб’єктивність не еквівалентна настроям, оскільки багато об’єктивних речень можуть означати думки, наприклад, «Ми купили машину минулого місяця, і склоочисник відпав». Дослідники також проаналізували речення, але рівня речення все ще недостатньо, наприклад, «Apple дуже добре себе почуває в цій паршивій економіці».

Рівень сутності та аспекту: аналіз як на рівні документа, так і на рівні речення не визначає, що саме сподобалося людям, а що ні. Рівень аспектів виконує більш детальний аналіз. Рівень аспектів раніше називався рівнем ознак, або «фіч» (визначення думок і узагальнення на основі ознак). Замість перегляду мовних конструкцій (документів, абзаців, речень, пунктів або фраз), аспектний рівень розглядає безпосередньо саму думку. Він заснований на ідеї, що думка складається з почуття (позитивного чи негативного) і цілі (думки). Думка без визначеної мети має обмежене використання. Усвідомлення важливості оцінки думок також допомагає нам краще зрозуміти проблему аналізу настроїв. Наприклад, хоча речення «хоча обслуговування не таке чудове, я все одно люблю цей ресторан» явно має позитивний відтінок, ми не можемо сказати, що це речення цілком позитивне. Насправді в реченні є позитивне ставлення до ресторану (підкреслено), але негативне щодо його обслуговування (не підкреслене). У багатьох програмах цільові думки описуються сутностями та/або їх різними аспектами. Таким чином, метою цього рівня аналізу є виявлення настроїв щодо сутностей та/або їх аспектів. Наприклад, речення «Якість дзвінків у iPhone хороша, але час роботи від батареї короткий» оцінює два аспекти iPhone (суб’єкта) — якість дзвінка та час роботи батареї. Настрої щодо якості дзвінків на iPhone позитивні, але настрої щодо часу автономної роботи негативні. Якість дзвінків і час автономної роботи iPhone є цілями оцінки. На основі цього рівня аналізу можна створити структуроване резюме думок про сутності та їх аспекти, яке перетворює неструктурований текст на структуровані дані та може використовуватися для всіх видів якісного та кількісного аналізу.

Класифікація як на рівні документа, так і на рівні речення вже є надзвичайно складною. З аспектним рівнем ще складніше. Щоб зробити речі ще більш цікавими та складними, існує два типи думок, тобто звичайні думки та порівняльні думки. Звичайна думка виражає настрої лише щодо певної сутності або аспекту сутності, наприклад, «Кола дуже смакує», що виражає позитивну думку щодо аспекту смаку напою Кока-кола. Порівняльний висновок порівнює кілька об'єктів на основі деяких їхніх спільних аспектів, наприклад, «Кока-Кола на смак краща, ніж Пепсі», у якому порівнюється Кока-Кола та Пепсі на основі їхніх смаків (аспект) і виражається перевага Кока-колі. Звичайна думка: звичайну думку в літературі часто називають просто думкою, і вона має два основні підтипи: Пряма думка: пряма думка відноситься до думки, висловленої безпосередньо щодо сутності або аспекту сутності, наприклад, «якість зображення чудова». Непряма думка: непряма думка – це думка, яка опосередковано висловлюється щодо сутності або аспекту сутності на основі його впливу на деякі інші сутності. Цей підтип часто зустрічається в медичній сфері. Наприклад, речення «Після ін'єкції препарату моїм суглобам стало гірше» описує небажаний вплив препарату на «мої суглоби», що опосередковано дає негативну думку або настрій щодо препарату. У цьому випадку суб'єктом є препарат, а аспектом – вплив на суглоби. Значна частина сучасних досліджень зосереджена на прямих думках. З ними простіше обходитися. З непрямими думками часто важче мати справу. Наприклад, у медицинській сфері потрібно знати, чи є якийсь бажаний і небажаний стан до чи після вживання медпрепарату. Наприклад, речення «Оскільки мої суглоби боліли, мій лікар призначив мені цей препарат» не виражає почуття чи думки щодо препарату, оскільки «болючі суглоби» (що є негативним) виникали перед використанням препарату.

Порівняльна думка: Порівняльна думка виражає відношення подібності чи відмінності між двома чи більше об'єктами та/або перевагу власника думки на основі деяких спільних аспектів об'єктів. Наприклад, речення «Кока-Кола смакує краще, ніж Пепсі» і «Кока-Кола смакує найкраще» виражають дві порівняльні думки. Порівняльна думка зазвичай виражається за допомогою

порівняльної або найвищої форми прикметника чи прислівника, хоча не завжди (наприклад, віддаю перевагу). Порівняльні думки також мають багато видів. Вони не тільки мають різні семантичні значення, але й мають різні синтаксичні форми. Наприклад, типове звичайне речення: «Якість голосу цього телефону чудова», а типове порівняльне речення: «Якість голосу телефонів Nokia краща, ніж iPhone». Це порівняльне речення не означає, що якість голосу будь-якого телефону хороша чи погана, а просто порівнює їх. Через цю різницю порівняльні думки вимагають різних методів аналізу. Подібно до звичайних речень, порівняльні речення можуть бути думними чи не думними. Порівняльне речення вище є думкою, оскільки воно прямо виражає порівняльні почуття свого автора, тоді як речення «iPhone на 1 дюйм ширший за звичайний телефон Nokia» не виражає ніяких думок.

Порівняльне речення виражає відношення за подібністю або

відмінності більш ніж однієї сутності. Існує кілька видів порівнянь. Їх можна згрупувати у дві основні категорії: порівняння, що оцінюється, і порівняння, що не піддається оцінці.

Порівняння, що оцінюються: Таке порівняння виражає впорядкований зв’язок об’єктів, що порівнюються. Він має три підвиди:

1. Нерівне градуюче порівняння: воно виражає відношення типу «більше» або «менше», ніж яке ранжує набір сутностей над іншим набором сутностей на основі деяких їхніх спільних аспектів, наприклад, «Кока-кола на смак краща, ніж Пепсі». Цей тип також включає вподобання, наприклад, «Я віддаю перевагу кока-колі, ніж пепсі».

2. Еквативне порівняння: воно виражає відношення типу рівності, яке стверджує, що дві або більше сутності є рівними на основі деяких їхніх спільних аспектів, наприклад, «Кола та Пепсі мають однаковий смак».

3. Чудове порівняння: воно виражає відношення типу «більше» або «менше, ніж усі інші», яке ставить одну сутність над усіма іншими, наприклад, «Кока-кола має найкращий смак серед усіх безалкогольних напоїв».

Порівняння що не оцінюється: таке порівняння виражає відношення двох або більше сутностей, але не оцінює їх. Є три основні підтипи:

1. Сутність А схожа на сутність В або відрізняється від неї на основі деяких спільних аспектів, наприклад, «Смак кока-коли відрізняється від пепсі».

2. Сутність А має аспект $a_1$, а сутність В має аспект $a_2$ ($a_1$ і $a_2$ зазвичай взаємозамінні), наприклад, «Настільні ПК використовують зовнішні динаміки, а ноутбуки — внутрішні динаміки».

3. Сутність А має аспект а, але сутність В не має, наприклад, «телефони Nokia постачаються з навушниками, а iPhone — ні».

Розглянемо детальніше підходи до сентиментного аналізу. Всі їх можна розділити на дві великі категорії — аналіз коротких текстів та аналіз довгих текстів. В цілому в цих категорій є багато спільного, але аналіз довгих текстів не підходить для аналізу дописів в соціальних мережах чи більшості відгуків, так як довгі тексти на практиці зустрічаються нечасто. Аналіз довгих текстів, крім спільних підходів, опирається на семантичний розбір тексту, що за низкою причин значно уповільнює обробку, збільшує вимоги до технічного оснащення та складність розробки. Аналіз великих текстів сильно покладається на методи обробки природньої мови (NLP) для аналізу синтаксису семантики тексту за великою кількістю введених вручну правил правопису та семантичної логіки.

Аналіз короткого тексту використовує простіші характеристики, які можна швидше обрахувати та компактніше зберегти в базі даних. Така обробка, зазвичай, потребує менше оперативної пам’яті та може бути ефективно розпаралелена. Існує три основних напрямки в аналізі коротких текстів:

- Аналіз на основі штучного інтелекту
- Аналіз соціальних мереж (SNA)
- Аналіз на основі лексикону

Аналіз на основі штучного інтелекту використовує дуже велику кількість інформації для тренування, тобто текстів, що були вручну оцінені людьми за досліджуваними критеріями. Всі ці тексти використовуються, щоб за допомогою

функції-оптимізатора натренувати нейронну мережу для оцінки певних параметрів тексту.

Аналіз соціальних мереж - це сімейство підходів, що, крім самого тексту, звертають увагу також на його контекст, тобто на автора допису, його інтересів, його коло спілкування, місця, в якому було залишено допис, інших дописів автора тощо. Проте як основу аналізу використовують методи аналізу на основі штучного інтелекту або на основі словника сентиментів.

Аналіз на основі словника аналізує самі слова в тексті, зазвичай, з мінімальною увагою до синтаксису чи зовсім без неї. В класичних підходах слова перевіряються за допомогою словників емоційного забарвлення слів. Проблеми для дослідників, які будують словники емоційного забарвлення слів стосуються як теоретичних аспектів, таких як об'єктивні закони вираження почуттів природною мовою, так і практичних аспектів, наприклад, аналіз оглядів споживчих товарів і послуг, моніторинг соціальних мереж, політичні дослідження Різні підходи та задачі потребують різних доповнень до цієї простої ідеї — врахування заперечень, порівнянь, прислівників міри та ступеню дій, тощо. Окремою задачею для будь якої системи, що спирається на словниковий аналіз є вибір або створення словника емоційного забарвлення (за яким потім аналізуються слова). Можна виділити два етапи створення словників: генерація списку сентиментальних слів, що містить кандидати у сентиментний словник, і
призначення «міток сентименту» цим словам, наприклад позитивний/негативни й/нейтральний. Обидва етапи виконуються вручну або автоматично. Одним зі старих підходів є створення словника за результатами опитування, в якому групу людей напряму запитують про те, як вони сприймають певне слово в кількох варіантах використання. Проте з популяризацією датамайнингу такий спосіб став відходити на другий план. Актуальні методи розглянемо в наступних главах.

## 2.2 Дослідження підходів до побудови сентиментного словника

Використання словника для складання слів-сентиментів є очевидним підходом, оскільки більшість словників наводять синоніми та антоніми для кожного слова. Таким чином, проста техніка в цьому підході полягає у використанні кількох початкових слів для завантаження на основі синонімічної та антонімічної структури словника. Зокрема, цей метод працює так: спочатку вручну збирається невеликий набір слів-сентиментів (зерен) із відомою позитивною чи негативною орієнтацією, що дуже легко. Алгоритм потім збільшує цей набір, шукаючи в WordNet або іншому онлайн-словнику їх синоніми та антоніми. Щойно знайдені слова додаються до початкового списку. Починається наступна ітерація. Ітеративний процес завершується, коли більше не знайдено нових слів. Після завершення процесу для очищення списку можна використати ручну перевірку.

Підсумовуючи, ми зазначимо, що перевага використання підходу, заснованого на словнику, полягає в тому, що можна легко і швидко знайти велику кількість слів з настроями та їх орієнтацією. Хоча отриманий список може містити багато помилок, можна виконати ручну перевірку, щоб очистити його, що забирає багато часу (не так погано, як люди думали, лише кілька днів для носія мови), але це лише одноразова спроба . Головним недоліком є те, що сентиментальна орієнтація слів, зібраних таким чином, є загальною або незалежною від домену та контексту. Іншими словами, важко використовувати підхід, заснований на словнику, щоб знайти залежну від домену або контексту орієнтацію слів настрою. Багато слів-сентиментів мають залежну від контексту орієнтацію. Наприклад, для телефону, якщо він тихий, то це зазвичай мінус. Однак для автомобіля, якщо він тихий, це позитивно. Сентиментальна орієнтація слова «тихий» залежить від домену чи контексту.

Не дивно, що найважливішими індикаторами настроїв є слова настроїв, які також називають словами думок. Це слова, які зазвичай використовуються для вираження позитивних чи негативних настроїв. Наприклад, «хороший», «чудовий» та «дивовижний» – це позитивні слова, а «неприємний», «поганий» та «жахливий» – негативні. Крім окремих слів, є також словосполучення та

ідіоми, наприклад, «Руку би віддав за цей продукт». Слова та фрази, що висловлюють настрої, є інструментом для аналізу настроїв із очевидних причин. Список таких слів і фраз називається словником настроїв (або сентиментним словником). Протягом багатьох років дослідники розробили численні алгоритми для складання таких словників.

Хоча слова та фрази, що виражають настрої, важливі для аналізу настроїв, лише їх використання недостатньо. Проблема набагато складніша. Іншими словами, ми можемо сказати, що сентиментний словник необхідний, але недостатній для аналізу настроїв. Розглянемо кілька проблем, що стосуються цього, нижче:

Речення, що містить слова настрою, може не виражати жодного настрою. Це явище часто трапляється в кількох типах речень. Питальні та умовні речення є двома важливими типами, наприклад, «Чи можете ви сказати мені, яка камера Sony хороша?» і «Якщо я знайду хорошу камеру в магазині, я її куплю». Обидва ці речення містять прикметник «хороший», але жодне не виражає позитивної чи негативної думки щодо будь-якої конкретної камери. Однак не всі умовні речення чи питальні речення не виражають почуття, наприклад, «Хтось знає, як відремонтувати цей жахливий принтер» або «Якщо ви шукаєте хорошу машину, купіть Toyota».

Багато речень без сентиментальних слів також можуть передбачати думки. Багато з цих речень насправді є об’єктивними реченнями, які використовуються для вираження певної фактичної інформації. Знову ж таки, існує багато типів таких речень. Тут ми наведемо лише два приклади. Речення «Ця пральна машина використовує багато води» означає негативне ставлення до пральної машини, оскільки вона використовує багато ресурсів (води). Речення «Після двох днів сну на матраці посередині утворилася долина» висловлює негативну думку про матрац. Це речення об’єктивне, оскільки в ньому констатується факт. У всіх цих реченнях немає сентиментальних слів.

В аналізі, що спирається на певні, заздалегідь підготовлені словники, дуже важливо правильно обрати спосіб його побудови, бо побудова точно

відкаліброваного словника займає дуже великий час в порівнянні з самим аналізом тексту та визначає шанс на успіх обраного підходу.

1. Метод побудови вручну

Один з підходів до побудови словника вже згадувався раніше. Це побудова через оцінку слів вручну. Група людей проглядає звичайний словник (без оцінок) і вручну виставляє словам оцінки відповідно до свого сприйняття слова. Очевидно, що такий підхід протирічить філософії автоматизації та спеціалізації праці, проте результати інколи оправдовують підхід, адже люди поки що справляються з розумінням природної мови значно краще за комп'ютери. Приклад використання цього методу можна знайти в роботі [4].

2. Метод синонімічного збагачення

Цей метод є прямим покращенням попереднього. Його ідея полягає в використанні тлумачного словника з вказаними синонімами слів, щоб після введення деякої малої кількості оцінок вручну можна було б ітеративно розширювати словник, доповнюючи його синонімами вже оцінених слів зі словника синонімів.

3. Метод контекстного доповнення (Corpus-based)

Іншим покращенням створення словника вручну є метод, при якому лише мала частина слів повинна бути оцінена вручну, а словник доповнюється в процесі аналізу великої кількості текстів. Метод контекстного доповнення, застосовують щоб досягти двох основних сценаріїв: (1) отримати вихідний список відомих (часто загального призначення) слів настрою, виявити інші слова настрою та їх орієнтації з корпусу домену та (2) адаптувати загальний -цільовий сентиментний словник на новий з використанням корпусу домену для програм аналізу настроїв у домені. Однак проблема є складнішою, ніж просто побудова сентиментного словника для конкретної області, тому що в тій самій області одне й те саме слово може бути позитивним в одному контексті та негативним в іншому. Зауважте, що хоча підхід, заснований на корпусі, також може бути використаний для створення словника настроїв загального призначення, якщо

доступний дуже великий і дуже різноманітний корпус, підхід, заснований на словнику, зазвичай більш ефективний для цього, оскільки словник містить усі слова.

Основна ідея заключається в тому, щоб для обраної мови виділити множину сполучників та категоризувати їх за співвідношенням слів, що пов'язані цими сполучниками. Наприклад, слова «і», «та», «також» і т.д. майже завжди пов'язують слова з однаковим емоційним забарвленням, а слово «але» пов'язує протилежно забарвлені слова. Тому, вписавши в словник кілька оцінок вручну, можна пройтись по всім наявним текстам та знайти словосполучення, де є вже введені слова. Наприклад записавши в словник слово «охайний» як позитивно забарвлене можемо знайти уривок «охайний ТА красивий» в тексті, з цього робимо висновок що слово «красивий» також позитивне і додаємо його в словник. А від уривку «красивий АЛЕ незручний» робимо висновок що «незручний» — негативне слово. Цей підхід зменшує кількість роботи, необхідної для побудови словника, але для нього потрібно мати деяку велику базу текстів, бажано різноманітних. Зокрема, не тільки слова можуть допомогти провести аналіз сентиментів Інтернет-повідомлення – безліч користувачів використовують смайлики або емоджі в своїх коментарях, як позитивні ":-)", ":)", "=)", ":D" абощо, так і негативні ":-(", ":(", "=(", ";(". Два типи зібраних масивів символів будуть використовуватися для навчання класифікатора по розпізнаванню позитивних і негативних настроїв. Це може бути не дуже ефективно при аналізі великих та середніх текстів, але водночас бути екстремально ефективним при аналізі мікроблогів, таких як в Twitter чи Facebook. Оскільки за правилами платформи Twitter мікроблогів кожне повідомлення не може перевищувати 140 символів, зазвичай воно складається з одного речення. Тому ми припускаємо, що смайлик у повідомленні представляє емоцію для всього повідомлення, і всі слова в повідомленні пов'язані з цією емоцією. Цей метод не тільки спрощує аналіз – його можна легко адаптувати під будь-яку мову, бо смайлики в абсолютній більшості випадків не залежать від

мови тексту. Детальніше про цей підхід та його сучасні покращення можна прочитати в роботі [5,9].

3. Статистичний підхід з оціненими словами

Заповнивши словник малим числом оцінок та маючи базу різноманітних текстів, можна спробувати використати статистичний підхід до побудови словника. Ідея схожа на ідею попереднього підходу, але замість орієнтування по зв'язуючим словам можна дивитись одразу на весь текст — якщо в тексті багато позитивно оцінених слів, то можна припустити, що інші слова у цьому тексті з великою ймовірністю також є позитивними Обравши якусь метрику, яка нормалізує, можна зробити припущення щодо всіх слів, які зустрічаються достатньо часто в текстах з відомими позитивними словами. Заповнивши словник частиною найбільш вірогідних слів, можна повторювати процес доти, доки на одній з ітерацій нові слова перестануть бути точними.

4. Статистичний підхід з оціненими текстами

Маючи велику кількість оцінених текстів, можна почати заповнення словника без ручної оцінки слів. Метод спирається на припущення, що слова в позитивному тексті, скоріше всього, теж позитивні. Знову ж таки, збираємо слова з якоюсь метрикою, що характеризує частоту входження слів в позитивні або негативні тексти, і обираємо всі слова, які мають достатньо високе значення цієї метрики (вище якогось порогу). Для цього, зазвичай, корисно використовувати логарифмічні метрики для меншої схильності процесу обирати популярні слова перед релевантними. Також припускаю, що в цьому методі корисно використати стабілізацію результатів за допомогою метрики TFIDF, яка описана в наступних розділах. Цей метод корисний, якщо немає можливості підготувати словник вручну, проте він потребує набору текстів та деякого часу для виконання комп'ютером. Детальніше про такий метод та його сучасні покращення можна прочитати в роботі [6]

5. Статистичний підхід з метрикою TFIDF

TFIDF (Term Frequency - Inversed Document Frequency, з англ. Частота терміну до оберненої частоти по документах) це метрика важливості слова в деякому наборі текстів, яка рахується за формулою:

$\text{tf-idf}(t, d, D) = \text{tf}(t, d) \times \text{idf}(t, D)$,де

$$\text{idf}(t, D) = \log \frac{|D|}{|\{ d_i \in D \mid t \in d_i \}|}$$, тут

в чисельнику – кількість документів в колекції, а в знаменнику – кількість документів, в яких зустрічається слово t.

$$\text{tf}(t, d) = \frac{n_t}{\sum_k n_k}$$ , тут

в чисельнику - кількість входжень слова t в документа, в знаменнику - кількість слів в документі.

Ця метрика дозволяє шукати важливі слова в тексті за простими обрахунками, виключаючи слова, що часто використовуються в будь-яких текстах (прислівники, займенники, сполучники тощо). Важливо при використанні цієї метрики мати в системі словник синонімів та асоціювати слова-синоніми, щоб в контексті обрахунку TF-IDF не рахувати майже ідентичні за змістовим навантаженням слова як відмінні.

7. Семантично та синтаксично свідомі підходи

Сімейство семантично свідомих підходів можна назвати покращенням синонімічного збагачення словника, в таких підходах аналізуються слова подібні за змістом, а не тільки прямі синоніми.

Методи обробки природньої мови (NLP) іноді використовуються разом з аналізом за лексиконом, щоб збагачувати словник за допомогою синтаксичної та семантичної інформації. До цього підходу відноситься врахування заперечень («не», «ні», «ніколи» тощо), прислівники ступеню («дуже», «трохи», «занадто» тощо). Використання цих підходів сильно збільшує точність аналізу, адже всі

емоційно забарвлені слова з запереченнями, які не були правильно оброблені, будуть діяти на оцінку прямо протилежно до очікуваного результату. А слова, на яких автор хотів наголосити, будуть зчитані як звичайні, якщо не врахувати прислівники, що стоять поруч.

### 2.3 Огляд допоміжних методів

1. Спеціалізація словників за контекстом

Відомо, що в різних контекстах слова можуть мати різні значення. Неоднозначність полярності стосується труднощів правильної інтерпретації почуття, пов’язаного з певними словами. Деякі слова можуть виражати позитивні настрої в одному контексті та негативні настрої в іншому, що ускладнює точну інтерпретацію настроїв моделям аналізу настроїв. Крім того, є також слова, які мають кілька тлумачень, що ускладнює для моделі визначення виражених настроїв. Щоб вирішити цю проблему, модель має бути навчена значною кількістю даних, щоб навчитися пов’язувати певні слова з відповідним почуттям. З цього випливає, що і емоційне забарвлення в них може бути різне. Наприклад, нейтральне, зазвичай, слово “олень” в жаргоні автомобілістів означає невмілого кермувальника і має негативне емоційне забарвлення. З огляду на це часто створюють окремі словники, які або є повними словниками сентиментних оцінок для обраного контексту, або доповнюють якийсь базовий словник термінів (накладаються на вже існуючі оцінки). Використовуючи такі словники можна добитись більшої точності в конкретних областях використання. Проте для створення таких словників потрібні спеціалізовані тексти, доступна кількість яких може бути обмежена.

2. Адаптація словників під інші мови

Маючи один заповнений словник сентиментних оцінок, можна спробувати перекласти його автоматичним перекладачем на інші мови, щоб не заповнювати словники для всіх необхідних мов вручну. При такому підході страждає точність оцінок та синонімічні зв’язки, не говорячи вже про помилки при перекладі, але,

з іншої сторони, так можна швидко підготувати словник для ручного перегляду оцінок та редагування, якщо необхідно.

3. Небінарні оцінки сентименту

Деякі старі методи побудови словників використовують оцінки виду позитивний / негативний (бінарна оцінка) або позитивний / нейтральний / негативний (тернарна оцінка). Такі оцінки, зазвичай, використовується лише при ручному заповненні словника. Такий підхід сильно зменшує точність аналізу. Очевидно, що деякі слова, навіть будучи одного емоційного забарвлення, можуть мати різну інтенсивність емоційного навантаження. Для прикладу візьмемо слова “непоганий” та “чудовий”. Обидва слова несуть позитивне емоційне забарвлення, проте очевидно, що “чудовий” — більш позитивне слово, а в системі бінарних оцінок це ніяк не відображується.

Покращенням в цьому питанні є підхід запису емоційного забарвлення в числовому еквіваленті — від -1.0 (найбільш негативне) до +1.0 (найбільш позитивне). Оцінки в числах з плаваючою точкою можуть мати набагато більшу точність. Для прикладу, в бінарній системі оцінок вислів “непоганий дизайн, але жахливий функціонал” мав би нейтральну оцінку (по одному негативному і позитивному слову), в той час як з оцінками в числах можна було б добитись більшої точності.

4. Нейронні мережі для бінарної класифікації

Щоб проаналізувати полярність від негативного до позитивного, можна використати машинне навчання. Використовуючи його, машини можуть визначати емоції будь-якого речення без участі людей, в основному, комп’ютери чи штучний інтелект можуть вивчати нові завдання без систематичної програми та виконувати їх ідеальним чином. Його також можна тренувати, і він зможе зрозуміти такі несвідомі для комп’ютера речі як сарказм, визначення, контекст і слова, які недоречні. Проіндексувавши всі слова для майбутнього словника, можна для будь якого тексту створити вектор значень, де для значення під номером кожного слова в словнику стоїть

1 якщо слово є в тексті або 0, якщо його нема. Якщо подати цей вектор на вхід в нейронну мережу, в якої кількість вхідних нейронів відповідає розмірності векторів, та подати як значення для оптимізації виводу числові значення, в яких закодована позитивність тексту, що відповідає заданому вектору, то можна натренувати нейронну мережу, що буде за наявністю певних слів передбачати емоційний настрій тексту. Позитивні сторони такого методу — нема необхідності в ручній оцінці слів; зі збільшенням кількості тренувальних тестів збільшується і точність оцінок. Негативні сторони — розробка такої нейронної мережі хоч і не складна, але потребує додаткового вивчення теорії нейронних мереж для створення ефективної архітектури; тренування займає тривалий час (що може вимірюватись годинами чи днями — в залежності від кількості даних та “епох” тренування); використання створеної моделі значно складніше ніж простий перегляд словника та займає більше часу як для підготовки, так і для виконання. На мою думку, використання нейронних мереж в задачах, з якими чудово справляються звичайні методи — недоцільне, проте за допомогою нейронних мереж можна робити і складніші речі в цій області. Наприклад, нейронні мережі с великою кількістю шарів нейронів можна натренувати для розпізнавання семантичних зв’язків в тексті, щоб отримати адаптивну NLP систему, за допомогою якої можна видобувати не тільки апроксимації емоційного навантаження, а й структуровані данні щодо будови текстів, їх стилістики та фактичного змісту. Приклад використання таких підходів описано в роботах Ruiz M, Srinivasan P. Hierarchical neural networks for text categorization [7] та Ng Hwee Tou, Goh Wei, Low Kok. Feature selection, perceptron learning, and a usability case study for text categorization. [8].

## 5. Використання н-грам в opinion mining

Розбиття тексту на н-грами дозволяє проводити статистичний аналіз не тільки окремо взятих слів, а і словосполучень, що є важливим елементом обробки натуральної мови (NLP). Для створення і збереження н-грам

доведеться пожертвувати значною кількістю місця в базі даних та часом виконання. Проте можливості, які відкриває цей підхід, на мою думку, незрівнянно більші, аніж затрати. Використання н-грам в датамайнингу добре описано в роботі Alexander Pak, Patrick Paroubek. Twitter as a Corpus for Sentiment Analysis and Opinion Mining [9]

### 2.4 Виклики Opinion Mining

1. Сарказм в текстах

Сарказм — це витончена форма мовленнєвого акту, у якій мовці або пишучі говорять або пишуть протилежне тому, що вони мають на увазі. Сарказм вивчають у лінгвістиці, психології та когнітивній науці. У контексті аналізу настроїв це означає, що коли хтось говорить щось позитивне, він/вона насправді має на увазі негативне, і навпаки. З сарказмом дуже важко боротися. Однією з найскладніших проблем в opinion mining є сарказм, розпізнати який в текстах неймовірно складно, адже, щоб комп'ютерній системі виявити його серед щирих дописів, їй потрібно мати дуже високий рівень розуміння тексту. Одним з можливих рішень є пошук високого контрасту в емоційних забарвленнях сусідніх слів. Для прикладу візьмемо вирази типу «я так обожнюю невчасну доставку!», в них поруч стоять слова «обожнюю» та «невчасну». Зрозуміло, що так можна знайти далеко не весь сарказм. Сарказм не так поширений у відгуках споживачів про продукти та послуги, але дуже сильно поширений у політичних дискусіях, через що важко мати справу з політичними думками.

Ще одним підходом є аналіз контексту даного тексту. Часто саркастичне висловлювання становить лише частину допису, тому буває корисно також перевіряти, чи не контрастує в емоційному забарвленні частина тексту з рештою допису.

2. Орфографічні помилки

Багато текстів мають певні орфографічні помилки, що ускладнює аналіз будь якими методами — слово з помилкою все рівно, що нове, не існуюче в словнику слово. Хоч похибка буде несуттєва для конкретного тексту, вона може наростати як сніжний шар в деяких методах побудови словників, заважати тренуванню, знецінювати спеціалізацію по контекстам тощо. Абревіатури та сленг можуть бути складними для точної інтерпретації моделі, оскільки висловлені почуття можуть бути неоднозначними або їх може бути нелегко вивести з використаних слів. Це пояснюється тим, що деякі абревіатури та сленг можуть мати кілька значень залежно від контексту та виражати різні почуття, навіть якщо вони використовуються в одному контексті. Наприклад, абревіатура «SMH» у соціальних мережах може означати «киваю головою»(«shaking my head») від розчарування або «стільки ненависті»(«so much hate») від гніву, причому почуття виражаються залежно від того, як воно використовується. Таким чином, інтерпретувати абревіатури та сленг може бути складно, і в ідеалі модель має бути навчена з достатнім обсягом даних, щоб мати можливість правильно визначити настрої. Можливе рішення такої проблеми — використання систем для пошуку та виправлення орфографічних помилок. Це трохи збільшує точність оцінок, проте, чи вартує мале покращення в точності значного збільшення часу аналізу — окреме питання для кожного практичного випадку.

3. Спам

Ключовою особливістю соціальних медіа є те, що вони дозволяють будь-якій людині з будь-якої точки світу вільно висловлювати свої погляди та думки, не розкриваючи свою справжню особу та не побоюючись небажаних наслідків. Таким чином, ці думки є дуже цінними. Однак ця анонімність також має свою ціну. Це дозволяє людям із прихованими цілями або зловмисними намірами легко обдурювати систему, щоб створити у людей враження, що вони є незалежними представниками громадськості, і публікувати фальшиві думки для просування або дискредитації цільових продуктів, послуг,

організацій або окремих осіб, не розкриваючи своїх справжніх намірів , або особа чи організація, на яку вони таємно працюють. Таких осіб називають спамерами думок, а їх діяльність — спамінгом (спамом) думок. Спам думок став серйозною проблемою. Крім окремих осіб, які висловлюють фальшиві думки в оглядах і обговореннях на форумах, існують також комерційні компанії, які займаються написанням фальшивих відгуків і фальшивих блогів для своїх клієнтів. У новинах повідомлялося про кілька гучних випадків фальшивих відгуків. Важливо виявляти таку спамерську діяльність, щоб переконатися, що думки в Інтернеті є надійним джерелом цінної інформації. На відміну від вилучення позитивних і негативних думок, виявлення спаму думок — це не лише проблема НЛП, оскільки вона передбачає аналіз поведінки людей, які публікують публікації. Таким чином, це також проблема інтелектуального аналізу даних. Спосіб запобігти ураженню аналітичних даних таким спамом був запропонований в [15]. Алгоритм знаходить групи спамерів, які, можливо, працювали в змові, просуваючи або знижуючи певні цільові організації. Це працює в два етапи:

Частий аналіз шаблонів: по-перше, він попередньо обробляє дані перегляду для створення набору транзакцій. Кожна транзакція представляє унікальний продукт і складається з усіх рецензентів (їхні ідентифікатори), які перевірили цей продукт. Використовуючи всі транзакції, він виконує частий аналіз шаблонів, щоб знайти набір частих шаблонів. Кожна модель — це, по суті, група рецензентів, які перевірили набір продуктів. Така група вважається потенційною спам-групою. Причина для частого аналізу шаблонів полягає в наступному: якщо група рецензентів працювала разом лише один раз, щоб просувати або знижувати один продукт, це може бути важко виявити на основі їх колективної поведінки. Однак ці фальшиві рецензенти (особливо ті, кому платять за написання) не можуть просто написати один відгук про один продукт, оскільки таким чином вони не зароблять достатньо грошей. Натомість вони працюють над багатьма продуктами, тобто пишуть багато

відгуків про багато продуктів, що також їх роздає. При частому аналізі шаблонів вони можуть працювати разом над кількома продуктами.

Ранжуйте групи на основі набору індикаторів групового спаму: групи, виявлені на кроці 1, можуть не бути справжніми групами спамерів. Чимало рецензентів об'єдналися в аналіз шаблонів просто випадково. Потім на цьому етапі спочатку використовується набір індикаторів для виявлення різних типів незвичайної поведінки груп та окремих членів. Ці показники включають спільне написання рецензій у короткий проміжок часу, написання рецензій одразу після запуску продукту, схожість контенту групового рецензування, відхилення групового рейтингу тощо [15]. Тоді була запропонована реляційна модель, яка називається GSRank (груповий спам-рейтинг), щоб використовувати зв'язки між групами, окремими членами групи та продуктами, які вони перевіряли, щоб класифікувати групи кандидатів на основі їхньої ймовірності бути групами спамерів. Потім для вирішення проблеми використовувався ітераційний алгоритм. Набір груп спамерів також було вручну позначено та використано для оцінки запропонованої моделі, яка показала обнадійливі результати. Одна слабка сторона цього методу полягає в тому, що через порогове значення частоти, яке використовується в аналізі шаблонів, якщо група не працювала разом багато разів (три або більше разів), вона не буде виявлена цим методом. Цей метод являє собою навчання без вчителя, оскільки він не використовує жодних вручну позначених даних для навчання.

Оскільки соціальні медіа все частіше використовуються організаціями та окремими особами для прийняття важливих рішень, спам думками стає все більш поширеним. Для багатьох підприємств публікація неправдивих думок або наймання інших для цього стало дешевим способом маркетингу та просування бренду, а особливо політичних поглядів (що навіть може застосовуватися у воєнних цілях). Незважаючи на те, що поточні дослідження щодо виявлення спаму думок ще знаходяться на ранній стадії, кілька

ефективних алгоритмів уже запропоновано та використовується на практиці. Спамери, однак, також стають більш досвідченими та обережними у написанні та публікації фальшивих думок, щоб уникнути виявлення. Фактично, ми вже бачили гонку озброєнь між алгоритмами виявлення та спамерами. Однак я оптимістично налаштований, що будуть розроблені більш складні алгоритми виявлення, щоб ускладнити спамерам публікацію фальшивих думок. Такі алгоритми, ймовірно, є цілісними підходами, які об'єднують усі можливі ознаки або підказки в процесі виявлення.

Нарешті, слід зазначити, що спам думок відбувається не лише в оглядах, але й в інших формах соціальних медіа, таких як блоги, обговорення на форумах, коментарі та публікації в Twitter. Однак поки що в цьому контексті мало досліджень.

4. Нечітко висловлені думки

Нездолання проблема opinion mining на текстах — текст, в якому автор не зміг донести свою думку. Залежності на великій відстані, такі як порядок слів у реченні, може бути важко правильно витлумачити через неоднозначність почуттів. Залежності на великій відстані стосуються порядку слів у реченні, і їх може бути важко правильно інтерпретувати, намагаючись визначити почуття, пов'язане з реченням. Це пояснюється тим, що настрій певних фраз може залежати від контексту сказаного перед ними або після них, тому порядок слів може бути важливим. Наприклад, фраза «непогано» може виражати позитивні настрої, коли вона розміщена безпосередньо перед негативним реченням, але більш негативна, коли вона розміщена відразу після позитивного речення. Таким чином, формулювання та порядок слів можуть бути дуже важливими для правильного визначення почуття, пов'язаного з реченням. Хоч і здається, що тут вже нічого не поробиш, все ж можна витягнути певну інформацію — сентимент тексту, тема допису, контекстна інформація.

Ця тема пов’язана з виявленням спаму в думках, але також відрізняється тим, що рецензії низької якості можуть не бути спамом або фальшивими рецензіями, а фальшиві рецензії можуть не сприйматися читачами як рецензії низької якості, тому що, як ми обговорювали в останньому абзаці, читання рецензій дуже важко помітити фальшиві відгуки. З цієї причини фальшиві відгуки також можуть розглядатися як корисні або високоякісні відгуки, якщо самозванці пишуть свої відгуки завчасно та добре їх створюють.

В такому випадку перед аналітиком постає задача визначити якість, корисність або практичність кожного огляду. Це важливе завдання, оскільки під час показу відгуків користувачеві бажано ранжувати відгуки на основі якості чи корисності, причому першими залишаються найкорисніші відгуки. Насправді, багато сайтів зі збору відгуків або хостингу практикують це роками. Вони отримують оцінку корисності або якості кожного огляду, просячи читачів надати відгуки щодо корисності кожного огляду. Наприклад, на amazon.com читач може вказати, чи вважає він відгук корисним, відповівши на запитання «Чи був відгук корисним для вас?» просто під кожним оглядом. Результати відгуків усіх тих, хто відповів, потім узагальнюються та відображаються безпосередньо перед кожним відгуком, наприклад, «15 із 16 людей вважають цей відгук корисним». Незважаючи на те, що більшість сайтів для розміщення відгуків уже надають цю послугу, автоматичне визначення якості кожного огляду все ще є корисним, оскільки для накопичення значної кількості відгуків користувачів може знадобитися багато часу. Ось чому багато відгуків мають мало лайків або взагалі їх не мають. Особливо це стосується нових оглядів.

Визначення якості рецензій зазвичай формулюють як задачу регресії. Вивчена модель призначає оцінку якості кожному відгуку, який можна використовувати для визначення рейтингу або рекомендації щодо огляду. У цій галузі досліджень основними правдивими даними, які використовуються як для навчання, так і для тестування, зазвичай є відгуки користувачів про

корисність, надані для кожного огляду, які, як ми обговорювали вище, надаються для кожного огляду на багатьох сайтах для розміщення оглядів. Отже, на відміну від виявлення підроблених відгуків, дані навчання та тестування тут не є проблемою. Щоб розв'язати подібну задачу регресії, можна включати в рішення такі набори функцій:

Характеристики структури: довжина огляду, кількість речень, відсоток питальних речень і окликів, а також кількість HTML-тегів жирного шрифту <b> і розривів рядків <br>.

Лексичні особливості: уніграми та біграми з вагами tf-idf.

Синтаксичні характеристики: відсоток проаналізованих лексем відкритого класу (тобто іменників, дієслів, прикметників і прислівників), відсоток лексем, які є іменниками, відсоток лексем, які є дієсловами, відсоток лексем, які є дієсловами, сполученими від першої особи. і відсоток лексем, які є прикметниками чи прислівниками.

Семантичні ознаки: товарні аспекти та настрої слів.

Характеристики метаданих: рейтинг відгуку (кількість зірочок).

Окрім вищеназваних, можна використовувати додаткові набори функцій, а саме: функції профілю рецензента, доступні на сайті рецензента, функції історії рецензента, які фіксують корисність його/її відгуків у минулому, а також набір функцій читабельності, тобто орфографічних помилок і читабельності показники дослідження читабельності.

Таким чином, визначення корисності огляду є важливою темою дослідження. Це особливо корисно для продуктів і послуг, які мають велику кількість відгуків. Щоб допомогти читачеві швидко отримати якісну думку, сайти з оглядами повинні надавати хороші рейтинги оглядів. Однак також варто було би додати деякі застереження. По-перше, рейтинг (рейтинги) огляду має відображати природний розподіл позитивних і негативних думок. Небажано класифікувати всі позитивні (або всі негативні) відгуки як найкращі

лише тому, що вони мають високі оцінки якості. Проблема надлишковості також викликає серйозне занепокоєння. На мою думку, важливі як якість, так і розповсюдження (з точки зору позитивної та негативної точок зору). По-друге, читачі, як правило, визначають, чи є рецензія корисною чи ні, виходячи з того, чи вона висловлює думку щодо багатьох аспектів продукту та виглядає справжньою. Спамер може задовольнити цю вимогу, ретельно створивши огляд, який буде схожий на звичайний корисний огляд. Таким чином, використання кількості відгуків про корисність для визначення якості огляду або лише як основної правди може бути проблематичним. Крім того, відгуки користувачів також можуть надсилатися спамом. Спам у зворотному зв’язку – це підпроблема шахрайства з кліками в пошуковій рекламі, коли людина або робот натискає деякі рекламні оголошення в Інтернеті, щоб створити враження реальних кліків клієнтів. Тут робот або спамер може натиснути кнопку відгуку про корисність, щоб підвищити корисність відгуку.

# РОЗДІЛ 3. РОЗРОБКА МОДЕЛІ АНАЛІЗУ ЕМОЦІЙНОГО ЗАБАРВЛЕННЯ ТЕКСТІВ

## 3.1 Підготовка даних

### 3.1.1 Архітектура бази даних

Для збереження відгуків та результатів н-грамізації потрібна проста база даних з трьома таблицями: тренувальні відгуки, тестові відгуки та н-грами. Таблиця н-грам має зовнішній ключ, що вказує на таблицю відгуків (GUID відгуку, з якого було створено н-граму). T-SQL (Transact-SQL) — це пропрієтарна процедурна мова, яка використовується для зв'язку з базами даних, особливо створеними Microsoft. Це де-факто стандартна мова, яка використовується для програмування Microsoft SQL Server, а також є основною мовою, що використовується в програмуванні Microsoft Access. Іноді кажуть, що T-SQL походить від стандарту Structured Query Language (SQL), але насправді він суттєво відрізняється. Хоча SQL є декларативним і зосереджується на описі того, що потрібно зробити, T-SQL розширює SQL, щоб стати процедурною мовою, і більше зосереджується на описі того, як виконати якісь дії. Код T-SQL можна запускати в інтерактивному середовищі за допомогою Microsoft SQL Server Management Studio (SSMS), а також за допомогою середовищ програмування, таких як Visual Studio Code, у поєднанні з такими інструментами, як Azure Data Studio. T-SQL забезпечує потужний контроль над базами даних Microsoft Access і SQL Server і використовується для широкого спектру завдань, включаючи вибір даних, вставлення, оновлення та видалення даних, а також забезпечення безпеки та цілісності баз даних. Для транзакцій (схема T-SQL) встановлено рівень ізоляції read commited, це потрібно для подальшого розпаралелення роботи з базою даних, що є одним з найдовших IO-процесів. Read uncommited — низький рівень ізоляції і дозволяє транзакціям читати дані з транзакцій, що ще не були оброблені до кінця. Це збільшує швидкість роботи, але якщо якась транзакція не буде завершена через помилку, то в інших можуть залишитись неправильні дані. Рівень ізоляції Read Uncommitted Isolation Level є найменш обмежуючим із чотирьох рівнів ізоляції та дозволяє транзакціям читати

дані, які ще не були зафіксовані іншими транзакціями. Це означає, що транзакції потенційно можуть отримати доступ до «брудних» даних, що в свою чергу означає, що дані можуть бути змінені іншою транзакцією до того, як їх буде зафіксовано. Це може призвести до неповторюваних зчитувань, тобто той самий запит повертатиме різні результати, якщо виконуватиметься кілька разів.

Read commited — більш високий рівень ізоляції. З таким рівнем транзакції можуть отримати дані тільки з уже завершених транзакцій. Швидкість буде трохи зменшена, проте так можна уникнути цілого сімейства помилок. Read commited ізоляції читання є більш обмежувальним, ніж рівень ізоляції Read uncommited, і гарантує, що запит не отримає брудних даних. Це досягається шляхом «блокування» даних і запобігання їх зміні іншою транзакцією, доки їх не буде зафіксовано.

Створено кластерний індекс по GUID відгуків та некластерний (окрема таблиця) по “центральним” словам н-грам для пришвидшення деяких подальших кроків обробки .

### 3.1.2 Відгуки для побудови моделі

Для побудови системи та її тестування візьмемо за основу тренувальних даних набір даних що складається з 800 тисяч невідсортованих, але оцінених відгуків з торгівельної платформи Amazon, що включають відгуки до книг, музики, фільмів, мультфільмів та настільних ігор. Дані отримано у вигляді текстового документу, в якому на кожен відгук відведено одну стрічку. Першим елементом кожної стрічки є маркер “ label 1”, якщо відгук позитивний і “label 2”, якщо відгук негативний. Найдовший відгук має довжину 1125 символів, найкоротший — 112.

```
__label__1 Failed after a few days: It was nice while it
__label__2 Worked as needed!: I used this for my son's P
__label__2 Good performance at a great price!: I bought
__label__1 Failed after a few uses: I bought this card f
__label__1 I bought FlshMemory Card for my daughter, but
__label__2 Transcend 8GB: I use this for my Canon 1000D
__label__2 Big SDHC Card: I bought this card so I could
```

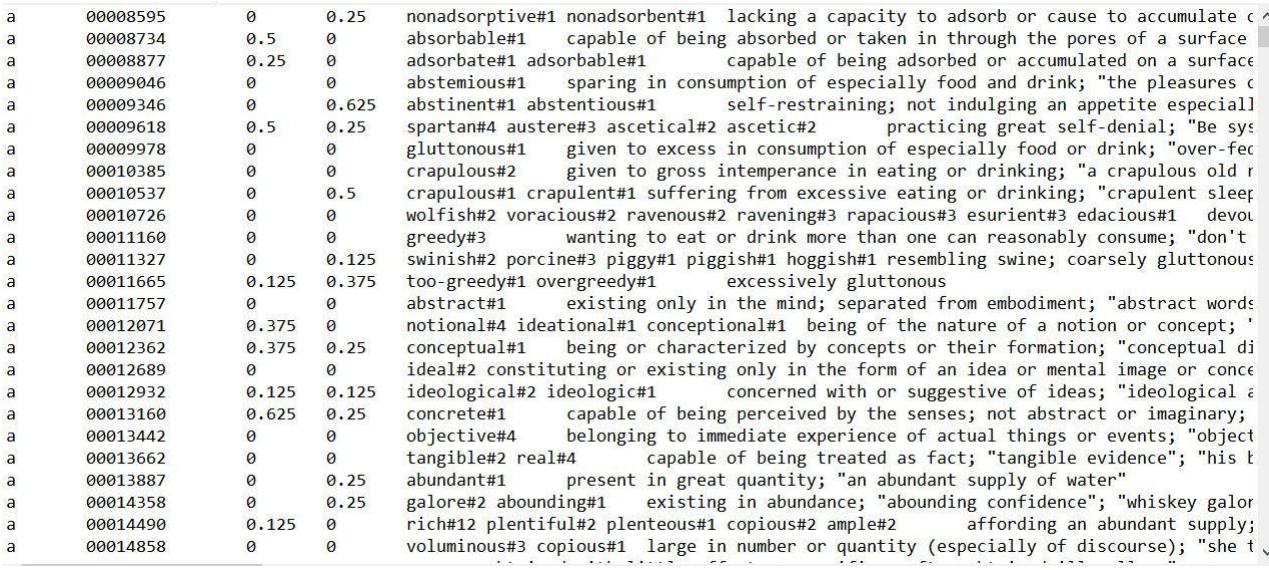
Рисунок 3.1 Тренувальні відгуки.

Для зручності користування цими відгуками було створено базу даних на основі СУБД MSSQL Server, виділено одну таблицю для розміщення відгуків у сирому вигляді з двома стовпчиками: один стовпчик текстового типу (nvarchar 1200) та один стовпчик булевого типу для збереження інформації про тип відгуку (true для позитивних, false для негативних).

### 3.1.3 Робота з лексиконом (словником)

Для тестування підходів, що покладаються на заповнений вручну словник, використаємо словник SentiWordNet 3 (англійський), що розповсюджується за ліцензією Attribution-ShareAlike 4.0 Unported (CC BY-SA 4.0), тобто вільний для використання та розповсюдження за умови вказання джерела та ліцензії. (Більше про ліцензію - за посиланням creativecommons.org/licenses/by-sa/4.0/.) SENTIWORDNET 3.0, лексичний ресурс, спеціально розроблений для підтримки програм класифікації настроїв і аналізу думок(сентиментів). SENTIWORDNET 3.0 є вдосконаленою версією SENTIWORDNET 1.0, лексичного ресурсу, загальнодоступного для дослідницьких цілей, на даний момент ліцензований понад 300 дослідницькими групами та використовується в різноманітних дослідницьких проектах по всьому світу. І SENTIWORDNET 1.0, і 3.0 є результатом автоматичного анотування всіх синсетів WORDNET відповідно до їхнього ступеня позитивності, негативності та нейтральності. SENTIWORDNET 1.0 і 3.0 відрізняються (a) версіями WORDNET, які вони анотують (WORDNET 2.0 і 3.0 відповідно), (б) алгоритмом, що використовується для автоматичного анотування WORDNET, який тепер включає (додатково до попереднього етапу напівконтрольованого навчання) ) крок випадкового блукання для уточнення балів. Тут ми обговорюємо SENTIWORDNET 3.0, особливо зосереджуючись на вдосконаленнях, що стосуються аспекту (b), які він втілює по відношенню до версії 1.0. Ми також повідомляємо про результати оцінки SENTIWORDNET 3.0 порівняно з фрагментом WORDNET 3.0, вручну анотованим щодо позитивності, негативності та нейтральності; ці результати вказують на підвищення точності приблизно на 20% відносно SENTIWORDNET 1.0[14]. Структура тексту цього

словника містить більше інформації, проте розпарсити (зчитати) його зовсім не складно. Для цього розбиваємо його на рядки стандартними засобами платформи .Net, після чого розбиваємо самі рядки на слова за пустими символами (пробіл, табуляція тощо) і додаємо в словник в пам’яті.



```
a	00008595	0	0.25	nonadsorptive#1 nonadsorbent#1  lacking a capacity to adsorb or cause to accumulate c
a	00008734	0.5	0	absorbable#1	capable of being absorbed or taken in through the pores of a surface
a	00008877	0.25	0	adsorbate#1 adsorbable#1	capable of being adsorbed or accumulated on a surface
a	00009046	0	0	abstemious#1	sparing in consumption of especially food and drink; "the pleasures c
a	00009346	0	0.625	abstinent#1 abstentious#1	self-restraining; not indulging an appetite especiall
a	00009618	0.5	0.25	spartan#4 austere#3 ascetical#2 ascetic#2	practicing great self-denial; "Be sys
a	00009978	0	0	gluttonous#1	given to excess in consumption of especially food or drink; "over-fec
a	00010385	0	0	crapulous#2	given to gross intemperance in eating or drinking; "a crapulous old r
a	00010537	0	0.5	crapulous#1 crapulent#1 suffering from excessive eating or drinking; "crapulent sleep
a	00010726	0	0	wolfish#2 voracious#2 ravenous#2 ravening#3 rapacious#3 esurient#3 edacious#1	devou
a	00011160	0	0	greedy#3	wanting to eat or drink more than one can reasonably consume; "don't
a	00011327	0	0.125	swinish#2 porcine#3 piggy#1 piggish#1 hoggish#1 resembling swine; coarsely gluttonous
a	00011665	0.125	0.375	too-greedy#1 overgreedy#1	excessively gluttonous
a	00011757	0	0	abstract#1	existing only in the mind; separated from embodiment; "abstract words
a	00012071	0.375	0	notional#4 ideational#1 conceptional#1  being of the nature of a notion or concept; "
a	00012362	0.375	0.25	conceptual#1	being or characterized by concepts or their formation; "conceptual di
a	00012689	0	0	ideal#2 constituting or existing only in the form of an idea or mental image or conce
a	00012932	0.125	0.125	ideological#2 ideologic#1	concerned with or suggestive of ideas; "ideological a
a	00013160	0.625	0.25	concrete#1	capable of being perceived by the senses; not abstract or imaginary;
a	00013442	0	0	objective#4	belonging to immediate experience of actual things or events; "object
a	00013662	0	0	tangible#2 real#4	capable of being treated as fact; "tangible evidence"; "his b
a	00013887	0	0.25	abundant#1	present in great quantity; "an abundant supply of water"
a	00014358	0	0.25	galore#2 abounding#1	existing in abundance; "abounding confidence"; "whiskey galor
a	00014490	0.125	0	rich#12 plentiful#2 plenteous#1 copious#2 ample#2	affording an abundant supply;
a	00014858	0	0	voluminous#3 copious#1  large in number or quantity (especially of discourse); "she t
```

Рисунок 3.2 Словник SentiWordNet 3.

Як видно на рисунку, в першій колонці вказана одна літера, що визначає частину мови слова, де a - adjective (прикметник), n - noun (іменник), v - verb (дієслово), r - adverb (прислівник). Наступний стовпчик — унікальний ідентифікатор слова. Третій стовпчик — оцінка позитивності слова по шкалі від 0 до 1 зі ступенями на відстані 0.125 (1/8). Четвертий стовпчик — те ж саме для негативної оцінки. Далі — слова та їх синоніми, останній стовпчик

— приклади використання слів. Для роботи з цим словником не підійде збереження в базі даних, бо для обробки кожного відгуку доведеться звернутись до словника від 30 до 200 разів, кожне звернення до бази даних займає від 50 мілісекунд, тобто аналіз кожного відгуку лише за словником (без статистики по відгукам) може розтягнутись на 1-7 секунд. І це ще без врахування збереження результатів да отримання самого відгуку. Для таких випадків доцільно вкомпактувати словник для збереження його цілком в оперативній пам’яті комп’ютера, звернення до якої займає наносекунди. Щоб вкомпактувати словник, можемо відразу відкинути приклади використання слів, які для подальшої роботи не мають ніякого практичного значення. Далі

можна відкинути позначення частини мови (при потребі краще використовувати більш повні словники та без зайвого навантаження). Для зручності розіб'ємо словник по синонімах на нові рядки (приведемо збережену в пам'яті таблицю в третю нормальну форму). Знову ж таки, при потребі краще знайти більш повний словник синонімів, тому цю інформацію зберігати не будемо. Ідентифікатор слова зберігати нам теж ні до чого — можна використовувати саме слово як унікальний ідентифікатор.

Про збереження обробленого словника на диску на цьому етапі можна не перейматись — зчитування словника з файлу займає менше двох секунд, проте для подальшого ітеративного покращення словника було додано можливість зберегти на диск та завантажити назад в пам'ять зчитаний в об'єкт C# словник. Таким чином не потрібно кожного разу заново "тренувати" словник перед роботою, достатньо відкрити збережений словник з диску.

При перегляді словника може виникнути питання, чому деякі слова мають і позитивну і негативну оцінку. Відповідь на це проста: деякі слова в різному контексті можуть означати як позитивні так і негативні речі. Інше питання — як це інтерпретувати? Банальний вихід — враховувати тільки найбільше значення з двох оцінок, а якщо вони рівні, то вважати слово нейтральним. Проте можна піти іншим шляхом і при аналізі текстів додати додаткові метрики, що експоненціально покладається на кількість слів, що мають позитивну чи негативну оцінку та стоять в тексті неподалік даного слова. Такі метрики можна враховувати при обрахуванні остаточного результату або відразу при аналізі слів для пошуку частин тексту, де автор висловлює тільки негативні думки, проте використовує слова, що мають обидва види оцінки.

### 3.1.4 Створення та робота з IDF словником

Необхідний для швидкого аналізу методом TF-IDF, IDF-словник являє собою список виду «пара-значення», в якому у відповідність до всіх слів в наявному лексиконі проставлено значення метрики IDF (inversed document

frequency), тобто оберненої частоти входжень в документи (відгуки). Ця метрика потрібна для визначення важливості слів у тексті, бо дуже низьке значення цієї метрики для слова означає, що це слово вживається не залежно від контексту, тобто всюди. Такі слова як «тому», «отож», «він» можна відкинути з пошуку ключових слів за низьким значенням IDF. А для слів, у яких значення IDF занадто високе, можна припустити, що ці слова є значно нетривіальними, рідкісними в текстах, що, частіше за все, означає, що слово не несе значного змістового навантаження, а у автора допису багатий словниковий запас.

TF-IDF означає частота термінів-зворотна частота документа, і є показником, який використовується в обробці природної мови для відображення того, наскільки важливим є слово для документа в корпусі. TF-IDF базується на частоті певних слів у документі, причому частіші слова мають вищу вагу, і враховує роль, яку окремі слова відіграють у всьому корпусі документів.

TF-IDF часто використовується, щоб допомогти визначити, які слова є найбільш релевантними в документі порівняно з рештою корпусу. Це робиться шляхом обчислення оцінки TF-IDF документа для кожного слова в документі. Оцінка обчислюється шляхом множення частоти слова в документі на зворотну частоту документа (IDF) цього слова.

IDF обчислюється шляхом ділення загальної кількості документів у колекції на кількість документів, що містять слово. Це допомагає переконатися, що слова, які є більш поширеними в корпусі, мають менше значення, надаючи більшій вазі унікальним словам у документі під час визначення релевантності вмісту.

Поєднуючи ці два показники, TF-IDF допомагає визначити, які слова є найбільш релевантними для документа, і може використовуватися в таких програмах, як класифікація тексту, кластеризація та пошукова оптимізація.

Створення IDF-словника технічно просте: потрібно для кожного слова поділити загальну кількість документів на кількість документів, де є дане слово. Проте побудова цього словника займає багато часу в порівнянні з самим аналізом тексту. Тому було додано можливість зберігати створені IDF-словники на випадок, якщо доведеться перервати роботу з якимось набором даних, для яких був створений цей словник (IDF-словник для якось набору текстів може бути нерелевантним для іншого набору текстів).

### 3.2 Побудова конвеєру для сентиментної класифікації

На етапі підготовки до побудови моделі та аналізу було розроблено архітектуру системи, що базується на понятті пайплайну, або конвеєру обробки, який збільшує модульність системи і, як наслідок, широту можливостей налаштування. На початку роботи система зчитує словник (з вихідного файла або з уже підготовленого та серіалізованого словника). Як стартовий елемент пайплайну використовується об’єкт-стратегія для зчитування відгуків з бази даних на основі фреймворку Entity Framework Core, що передає далі по конвеєру об’єкти відгуків (текст та оцінка для тренування моделі). Всі наступні елементи пайплайну — модульні, іншими словами необов’язкові, та можуть бути підключені в різних варіаціях та комплектаціях або з різними параметрами. Здійснюється це за допомогою уніфікації вхідних та вихідних об’єктів у вигляді об’єкту класу Report та кількох наслідуючих класів (ShingledReport, NegationReport, ComparisonReport для н-грамізації, заперечень і ступеней відповідно). В цій роботі імплементовано такі елементи пайплайну:

- Нормалізатор формату — знаходить в тексті невідповідності по формату, зайві для подальшої обробки символи та переводить всі слова в lowercase, тобто текст без великих літер (це потрібно для того щоб не робити це в кожній подальшій перевірці тексту).
- “Грубий” зчитувач знаходить в тексті слова, що відповідають наявним в підготовленому словнику, та з окремих слів будує об’єкт

грубого звіту, що містить впорядковані розділені слова зіставлені з їх грубою оцінкою (не свідомою семантично чи синтаксично).

- Н-грамізатор — частина пайплайну, що розбиває слова в отриманому об’єкті звіту на н-грами з оцінкою сентименту та звертається до статичного репозиторію для збереження результатів у базу даних. Цей етап потрібен для подальшого аналізу тексту окрім емоційного.
- Обробник заперечень — знаходить в отриманому об’єкті звіту слова-заперечення (типу «не», «ані», «ніскільки») та обертає оцінку емоційного забарвлення слів на заданій відстані від даного вперед чи назад, що встановлюється відповідними параметрами negationStart і negationEnd.
- Обробник прислівників ступеню — знаходить в отриманому об’єкті звіту слова типу «найбільш», «дуже», «трохи» тощо та за передвстановленими коефіцієнтами змінює оцінку емоційного забарвлення слів на заданій відстані від даного вперед чи назад та з певним глобальним коефіцієнтом (для лінійної оптимізації), що встановлюється відповідними параметрами.
- Генератор кінцевого підсумку — за отриманим об’єктом звіту зіставляє текстовий підсумок аналізу за заданими параметрами. Цей підсумок потрібен лише для виводу в інтерфейсі, тому з цього кроку конвеєра повертається як простий об’єкт з текстом та числовими результатами, а не у вигляді уніформного для конвеєру звіту.


Рисунок 3.3 Конвеєр (пайплайн) обробки текстів з модульним підходом

Таким є базовий конвеєр для пошуку сентименту. Окремою важливою частиною є лінійна оптимізація оцінки, тобто вибір зсуву кінцевої оцінки, що максимізує точність на етапі валідації.

### 3.3Побудова валідатору для перевірки та ітеративного покращення моделі

Перевірка, або валідація моделі – це процес визначення того, чи модель точно відображає поведінку системи. Валідність моделі слід оцінювати як операційно (тобто, визначаючи, чи результати моделі узгоджуються з даними спостереження), так і концептуально (тобто, визначаючи, чи є теорія та припущення, що лежать в основі моделі, виправданими). Моделі можна перевірити шляхом порівняння вихідних даних із незалежними польовими або експериментальними наборами даних, які відповідають змодельованому сценарію. Однак важливо враховувати якість даних (наприклад, рівень похибки вимірювання), чи справді вони представляють систему та чи є вони найкращим тестом моделі. Оперативна валідація моделі з використанням незалежних даних може бути неможливою, якщо змодельований сценарій виходить за межі спостережуваних умов (наприклад, прогнозування реакцій на майбутню зміну клімату) або коли використовуються імовірнісні прогнози (тобто такі, які включають невизначеність у системні процеси). В останньому випадку рішення

між використанням детермінованої чи імовірнісної системи залежить від компромісу між точністю та точністю. Загалом детерміновані моделі демонструють вищу точність, але менш точні, ніж ті, які включають невизначеність. Однак, незалежно від типу моделювання, концептуальна перевірка завжди можлива. Виконання аналізу чутливості є ще однією важливою частиною процесу перевірки моделі. Метою виконання аналізу чутливості є визначення відносного впливу параметрів, початкових умов і альтернативних припущень на результати моделі. Процес ітераційний, забезпечує зворотний зв'язок, який може покращити модель. Аналіз чутливості порівнює змінні відповіді з кількох прогонів моделі. У кожному з порівняльних прогонів усі параметри залишаються постійними, за винятком параметра, який досліджується. Якщо спостерігається надмірний вплив параметра моделі на результат моделювання, який не відображає реальність, характеристика моделі повинна бути переоцінена. Проведення обширного аналізу чутливості, щоб зрозуміти, як кожен параметр впливає на поведінку моделі, є важливою частиною процесу моделювання. Зрештою, перевірка моделі посилює підтримку моделі та надійність її результатів. Побудова корисної симуляції вимагає побудови моделі, яка є досить точним представленням біологічного явища, яке розглядається. Хоча жодна модель не може бути «доведена правильною», валідація стосується перевірки надійності та правдоподібності продуктивності моделі.

Валідація моделі по ідеї — простий процес, в якому потрібно просто запустити конвеєр аналізу на всіх відгуків що були відведені для тестування (не використовувались для тренування), а потім порахувати просту метрику (наприклад лінійне співвідношення позитивних та негативних термінів до загальної кількості слів. Проте без додаткової оптимізації це може зайняти дуже багато часу (при хорошій кількості даних). В цій роботі було використано 5% (40 тисяч) відгуків як тренувальних. Повна валідація на таких даних займає до хвилини без оптимізації. Може здатись, що це не великий час, проте для ітеративного покращення моделі потрібно буде часто викликати валідацію (від 1

разу кожні кілька кроків до 1 разу на ітерацію в залежності від методу), тому необхідно її прискорити. Для цього було розпаралелено задачу валідації по відгукам. Конкретніше — з ThreadPool викликаються створені наперед потоки обробки (створення і видалення потоків - дуже дорога по часу операція, тому стандартна практика — використання пулів, в яких зберігаються не зайняті потоки. Після закінчення мікро-задачі потоки повертаються в пул). ThreadPool може бути корисним у перевірці моделі, дозволяючи одночасне виконання кількох завдань перевірки моделі. Це забезпечує швидку перевірку моделей і дозволяє виконувати декілька завдань перевірки паралельно. Це може бути особливо корисним для завдань із інтенсивним використанням даних, таких як підрахунок балів моделі, тестування точності або налаштування гіперпараметрів, оскільки всі завдання, пов’язані з перевіркою моделі, можна швидко й ефективно обробляти паралельно. Завдяки збільшеній thread-safety (потоко-безпеці) звязку між Entity Framework та MSSQL можемо реалізувати паралельні звернення в різних працівниках. Також для комп’ютерів з великою кількістю оперативної памяті (такою, що вміщує всі тренувальні відгуки) додано можливість витягнути відгуки в оперативну пам’ять для збільшення швидкості обробки між різними заходами валідації. Тобто Entity Framework потрібно буде відправити лише один запит до бази даних перед початком валідації. Таким чином маємо покращення в швидкості:

- Без паралельності на 40 тисячах відгуків — 18-20 секунд
- Паралельність, 4 “працівники» — 16-17 секунд
- Паралельність, максимум потоків із пула — 12-14 секунд
- Паралельність, максимум потоків із пула + кешування відгуків — 9-11 секунд якщо відгуки вже закешовано

Результат роботи валідатора — процент правильно оцінених відгуків та масив неправильно оцінених відгуків (для подальшого тренування за цими результатами).

Маючи валідатор, можемо спробувати додати просту лінійну оптимізацію методом дихотомії. Отримуємо результат: на тренувальних данних найбільше

зростання точності досягається лінійним зсувом на -0.152 (за шкалою сентиментних оцінок від -1 до 1).

Також можна встановити експоненційну залежність оцінки тексту від кількості слів того чи іншого забарвлення. Експериментальним шляхом було знайдено оптимальну основу, 1.05. Тобто з кожним знайденим позитивним словом вся позитивна оцінка тексту буде зростати на 5%, те ж саме і з негативною оцінкою.

### 3.4 Перевірка моделі перед використанням ітеративного покращення

При поетапному підключенні всіх елементів пайплайну можемо спостерігати таку динаміку точності та часу аналізу (сірим виділено відкинуті через недоцільність варіанти):

Таблиця 3.1 – ПОРІВНЯННЯ ЕФЕКТИВНОСТІ ОБРОБКИ РІЗНИХ ЕТАПІВ

| Етап пайплайну | Точність | Час валідації на 40 тисячах відгуків |
|---|---|---|
| Базовий етап (грубий аналіз) | 55.56% | 5-6 секунд |
| Обробка заперечень з мінімальним радіусом (1 терм вперед) | 61.31% | 7-8 секунд |
| Обробка заперечень з великим радіусом (3 терма вперед) | 63.44% | 7-8 секунд |
| Обробка заперечень в обидві сторони (1 терм назад, 3 терма вперед) | 63.55% | 7-8 секунд |

| | | |
|---|---|---|
| Обробка заперечень в обидві сторони в великому радіусі (3 терми назад, 5 термів вперед) | 62.51%<br>Не таке значне покращення. Радіус обробки треба збільшувати в міру. | 7-8 секунд |
| Обробка прислівників ступеню та способу дії (малий радіус -1:+1) | 64.98% | 9-11 секунд |
| Обробка прислівників ступеню та способу дії (великий радіус -2:+3) | 66.39% | 9-11 секунд |
| Лінійна оптимізація | 67.40% | 9-11 секунд |

Щодо заперечень, спостерігаємо різну ефективність при різних діапазонах обробки заперечень. Розумно припустити, що ефективність залежить від мови (наприклад в німецькій заперечення ставляться в кінці речення і це теж треба враховувати).

Щодо часу виконання — після першого виконання валідації .NET CLR (Common Language Runtime) оптимізує виконання цієї задачі на рівні компілятора в IL (Intermediate Language), тому всі подальші виконання виконуються за меншим вказаним часом. Common Language Runtime (скорочено CLR — «загальномовне середовище виконання») — це компонент пакета Microsoft .NET Framework, віртуальна машина, на якій виконуються всі мови платформи .NET Framework. CLR перетворює вихідний код у байт-код IL, скомпільовану реалізацію Microsoft під назвою MSIL, і надає програмам MSIL (і, отже, програмам, написаним на мовах високого рівня, які підтримують .NET Framework) доступ до бібліотеки класів .NET Framework. NET Framework, або так званий .NET FCL. Середовище CLR є реалізацією специфікації CLI (English Common Language Infrastructure), специфікації загальномовної інфраструктури Microsoft. Віртуальна машина CLR дозволяє програмістам забути про багато деталей про конкретний процесор, на якому буде працювати програма. CLR також надає такі важливі послуги, як: управління пам'яттю; управління

потоками; обробка винятків; збір сміття; безпека виконання. Компонент Common Language Runtime розташований поверх служб операційної системи, якою зараз є операційна система Windows, але в майбутньому це може бути майже будь-яка програмна платформа. Основною метою CLR є виконання програм, підтримка всіх програмних залежностей, керування пам'яттю, забезпечення безпеки, інтеграція з мовами програмування тощо. Середовище виконання надає багато послуг, які полегшують створення та реалізацію програм і значно підвищують надійність останній.

Розробники не взаємодіють із Common Language Runtime безпосередньо: усі служби надаються уніфікованою бібліотекою класів, яка розташована поверх CLR. Ця бібліотека містить більше 1000 класів для вирішення різних програмних завдань - від взаємодії зі службами операційної системи до роботи з даними і XML-документами. Common Language Runtime забезпечує середовище виконання. програми NET. Серед функцій, які надає це середовище, слід відзначити обробку виняткових ситуацій, забезпечення безпеки, інструменти налагодження для підтримки версій. Усі ці функції доступні з будь-якої мови програмування відповідно до специфікації загальної мови. Microsoft пропонує три мови програмування, які можуть використовувати CLR: Visual Basic. NET, Visual C#. NET і Visual C++ із керованими розширеннями. Крім того, працює ряд сторонніх компаній. NET-версії таких мов програмування як Perl, Python і COBOL.

Код, скомпільований компілятором для CLR, називається керованим кодом(managed code). Керований код використовує переваги середовища виконання та, крім самого коду, містить метадані, які створюються під час процесу компіляції та містять інформацію про типи, члени та посилання, що використовуються в коді. Метадані використовуються середовищем виконання. Рантайм також відстежує час життя об'єктів. У COM/COM+ для цього використовувалися спеціальні лічильники (reference counter). CLR також використовує лічильники, а об'єкти видаляються з пам'яті за допомогою процесу, який називається збиранням сміття(garbage collection).

Common Language Runtime також визначає загальну систему типів, яка використовується всіма мовами програмування. Це означає, наприклад, що всі мови програмування працюватимуть з цілочисельними даними або даними з плаваючою комою однакового формату та довжини, а представлення рядків також буде однаковим для всіх мов програмування. Завдяки єдиній системі типів досягається більш легка інтеграція компонентів і коду, написаного на різних мовах програмування. На відміну від технології COM, яка також базується на наборі стандартних типів, але які надаються у двійковій формі, CLR дозволяє інтегрувати код (який може бути написаний різними мовами програмування) у режимі розробки, а не в режимі виконання. Після компіляції керований код містить метадані, що описують сам компонент, а також компоненти, які використовуються для створення коду. Середовище виконання перевіряє наявність усіх необхідних ресурсів. Використання метаданих дозволяє відмовитися від необхідності зберігати інформацію про компоненти в реєстрі. Таким чином, коли ви переміщуєте компонент на інший комп'ютер, вам більше не потрібно реєструвати цей компонент, а видалення компонента є простим питанням видалення збірки, яка його містить.

Також слід зазначити, що зі зростанням бази текстів час валідації зростатиме лінійно, а час пошуку ключових аспектів —швидше ніж лінійно, але повільніше ніж квадратично, тобто лінійно-логарифмічно, O(n log n).

### 3.5 Ітеративне покращення моделі

Розглянемо кілька стратегій покращення моделі (словника):

1. Побудова індексу слів, що зустрічаються у неправильно оцінених тренувальних відгуках, за їх частотою та поетапне видалення слів зі "страховкою". Тобто перед першою ітерацією запускаємо валідацію на тренувальних відгуках та знаходимо контрольну точність. Також скануємо всі тренувальні відгуки, що були оцінені невірно при валідації, та зберігаємо кількість входження слів (це одна із найдовших операцій, для 40 тисяч запитів по часу виконується на рівні з валідацією, проте складність лише O(n) тобто зі зростанням кількості тренувальних відгуків буде ставати

швидшим за валідацію). Сортуємо слова по кількості входжень і видаляємо службові слова (заперечення і прислівники ступеню). Після цього видаляємо перші слова по списку зі словника і знову валідуємо на всіх відгуках (не тільки не вірно оцінених). Якщо точність збільшилась, то залишаємо зміни, якщо не змінилась чи зменшилась — відновлюємо стару версію словника і відмічаємо, що це слово видаляти не слід (щоб пропускати його на наступних ітераціях). Цю інформацію записуємо разом зі словником, щоб в разі, якщо треба перервати роботу, можна було продовжити ітеративне покращення без втрати переліку слів, що не слід видаляти.

2. Пошук слів, у яких різниця між їх оцінкою та середнім сентиментом правильно оціненого тексту, в якому вони присутні, найбільш не збігається та їх видалення. Для кожного слова розраховуємо середній сентимент вірно оцінених відгуків, в яких воно присутнє, Записуємо різницю між цим словом та середнім сентиментом. Перевіряємо точність до і після видалення слова зі словника. Якщо зміна негативна, то відновлюємо стару версію. Зберігаємо список слів, що не слід видаляти. Складність O(n) по відгукам, якщо візьмемо що кількість різних слів в текстах — константа. Якщо припустити, що кількість різних слів в текстах, в залежності від кількості текстів, зростає логарифмічно, то маємо O (n log n).
3. Пошук слів з низьким показником TF-IDF в текстах з неправильною оцінкою та їх видалення. За метрикою TF-IDF описаною в розділі 2.3.6 можна обрати слова для видалення зі словника. Низький показник означає, що слово не дуже важливе в тексті і часто зустрічається серед усіх текстів. Знову ж таки, для ефективного використання цього підходу потрібен хороший словник синонімів та дуже багато обчислень. Процедура така ж, як і в попередніх підходах — збереження стану словника на випадок падіння точності і відновлення за потреби. Складність O(n log n), але з великою кількістю обчислень для кожного слова.

4. Пошук та видалення слів, що стоять поруч з позитивними та негативними словами зі схожою частотою. Тобто, якщо якесь слово з приблизно однаковою вірогідністю може зустрічатись і в позитивному, і в негативному безпосередньому оточенні, то можна припустити, що його оцінка не грає ролі. Складність O(n log n).

Протестуємо кожен з цих підходів на 10 хвилинах покращення та порівняємо результати.

Таблиця 3.2 – ПОРІВНЯННЯ ПІДХОДІВ ПОКРАЩЕННЯ СЛОВНИКА

| Підхід | Кількість ітерацій за 10 хвилин | Покращення точності | Середнє покращення за ітерацію |
|---|---|---|---|
| Часті слова в неправильно оцінених відгуках | 30 | 3.52% | 0.117%. Покращення вповільнюється на 75.55% |
| Слова з великою середньою різницею між їх сентиментом та сентиментом тексту | 27 + перший запуск на базі текстів займає 2 хвилини на обрахування | 1.15% | 0.04%. Багато ітерацій неефективні. |
| Слова з низьким TF-IDF в неправильно оцінених текстах | 12, майже в три рази повільніше | 0.721% | 0.06%. |
| Слова що часто бувають в відмінних контекстах | 22 | 0.136% | 0.06% Покращення майже зупиняється на 71%, тобто лише кілька слів слід видалити |

Розглянемо окремо результати тестів.

Перший підхід показав себе найкраще, проте очевидно, що видалення з таким наївним підходом легко можна перетренувати (overtrain) модель на даній базі текстів, тобто зробити її дуже ефективною для тренувальних текстів, проте менш ефективною в цілому. Тому потрібно періодично звірятись з тестами на зовнішніх текстах, щоб запобігти цьому явищу.

Другий підхід має гірші показники, що свідчить про те, що обрана метрика слабо корелює з точністю оцінки слова. Про це також можна здогадатись,

переглянувши логи ітерацій: більше третини ітерацій відпрацювали впусту, не збільшивши точності моделі, а тому були відкинуті.

Третій підхід має право на життя, проте потребує сильної оптимізації.

Четвертий підхід трохи збільшує точність моделі, проте самостійно видає дуже слабке покращення. Можливо, його слід використовувати разом з іншими підходами.

Для проведення подальших тестів перший та четвертий підхід використовуються разом (почергово). Після 30 хвилин покращення моделі отримуємо 76% точності.

Проведено ручний аналіз неправильно оцінених відгуків. З випадково взятої сотні неправильно оцінених відгуків 44 не мали чітко вираженої думки. Тобто короткі неінформативні відгуки типу «Непогана книга» [негативний], в яких справжня оцінка не сильно корелює з текстом. 35 неправильно оцінених відгуків мали чітко виражену думку, але не були правильно оцінені аналізатором. 27 неправильно оцінених відгуків містили сарказм в тій чи іншій формі.

Тобто приблизно 20% неправильно оцінених відгуків теоретично можна оцінити правильно, якщо мати кращий словник.

### 3.6 Побудова нового словника

Для того, щоб не спиратись на створений іншими людьми словник (який очевидно має багато невідповідностей кожному конкретному полю використання), можна використати один зі статистичних підходів датамайнингу для побудови словника. Найпростіший спосіб, покращенням якого є більшість інших - це збір статистик про частоту використання всіх слів у позитивних та негативних відгуках, нормалізація даних, тобто зведення абсолютних частот до відносних, та до рівнів від -1 до 1 в залежності від сентименту. Маючи простий словник такого типу, потрібно відкинути слова, які однаково часто зустрічаються в усіх відгуках. Це різні службові та нейтральні слова. Наприклад сполучники, займенники, деякі додатки, нейтральні дієслова, тощо. Вони хоч і зустрічаються часто, все ж не несуть ніякого емоційного навантаження. Після

цього словником можна користуватись. Проте все ж краще також прогнати його якимось зі способів ітеративного покращення. На виході отримаємо словник, спеціалізований на конкретному полі використання, тобто на тому, звідки були взяті тексти для тренування, як описано в розділі 2.3.8.

Для тестування підходу було оцінено кілька слів таким способом, їх оцінки відрізняються від оцінених вручну в словнику SentiWordNet3:

Таблиця 3.3 – ПОРІВНЯННЯ СТАРОГО І НОВОГО СЛОВНИКІВ

| Слово | SentiWordNet3 | Статистичний збір на відгуках з Amazon: |
|---|---|---|
| Disappointed (розчарований) | -0.5 | -0.65 |
| Lame (в різних контекстах: кульгавий, недолугий, нецікавий) | 0 (нейтральне) | -0.51 |
| Boring (нудний) | -0.25 | -0.76 |
| Solid (в різних контекстах: твердий, достатній, солідний) | +0.25 | +0.52 |

Для тестування було замінено оцінки цих слів у словнику на нові значення, точність оцінки на тестувальних текстах (не тих, з яких збиралась статистика) зросла відразу на 0.5% всього від 4 слів. Можна припустити, що значна кількість неправильно оцінених відгуків буде оцінена правильно після повного збору словника.

Побудова повного словника може зайняти дуже довгий час на звичайному робочому комп’ютері. В англійській мові більше мільйона слів. Всі їх треба шукати в кожному з відгуків — неймовірна кількість роботи комп’ютера. З користувацьким процесором AMD Ryzen 3 2200U обробка одного слова на 760 тисячах відгуків займає 8 секунд, що для мільйону слів означає більше трьох місяців роботи.

Замість цього було додано можливість поступово переписувати старий словник новими даними по ходу їх збору суто для демонстрації.

Після комбінації різних методів покращення та скорочення словника з часом роботи 30 хвилин отримано точність 77.32%.

# РОЗДІЛ 4. РОЗРОБКА ПРОГРАМНОГО ЗАСОБУ

## 4.1 Вибір моделі та платформи розробки програмного забезпечення

### 4.1.1 Вибір моделі

Дослідження також потребувало створення програмного засобу для тестування обраних підходів, з якого в майбутньому можна буде запозичити основний функціонал для використання обраного підходу в інших умовах (наприклад інший вид даних, інші вимоги до створення висновків, тощо).

Розробка відбувалась при використанні так званої спіральної моделі розробки програмних засобів, в основі якої лежить ітеративне покращення продукту, постійна робота над помилками, та ситуативне переозначення пріоритетів розробки. Спіральна розробка програмного забезпечення — це перевірена часом та ефективна техніка для створення програмного забезпечення найвищої якості та стандартів. Це гібрид моделі Agile та Waterfall, орієнтованої на ітераційну розробку та управління ризиками. Підхід Spiral Software Development дозволяє розробникам неодноразово надбудовувати поточний код, гарантуючи, що кожна версія є надійнішою, ніж попередня. Він починається з невеликого набору вимог і проходить кілька циклів планування, аналізу ризиків, проектування, розробки та оцінки, щоразу додаючи більше функціональних мо жливостей і знижуючи ризики. Модель називається «спіраль», тому що вона приймає форму спіралі в міру просування з кожною спіральною ітерацією до кінцевого рішення.

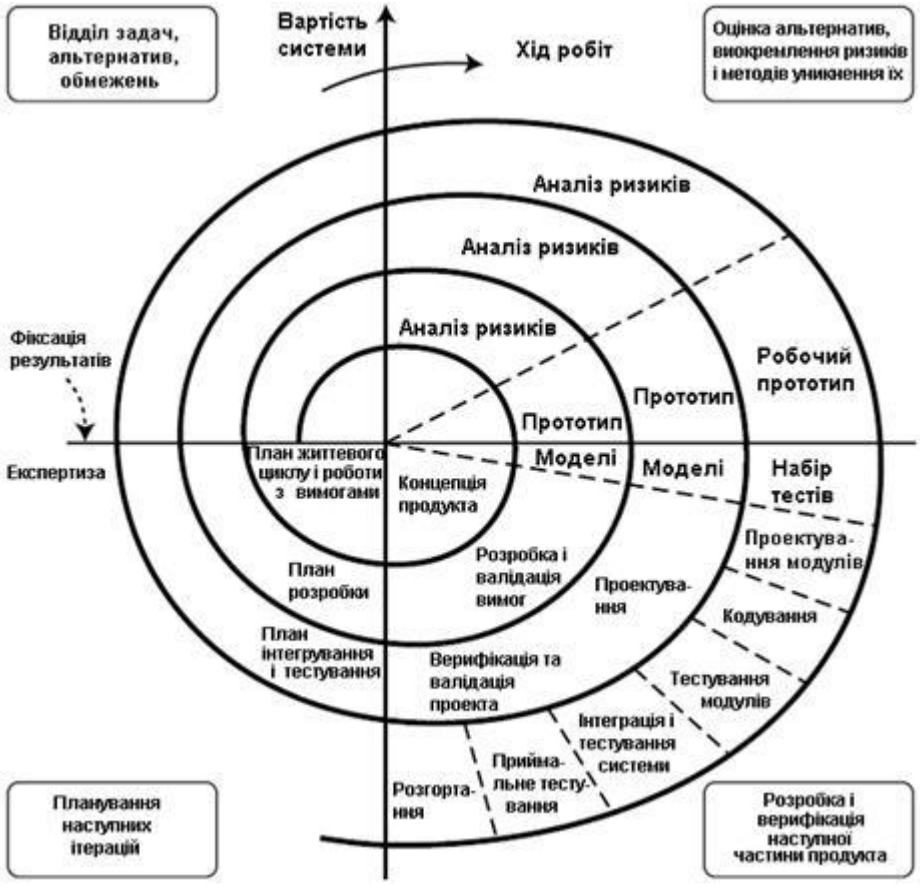



Рисунок 4.1 Спіральна модель розробки програмного забезпечення

Кожен цикл спіралі починається з визначення цілей частини продукту, що розробляється (продуктивність, функціональність, здатність вносити зміни тощо); альтернативні засоби реалізації цієї частини продукту (дизайн А, дизайн В, повторне використання, покупка тощо); і обмеження, накладені на застосування альтернатив (вартість, розклад, інтерфейс тощо). Наступним кроком є оцінка альтернатив щодо цілей і обмежень. Часто цей процес визначає зони невизначеності, які є значними джерелами ризику проекту. Якщо так, то наступним кроком має бути розробка економічно ефективної стратегії для усунення джерел ризику. Це може включати створення прототипів, моделювання, порівняльний аналіз, перевірку посилань, адміністрування анкет користувачів, аналітичне моделювання або комбінації цих та інших методів вирішення ризиків. Після того, як ризики оцінені, наступний крок визначається відносними ризиками, що залишилися. Якщо продуктивність або ризики

інтерфейсу користувача значною мірою домінують над ризиками розробки програми або внутрішнього контролю інтерфейсу, наступним кроком може бути еволюційний розвиток: мінімальні зусилля для визначення загальної природи продукту, план для наступного рівня прототипування та розробка більш детального прототипу для продовження вирішення основних проблем ризику.

Якщо цей прототип є експлуатаційно корисним і достатньо надійним, щоб служити основою з низьким рівнем ризику для майбутньої еволюції продукту, наступними кроками, керованими ризиком, буде еволюційна серія еволюційних прототипів, що йдуть праворуч на рисунку 1. У цьому випадку варіант написання специфікацій буде розглядатися, але не здійснюватися. Таким чином, міркування щодо ризику можуть призвести до реалізації проекту лише підмножини всіх потенційних кроків у моделі.

З іншого боку, якщо попередні спроби створення прототипів уже усунули всі ризики, що загрожують продуктивності або інтерфейсу користувача, а ризики розробки програми або контролю інтерфейсу домінують, наступний крок слідує базовому каскадному підходу (концепція роботи, вимоги до програмного забезпечення, попередній дизайн), і т.д. на рисунку 1), модифікований за потреби для включення поступового розвитку. Після кожного рівня специфікації програмного забезпечення на малюнку слідує етап перевірки та підготовка планів для наступного циклу. У цьому випадку параметри прототипування, моделювання, моделювання тощо розглядаються, але не реалізуються, що призводить до використання іншої підмножини кроків.

Ця підгрупа кроків спіральної моделі, керованих ризиком, дозволяє моделі вмістити будь-яку відповідну суміш орієнтованих на специфікацію, орієнтованих на прототип, орієнтованих на моделювання, орієнтованих на автоматичне перетворення, або інших підходів до розробки програмного забезпечення. У таких випадках вибирається відповідна змішана стратегія, враховуючи відносну величину програмних ризиків і відносну ефективність різних методів для усунення ризиків. Подібним чином міркування щодо управління ризиками можуть визначати кількість часу та зусиль, які слід

приділити таким іншим видам діяльності проекту, як створення плану, управління конфігурацією, забезпечення якості, формальна перевірка та тестування. Зокрема, керовані ризиком специфікації (як обговорюється в наступному розділі) можуть мати різний ступінь повноти, формальності та деталізації, залежно від відносних ризиків виконання занадто малого або занадто великого обсягу специфікацій.

Важливою особливістю спіральної моделі, як і більшості інших моделей, є те, що кожен цикл завершується оглядом за участю основних людей або організацій, зацікавлених у продукті. Цей огляд охоплює всі продукти, розроблені протягом попереднього циклу, включаючи плани на наступний цикл і ресурси, необхідні для їх виконання. Основна мета перегляду полягає в тому, щоб переконатися, що всі зацікавлені сторони взаємно зобов'язані підходити до наступного етапу.

Плани для наступних етапів можуть також включати поділ продукту на інкременти для послідовного розвитку або компоненти, які будуть розроблені окремими організаціями чи особами. Для останнього випадку візуалізуйте серію паралельних спінальних циклів, по одному для кожного компонента, додаючи третій вимір концепції, представленій на рисунку 1. Наприклад, окремі спіралі можуть розвиватися для окремих програмних компонентів або приростів. Таким чином, етап перегляду та прийняття зобов'язань може варіюватися від індивідуального проходження дизайну окремого компонента програміста до перегляду основних вимог за участю розробника, клієнта, користувача та організацій з обслуговування.

Спіральна модель особливо добре підходить для великих, складних програмних проектів із значною невизначеністю щодо вимог або технології. Така модель добре підходить для  розробки програмних засобів одноосібно, бо вона не жертвує ефективністю індивідуальних розробників для покращення командної співпраці, що при роботі самотужки не актуально. Перегляд, тестування, та доповнення специфікації повторюються з різними пріоритетностями на кожній ітерації розробки, що також добре пасує таким

умовам розробки, тому що допомагає швидко переключатись між різними аспектами розробки за необхідності.

Нижче наведено деякі з переваг спіральної моделі для розробників програмного забезпечення та керівників проектів груп розробників програмного забезпечення:

1. Управління ризиками: спіральна модель наголошує на управлінні ризиками, яке допомагає виявити та вирішити потенційні проблеми на ранніх етапах процесу розробки.

2. Гнучкість: модель допускає ітерації та еволюцію, що робить її гнучкою та адаптованою до мінливих вимог і ризиків.

3. Поступова розробка: спіральна модель надає програмний продукт поступово, що дозволяє проводити часті оцінки та коригування курсу.

4. Ретельне тестування: модель включає формальний етап оцінювання, який гарантує, що програмна система ретельно протестована та відповідає всім вимогам.

### 4.1.2 Вибір платформи

Розробка програмного засобу була націлена на платформу .NET Core 3.1 від Microsoft, що є однією з найбільш популярних та найкраще підтримуваних платформ у світі станом на наш час. Платформа має відкритий джерельний код, підтримку багатьох операційних систем (Windows, Linux, MacOS), та може бути портована до інших систем, підтримку багатьох мов (C#, C++, F#, VB), та перебуває в активній розробці. В якості середовища розробки було обрано Visual Studio 2019, порівняно із минулими версіями в ньому з’явилась повна підтримка .NET Core 3.1, розширені можливості інтеграції сторонніх інструментів, та низка інших покращень. Мовою розробки було обрано C#. Вибір C# для opinion mining і аналізу сентиментів може бути корисним з широкого спектру причин. По-перше, C# — це мова загального призначення, яка підходить для різноманітних складних завдань, зокрема для аналізу думок і настроїв. Це об’єктно-орієнтована мова, яка дозволяє повторно використовувати код і з якою легше працювати для більш масштабних проектів, які можуть включати складніші набори даних і

алгоритми. Крім того, C# включає в себе набір бібліотек доступних інструментів, які можна використовувати для обробки тексту та аналізу настроїв, а також різноманітні можливості машинного навчання та штучного інтелекту. Нарешті, C# має добре відоме співтовариство та різноманітні онлайн-ресурси для вказівок і допомоги, що полегшує розробникам і дослідникам створення складних програм для аналізу думок і настроїв. Для збереження та обробки великої кількості інформації (в цій роботі — до 1 мільйона відгуків та 5-10 мільйонів н-грам) було використано систему управління реляційними базами даних MSSQL, що, завдяки багатьом низькорівневим оптимізаціям, дозволяє значно зменшити час виконання операцій над даними в порівнянні зі збереженням інформації в необробленому вигляді. MSSQL, також відома як Microsoft SQL Server, — це система керування реляційними базами даних, розроблена Microsoft. Вона розроблена для ефективної роботи з інтенсивними робочими навантаженнями, допомагаючи користувачам спростити складність керування та захисту даних, що зберігаються з багатьох джерел. Це дозволяє користувачам створювати масштабовані програми та великомасштабні сховища даних. Вона(система) також включає розширені аналітичні можливості, такі як аналізи на основі штучного інтелекту, які допомагають розкривати приховані шаблони даних за допомогою прогнозної аналітики та машинного навчання. Деякі з її інших функцій включають оптимізацію використання ресурсів і зберігання, сприяння безпеці та зниження витрат на обслуговування.

### 4.2 Пошук ключових особливостей з тексту

Для пошуку ключових особливостей для порівняння застосовано кілька підходів швидкої розгортки, що не потребують обчислювальних потужностей рівня великого підприємства. На цьому етапі найбільше використовуються н-грами, що можна створити при роботі пайплайну з включеним модулем н-грамізації.

1. Пошук слів, що часто стоять поруч з емоційно забарвленими словами — досить наївний метод, проте потребує найменше обчислень. Потрібно

знайти всі н-грами, що мають в собі емоційно забарвлені слова (ця операція значно прискорюється наявністю некластерного індексу по центральним словам). Серед цих н-грам шукаються часті слова. Звісно, найчастішими словами будуть службові слова. Теоретично можна відкидати їх підбором проценту найчастіших слів, що зустрічаються. Проте так можна відкинути важливі слова.

2. TF-IDF метрика для пошуку ключових особливостей. За високим значенням TF-IDF для слова можна припустити, що слово важливе для конкретного тексту. Можемо зібрати значення метрики для кожного слова, для кожного тексту, але це займе багато часу. Замість цього було використано лише IDF. Так для кожного слова ми збираємо лише цю частину метрики і другий прохід по текстам не потрібен. Більш того, для коротких текстів шанс повторення слів не досить великий щоб використовувати TF-IDF, а більша частина відгуків коротша за 400 символів, тому обмежимось IDF. Якщо відсортувати слова по спаданню лише IDF, зверху буде багато рідкісних слів, що можуть не нести змістової нагрузки, тобто таких, що просто рідко зустрічаються в повсякденній мові. Тому потрібно сортувати слова по комбінованій метриці, що враховує частоту вживання слова в цілому. Наприклад IDF помножена на частоту слова по всіх текстах. Знову ж таки, генерація IDF-словника займає багато часу. Її слід провести один раз для кожної бази текстів і зберегти результат.
3. Word2Vec. Word2Vec — це нещодавній прорив у світі програмування натуральних мов. Вбудовування слів(embedding) є невід'ємною частиною вирішення багатьох проблем у програмуванні натуральних мов. Вони зображують машині, як люди розуміють мову. Ви можете уявити це як векторизоване представлення тексту. Word2Vec, поширений метод генерації вбудованих слів, має різноманітні застосування, такі як подібність тексту, системи рекомендацій, аналіз настроїв тощо. Перш ніж перейти до word2vec, давайте розберемося, що таке вбудовування слів. Це важливо знати, оскільки загальний результат і вихід word2vec будуть вбудованими,

пов’язаними з кожним унікальним словом, яке пройшло через алгоритм. Вбудовування слів — це техніка, за якої окремі слова перетворюються на числове представлення слова (вектор). Коли кожне слово зіставляється з одним вектором, цей вектор вивчається у спосіб, який нагадує нейронну мережу. Вектори намагаються вловити різні характеристики цього слова щодо загального тексту. Ці характеристики можуть включати семантичний зв’язок слова, визначення, контекст тощо. За допомогою цих числових представлень ви можете робити багато речей, як-от визначати схожість чи відмінність між словами. Очевидно, що вони є невід’ємною частиною різних аспектів машинного навчання. Машина не може обробляти текст у його необробленому вигляді, тому перетворення тексту на вбудований дозволить користувачам передавати вбудований текст у класичні моделі машинного навчання. Найпростішим вбудовуванням було б одне гаряче кодування текстових даних, де кожен вектор буде зіставлено з категорією. Проте є численні обмеження таких простих вставок, як це, оскільки вони не фіксують характеристики слова, і вони можуть бути досить великими залежно від розміру корпусу. Ефективність Word2Vec полягає в його здатності групувати разом вектори подібних слів. Враховуючи достатньо великий набір даних, Word2Vec може зробити точні оцінки значення слова на основі його появи в тексті. Ці оцінки дають асоціації слів з іншими словами в корпусі. Наприклад, такі слова, як «король» і «королева», будуть дуже схожі одне на одне. Виконуючи алгебраїчні операції над вкладеннями слів, ви можете знайти близьке наближення подібності слів. Наприклад, двовимірний вектор вбудовування "короля" - двовимірний вектор вбудовування "чоловіка" + двовимірний вектор вбудовування "жінки" дали вектор, який дуже близький до вектора вбудовування "королеви". Зауважте, що наведені нижче значення вибрано довільно. Цей метод використовує широку нейронну мережу (мало шарів, велика розмірність шарів), на вхід якої передаються вектори присутності різних слів в конкретних текстах, а на виході очікуються вектори присутності особливостей. Тобто на етапі

підготовки даних створюються вектори присутності слів з розмірністю, що дорівнює кількості унікальних слів у текстах, де для кожного слова ставиться у відповідність 1, якщо слово є в тексті, що відповідає даному вектору, або 0, якщо його нема. Також для кожного тренувального тексту потрібно мати ручну оцінку з приводу наявності певних особливостей в тексті. Тренування нейронної мережі відбувається як завжди, проте доцільним майже завжди є наявність певного dropout, тобто відкидання випадкових клітин у кожну епоху для того щоб нейронна мережа не перетренувалась на певних словах і була більш всебічною. Такий підхід можна натренувати лише для розпізнавання вже знайдених ключових особливостей і період розгортання (налаштування) моделі в рази більший ніж у попередніх методів, проте ефективність для пошуку заздалегідь визначених особливостей інколи виправдовує такий підхід.

Нажаль, хоч всі три методи хоч і вирішують схожі задачі, все ж об’єктивно оцінити їх ефективність дуже складно, бо ніякої спільної числової характеристики для них не існує. Тому внизу наведено лише результати перевірки на практиці та загальні характеристики методів.

Таблиця 4.1 – ПОРІВНЯННЯ МЕТОДІВ ПОШУКУ КЛЮЧОВИХ ОСОБЛИВОСТЕЙ

| Метод | Переваги | Недоліки |
|---|---|---|
| Часті слова поруч з емоційно забарвленими | Висока швидкість розгортання, пошук нових особливостей, більшість робити перекладена на створення словника та не вповільнює сам аналіз | Результати повністю залежать від словника, потребує збереження н-грам з даними про сентимент, потребує доопрацювання результатів вручну |
| Базовані на TF-IDF метрики | Мінімізує засмічення результатів службовими словами, краще працює для великих текстів, середня швидкість розгортання, може адаптуватись під різні контексти | Для кожного контексту потрібен новий IDF словник, що створюється довго, не зовсім доцільно для коротких текстах |
| Word2Vec | Добре знаходить вже відомі особливості, з правильно натренованою мережею можна отримати досить стійку до аномалій модель | Потребує дуже великої кількості специфічно оцінених даних (вручну), знаходження нових особливостей неможливе |

Можна зробити висновок, що для системи швидкого розгортання, що могла б швидко змінити область використання, доцільніше використовувати перший метод. Нижче приведено приклад результатів роботи цього методу для загального пошукового запиту “book” (книга):

Таблиця 4.2 – РЕЗУЛЬТАТИ ПОШУКУ СЛІВ, ЩО СТОЯТЬ ПОРУЧ ІЗ «BOOK»

| Слово англійською | Переклад | Кількість входження коло емоційних слів |
|---|---|---|
| interesting | Цікавий(-а) | 145 |
| boring | Нудний(-а) | 118 |
| story | Історія (сюжет) | 84 |
| fun | Веселий(-а) | 65 |
| characters | Персонажі | 60 |
| character | Персонаж (однина) | 30 |
| short | Короткий(-а) | 65 |
| long | Довгий(-а) | 57 |

Зауваження: всі службові слова відфільтровано. Кількість відгуків, що було просканована — 2200.

З цих результатів можна зробити висновок, що власники книг у відгуках часто звертають увагу на сюжет, персонажів та довжину книги.

Серед цих відгуків знайдемо, на що звертають увагу, коли пишуть про сюжет:

Таблиця 4.3 – РЕЗУЛЬТАТИ ПОШУКУ СЛІВ, ЩО СТОЯТЬ ПОРУЧ ІЗ «STORY»

| Слово англійською | Переклад | Кількість входження коло емоційних слів |
|---|---|---|
| original | Оригінальний | 26 |
| entertaining | Такий що розважає | 22 |
| boring | Нудний | 21 |
| moving / touching | Зворушливий | 19+16 |
| fascinating | Захоплюючий | 17 |
| kind | Добрий | 15 |
| character | Персонаж | 15 |
| intriguing | Інтригуючий | 14 |
| adventure | Пригода | 12 |
| romance | Роман | 11 |
| realistic | Реалістичний | 9 |

| | | |
|---|---|---|
| mystery | Таємничість | 8 |
| believable | Правдоподібний | 7 |
| deep | Глибокий | 6 |

З таких результатів можна зробити ще ряд висновків і далі заглиблюватись в аналіз деталей за необхідності.

### 4.3 Огляд інтерфейсу

Інтерфейс програмного засобу включає вхідні точки для всіх аспектів роботи програми від первинної обробки з файлів до бази даних, до всіх видів аналізу.

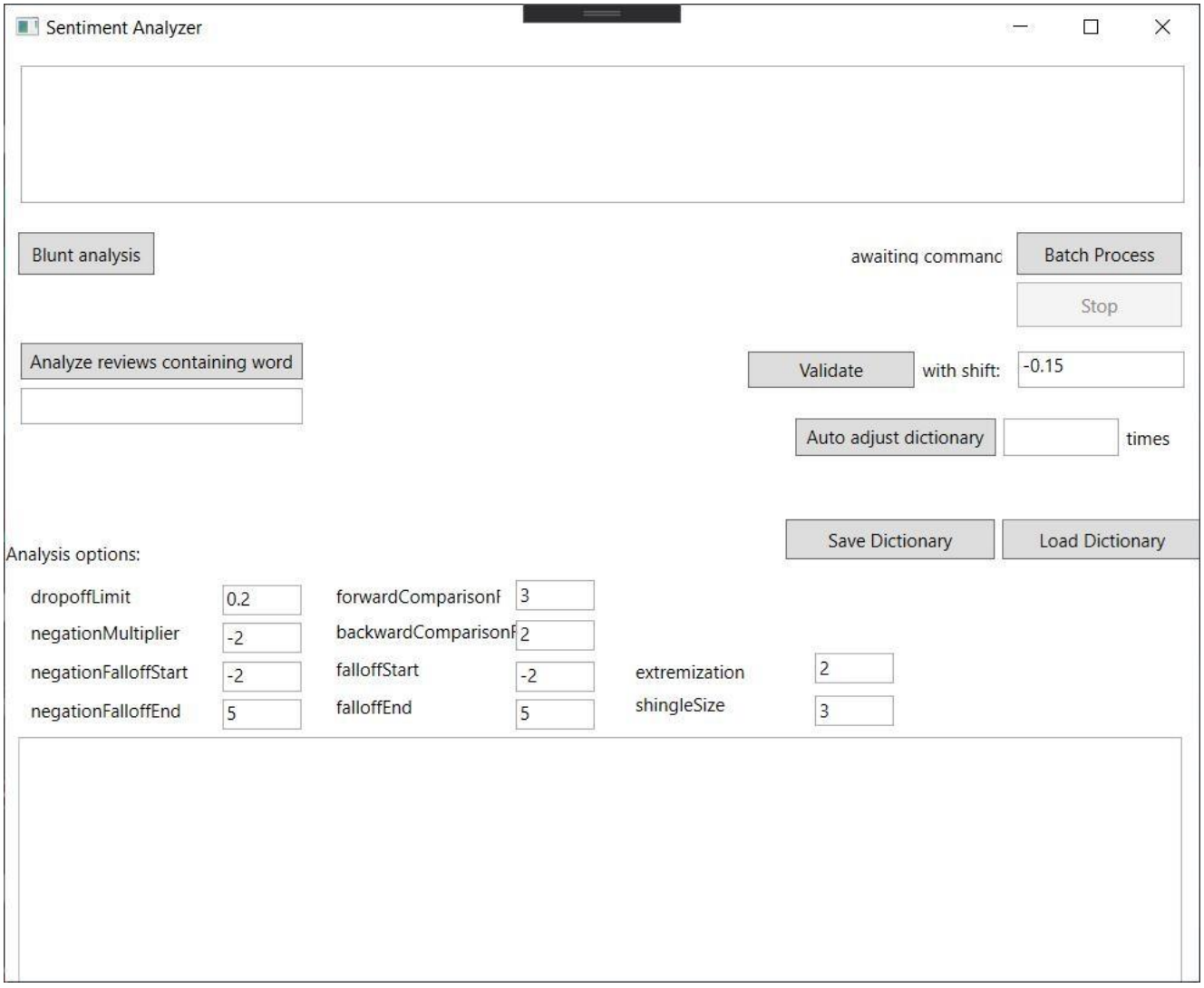



Рисунок 4.2 Базовий інтерфейс програми

Панель внизу інтерфейсу слугує для виводу результатів різних видів робіт

— включно з роботою з базою даних, валідацією моделі, аналізу одиничних словників, виводу неправильно проаналізованих текстів (для тестування моделі), пошуком ключових особливостей, збереження за завантаження словників з файлів та інших. Всі параметри аналізу автоматично завантажуються з передвстановлених та їх можна міняти безпосередньо в програмі.

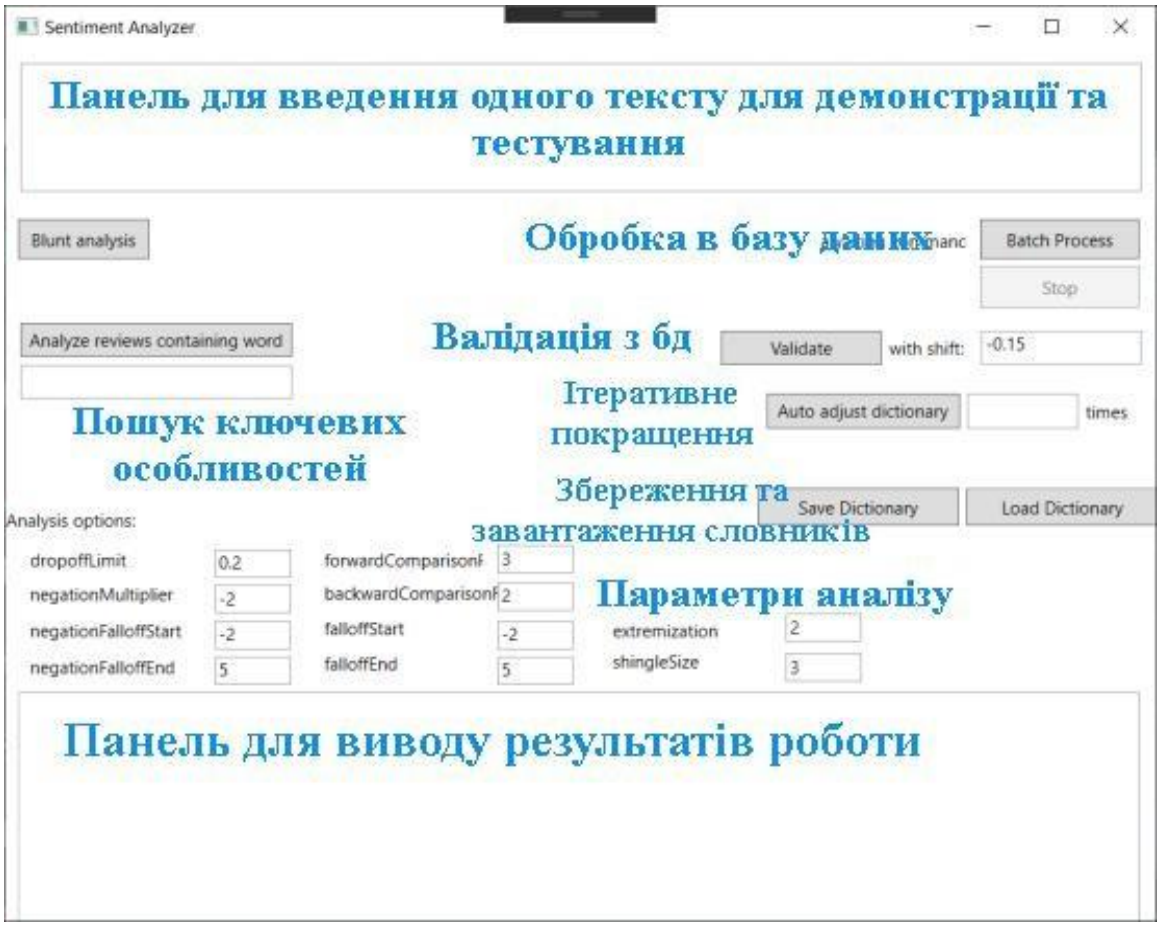



Рисунок 4.3 Пояснення щодо елементів інтерфейсу


Рисунок 4.4 Приклад виводу валідації моделі

# ВИСНОВКИ

В ході роботи було проаналізовано, протестовано та порівняно підходи до opinion mining, оцінки емоційного навантаження тексту, пошуку важливих особливостей в тексті для великих баз дописів з інтернету, проаналізовано та протестовано різноманітні підходи то побудови сентиментного словника та метрик для ефективного збору інформації. Створено програмний засіб, що дозволяє швидко розгонути систему для аналізу текстів та пошуку ключових особливостей у великих базах текстів за різними підходами. Програмний засіб також має можливість ітеративного покращення сентиментних словників на базі оцінених текстів, а також, за наявності великих обчислювальних потужностей, дозволяє побудувати сентиментний словник під свій конкретний контекст текстів кількома статистичними методами.

## ПЕРЕЛІК ВИКОРИСТАНИХ ІНФОРМАЦІЙНИХ ДЖЕРЕЛ


1. Usama Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth From Data Mining to Knowledge Discovery in Databases

2. Haldorai, Anandakumar, Ramu, Arulmurugan Cognitive Social Mining Applications in Data Analytics and Forensics

3. Gellman, Barton; Poitras, Laura (6 Червня, 2013). "US Intelligence Mining Data from Nine U.S. Internet Companies in Broad Secret Program". The Washington Post.

4. Kotelnikov E. V., Bushmeleva N. A., Razova E. V., Peskisheva T. A., Pletneva M. V, Manually Created Sentiment Lexicons: Research and Development, 2016

5. Vasileios Hatzivassiloglou, Kathleen R. McKeown, Predicting the Semantic Orientation of Adjectives

6. Read J, Carroll J. Weakly supervised techniques for domain-independent sentiment classification. In: Proceeding of the 1st international CIKM workshop on topic-sentiment analysis for mass opinion; 2009. ст. 45-52

7. Ruiz M, Srinivasan P. Hierarchical neural networks for text categorization.

8. Ng Hwee Tou, Goh Wei, Low Kok. Feature selection, perceptron learning, and a usability case study for text categorization. Конференція ACM SIGIR;

9. Alexander Pak, Patrick Paroubek. Twitter as a Corpus for Sentiment Analysis and Opinion Mining (2010)

10. Carenini, G., Ng, R. and Zwart, E. “Extracting Knowledge from Evaluative Text”. Proceedings of the Third International Conference on Knowledge Capture (K-CAP’05), 2005.

11. Ayesha Rashid, Naveed Anwer, Dr. Muddaser Iqbal, Dr. Muhammad Sher"A Survey Paper: Areas, Techniques and Challenges of Opinion Mining",International Journal of Computer Science, November 2013

12. McKinsey&Company. “Marketing & Sales Big Data, Analytics, and the Future of Marketing & Sales” (March 2015).

13. Michael Svilar, Arnab Chakraborty and Athina Kanioura “From hype to real help: Finding valuable consumer insight in a stream of data”.

14. Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani “SENTIWORD NET 3.0: An Enhanced Lexical Resourcefor Sentiment Analysis and Opinion Mining”

15. Mukherjee, Liu and Glance, “Spotting Fake Reviewer Groups in Consumer Reviews” (2012)

16. Parrott, W.G. “Emotions in Social Psychology: Volume Overview.” (2001)

17. Pyle D. “Data Preparation for Data Mining” (1999)

18. M.Bharati. “Data mining techniques and applications”

19. Bing Liu. “Sentiment Analysis and Opinion Mining”, Morgan & Claypool Publishers (May 2012).

20. Веб-сайт. URL: https://www.javatpoint.com/types-of-data-mining

21. Lecture Notes on Data Mining & Data Warehousing, Veer Surendra Sai University of Technology

22. Веб-сайт. URL: www.geeksforgeeks.org

23. Веб-сайт. URL: https://towardsdatascience.com/word2vec-explained-49c52b4ccb71

24. Kevin Burton, Sams. “NET Common Language Runtime Unleashed” (2002)

25. Sindhura Kannappan. “Sentiment analysis using natural language processing and machine learning” Shu Ju Cai Ji Yu Chu Li/Journal of Data Acquisition and Processing 38(2):520-526 (April 2023)

26. Yanqing Chen, Steven Skiena. “Building Sentiment Lexicons for All Major Languages” (June 2014)

27. Swati Kunwar, Neetu Bansla “Opinion Mining Analysis: A Framework” (2019)

28. Barry W. Boehm “A Spiral Model of Software Development and Enhancement”

29. Веб-сайт. URL: https://www.developer.com/project-management/spiral-software-development/

30. Hulth A. Beata B. Megyesi “A Study on Automatically Extracted Keywords in Text Categorization”

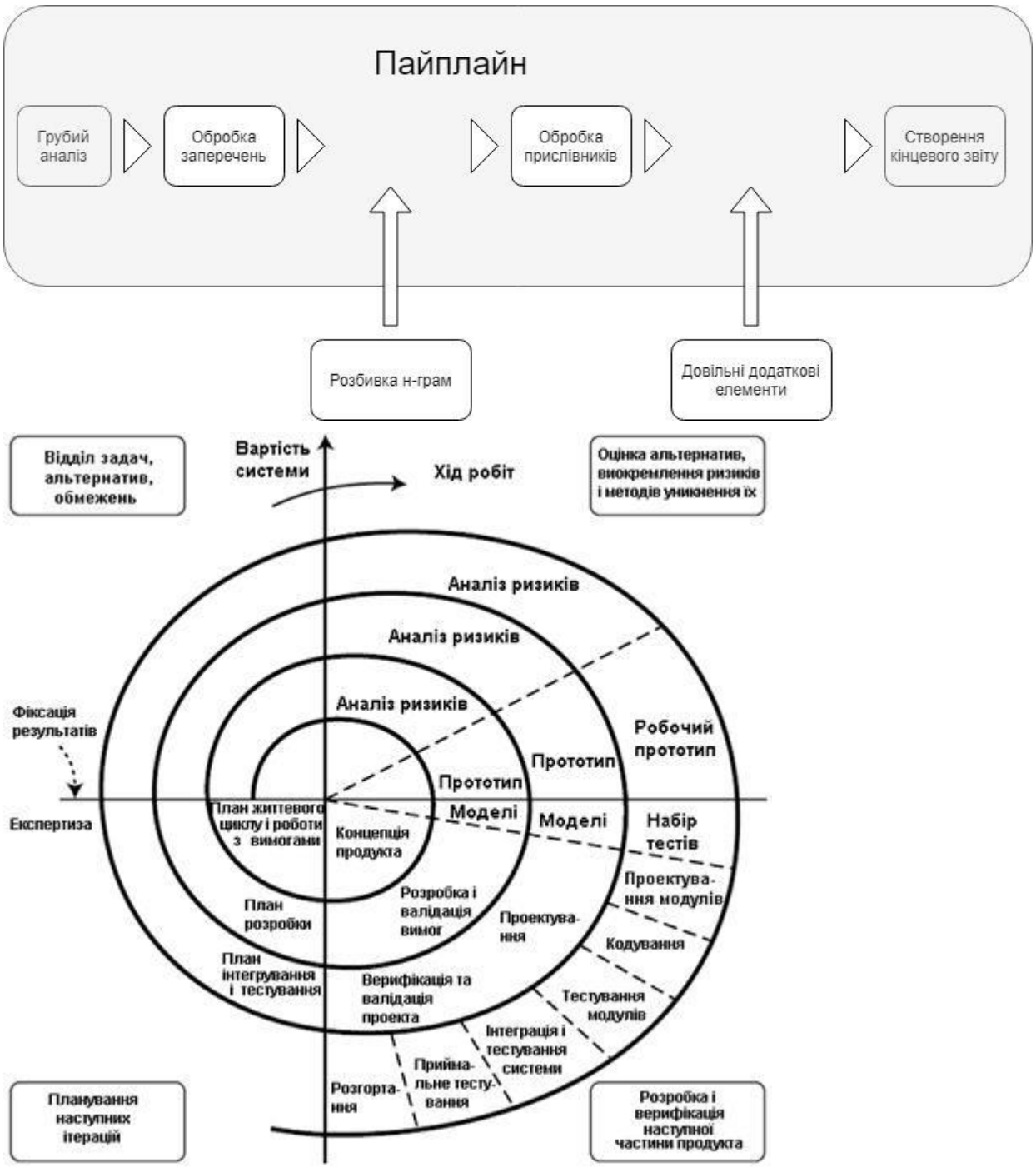
# ДОДАТКИ

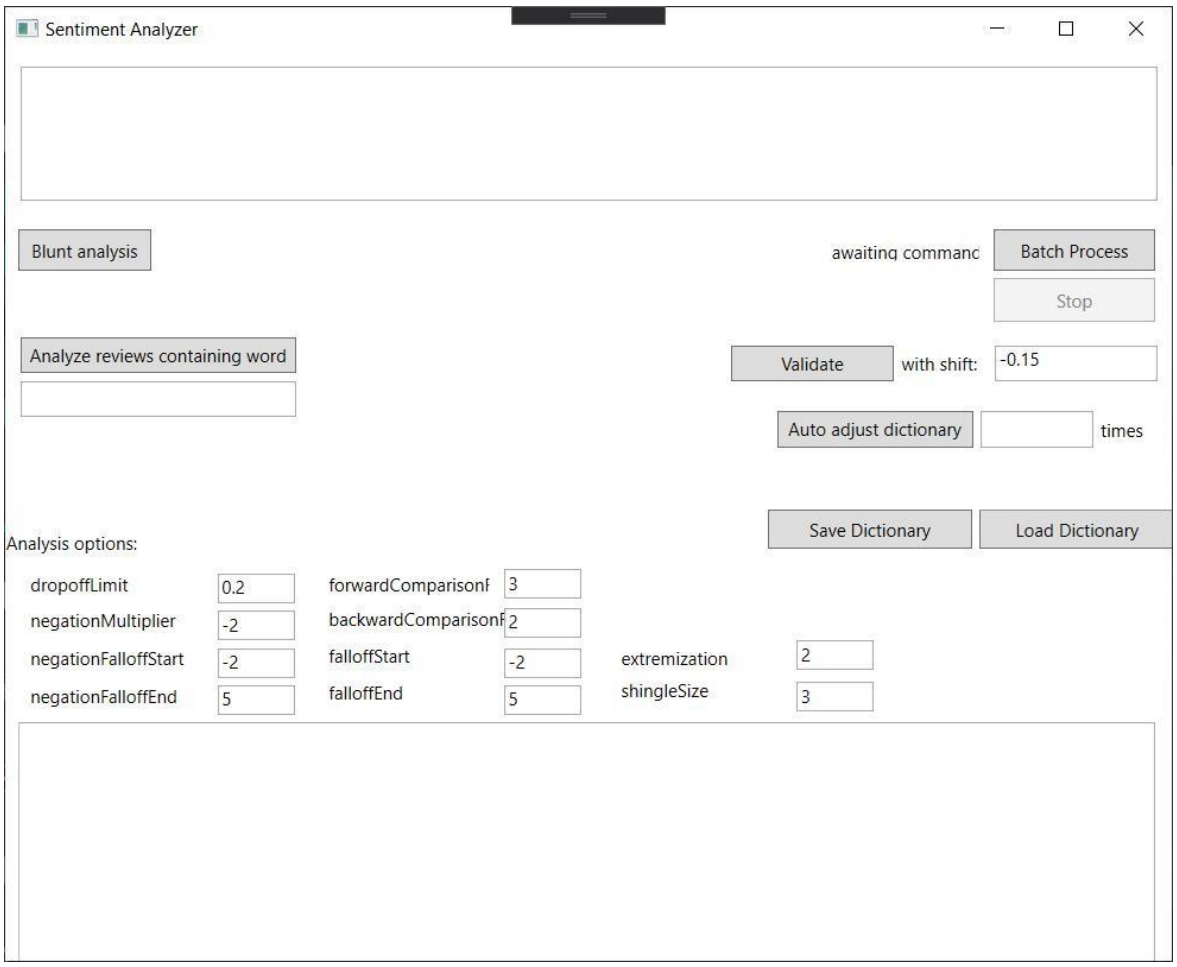
## Додаток А. Графічний матеріал

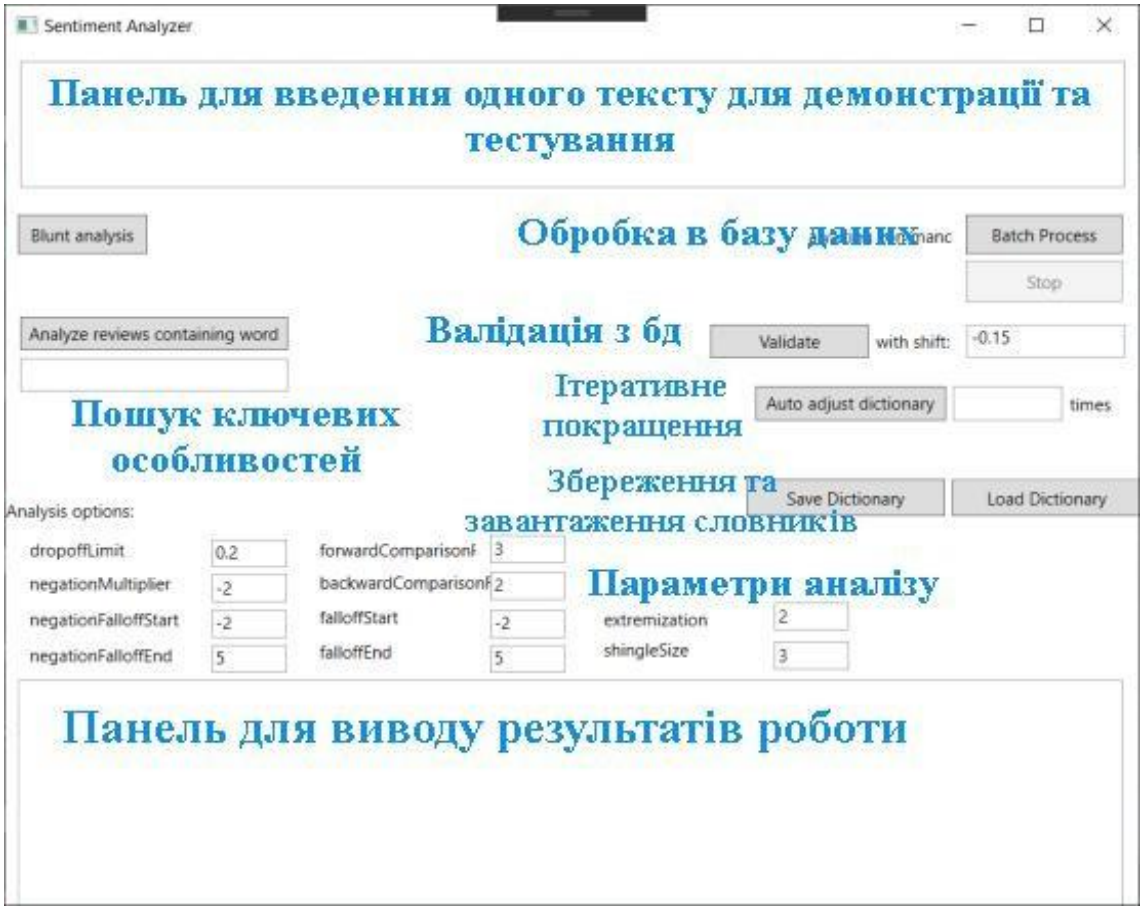


```
__label__1 Failed after a few days: It was nice while it
__label__2 Worked as needed!: I used this for my son's F
__label__2 Good performance at a great price!: I bought
__label__1 Failed after a few uses: I bought this card f
__label__1 I bought FlshMemory Card for my daughter, but
__label__2 Transcend 8GB: I use this for my Canon 1000D
__label__2 Big SDHC Card: I bought this card so I could
```

```
a       00008595        0       0.25    nonadsorptive#1 nonadsorbent#1  lacking a capacity to adsorb or cause to accumulate c
a       00008734        0.5     0       absorbable#1    capable of being absorbed or taken in through the pores of a surface
a       00008877        0.25    0       adsorbate#1 adsorbable#1        capable of being adsorbed or accumulated on a surface
a       00009046        0       0       abstemious#1    sparing in consumption of especially food and drink; "the pleasures c
a       00009346        0       0.625   abstinent#1 abstentious#1       self-restraining; not indulging an appetite especiall
a       00009618        0.5     0.25    spartan#4 austere#3 ascetical#2 ascetic#2       practicing great self-denial; "Be sys
a       00009978        0       0       gluttonous#1    given to excess in consumption of especially food or drink; "over-fec
a       00010385        0       0       crapulous#2     given to gross intemperance in eating or drinking; "a crapulous old r
a       00010537        0       0.5     crapulous#1 crapulent#1 suffering from excessive eating or drinking; "crapulent sleep
a       00010726        0       0       wolfish#2 voracious#2 ravenous#2 ravening#3 rapacious#3 esurient#3 edacious#1   devou
a       00011160        0       0       greedy#3        wanting to eat or drink more than one can reasonably consume; "don't
a       00011327        0       0.125   swinish#2 porcine#3 piggy#1 piggish#1 hoggish#1 resembling swine; coarsely gluttonous
a       00011665        0.125   0.375   too-greedy#1 overgreedy#1       excessively gluttonous
a       00011757        0       0       abstract#1      existing only in the mind; separated from embodiment; "abstract words
a       00012071        0.375   0       notional#4 ideational#1 conceptional#1  being of the nature of a notion or concept; "
a       00012362        0.375   0.25    conceptual#1    being or characterized by concepts or their formation; "conceptual di
a       00012689        0       0       ideal#2 constituting or existing only in the form of an idea or mental image or conce
a       00012932        0.125   0.125   ideological#2 ideologic#1       concerned with or suggestive of ideas; "ideological a
a       00013160        0.625   0.25    concrete#1      capable of being perceived by the senses; not abstract or imaginary;
a       00013442        0       0       objective#4     belonging to immediate experience of actual things or events; "object
a       00013662        0       0       tangible#2 real#4       capable of being treated as fact; "tangible evidence"; "his t
a       00013887        0       0.25    abundant#1      present in great quantity; "an abundant supply of water"
a       00014358        0       0.25    galore#2 abounding#1    existing in abundance; "abounding confidence"; "whiskey galor
a       00014490        0.125   0       rich#12 plentiful#2 plenteous#1 copious#2 ample#2       affording an abundant supply;
a       00014858        0       0       voluminous#3 copious#1  large in number or quantity (especially of discourse); "she t
```


Пайплайн
Грубий аналіз
Обробка заперечень
Обробка прислівників
Створення кінцевого звіту
Розбивка н-грам
Довільні додаткові елементи
Відділ задач, альтернатив, обмежень
Вартість системи
Хід робіт
Оцінка альтернатив, виокремлення ризиків і методів уникнення їх
Аналіз ризиків
Аналіз ризиків
Аналіз ризиків
Фіксація результатів
Робочий прототип
Прототип
Прототип
Експертиза
План життєвого циклу і роботи з вимогами
Концепція продукта
Моделі
Моделі
Набір тестів
Проектування модулів
План розробки
Розробка і валідація вимог
Проектування
Кодування
План інтегрування і тестування
Верифікація та валідація проекта
Тестування модулів
Інтеграція і тестування системи
Прийма-льне тестування
Розгортання
Планування наступних ітерацій
Розробка і верифікація наступної частини продукта


Sentiment Analyzer
Blunt analysis
awaiting commanc
Batch Process
Stop
Analyze reviews containing word
Validate
with shift:
-0.15
Auto adjust dictionary
times
Save Dictionary
Load Dictionary
Analysis options:
dropoffLimit
0.2
forwardComparisonF
3
negationMultiplier
-2
backwardComparisonF
2
negationFalloffStart
-2
falloffStart
-2
extremization
2
negationFalloffEnd
5
falloffEnd
5
shingleSize
3


Sentiment Analyzer
Панель для введення одного тексту для демонстрації та тестування
Blunt analysis
Обробка в базу даних
Batch Process
Stop
Analyze reviews containing word
Валідація з бд
Validate
with shift:
-0.15
Ітеративне покращення
Auto adjust dictionary
times
Пошук ключевих особливостей
Збереження та завантаження словників
Save Dictionary
Load Dictionary
Analysis options:
dropoffLimit
0.2
forwardComparisonF
3
negationMultiplier
-2
backwardComparisonF
2
Параметри аналізу
negationFalloffStart
-2
falloffStart
-2
extremization
2
negationFalloffEnd
5
falloffEnd
5
shingleSize
3
Панель для виводу результатів роботи

| | | | | | |
|---|---|---|---|---|---|
| negationMultiplier | -2 | backwardComparison | 2 | | |
| negationFalloffStart | -2 | falloffStart | -2 | extremization | 2 |
| negationFalloffEnd | 5 | falloffEnd | 5 | shingleSize | 3 |

Validation completed. Correctness: Value: 78.56%.

Examples of incorrectly rated reviews:

Positiveness: -0.0009748750357923552

captivating one of king's best so far i was litteraly sucked into jessie's horrifying world king kept me up all night i just couldn't put the book down

Positiveness: -0.19141426927818445