ДОДАТОК А

Перелік джерел посилання за науковими напрямами керівника та науковців кафедри програмної інженерії

10. Models of adaptive integration of weighted interval data in tasks of predictive expert assessment / I. Ruban et al. Eastern-European Journal of Enterprise Technologies. 2022. Vol. 5, no. 4(119). P. 6–15. URL: https://doi.org/10.15587/1729-4061.2022.265782 (дата звернення 19.04.2024).

11. Hacimahmud A. V., Khakhanova H., Litvinova E. Vector Logic Analysis of Big Data. 2023 IEEE East-West Design & Test Symposium (EWDTS), Batumi, Georgia, 22–25 September 2023. 2023. URL: https://doi.org/10.1109/ewdts59469.2023.10297032 (дата звернення 19.04.2024).

12. Evaluation and Analysis of the NLP Model Zoo for Ukrainian Text Classification / D. Panchenko et al. Information and Communication Technologies in Education, Research, and Industrial Applications. Cham, 2022. P. 109–123. URL: https://doi.org/10.1007/978-3-031-20834-8_6 (дата звернення 19.04.2024).

ДОДАТОК Б

Слайди презентації

1

Дослідження методів аналізу сентименту на основі неструктурованих текстових відгуків

Виконав: ст. гр. ІПЗм-

Науковий керівник:

2

Дослідження

- **Актуальність:** Дослідження методів аналізу сентименту на основі неструктурованих текстових відгуків є важливим для розуміння громадської думки та автоматизації обробки великого обсягу текстової інформації в різних галузях.
- **Визначення напряму дослідження:** Вивчення ефективності та точності різних алгоритмів і підходів до аналізу сентименту в неструктурованих текстових відгуках.

3

Постановка задачі

- Огляд методів: Провести огляд існуючих методів аналізу сентименту та визначити їх переваги і недоліки.
- Критерії оцінки: Встановити критерії для оцінки ефективності методів аналізу сентименту.
- Підготовка даних: Зібрати та підготувати неструктуровані текстові відгуки для аналізу.
- Моделювання: Розробити та реалізувати моделі аналізу сентименту, використовуючи різні підходи.
- Аналіз результатів: Порівняти продуктивність моделей та виявити фактори, що впливають на точність.
- Аналіз можливостей покращення: Проаналізувати можливості розробки і валідації моделі яка ефективно комбінує результати підходів для підвищення точності виявлення сентименту у текстах.

4

Технології дослідження

- **Python.**
- **VADER (Valence Aware Dictionary and sEntiment Reasoner):** VADER - це лексикон і правило-базований метод аналізу сентименту, оптимізований для соціальних мереж, який визначає полярність (позитивний, негативний, нейтральний) і інтенсивність (наскільки сильно) сентименту в тексті.
- **TextBlob:** TextBlob - це проста у використанні бібліотека для обробки природної мови в Python, яка надає інструменти для різних завдань NLP, включаючи аналіз сентименту на основі правила-базованого підходу, який повертає полярність і суб'єктивність тексту.

5

Вибір методів для реалізації і дослідження

- Прийнято рішення зосередитись на методах, які дозволяють досягти гарних результатів при менших затратах та спрощують процес інтеграції та використання в різних проектах.
- Отже надалі буде порівняно підходи більш прості в реалізації, а саме підхід машинного навчання та лінгвістичний метод.
- Ми порівнюємо два підходи(машинне навчання та лінгвістичні моделі) на двох бібліотеках які їх реалізують vader vs textblop.Варто пояснити що ці бібліотеки є провідними бібліотеками з відкритим вихідним кодом для свого метода на даний час, тож ефективність методів в рамках дослідження можна прив'язати до ефективності бібліотек.

6

Порівняння бібліотек VADER і TextBlob

VADER і TextBlob є одними з найвідоміших та найбільш використовуваних бібліотек для аналізу сентименту з відкритим кодом. Вони представляють два різних підходи до розуміння тексту:

VADER (Valence Aware Dictionary and sEntiment Reasoner) є спеціалізованою бібліотекою, яка розроблена для роботи з соціальними медіа та іншими веб-текстами, де часто використовується неформальна мова, сленг та емодзі. Це лінгвістична модель, яка використовує набір заздалегідь визначених правил для аналізу сентименту.

TextBlob пропонує підхід, заснований на машинному навчанні, з використанням класифікаційних та регресійних алгоритмів. Ця бібліотека більш універсальна та придатна для широкого спектра застосувань від простого аналізу сентименту до складних задач обробки природної мови.

7

Тестовий датасет

Для аналізу, представленого на графіку у зображенні на наступному слайді, було використано датасет відгуків про готелі, який містить думки клієнтів, які зупинялися в готелі.

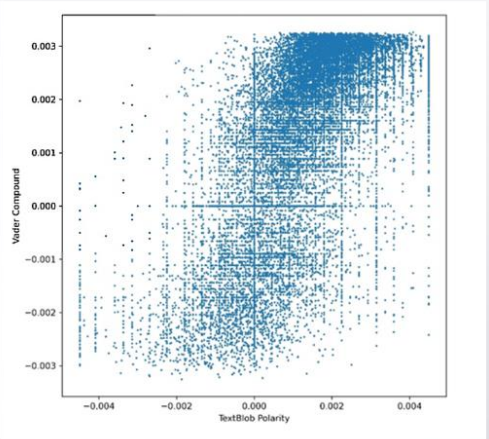
Цей датасет включав текстові описи відгуків, які були аналізовані за допомогою бібліотек TextBlob та VADER для визначення сентименту цих текстів. Кожен відгук був оброблений для отримання оцінок полярності (від -1 до +1, де -1 означає негативний сентимент, а +1 — позитивний) та суб'єктивності (від 0 до 1, де 0 означає об'єктивний текст, а 1 — суб'єктивний).

Review_ID	Review_Text	Polarity	Subjectivity
1	"The hotel was fantastic! The staff were friendly and the rooms were very clean. I would definitely stay here again."	0.8	0.75
2	"Terrible experience. The room was dirty and the service was slow. Not worth the money."	-0.9	0.8
3	"Average stay. The location was convenient, but the amenities were lacking."	0.1	0.5
4	"Loved the breakfast! The variety of food was great and everything was fresh."	0.7	0.6
5	"The hotel was okay. Nothing special, but nothing particularly bad either."	0.2	0.4

8

Результат випробування методів на датасеті

Діаграма розсіювання вище ілюструє коефіцієнт кореляції Пірсона між VADER та TextBlob. Аналізуючи графік, бачимо, що хоча VADER класифікував певні пропозиції як негативні, TextBlob ідентифікував їх переважно як позитивні. У першому та третьому квадрантах обидва алгоритми показують згоду. Однак у другому та четвертому квадранті спостерігається помітна неузгодженість, особливо у четвертому квадранті, де переважають суперечливі дані. TextBlob позначає пропозиції як позитивні, тоді як VADER класифікує їх як негативні.



9

Створення даних для експерименту

Тепер проведемо експеремент на реальних датасетах які будуть відображати різні характеристики неструктурованих текстових відгуків.

Створимо список із 3 речень, у кожному з яких буде одне позитивне, одне негативне та одне нейтральне речення. І тому ми можемо побачити, як працюють VADER і TextBlob на кожному з них.

```
text_list = ["This is my first ever post on the internet.",
             "I am very excited to write this post.",
             "It's not good to work late hours."]
```

10

Проведення експерименту на звичайному тексті

```
sentence: This is my first ever post on the internet.
 VADER sentiment score: 0.0
 TextBlob score: 0.25
==============================
sentence: I am very excited to write this post.
 VADER sentiment score: 0.4005
 TextBlob score: 0.48750000000000004
==============================
sentence: It's not good to work late hours.
 VADER sentiment score: -0.3412
 TextBlob score: -0.32499999999999996
==============================
```

З рисунку вище ми можемо зробити висновок, що VADER ідеально визначив перше речення як нейтральне речення, де TextBlob не такий вже й далекий від нього.

Тоді для другого речення VADER дає позитивну оцінку, але TextBlob дає нам більш позитивну оцінку. І для останнього речення VADER дає більш негативний бал, ніж TextBlob.

11

Проведення експерименту з доданням знаків оклику в тексті

```
sentence: This is my first ever post on the internet!
 VADER sentiment score: 0.0
 TextBlob score: 0.3125
==============================
sentence: I am very excited to write this post!
 VADER sentiment score: 0.4561
 TextBlob score: 0.609375
==============================
sentence: It's not good to work late hours!
 VADER sentiment score: -0.4015
 TextBlob score: -0.3625
==============================
```

Тепер, з наведеної вище комірки, ми можемо сказати, що знак оклику справді покращує нашу оцінку в усіх реченнях. Але для нашого нейтрального речення TextBlob відривається.

12

Проведення експерименту з доданням капіталізації в тексті

```
sentence: This is my FIRST EVER post on the internet!
 VADER sentiment score: 0.0
 TextBlob score: 0.3125
==============================
sentence: I am very EXCITED to write this post!
 VADER sentiment score: 0.5744
 TextBlob score: 0.609375
==============================
sentence: It's NOT GOOD to work late hours!
 VADER sentiment score: -0.5007
 TextBlob score: -0.3625
==============================
```

Тепер ми можемо сказати, що наша оцінка VADER покращилася, але оцінка TextBlob залишилася незмінною.

Причина в тому, що VADER вважає, що версія з великими літерами має сильніший настрій і збільшила оцінку настрою. У той же час TextBlob не розрізняв настрої між верхнім і нижнім регістром слова.

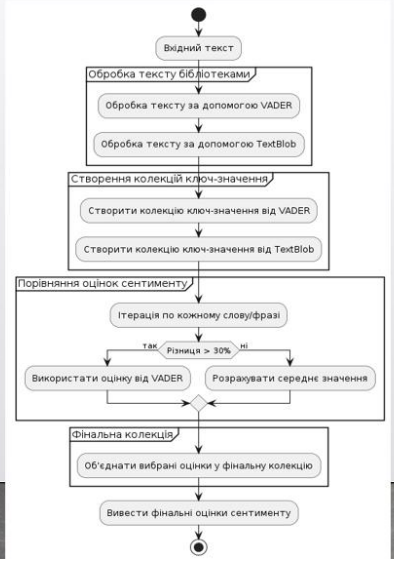
13

Розробка алгоритма поєднання методів

Спочтку ми зберемо дані з однакового тексту та проженем по двох цих бібліотеках, після цього ми створемо з двох результатів - дві колекції ключ-значення де значенням буде оцінка кожного слова кожною відповідною бібліотекою, після цього ми пройдемось по кожному слову(в обох колекціях) порівнюючи одні й ті самі елементи та відберем такі - різниця між якими - достатньо велика(30%), так само ми зробимо із комбінаціями слів (комбінація токенів або фраз), після цього оскільки ми вже з'ясували що Vader краще з'ясовує емоційну забарвленість - ми замінимо в фінальній спільній коллекції(яку зробимо шляхом поєднання двох де перерахуємо ключі всим словам) такі слова(комбінація токенів або фраз) із TextBlob на слова із VADER, а всі інші слова де різниця менше 30% - візьмемо середне між двома колекціями.

14

Розробка алгоритма поєднання методів

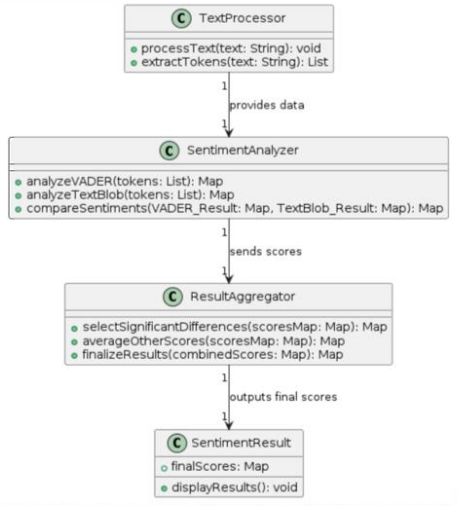


Наш підхід до поєднання результатів аналізу сентименту від VADER та TextBlob досить інноваційний та зосереджений на усуненні відмінностей у враженнях сентименту між цими двома бібліотеками.

Такий підхід може допомогти максимально використати сильні сторони кожної бібліотеки, забезпечуючи більш точне та надійне розуміння емоційного забарвлення тексту.

15

Розробка системи класів



Ця система класів спрямована для гнучкого та ефективного аналізу сентименту, де кожен компонент виконує свою специфічну задачу, але разом вони формують інтегровану систему аналізу сентименту.

16

Приклад фрагментів коду основного класу поєднання

```
class SentimentAnalyzer:

  def __init__(self):

    self.vader_analyzer = SentimentIntensityAnalyzer()


  def analyze_with_vader(self, tokens):

    scores = {}

    for token in tokens:

      score = self.vader_analyzer.polarity_scores(token)['compound']

      scores[token] = score

    return scores


  def analyze_with_textblob(self, tokens):

    scores = {}

    for token in tokens:

      score = TextBlob(token).sentiment.polarity

      # Normalize TextBlob scores to match VADER's scale

      normalized_score = (score + 1) / 2

      scores[token] = normalized_score

    return scores
```

```
class ResultAggregator:

  def combine_results(self, vader_scores, textblob_scores):

    final_scores = {}

    for token in vader_scores:

      if abs(vader_scores[token] - textblob_scores.get(token, 0)) > 0.3:

        final_scores[token] = vader_scores[token] if vader_scores[token] >
textblob_scores.get(token, 0) else textblob_scores[token]

      else:

        final_scores[token] = (vader_scores[token] + textblob_scores.get(token, 0)) / 2

    return final_scores

class SentimentAnalysisSystem:

  def __init__(self, text):

    self.text = text

    self.processor = TextProcessor()

    self.analyzer = SentimentAnalyzer()

    self.aggregator = ResultAggregator()

  def perform_analysis(self):

    tokens = self.processor.tokenize_text(self.text)

    vader_scores = self.analyzer.analyze_with_vader(tokens)

    textblob_scores = self.analyzer.analyze_with_textblob(tokens)

    final_scores = self.aggregator.combine_results(vader_scores, textblob_scores)

    return final_scores
```

Опис коду:
textprocessor: токенізує текст для подальшого аналізу;
sentimentanalyzer: аналізує токени, використовуючи vader та textblob, і повертає словник з оцінками сентименту;
resultaggregator: об'єднує результати від двох аналізаторів. якщо різниця оцінок більше 30%, вибирається більша оцінка. в інших випадках обраховується середнє;
sentimentanalysissystem: керує процесом аналізу від початку до кінця, від токенізації тексту до отримання кінцевих оцінок сентименту.

Висновки

•Результати аналізу: Інтеграція результатів VADER і TextBlob підвищує точність у порівнянні з окремим використанням кожної бібліотеки.

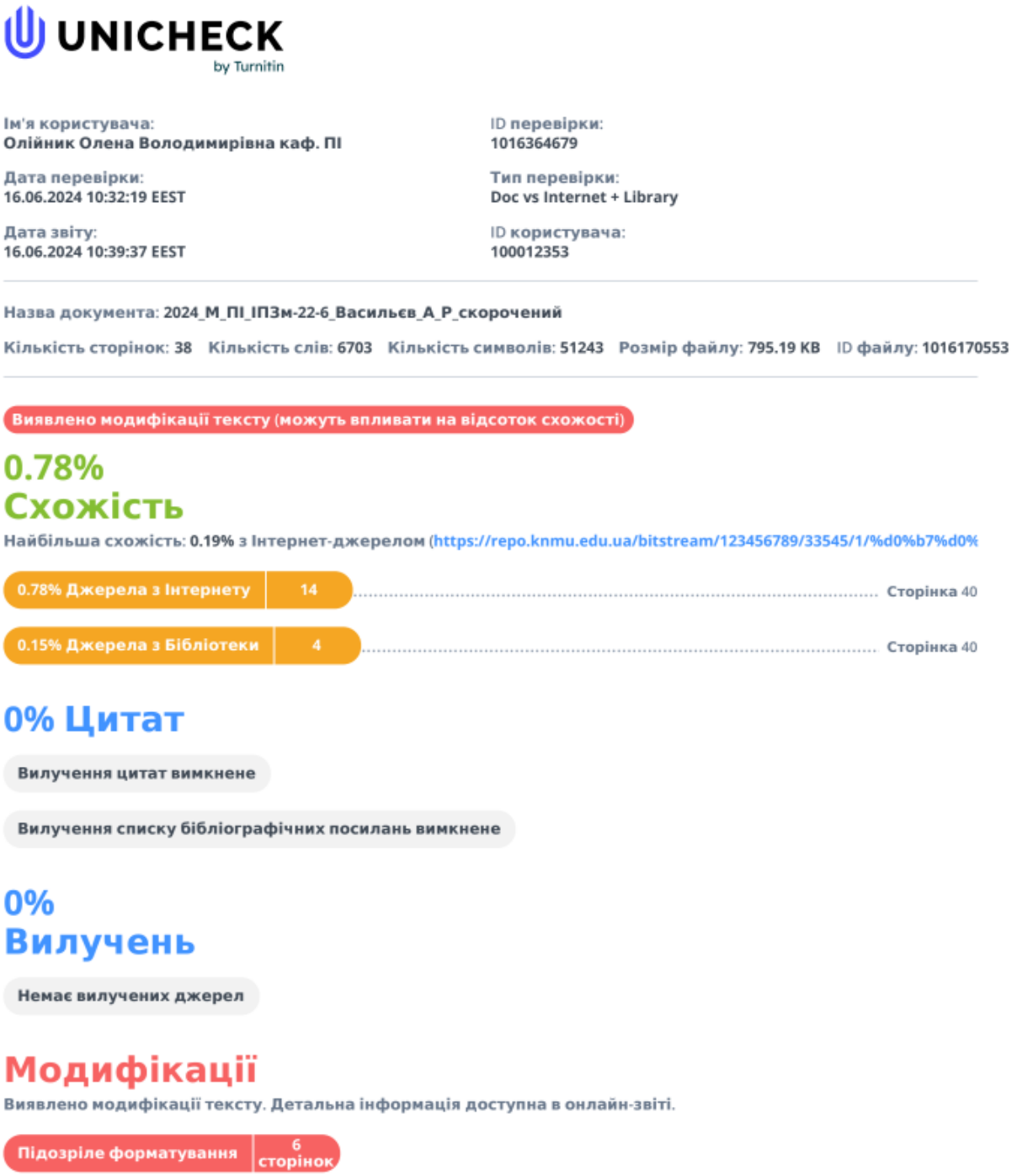
•Особливості бібліотек:
•VADER краще виявляє емоційну забарвленість текстів з іронією та сленгом.
•TextBlob показує кращі результати у стандартних висловлюваннях.

•Переваги інтеграції: Поєднання сильних сторін обох бібліотек дозволяє адекватно обробляти ширший спектр текстових форматів.

•Практичне застосування: Модель корисна для швидкого та точного розуміння сентименту в маркетингу, аналізах соціальних медіа, підтримці клієнтів та виробничих системах.

ДОДАТОК В

Результат проходження на академічний плагіат



UNICHECK
by Turnitin

Ім'я користувача:
Олійник Олена Володимирівна каф. ПІ

ID перевірки:
1016364679

Дата перевірки:
16.06.2024 10:32:19 EEST

Тип перевірки:
Doc vs Internet + Library

Дата звіту:
16.06.2024 10:39:37 EEST

ID користувача:
100012353

Назва документа: 2024_М_ПІ_ІПЗм-22-6_Васильєв_А_Р_скорочений

Кількість сторінок: 38 Кількість слів: 6703 Кількість символів: 51243 Розмір файлу: 795.19 KB ID файлу: 1016170553

Виявлено модифікації тексту (можуть впливати на відсоток схожості)

0.78%
Схожість

Найбільша схожість: 0.19% з Інтернет-джерелом (https://repo.knmu.edu.ua/bitstream/123456789/33545/1/%d0%b7%d0%

0.78% Джерела з Інтернету 14 ..... Сторінка 40

0.15% Джерела з Бібліотеки 4 ..... Сторінка 40

0% Цитат

Вилучення цитат вимкнене

Вилучення списку бібліографічних посилань вимкнене

0%
Вилучень

Немає вилучених джерел

Модифікації

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

Підозріле форматування 6 сторінок

ДОДАТОК Г

Апробація результатів роботи

ДОДАТОК Д

Експертний висновок результатів перевірки кваліфікаційної роботи на відповідність оформлення вимогам ДСТУ 3008:2015

Експертний висновок результатів перевірки кваліфікаційної роботи

студент (посада) | програмної інженерії (кафедра) | ІПЗм-22-6 (група)

Васильєв Артем Ростиславович
(прізвище, ім'я, по батькові)

Зауваження

Пункт ДСТУ 3008-2015	Зміст пункту	Сторінка кваліфікаційної роботи
1	2	3
	7.1 Загальні положення	
	7.3 Нумерація сторінок звіту	
	7.4 Нумерація розділів, підрозділів, пунктів, підпунктів	
	7.5 Рисунки	
	7.6 Таблиці	
	7.7 Переліки	
	7.8 Примітки	
	7.9 Виноски	
	7.10 Формули та рівняння	
	7.11 Посилання	
	7.13 Список авторів	
	7.14 Скорочення та умовні познаки	
	7.15 Додатки	

зауважень немає

Експерт _____ (підпис) Олена ОЛІЙНИК (прізвище, ініціали)

16.06.2024