Laboratorium 2: Zrozumienie danych

Michał Binda, Mikołaj Mróz

7 listopada 2024

Propozycja rozwiązania: ReviewRadar - analiza sentymentu

Spis treści

Tabela rewizji

Wersja	Data	Opis zmian	Odpowiedzialny
1.0	28.10.2024	- Zrobiona pierwsza wersja dokumentu: - Legenda dostępnych danych - Wprowadzenie - Ocena narzędzi	Michał Binda, Mikołaj Mróz
1.1	03.11.2024	- Raporty i wykresy dot. danych - Opis modułów aplikacji - Napisano sekcję "Propozycja rozwiązania" - Dodano diagram architektury systemu	Michał Binda, Mikołaj Mróz
1.2	04.11.2024	- Rozszerzenie oceny narzędzi i technik - Dodanie projektu GUI	Michał Binda, Mikołaj Mróz

1 Wprowadzenie

Aplikacja ReviewRadar jest narzędziem do analizy sentymentu komentarzy użytkowników aplikacji mobilnych dostępnych na platformie Google Play. Celem dokumentu jest przedstawienie wstępnych założeń oraz propozycji narzędzi i technik, które zostaną użyte w analizie danych i modelowaniu sentymentu.

2 Propozycja rozwiązania

Dokładny dobór narzędzi i technik ma kluczowe znaczenie dla jakości analizy oraz szybkości działania aplikacji. Niniejsza sekcja przedstawia wybrane technologie oraz algorytmy, które będą stosowane w analizie komentarzy użytkowników.

2.1 Ocena narzędzi i technik

- **Języki programowania**:
  - **Python**: główny język do analizy danych, przetwarzania tekstu i implementacji narzędzi służących do analizy sentymentu oraz budowy interfejsu użytkownika.
  - **SQL**: używany do zapytań i przetwarzania danych w bazie PostgreSQL, umożliwiający szybkie filtrowanie i agregowanie danych.
- **psycopg2**: Biblioteka Python umożliwiająca łączenie się z bazą danych PostgreSQL oraz wykonywanie zapytań SQL. Dzięki `psycopg2` można dynamicznie pobierać, aktualizować i zapisywać dane bezpośrednio z poziomu skryptów w Pythonie.
- **google-play-scraper**: Pakiet używany do automatycznego pobierania recenzji oraz szczegółowych danych o aplikacjach z Google Play. Umożliwia gromadzenie danych, takich jak opinie użytkowników, oceny oraz dodatkowe informacje o wersji aplikacji, które następnie mogą być wykorzystane do analizy sentymentu oraz innych statystyk.
- **Streamlit**: Biblioteka Python służąca do szybkiego budowania interaktywnych aplikacji webowych. Streamlit pozwala na wizualizację i analizowanie wyników analizy sentymentu w przejrzystym interfejsie użytkownika, umożliwiając również interakcje, takie jak filtrowanie danych według oceny, języka recenzji czy daty.
- **Baza danych PostgreSQL**: Wykorzystujemy bazę danych PostgreSQL do przechowywania danych pobranych z Google Play. PostgreSQL jest wysoce skalowalną, bezpieczną bazą danych o otwartym kodzie źródłowym, która wspiera automatyczne kopie zapasowe oraz ułatwia zarządzanie i zwiększa dostępność danych. Chociaż jesteśmy w stanie uruchomić instancję PostgreSQL na platformie Google Cloud, obecnie nie posiadamy wystarczających środków na utrzymanie jej przez kilka miesięcy, więc rozważymy jej uruchomienie w końcowej fazie projektu. W tej chwili baza danych jest uruchomiona lokalnie na komputerze użytkownika. Narzędzie `psycopg2` umożliwia bezpośrednią komunikację z lokalną bazą PostgreSQL w celu wykonywania operacji na danych.
- **Analiza sentymentu**:

- **Natural Language Toolkit (NLTK)**: Popularna biblioteka do NLP z narzędziem Vader do analizy sentymentu w tekstach z mediów społecznościowych.
- **TextBlob**: Prosta w użyciu biblioteka NLP z wbudowaną analizą sentymentu i oceną polaryzacji tekstu.
- **spaCy**: Szybka biblioteka NLP, wspierająca analizę sentymentu oraz zaawansowane zadania, jak rozpoznawanie nazw własnych.
- **Transformers**: Biblioteka od Hugging Face z modelami jak BERT i GPT-2, zapewniająca zaawansowaną analizę sentymentu.
- **VADER Sentiment Analysis**: Narzędzie leksykonowe do analizy sentymentu, idealne do krótkich tekstów w mediach społecznościowych.

2.2 Założenia technik modelowania

Podczas modelowania sentymentu wykorzystane zostaną poniższe założenia:

- Analiza sentymentu obejmuje trzy klasy: pozytywny, neutralny, negatywny.
- Komentarze będą tokenizowane i przetwarzane w celu identyfikacji słów kluczowych.
- Do modelowania zostaną użyte wybrane modele, które zostaną porównane pod kątem skuteczności i wydajności.

3 Raporty danych

3.1 Legenda

W danych występują następujące kolumny:

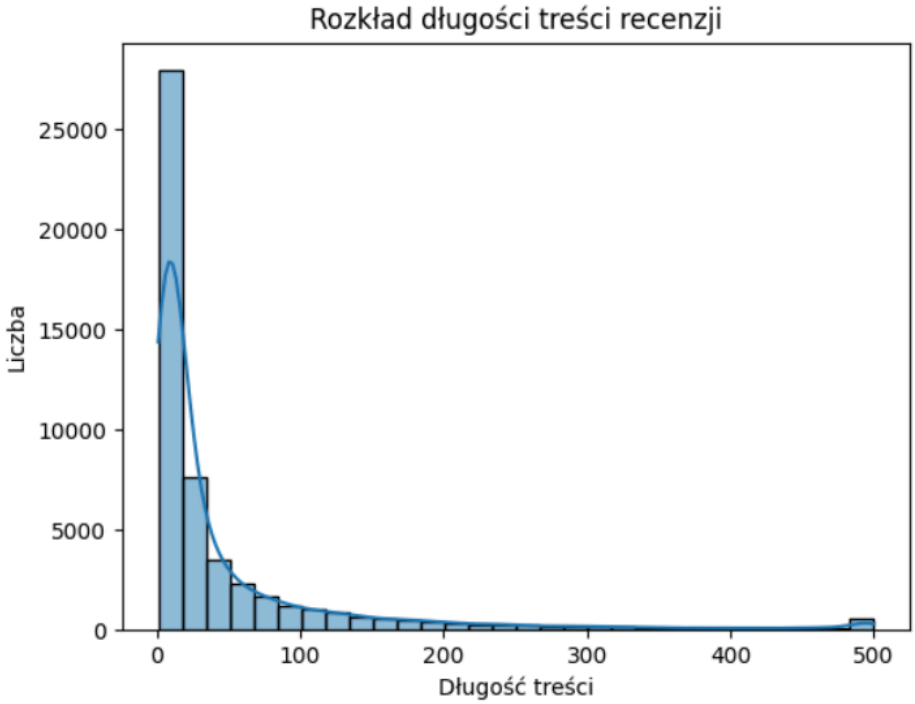
- **reviewId** - Unikalny identyfikator recenzji.
- **userName** - Nazwa użytkownika, który opublikował recenzję.
- **userImage** - URL do zdjęcia profilowego użytkownika.
- **content** - Treść recenzji, napisana przez użytkownika.
- **score** - Ocena przypisana przez użytkownika (w skali od 1 do 5).
- **thumbsUpCount** - Liczba polubień (thumbs up) dla danej recenzji.
- **reviewCreatedVersion** - Wersja aplikacji, w której utworzono recenzję.
- **at** - Data i czas publikacji recenzji w formacie `YYYY-MM-DD HH:MM:SS`.
- **replyContent** - Treść odpowiedzi aplikacji na recenzję użytkownika (jeśli istnieje).
- **repliedAt** - Data i czas odpowiedzi na recenzję, jeśli odpowiedź istnieje.
- **appVersion** - Wersja aplikacji, której dotyczy recenzja.
- **appName** - Nazwa aplikacji, której dotyczy recenzja.

3.2 Tabela Informacyjna

Tabela 1: Tabela z liczbą unikalnych wartości, typem danych oraz procentem braków w każdej kolumnie.

Kolumna	**Typ danych**	**Liczba unikalnych wartości**	**Procent braków**
reviewId	string	51 740	0%
userName	string	48 863	0%
userImage	string	50 928	0%
content	string	37 129	0,77%
score	int64	5	0%
thumbsUpCount	int64	249	0%
reviewCreatedVersion	string	1 573	20,02%
at	string	48 684	0%
replyContent	string	132	99,71%
repliedAt	string	152	99,71%
appVersion	string	1 573	20,02%
appName	string	10	0%

3.3 Wykresy Analizy Danych

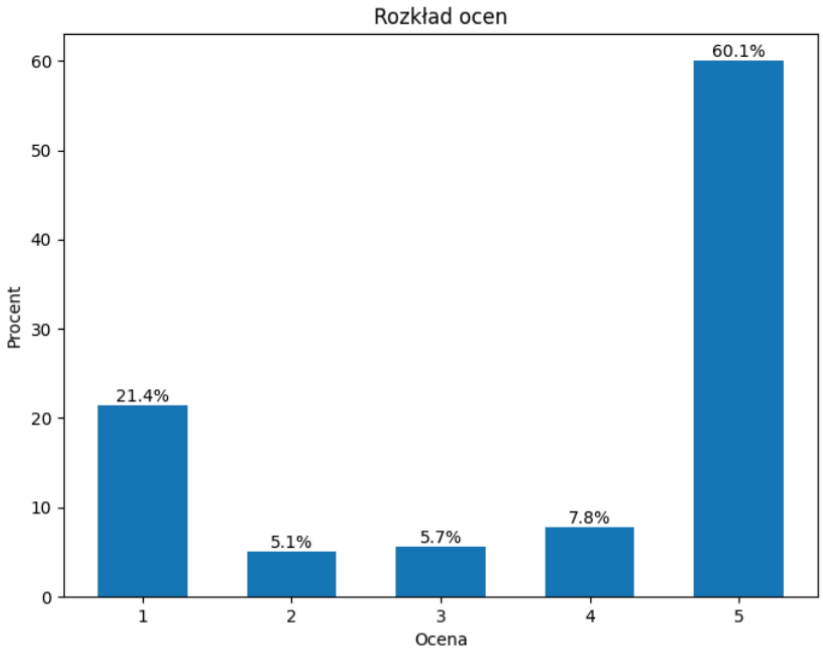


Rysunek 1: Rozkład długości treści recenzji

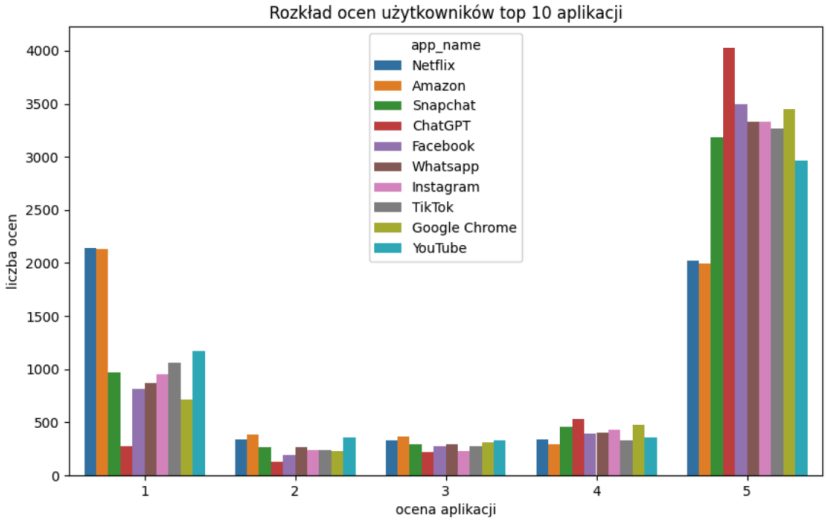
- **Średnia długość recenzji**: 51,40 znaków.
- **Maksymalna długość recenzji**: 500 znaków.

Tabela 2: Najczęściej występujące treści w kolumnie `content`.

Treść	**Procentowy udział (%)**
Good	4,90
Nice	2,08
good	1,66
nice	0,79
Very good	0,68
Ok	0,62
Good app	0,55
Excellent	0,54
Nice app	0,51
Best	0,44



Rysunek 2: Rozkład ocen



Rysunek 3: Rozkład ocen użytkowników dla top 10 aplikacji

3.4 Podsumowanie

Analiza danych recenzji aplikacji ujawnia kilka kluczowych obserwacji:

- **Liczba kolumn**: 12.
- **Typy danych kolumn**: 7 kolumn tekstowych (string), 3 kolumny numeryczne (int64), 2 kolumny daty i czasu (string).
- **Długość recenzji:** Średnia długość wynosi 51.4 znaki, z maksymalną długością 500 znaków. Większość recenzji jest krótka, co potwierdza histogram.
- **Rozkład ocen:** Dominują oceny skrajne, głównie 1 oraz 5, co sugeruje, że użytkownicy oceniają aplikacje głównie w przypadku skrajnych doświadczeń.
- **Najczęstsze treści recenzji:** Krótkie, pozytywne wyrażenia, takie jak Good i Nice, przeważają, wskazując na zwięzłość w ocenach.
- **Jakość danych:** Braki danych występują głównie w kolumnach replyContent i repliedAt (99.71%), oraz reviewCreatedVersion i appVersion (20.02%). Kolumny te mogą wymagać dodatkowego przetwarzania lub uzupełnienia dla pełniejszej analizy.

4 Opcjonalne komponenty

4.1 Opis modułów aplikacji

Aplikacja ReviewRadar została podzielona na kilka modułów, które umożliwiają jej sprawne działanie i łatwość w dalszym rozwoju:

- **Moduł pobierania danych** – Ten moduł łączy się z Google Play API, pobiera recenzje użytkowników oraz zapisuje je do lokalnej bazy danych, umożliwiając ich przetwarzanie w kolejnych krokach.

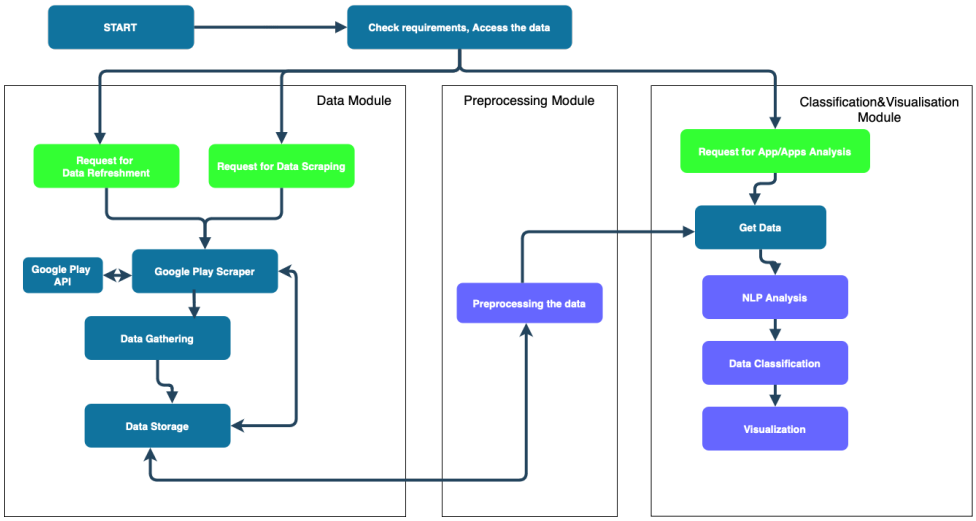
- **Moduł preprocessingu** – Odpowiada za przetwarzanie językowe, takie jak tokenizacja, lematyzacja oraz usuwanie stop-słów, co pozwala na wyodrębnienie kluczowych słów i fraz.
- **Moduł klasyfikacji i wizualizacji** Analizuje sentyment komentarzy przy pomocy wybranych modeli NLP. Klasyfikuje teksty do odpowiedniej grupy (negatywne, neutralne, pozytywne). Generuje raporty i wykresy, przedstawiając kluczowe wskaźniki aplikacji, takie jak rozkład ocen, wykresy sentymentu oraz najczęściej występujące słowa kluczowe.

4.1.1 Moduł danych

Moduł danych jest odpowiedzialny za pozyskiwanie i aktualizację danych niezbędnych do analizy aplikacji. Składa się z kilku kluczowych komponentów:

- **Request for Data Refreshment** – Funkcja, która dla każdej aplikacji sprawdza ostatnią datę, dla której są widoczne komentarze, a następnie ściąga wszystkie nowe dane począwszy od tej daty.
- **Request for Data Scraping** – Funkcja służąca do pobrania nowych danych w podanym przez użytkownika przedziale czasowym. Funkcja ta pobiera tylko dane, których nie ma jeszcze w bazie danych (na podstawie klucza reviewId)
- **Data Gathering** – Zajmuje się gromadzeniem pobranych danych, które następnie przechowywane są w systemie.
- **Storage Collect** – Przechowuje wszystkie pozyskane dane, przygotowując je do kolejnych etapów analizy.

4.2 Architektura Systemu



Rysunek 4: Proponowany model systemu

4.2.1 Moduł preprocessingu

Moduł wstępnego przetwarzania danych odpowiada za przygotowanie danych do analizy sentymentu i klasyfikacji. Dzięki temu etapy analizy są bardziej precyzyjne i efektywne.

- **Data Storage** – Magazynuje zgromadzone dane, które będą wykorzystywane w procesie przetwarzania.
- **Preprocessing the Data** – Moduł, który zajmuje się czyszczeniem, standaryzacją i transformacją danych. Usuwamy zbędne informacje i odpowiednio przetwarzamy teksty, aby były gotowe do analizy sentymentu.

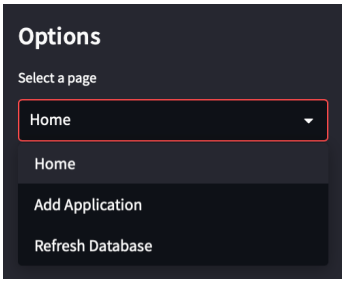
4.2.2 Moduł klasyfikacji i wizualizacji

Moduł klasyfikacji i wizualizacji odpowiada za analizę danych, ich klasyfikację oraz wizualizację wyników, co umożliwia łatwe zrozumienie wyników przez użytkownika.

- **Request for App/Apps Analysis** – Obsługuje zapytania użytkowników dotyczące analizy wybranych aplikacji lub zestawu aplikacji.
- **Get Data** – Pobiera dane potrzebne do analizy na podstawie żądania użytkownika.
- **Data Labeling** – Klasyfikuje dane, przypisując im etykiety na podstawie analizy sentymentu. Każdy komentarz zostaje oznaczony jako pozytywny, negatywny lub neutralny.
- **Visualization** – Tworzy graficzne przedstawienie wyników analizy, co pozwala użytkownikowi szybko zrozumieć ogólny sentyment i nastroje użytkowników wobec analizowanych aplikacji.

4.3 Projekt GUI

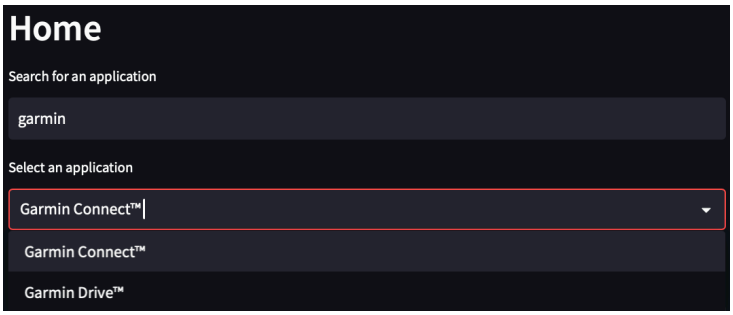
Interfejs użytkownika aplikacji ReviewRadar zaprojektowany został z myślą o intuicyjności i wygodzie użytkownika, zapewniając szybki dostęp do najważniejszych funkcji. Aplikacja zostanie podzielona na kilka stron. W poniższym rozdziale pokazane są wstępne wizualizacje interfejsu użytkownika, a nie docelowy produkt. Zaprezentowane strony pokazują moduł pobierania danych, a jeszcze nie moduł wizualizacji i analizy danych.



Rysunek 5: Menu Aplikacji Radar Review

4.3.1 Ekran główny - Home

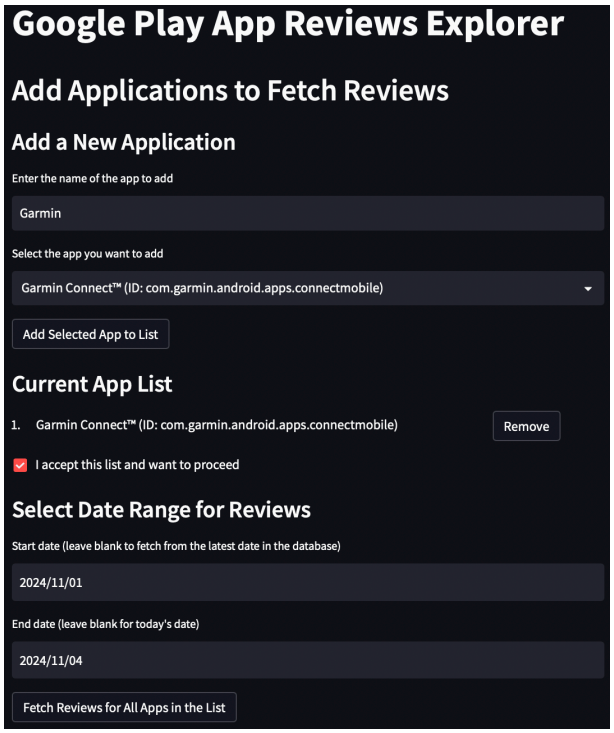
Umożliwia użytkownikowi wyszukiwanie aplikacji poprzez nazwę lub identyfikator oraz wizualizuje statystyki i analize sentymentu opinii.



Rysunek 6: Pierwsza strona Aplikacji Radar Review

4.3.2 Ekran do pobierania nowych Aplikacji

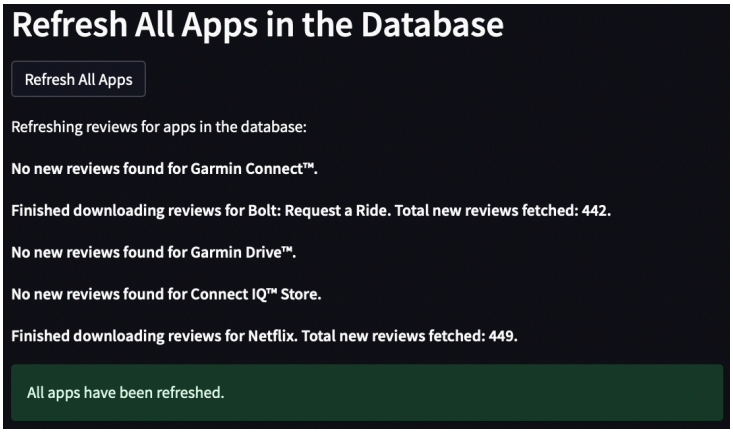
W przypadku gdy Aplikacja nie jest jeszcze w bazie danych, użytkownik może ręcznie pobrać dane wybierając aplikację, bądź listę aplikacji, następnie przedział czasowy dla jakiego chce pobrać dane, a na koniec potwierdza swój wybór:



Rysunek 7: Druga strona Aplikacji Radar Review

4.3.3 Ekran do uaktualniania danych

W przypadku gdy użytkownik chce się upewnić, że dane są aktualne użytkownik może ręcznie zaaktualizować bazę danych aplikacji:



Rysunek 8: Trzecia strona Aplikacji Radar Review