Raport z projektu

Tytuł projektu: Analiza sentymentu recenzji

Autor: **Krystian Jachna**

Cel projektu:

Implementacja binarnego klasyfikatora analizy sentymentu recenzji, który ma na celu określenie, czy recenzje są negatywne czy pozytywne.

Realizacja projektu:

a) Znalezienie danych do wytrenowania modelu
b) Analiza danych
c) Stworzenie pipeline'u przygotowującego dane do uczenia
d) Eksperymenty z różnymi klasyfikatorami i dopasowanie hiperparametrów
e) Wytrenowanie i ewaluacja modelu
f) Stworzenie prostego narzędzia uruchomieniowego

Dane treningowe:

Dane pochodzą z recenzji Amazon i są używane do analizy sentymentów. Składają się z 3,000,000 próbek treningowych i 650,000 próbek testowych. Recenzje są oznaczone liczbą od 1 do 5, reprezentującą liczbę gwiazdek przyznanych przez recenzenta. Surowe dane zawierają trzy kolumny: indeks klasy, tytuł recenzji, treść recenzji.

Podczas analizy danych, sprawdzone zostały brakujące wartości, rozkład klas (test, train), usunięto dane zaklasyfikowane jako 3, a pozostałe zostały podzielone na pozytywne (dla klasy 4 i 5) i negatywne (dla klasy 1, 2) zgodnie z opisem autora.

Przetwarzanie wstępne danych

Preprocessing pipeline składa się z następujących kroków:

1. Czyszczenie danych:
   - Konwersja tekstu na małe litery
   - Usunięcie „stopwords" (tj. słów bez znaczenia) , znaków interpunkcyjnych, URL-i, oznaczeń, emotikon i dodatkowych spacji

2. „Stemming":
   - Redukcja słów do ich podstawowej formy

3. Wektoryzacja:
    - Konwersja tekstu na macierz liczby tokenów z uwzględnieniem n-gramów (1, 2)
    - Ograniczenie liczby tokenów do najczęściej występujących

4. Transformacja TF-IDF:
    - Przekształcenie macierzy liczby tokenów do znormalizowanej reprezentacji TF-IDF

Wybór klasyfikatora

Aby wybrać odpowiedni klasyfikator, przeprowadzono naukę trzech modeli na części danych i porównano wyniki każdego z nich. Modele te to Logistic Regression, SVM oraz Random Forest. Wyniki porównania przedstawiono w tabeli:

Klasyfikator	Precyzja	Czułość	F1 Wynik	Dokładność	ROC AUC Wynik
Logistic Regression	0.84	0.84	0.84	0.84	0.92
SVM	0.69	0.85	0.76	0.73	0.83
Random Forest	0.78	0.83	0.80	0.80	0.89

Wybrano model Logistic Regression, ponieważ uzyskał najlepsze wyniki w praktycznie wszystkich kategoriach. Dodatkowo, Logistic Regression trenowała się najszybciej spośród wszystkich modeli.

Trenowani i uruchamianie modelu

Ostatecznie model został wytrenowany na pełnym zbiorze danych treningowych i przetestowany na zbiorze danych testowych. Model uzyskał 88% dla każdej z metryk: dokładność, precyzja, czułość i F1 wynik. Poniżej przedstawiono macierz pomyłek:

	Przewidywane Negatywne	Przewidywane Pozytywne
Faktycznie Negatywne	230476	29524
Faktycznie Pozytywne	28821	231179

W projekcie wykorzystano narzędzie Make, które umożliwia łatwe pobieranie danych, trenowanie modelu, ładowanie modelu i dokonywanie predykcji z konsoli, a także przy użyciu biblioteki Gradio z prostym interfejsem graficznym w przeglądarce.