Analiza danych niestrukturalnych
Kinga Szczepańczyk, Marzena Nowakowska, Wydział Zarządzania i Modelowania Komputerowego
Politechnika Świętokrzyska

Analiza sentymentu do opinii internautów na temat samochodów

Instrukcja na podstawie pracy magisterskiej Marty Lisińskiej pt. „Profilowanie opinii klientów o wybranym produkcie z wykorzystaniem analizy sentymentu"; opieka merytoryczna: M. Nowakowska

Zadanie 1.

Wykonać analizę sentymentu opinii internatów na temat samochodów wykorzystując modele transformerowe. Przeanalizować i wykorzystać program *S3-01_TRANSFORMERS.py*.
Źródłem danych jest plik *Dane_samochody-oczyszczone.xlsx*.
Plik wynikowy *Dane_samochody_sentyment.xlsx* zawiera wszystkie kolumny z pliku wejściowego oraz dodatkowo kolumny z wartościami sentymentu.
Sprawdzić wynik przetwarzania.

Uwaga 1. Rodzaje sentymentu

W modelu *nlptown/bert-base-multilingual-uncased-sentiment* etykiety są wyrażone jako liczba gwiazdek, np. "1 star", "2 stars", "3 stars", "4 stars", "5 stars". Dlatego w programie zastosowano dodatkową konwersję: 1–2 gwiazdki oznaczają sentyment negatywny, 3 gwiazdki sentyment neutralny, a 4–5 gwiazdek sentyment pozytywny. Model RoBERTa wykorzystywany w programie zwraca bezpośrednio klasy sentymentu podstawowego: "negative", "neutral", "positive".

Uwaga 2. Porównanie metody i opinii

W programie przewidziano porównanie wyników dla dwóch modeli oraz dwóch wersji tekstu opinii.
Modele
Model BERT wielojęzyczny dostarcza wartości emocji w postaci gwiazdek. Model RoBERTa jest dostrojony do krótkich tekstów internetowych i zwraca podstawowe klasy sentymentu. Porównanie tych metod pozwala sprawdzić, na ile wybór modelu wpływa na ocenę sentymentu.
Opinie
Porównanie dotyczy wersji tekstu: *Opinion* oraz *Opinion_norm*. Kolumna *Opinion* zawiera tekst pierwotny, natomiast *Opinion_norm* zawiera tekst po normalizacji, czyli po wcześniejszym oczyszczeniu i ujednoliceniu. Porównanie wyników na obu kolumnach pozwala sprawdzić, czy czyszczenie tekstu poprawia stabilność klasyfikacji, czy przeciwnie usuwa elementy ważne dla sentymentu, np. emocje, skróty lub wyrażenia wartościujące.
Wynikiem programu są cztery pary kolumn:

- *sentiment_BERT* oraz *score_BERT* - wynik modelu BERT dla opinii pierwotnej,
- *sentiment_norm_BERT* oraz *score_norm_BERT* - wynik modelu BERT dla opinii znormalizowanej,
- *sentiment_BERTa* oraz *score_BERTa* - wynik modelu RoBERTa dla opinii pierwotnej,
- *sentiment_norm_BERTa* oraz *score_norm_BERTa* - wynik modelu RoBERTa dla opinii znormalizowanej.

Zadanie 2.

Na podstawie programu *S3-01_TRANSFORMERS.py* przygotować nowy program *S3-02_TRANSFORMERS.py* , tak aby była wykonana analiza sentymentu dla wpisów w języku polskim zawartych w kolumnie *Opinia*. Źródłem danych jest *Dane_samochody_sentiment.xlsx.* Wyniki zapisać w pliku *Dane_samochody_sentymenty.xlsx*. Plik zawiera wszystkie kolumny z pliku wejściowego oraz dodatkowe kolumny:

- *sentimentPL_BERT* oraz *scorePL_BERT* - wynik modelu BERT dla opinii w języku polskim,
- *sentimentPL_BERTa* oraz *scorePL_BERTa* - wynik modelu RoBERTa dla opinii w języku polskim.

Sprawdzić wynik przetwarzania.

Zadanie 3.

Wykonać analizę porównawczą otrzymanych wyników analizy sentymentu, uwzględniają cocene ilościową internautów. Sformułować wnioski.