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

Obszary tematyczne w opiniach użytkowników o pojazdach

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

Źródłem danych do programów jest plik: *Dane_samochody_sentyment.xlsx*.

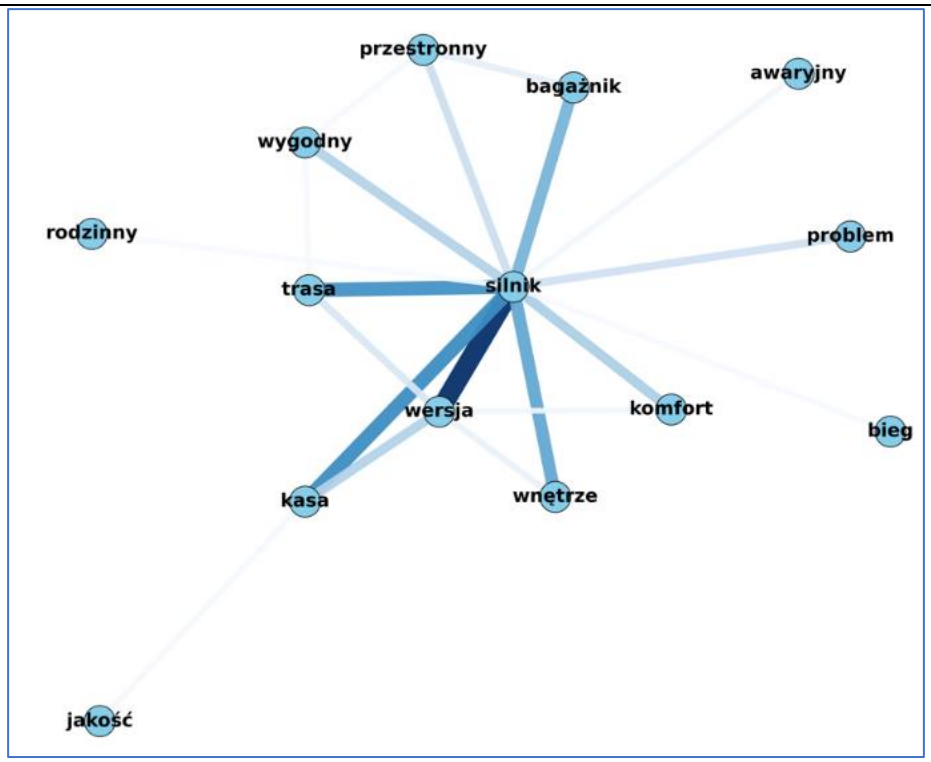
Zadanie 1.
Opracować charakterystykę poszczególnych sentymentów uzyskanych z opinii internatów w celu zrozumienia, jakie konkretnie aspekty były najczęściej komentowane oraz które elementy odnoszące się do samochodów wzbudzały poszczególne emocje. Przeprowadzić analizę opartą na ekstrakcji słów kluczowych. Słowa kluczowe pobierać z kolumn zawierających listy synonimów, np. ['renault', 'skoda', 'skoda', 'box'] z kolumny *Nouns_new*.
Dla każdego sentymentu wygenerować worek słów (*bag of words*) a następnie zwizualizować go w postaci chmury słów (*word cloud*) zwracając szczególną uwagę na najczęściej pojawiające się pojęcia w opiniach użytkowników – por. rys. 1. Ujednolicić hiperparametry (np. najmniejsza waga słowa w chmurze, liczba słów uwzględnionych w wizualizacji) dla wszystkich sentymentów.
Przeanalizować i wykorzystać program *S4-01_ChmrSlw.py*.

> Uwaga: zadecydować, które części mowy będą uwzględniane w ostatecznej charakterystyce sentymentów (metoda prób i błędów).



Rys. 1. Przykładowa chmura słów sentymentu pozytywnego.

Zadanie 2.
Dla każdego sentymentu wygenerować grafy współwystępowania, w celu zidentyfikowania, jakie pojęcia najczęściej pojawiają się razem w obrębie jednej opinii. Do utworzenia grafów współwystępowania wykorzystać ten sam zestaw wyrazów, które tworzą ostateczny worek słów.
Przykładowy graf dla sentymentu pozytywnego przedstawiono na rys. 2. Im ciemniejsza i grubsza linia łącząca dane pojęcia, tym silniejsze jest ich współwystępowanie, co oznacza, że dane słowa częściej pojawiały się razem.
Przeanalizować i wykorzystać program *S4-02_Grafy.py*.
Poprawić czytelność wykresu, wprowadzając większe litery do etykiet i lepszy kontrast miedzy węzłami i ich etykietami.



Rys. 2. Przykładowy graf współwystępowania słów sentymentu pozytywnego.

Zadanie 3.

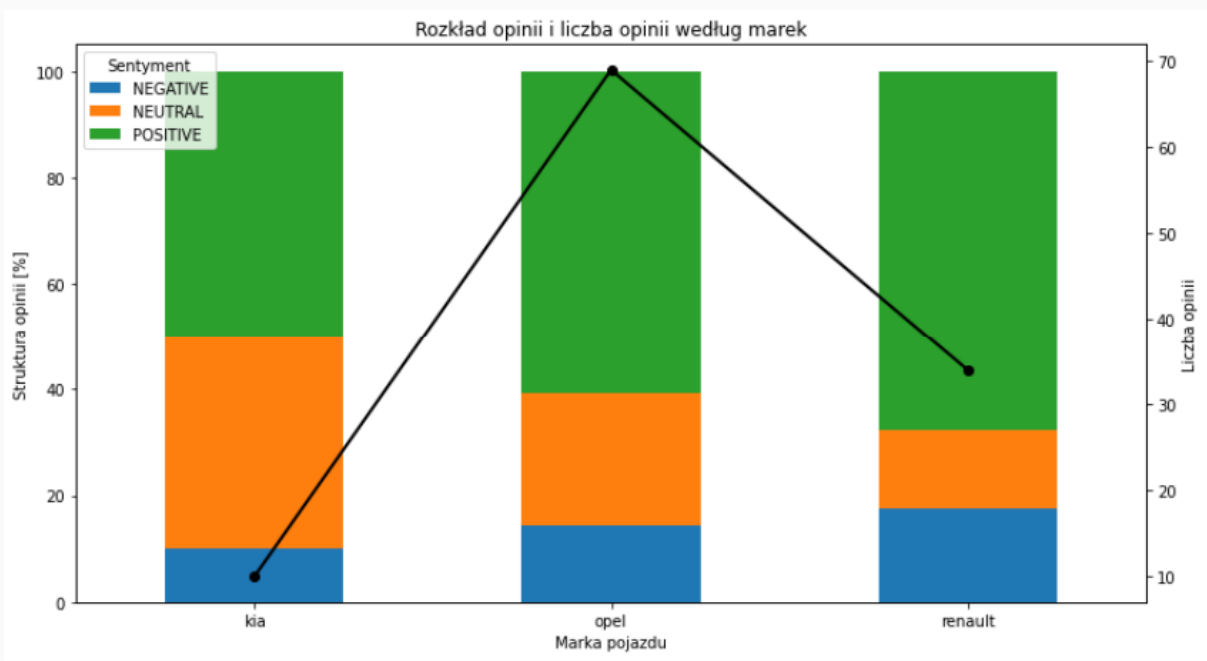
Zbadać relację między marką pojazdu a wartościami sentymentu i ocenić, jakie występują różnice w postrzeganiu poszczególnych marek i które z nich są częściej oceniane pozytywnie a które częściej negatywnie przez użytkowników pojazdów.

W celu zbadania tego związku opracować tablicę dwudzielczą zestawiającą częstości występowania poszczególnych marek o poszczególnych sentymentach. Fragment takiej tablicy dla marki Ford przestawia tabela 1. Przeanalizować i wykorzystać program *S4-03_MarkaSentyment.py*.

Tabela 1. Częstości występowania sentymentu względem marek pojazdów – wycinek przykładowej tablicy dwudzielczej

Marka pojazdu	Wartość sentymentu: pozytywna	Wartość sentymentu: neutralna	Wartość sentymentu: negatywna	Razem
Ford	948 82%	131 11%	75 6%	1154

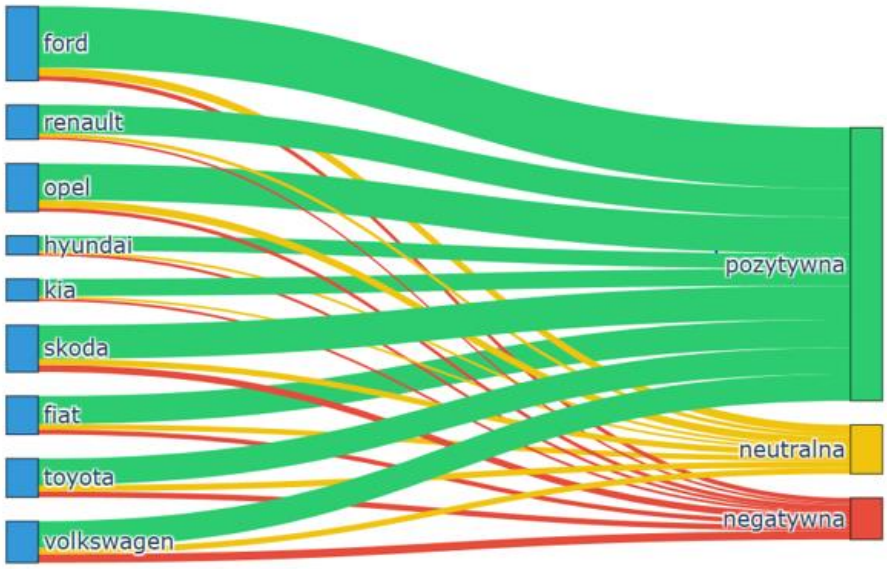
Zilustrować na łączonym wykresie liniowo-słupkowym rozkład opinii i ich strukturę wg marek pojazdów - por. rys. 2. Przeanalizować i wykorzystać program *S4-04_MarkaSentyment_Wykres.py*.



Rys. 2. Rozkład opinii wg przykładowych arek

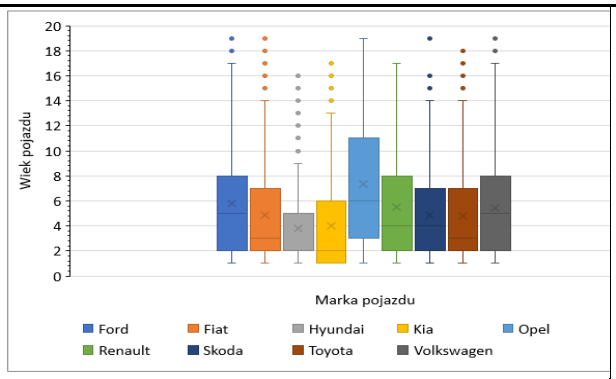
Wykorzystując wykres Sankeya (np. rys. 3) przedstawić przepływy emocji pomiędzy markami pojazdów (wstęgi reprezentują wartości sentymentów w procentach).

Przeanalizować i wykorzystać program *S4-05_MarkaSentyment_SankeyProc*.py. Wykonać eksperymenty z kolorami, tak aby ustalić ostateczną kolorystykę wykresu wg własnego uznania.

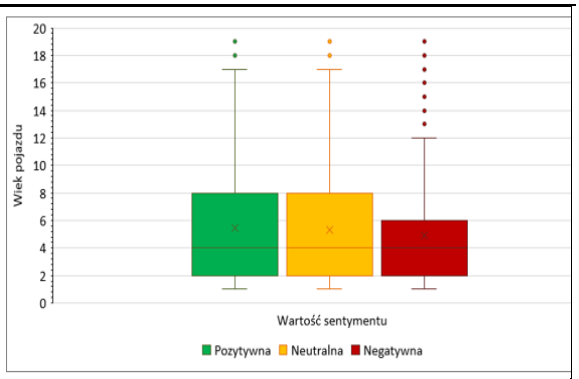


Rys. 3. Powiązanie między markami pojazdów a wartościami sentymentu – diagram Sankeya

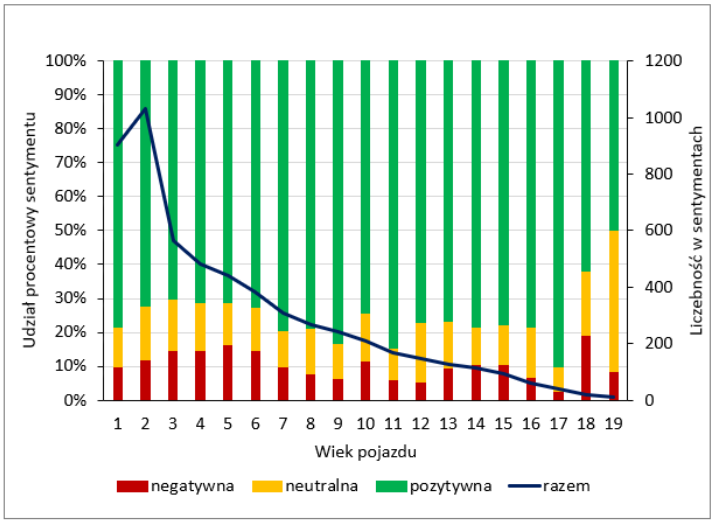
Zadanie 4.
Zbadać sentyment względem wieku pojazdu. Zilustrować jak na rys. 4-6. Przeanalizować i wykorzystać program *S4-06_WiekSentyment.py*.



Rys. 4. Rozkład wieku pojazdów według marki



Rys. 5. Rozkład wieku pojazdu wg sentymentu



Rys. 6. Rozkład sentymentu wg wieku pojazdu