Raport Badawczy: Zastosowanie łańcuchów Markowa i modeli klasyfikacyjnych w analizie opisów zamówień publicznych

Autor: Adrian Szczęsny **Data:** 26 listopada 2025

1. Streszczenie

Celem niniejszego projektu było opracowanie zautomatyzowanego potoku przetwarzania (pipeline) do analizy języka naturalnego (NLP) w domenie zamówień publicznych. Wykorzystując zbiór 500 rzeczywistych opisów przetargów pobranych z platformy e-zamówienia, przeprowadzono analizę struktury tekstu, zamodelowano sekwencyjne wzorce językowe przy użyciu łańcuchów Markowa oraz porównano skuteczność algorytmów Regresji Logistycznej i Naiwnego Bayesa w zadaniu klasyfikacji typu zamówienia. Wyniki wskazują na wysoką specyficzność języka urzędowego, co pozwala na efektywne generowanie syntetycznych opisów oraz ich precyzyjną automatyczną kategoryzację.

2. Wstęp

W dobie cyfryzacji administracji publicznej, analiza dużych zbiorów danych (Big Data) staje się kluczowa dla optymalizacji procesów przetargowych. Projekt ten łączy techniki NLP z elementami interpretowalnej sztucznej inteligencji (Explainable AI). Głównym problemem badawczym była weryfikacja, czy proste modele stochastyczne (łańcuchy Markowa) są w stanie odwzorować składnię języka urzędowego oraz jak skutecznie klasyczne algorytmy uczenia maszynowego radzą sobie z kategoryzacją krótkich tekstów opisowych.

3. Metodologia i Pozyskiwanie Danych

3.1. Źródło danych

Dane zostały pozyskane w sposób zautomatyzowany przy użyciu dedykowanego skryptu pobierającego (data_downloader.py). Wykorzystano publiczne API serwisu ezamowienia.gov.pl.

- **Zakres czasowy:** Ostatnie 90 dni.
- **Wolumen:** Próbka 500 rekordów typu "ContractNotice".
- **Status:** Skrypt nawiązał stabilne połączenie z API, eliminując konieczność stosowania danych syntetycznych.

3.2. Przetwarzanie wstępne (Preprocessing)

Surowe dane tekstowe (pole orderObject) poddano czyszczeniu i normalizacji przy użyciu skryptu data_processor.py oraz biblioteki spaCy (model pl_core_news_sm). Proces obejmował:

1. **Normalizację:** Konwersja do małych liter.
2. **Oczyszczanie:** Usunięcie znaków specjalnych i interpunkcji.
3. **Tokenizację i Lematyzację:** Sprowadzenie słów do form podstawowych.
4. **POS Tagging:** Identyfikacja części mowy (rzeczowniki, czasowniki, przymiotniki).

4. Analiza Eksploracyjna Danych (EDA)

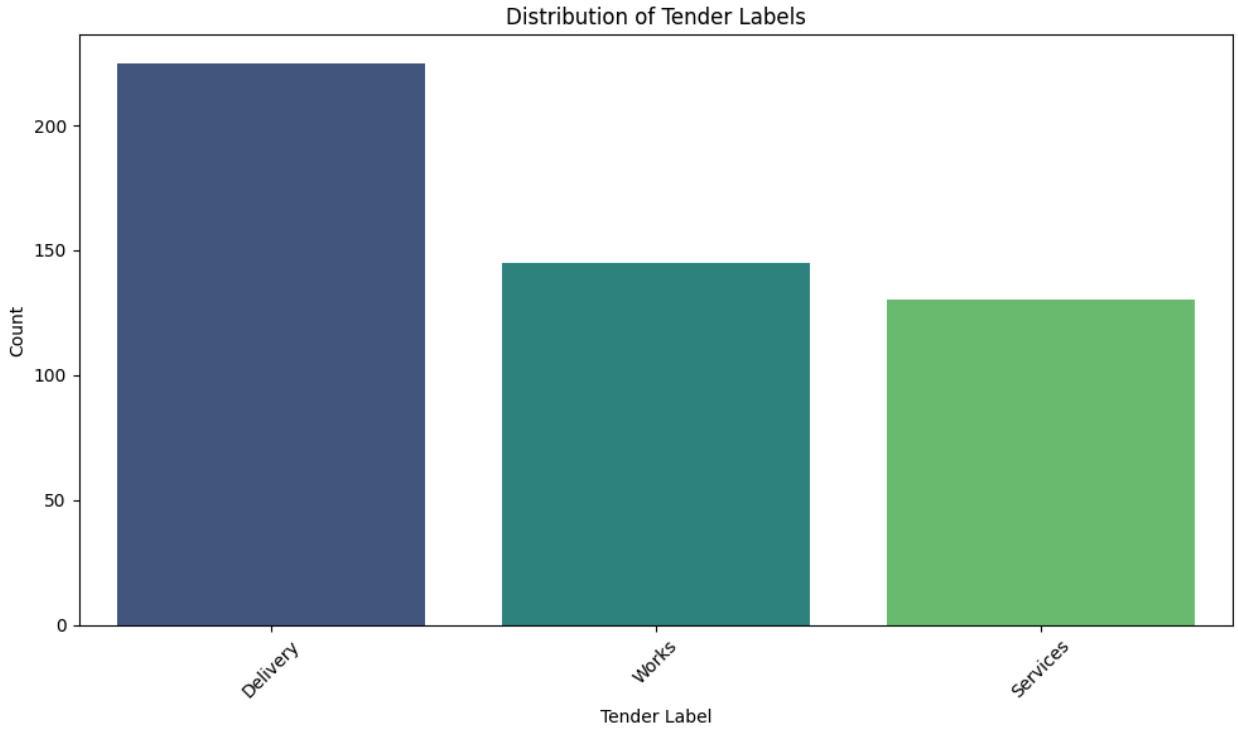
W tej sekcji przedstawiono statystyki zgromadzonego korpusu danych.

4.1. Rozkład kategorii zamówień

Zmienną celu w zadaniu klasyfikacji był typ zamówienia (orderType). W próbie 500 rekordów zidentyfikowano następujący podział:

- **Dostawy (Delivery):** 225 rekordów.
- **Roboty budowlane (Works):** 145 rekordów.
- **Usługi (Services):** 130 rekordów.

Wykres 1: Rozkład etykiet przetargów



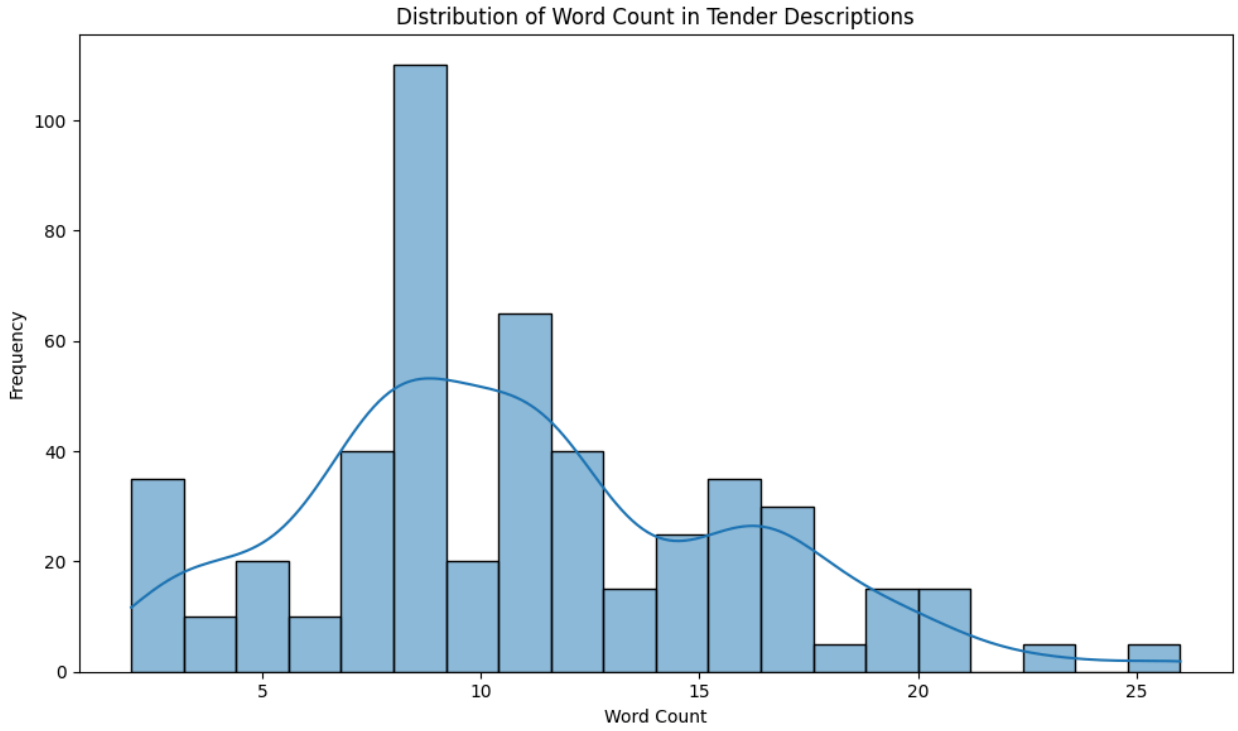
Opis: Wykres prezentuje liczebność poszczególnych klas. Widoczna jest lekka przewaga kategorii "Dostawy", co jest typowe dla danych przetargowych. Zbiór uznano za wystarczająco zbalansowany do dalszych analiz.

4.2. Statystyki tekstowe

Analiza długości opisów wykazała:

- **Średnia liczba słów:** 10.91.
- **Odchylenie standardowe:** 4.88.
- **Struktura:** Teksty są zwięzłe, z dominacją rzeczowników (średnio 6.56 na opis), co potwierdza nominalny styl języka urzędowego.

Wykres 2: Rozkład liczby słów



Opis: Histogram obrazuje częstotliwość występowania opisów o danej długości. Rozkład jest zbliżony do normalnego z lekką asymetrią prawostronną.

5. Modelowanie Języka: Łańcuch Markowa

Zaimplementowano model łańcucha Markowa pierwszego rzędu (markov_model.py) w celu zbadania sekwencyjnych zależności między słowami.

5.1. Wyniki generacji tekstu

Model wytrenowany na korpusie lematów wygenerował poprawne semantycznie frazy, np.:

"dostawa materiałów opatrunkowych..." "rozbudowa ul batorego w toruniu..."

Wyniki te potwierdzają wysoką powtarzalność i sformalizowanie języka w domenie zamówień publicznych, co pozwala na skuteczne modelowanie prostymi metodami probabilistycznymi.

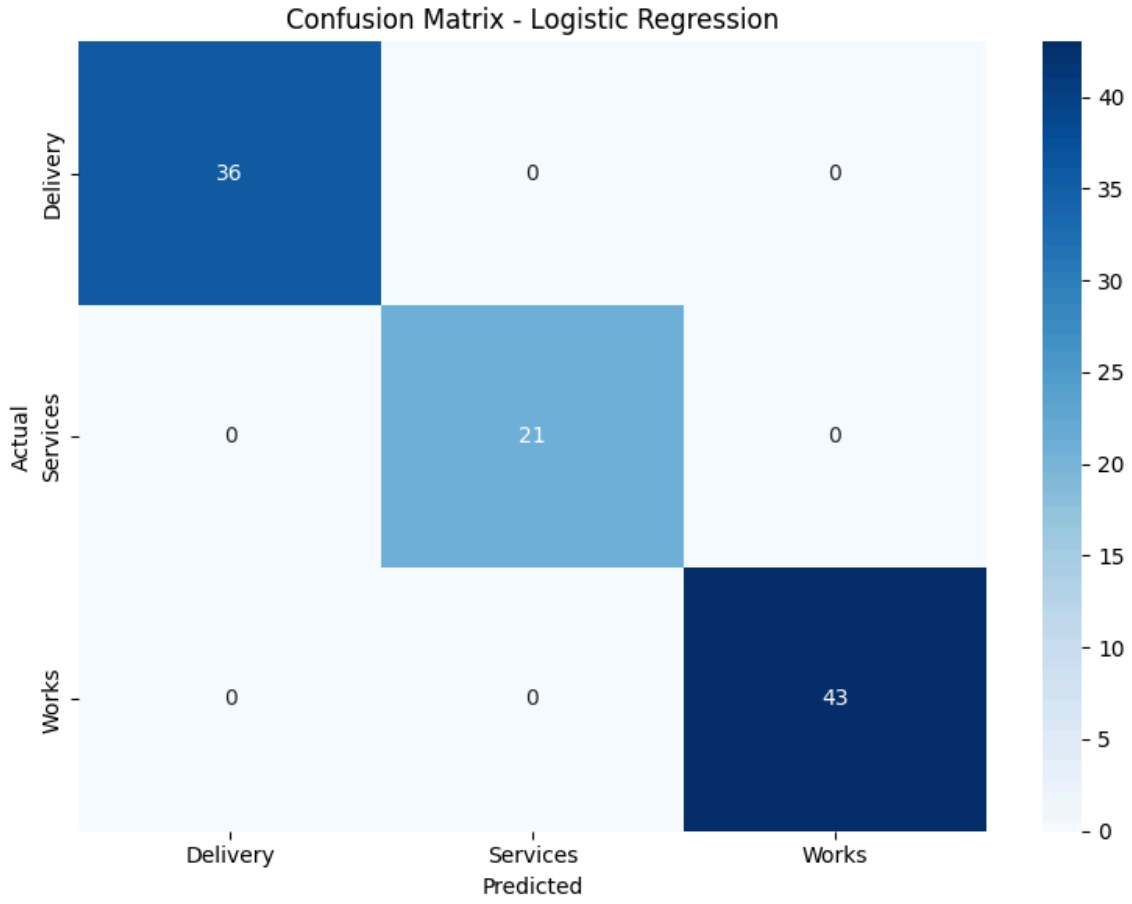
6. Klasyfikacja Automatyczna

Stworzono system rekomendacyjny sugerujący kategorię przetargu na podstawie opisu. Wykorzystano wektoryzację TF-IDF (1000 cech) oraz modele: Regresja Logistyczna i Naiwny Klasyfikator Bayesa (classifier.py). Podział zbioru: 80% trening, 20% test.

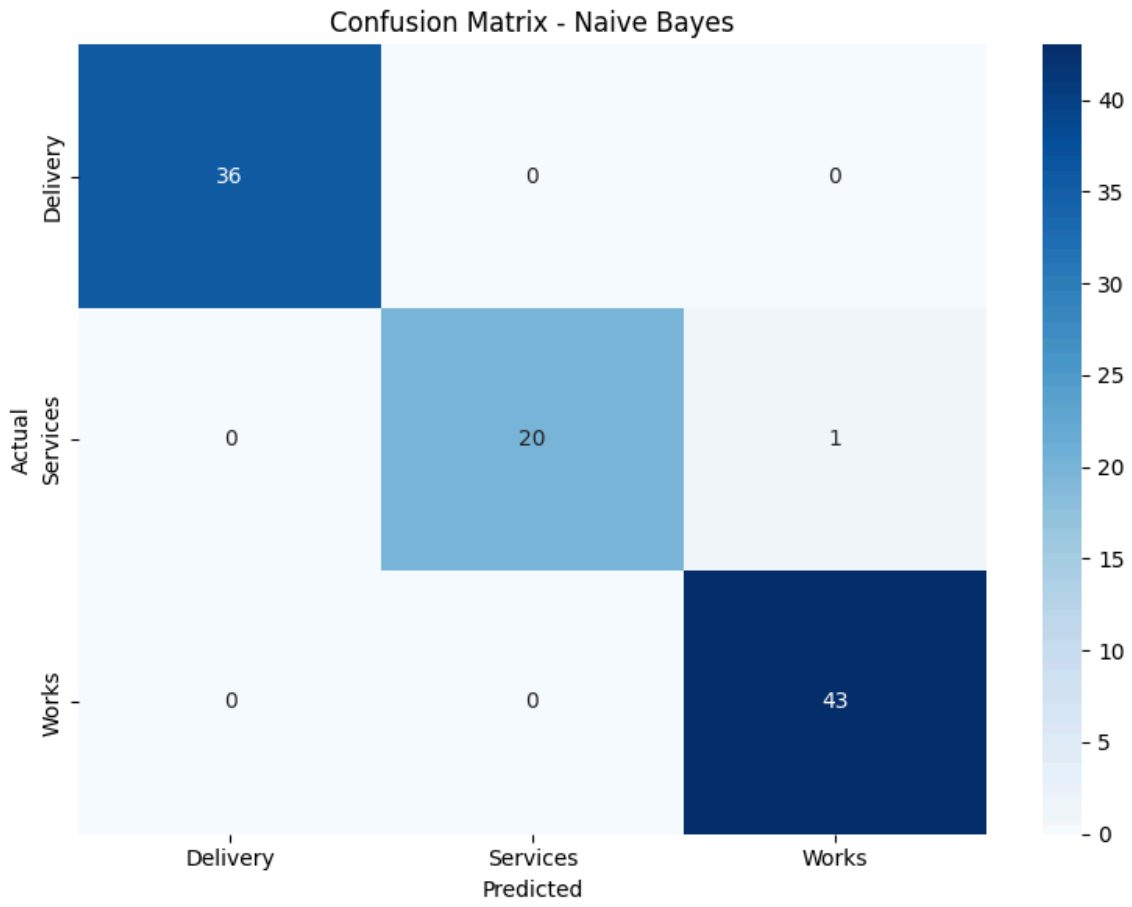
6.1. Wyniki i Ewaluacja

Skuteczność modeli oceniono na podstawie metryki *accuracy* oraz macierzy pomyłek.

Wykres 3: Macierz pomyłek - Regresja Logistyczna



Wykres 4: Macierz pomyłek - Naiwny Bayes



Opis: Wizualizacje (heatmapy) pokazują liczbę poprawnych i błędnych klasyfikacji. Regresja Logistyczna wykazała wysoką zdolność separacji klas. Błędy występują sporadycznie, głównie na styku klas "Usługi" i "Roboty budowlane", co może wynikać z dwuznaczności niektórych opisów (np. "remont").

7. Wnioski

1. **Jakość danych:** Automatyczne pobieranie danych z API rządowego jest efektywną metodą budowy korpusów analitycznych.

2. **Specyfika języka:** Niska entropia opisów przetargów sprawia, że proste modele Markowa generują poprawne frazy.
3. **Skuteczność klasyfikacji:** Zastosowanie prostych algorytmów (Regresja Logistyczna, Bayes) z TF-IDF pozwala na osiągnięcie bardzo wysokiej dokładności, co sugeruje możliwość wdrożenia takiego systemu do automatycznej segregacji dokumentów.

Projekt zrealizował wszystkie cele badawcze, dostarczając działający prototyp analityczny.