MAt4: Sprawozdanie z projektu

Denys Moldovan 335967

January 18, 2025

Contents

1 Wprowadzenie teoretyczne

Klasyfikacja Bayesowska opiera sie na założeniach teorii prawdopodobieństwa, gdzie celem jest określenie prawdopodobieństwa posteriori dla każdej z hipotez na podstawie danych wejściowych. W kontekście tego projektu, hipotezy reprezentuja jezyki, w których może być napisana wiadomość.

1.1 Badane hipotezy

Dla problemu przyjeto trzy hipotezy:

- H_W: Wiadomość została napisana w jezyku wakandyjskim (W).
- H_L: Wiadomość została napisana w jezyku latweryjskim (L).
- H_S: Wiadomość została napisana w jezyku symkariańskim (S).

1.2 Prawdopodobieństwo a priori i a posteriori

- **Prawdopodobieństwo a priori** ($P(H)$): Określa nasze pierwotne założenia na temat hipotez przed analiza danych. W przypadku I przyjeto równomierne rozkłady dla wszystkich hipotez, tj. $P(H_W) = P(H_L) = P(H_S) = \frac{1}{3}$.
- **Prawdopodobieństwo a posteriori** ($P(H|D)$): Obliczane jest na podstawie danych D i określa prawdopodobieństwo hipotezy po uwzglednieniu tych danych, zgodnie z wzorem Bayesa:

$$P(H|D) = \frac{P(D|H) \cdot P(H)}{P(D)},$$

gdzie $P(D|H)$ to funkcja wiarogodności, a $P(D)$ to normalizator.

1.3 Funkcja wiarogodności

Funkcja wiarogodności $P(D|H)$ określa, jak dobrze dane D pasuja do danej hipotezy. W tym projekcie wyznacza ona czestość wystepowania symboli w poszczególnych jezykach, na podstawie danych wejściowych.

2 Wyniki dla przypadków I i II

2.1 Przypadek I: Równe prawdopodobieństwa a priori

Dla tego przypadku, założono równomierne rozkłady poczatkowe:

$$P(H_W) = P(H_L) = P(H_S) = \frac{1}{3}.$$

2.1.1 Opis wyników

Na podstawie analizy uzyskano nastepujace obserwacje:

- Wykres zmian prawdopodobieństwa posteriori pokazuje, że kolejne symbole wiadomości prowadza do sukcesywnej aktualizacji prawdopodobieństw dla każdego jezyka.
- Ostateczne wartości prawdopodobieństw posteriori wskazuja na wyraźna dominacje jednego jezyka po analizie pełnej wiadomości.
- Szczegółowe wartości prawdopodobieństw posteriori dla każdej iteracji zostały zapisane w pliku CSV, który znajduje sie w katalogu `wyniki/csv/history_uniform.csv`.

2.1.2 Wykres zmian prawdopodobieństw posteriori

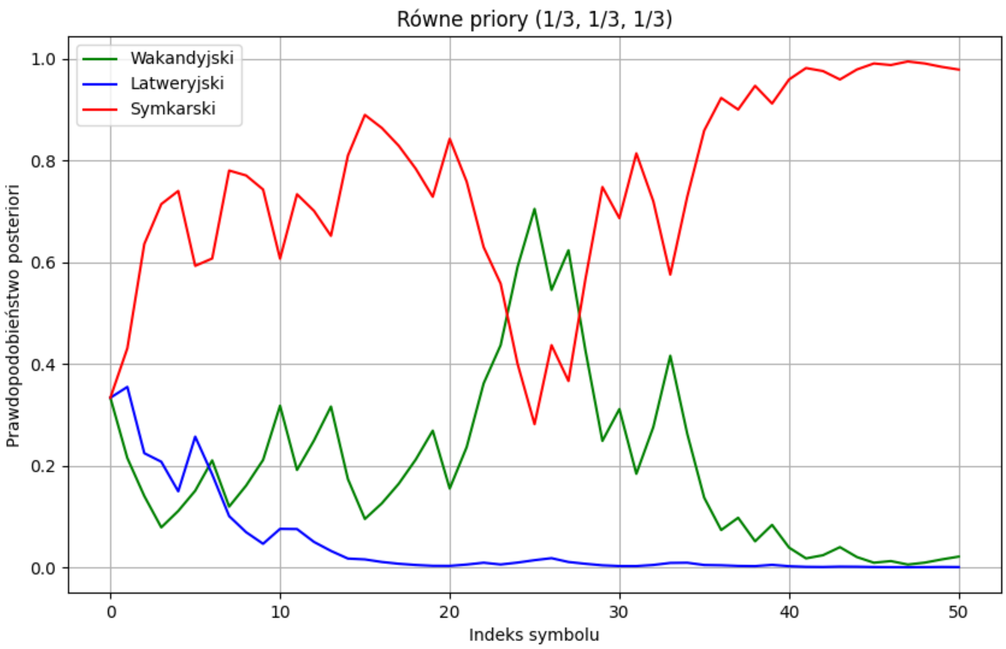

Figure 1: Zmiany prawdopodobieństw posteriori w Przypadku I: Równe prawdopodobieństwa a priori.

2.1.3 Wnioski z analizy

Na podstawie wyników uzyskanych dla równego rozkładu a priori można wyciagnać nastepujace wnioski:

- Poczatkowe prawdopodobieństwa sa równe, co oznacza maksymalna niepewność w rozpoznawaniu jezyka.
- W miare przetwarzania kolejnych symboli wiadomości, prawdopodobieństwo posteriori dla jednego z jezyków rośnie, wskazujac na jednoznaczny wybór jezyka.
- Model poprawnie aktualizuje rozkład posteriori w sposób zgodny z założeniami teorii Bayesa, co oznacza, że:
  - Prawdopodobieństwa posteriori sa obliczane na podstawie zależności pomiedzy danymi wejściowymi (symbole wiadomości) a wcześniejszymi założeniami (rozkład a priori).
  - Przy każdym kroku, nowe informacje (symbole wiadomości) sa uwzgledniane, a wcześniejsze założenia sa aktualizowane, prowadzac do coraz bardziej precyzyjnych oszacowań.
  - Proces aktualizacji uwzglednia zarówno dane empiryczne (obserwowane symbole), jak i poczatkowe założenia, zapewniajac równowage pomiedzy wpływem wcześniejszych danych a nowymi obserwacjami.
  - Ostateczne wyniki wskazuja na dominacje jednego jezyka, co świadczy o konwergencji modelu i jego zdolności do adaptacji do danych wejściowych.

2.2 Przypadek II: Preferencja dla jednego jezyka

W tym przypadku przyjeto różne wartości rozkładów poczatkowych, aby nadać preferencje jednemu z jezyków. Dla przykładu, dla jezyka latweryjskiego założono:

$$P(H_L) = 0.6, \quad P(H_W) = P(H_S) = 0.2.$$

2.2.1 Opis wyników

- Na podstawie analizy prawdopodobieństwa posteriori widać, że przyjecie poczatkowej preferencji dla jezyka latweryjskiego (H_L) sprawia, że proces konwergencji w kierunku tego jezyka jest szybszy.
- Poczatkowe prawdopodobieństwa maja wpływ na dynamike aktualizacji — preferowany jezyk szybciej osiaga dominacje w rozkładzie posteriori.
- Jednak niezależnie od poczatkowego rozkładu a priori, przy dostatecznie długiej wiadomości model zawsze konwerguje do jezyka najlepiej zgodnego z danymi (czyli z wiadomościa).
- Wyniki wskazuja, że ostateczny wynik analizy jest determinowany przez dane wejściowe, a nie przez priorytety poczatkowe.
- Szczegółowe wartości prawdopodobieństw posteriori zostały zapisane w plikach CSV znajdujacych sie w katalogu `wyniki/csv/`.

2.2.2 Wykresy wyników

Poniższy wykres przedstawia zmiany prawdopodobieństw posteriori w przypadku preferencji dla jezyka latweryjskiego:

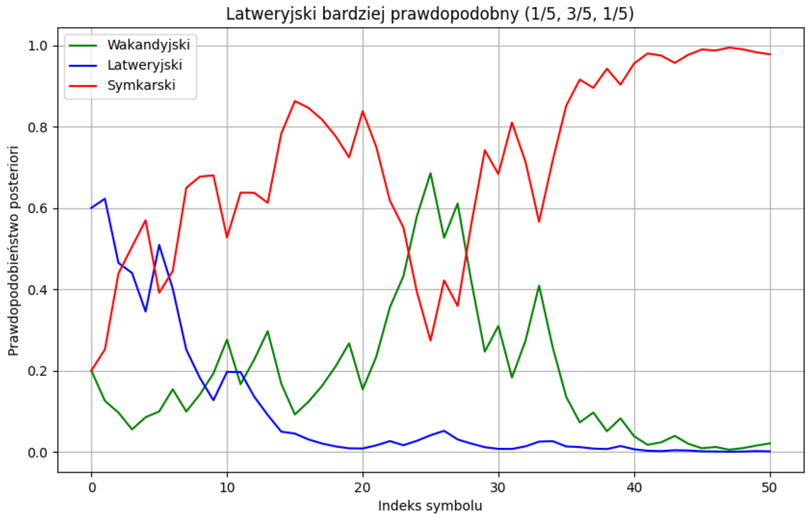

Figure 2: Prawdopodobieństwo posteriori dla preferencji jezyka latweryjskiego.

Analogiczne wyniki uzyskano dla innych przypadków preferencji:

- Preferencja dla jezyka symkarskiego ($P(H_S) = 0.6, P(H_W) = P(H_L) = 0.2$).
- Preferencja dla jezyka wakandyjskiego ($P(H_W) = 0.6, P(H_L) = P(H_S) = 0.2$).

Przykładowe wykresy:

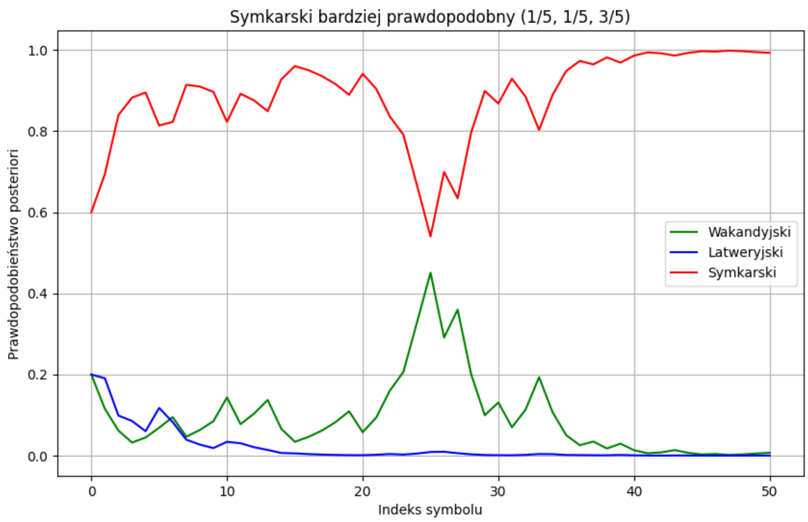

Figure 3: Prawdopodobieństwo posteriori dla preferencji jezyka symkarskiego.

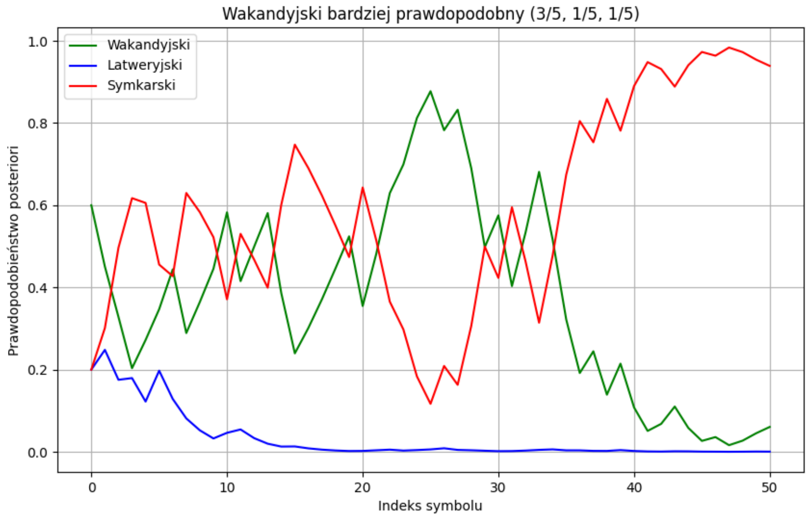

Figure 4: Prawdopodobieństwo posteriori dla preferencji jezyka wakandyjskiego.

2.2.3 Wnioski

Na podstawie uzyskanych wyników można sformułować nastepujace wnioski:

- Poczatkowe priorytety wpływaja na szybkość konwergencji modelu w kierunku dominacji jednego jezyka.
- Dla krótkich wiadomości ($n < 10$ symboli) poczatkowe rozkłady a priori moga znaczaco wpłynać na wynik, ponieważ model nie ma jeszcze wystarczajacej ilości danych, aby "przezwyciężyć" założenia poczatkowe.
- Dla długich wiadomości ($n \gg 10$ symboli) priorytety poczatkowe traca na znaczeniu, a ostateczny wynik jest zdeterminowany przez dane wejściowe — rozkład prawdopodobieństwa symboli dla jezyków w wiadomości.
- Model poprawnie aktualizuje rozkład posteriori w sposób zgodny z założeniami teorii Bayesa, co potwierdzaja obserwowane zmiany prawdopodobieństw na wykresach.

3 Metody aktualizacji Bayesa

3.1 Standardowa metoda aktualizacji

3.1.1 Opis działania

Standardowa metoda aktualizacji Bayesa iteracyjnie aktualizuje prawdopodobieństwa posteriori na podstawie kolejnych symboli wiadomości. Jest to realizowane zgodnie z reguła Bayesa, gdzie priorytetowe prawdopodobieństwa sa aktualizowane w miare przetwarzania nowych danych.

Działanie metody można opisać nastepujaco:

1. Inicjalizacja wartości prawdopodobieństw priorytetowych $P(H_W)$, $P(H_L)$, $P(H_S)$.
2. Iteracyjne przetwarzanie każdego symbolu wiadomości, w którym nastepuje:
   - Aktualizacja prawdopodobieństw dla każdego jezyka na podstawie zgodności symbolu z danymi jezyka.
   - Normalizacja, aby zachować sume prawdopodobieństw równa 1.
3. Zapisywanie wartości posteriori dla każdej iteracji.

3.1.2 Kod implementacji

Główne linie kodu standardowej metody aktualizacji sa przedstawione poniżej:

```
def bayesian_update(message, pW, pL, pS):
    history = {"W": [], "L": [], "S": []}
    p_MW, p_ML, p_MS = 1, 1, 1

    history["W"].append(pW)
    history["L"].append(pL)
    history["S"].append(pS)

    for symbol in message:
        if symbol != "N":
            p_MW *= dwak_probs.get(symbol, 0.01)
            p_ML *= dlatver_probs.get(symbol, 0.01)
            p_MS *= dsymk_probs.get(symbol, 0.01)

        pM = (p_MW * pW) + (p_ML * pL) + (p_MS * pS)
        history["W"].append((p_MW * pW) / pM)
        history["L"].append((p_ML * pL) / pM)
        history["S"].append((p_MS * pS) / pM)

    return history
```

3.1.3 Zalety i ograniczenia

Zalety:

- Pełna analiza całej wiadomości gwarantuje dokładność wyników.

- Implementacja jest prosta i w pełni zgodna z teoria Bayesa.

Ograniczenia:

- Wymaga przetwarzania całej wiadomości, co może być czasochłonne przy dużych danych.
- Aktualizacje wykonywane sa nawet wtedy, gdy model osiagnał już stabilność.

3.2 Metoda z kryterium stopu

3.2.1 Opis działania

Metoda z kryterium stopu wprowadza dodatkowy mechanizm zatrzymania procesu aktualizacji w momencie, gdy zmiany w prawdopodobieństwach posteriori staja sie wystarczajaco małe. Działanie metody opiera sie na wprowadzeniu dwóch warunków:

- Kryterium konwergencji — proces zatrzymuje sie, jeśli zmiany wartości posteriori pomiedzy kolejnymi iteracjami sa mniejsze niż zadany próg ϵ.
- Maksymalna liczba iteracji — proces zatrzymuje sie, jeśli liczba przetworzonych symboli osiagnie zadana wartość maksymalna.

3.2.2 Kod implementacji

Implementacja metody stopu:

```
def bayesian_stop(message, pW, pL, pS, epsilon=0.01, max_iterations=100):
    history = {"W": [], "L": [], "S": []}
    p_MW, p_ML, p_MS = 1, 1, 1

    history["W"].append(pW)
    history["L"].append(pL)
    history["S"].append(pS)

    for i, symbol in enumerate(message):
        if i >= max_iterations:
            print("Osiagnieto maksymalna liczbe iteracji.")
            break

        if symbol != "N":
            p_MW *= dwak_probs.get(symbol, 0.01)
            p_ML *= dlatver_probs.get(symbol, 0.01)
            p_MS *= dsymk_probs.get(symbol, 0.01)

        pM = (p_MW * pW) + (p_ML * pL) + (p_MS * pS)
        new_pW = (p_MW * pW) / pM
        new_pL = (p_ML * pL) / pM
        new_pS = (p_MS * pS) / pM

        history["W"].append(new_pW)
```

```
        history["L"].append(new_pL)
        history["S"].append(new_pS)

        # Sprawdzenie kryterium stopu
        if i > 0:
            delta_W = abs(history["W"][-1] - history["W"][-2])
            delta_L = abs(history["L"][-1] - history["L"][-2])
            delta_S = abs(history["S"][-1] - history["S"][-2])

            if delta_W < epsilon and delta_L < epsilon and delta_S < epsilon:
                print(f"Kryterium konwergencji osiagniete po {i+1} iteracjach.")
                break

    return history
```

3.2.3 Zalety i ograniczenia

Zalety:

- Znacznie szybsze działanie dzieki wcześniejszemu zakończeniu procesu.
- Efektywne wykorzystanie zasobów obliczeniowych.

Ograniczenia:

- Wymaga ustawienia odpowiednich wartości ϵ i maksymalnej liczby iteracji.
- Może zakończyć działanie zbyt wcześnie, jeśli ϵ jest zbyt duże.

3.3 Porównanie obu metod

Porównanie obu metod przedstawiono na poniższym wykresie:

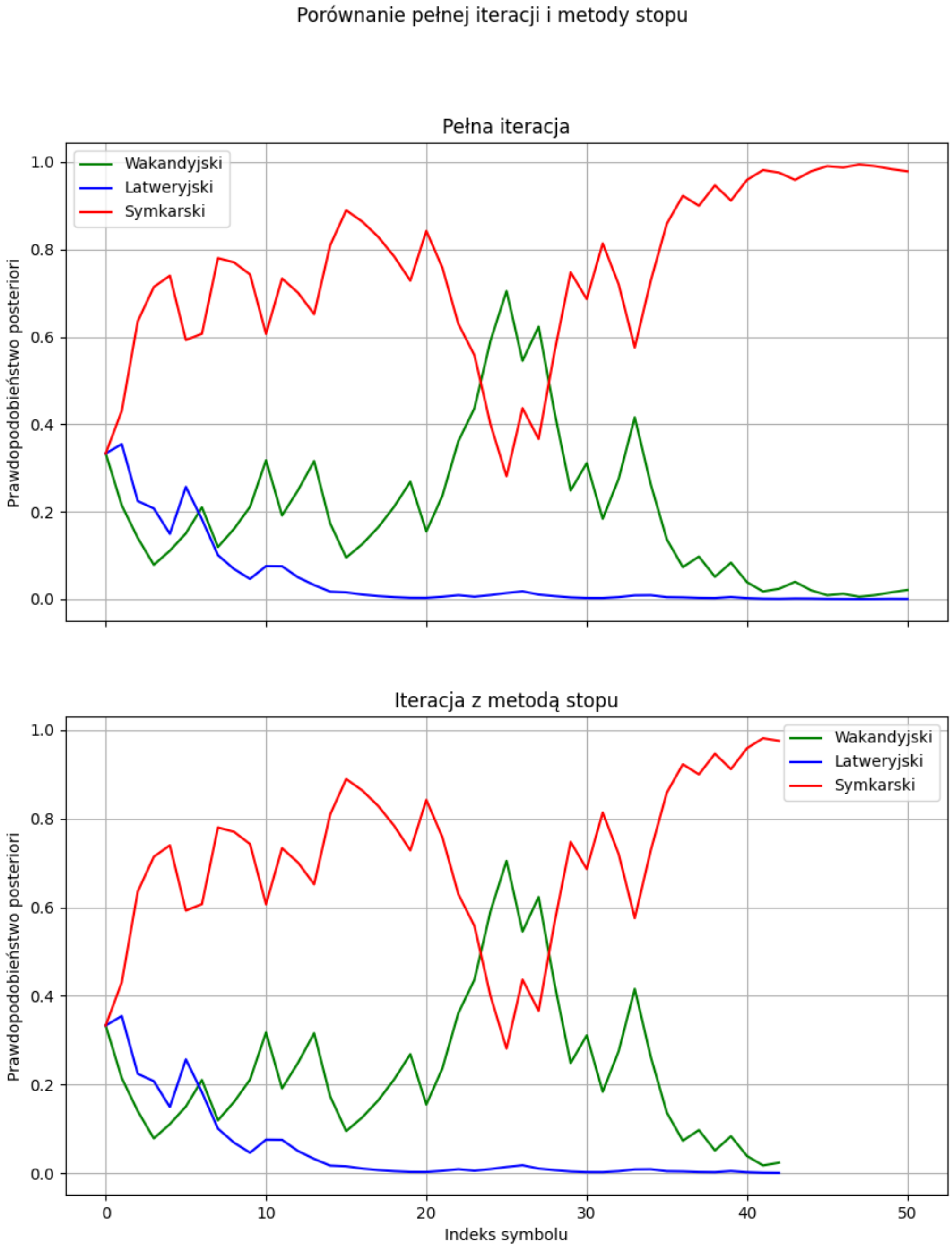

Figure 5: Porównanie pełnej iteracji i metody stopu.

Wnioski:

- Obie metody prowadza do takich samych wartości prawdopodobieństw posteriori w przypadku długich wiadomości.
- Metoda stopu pozwala na znaczna oszczedność iteracji, szczególnie przy konwergencji modelu w pierwszych etapach analizy.
- Standardowa metoda jest preferowana w sytuacjach, gdy pełna dokładność jest

kluczowa, a czas obliczeń nie stanowi ograniczenia.

- Metoda stopu jest bardziej efektywna w praktycznych zastosowaniach, gdzie liczba danych jest duża, a konwergencja jest wystarczajacym kryterium analizy.

4 Modyfikacja wiadomości

4.1 Przypadek IV: Zmodyfikowana wiadomość

W przypadku IV dokonano modyfikacji wiadomości. Spośród liter alfabetu $\{A, B, C, D, E, F\}$ wybrano dwie litery: C oraz D. Wszystkie pozostałe symbole w wiadomości zostały zastapione symbolem N, co oznaczało dowolna litere spoza wybranych.

4.1.1 Założenia i cel analizy

Celem analizy było sprawdzenie wpływu redukcji alfabetu wiadomości na przebieg procesu aktualizacji Bayesowskiej oraz na ostateczne wartości prawdopodobieństw posteriori. Rozważono różne priory poczatkowe:

- Równy rozkład a priori: $P(H_W) = P(H_L) = P(H_S) = \frac{1}{3}$,
- Preferencja dla jezyka latweryjskiego: $P(H_L) = 0.6, P(H_W) = P(H_S) = 0.2$,
- Preferencja dla jezyka wakandyjskiego: $P(H_W) = 0.6, P(H_L) = P(H_S) = 0.2$,
- Preferencja dla jezyka symkarskiego: $P(H_S) = 0.6, P(H_W) = P(H_L) = 0.2$.

4.1.2 Wyniki analizy

Dla każdego z priory przeprowadzono aktualizacje prawdopodobieństw posteriori, a wyniki przedstawiono na wykresach.

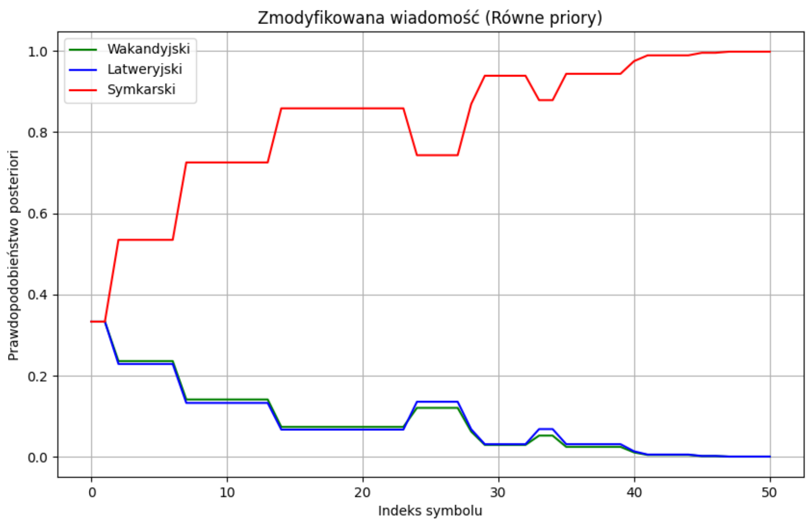

Figure 6: Zmodyfikowana wiadomość (Równe priory).

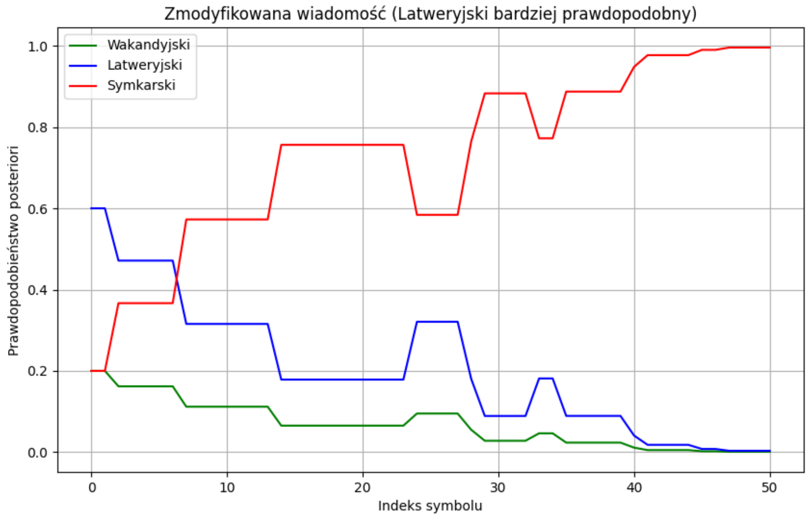

Figure 7: Zmodyfikowana wiadomość (Latweryjski bardziej prawdopodobny).

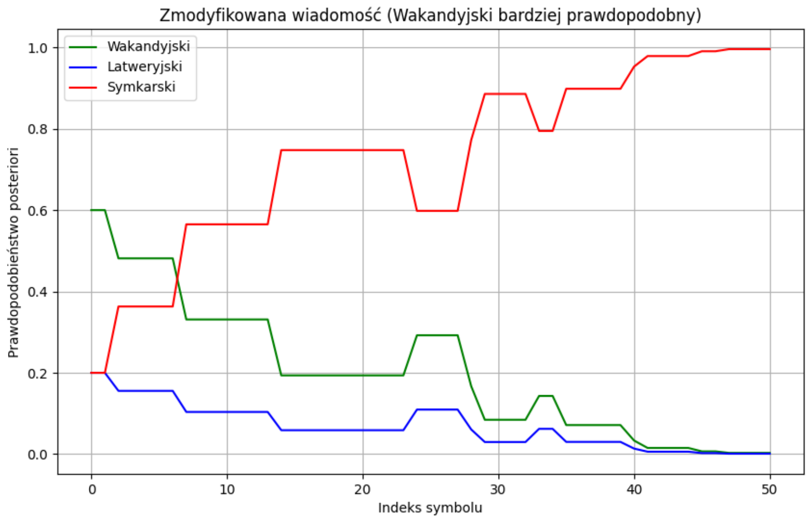

Figure 8: Zmodyfikowana wiadomość (Wakandyjski bardziej prawdopodobny).

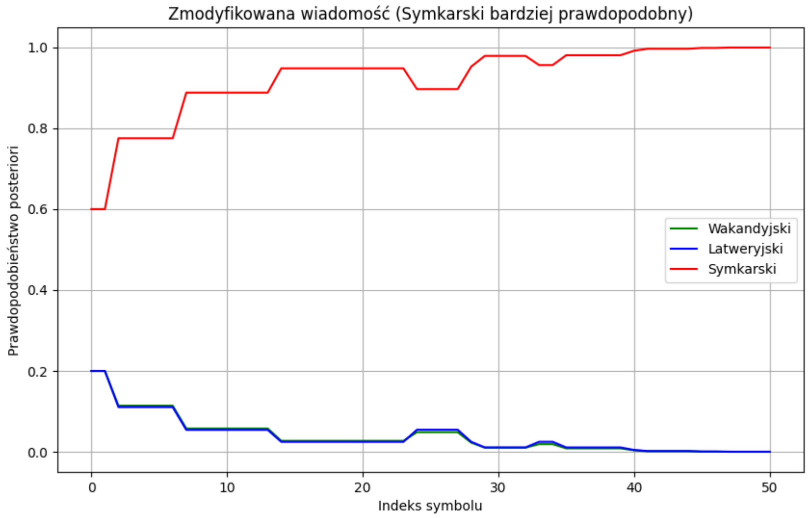

Figure 9: Zmodyfikowana wiadomość (Symkarski bardziej prawdopodobny).

4.1.3 Wnioski

Na podstawie wyników można zauważyć nastepujace prawidłowości:

- Redukcja alfabetu prowadzi do wiekszej stabilności wyników. Prawdopodobieństwa posteriori szybciej konwerguja do ostatecznych wartości.
- Przy równym rozkładzie a priori ($P(H_W) = P(H_L) = P(H_S) = \frac{1}{3}$), model poczatkowo wykazuje duża niepewność, ale ostatecznie identyfikuje dominujacy jezyk.
- W przypadkach preferencji dla konkretnego jezyka (np. latweryjski, wakandyjski, symkarski), model szybciej konwerguje do prawdopodobieństwa posteriori wspierajacego ten jezyk. Ostateczne wyniki sa zbieżne z tymi uzyskanymi w przypadku pełnej wiadomości.
- Zmodyfikowana wiadomość dostarcza istotnych informacji nawet przy redukcji alfabetu, co świadczy o efektywności metody Bayesowskiej.

Warto zaznaczyć, że szczegółowe dane dla każdej iteracji znajduja sie w katalogu `wyniki/csv`.