Informatyka, studia dzienne, I st. semestr VI

Komputerowe systemy rozpoznawania **2019/2020**

Prowadzący: dr. inż. Marcin Kacprowicz wtorek, 14:00

Data oddania: Ocena:

Adam Krzanowski 216815
Gracjan Grala 210192

Projekt 1: Ekstrakcja cech, miary podobieństwa, klasyfikacja

1. Cel

Celem zadania było stworzenie aplikacji do klasyfikacji zbioru artykułów z wykorzystaniem metody k-NN, zawierającej moduł ekstrakcji cech oraz moduł klasyfikatora k-NN. Dodatkowo należało zbadać wpływ poszczególnych parametrów klasyfikatora k-NN na wyniki otrzymywane w trakcie przeprowadzania eksperymentów.

2. Wprowadzenie

Głównym celem aplikacji była klasyfikacja tekstu, a właściwie jego przyporządkowanie do odpowiedniego państwa - etykiety “places”. Zagadnienie idealnie wpasowuje się do zagadnień sztucznej inteligencji, opiera się jednak na algorytmie statycznym, którego zadaniem jest dopasowanie tekstu do odpowiedniej etykiety. Wykorzystując metodę k-NN należało, więc wyekstrahować zbiór cech dla każdego z tekstów.

Zaimplementowane cechy skupiały się, więc na słowach kluczowych w danym tekście, co pozwoliło zaprezentować każdy z artykułów w postaci wektora w przestrzeni n-wymiarowej. Konstruując każdą z cech należało postawić nacisk na to, aby żadna z cech nie miała powiązania z długością tekstu, zdań czy samych wyrazów, a jedynie skupiała się na samej treści zawartej w artykułach. Na potrzeby implementacji dodatkowo podzielono zbiór słów kluczowych na dwa zbiory:

- słowa kluczowe unikalne (występujące w artykule tylko jeden raz)
- słowa kluczowe pospolite (występujące w artykule więcej niż jeden raz)

Wykorzystując wcześniej wyliczone cechy dla każdego z artykułów do klasyfikacji tekstów zastosowano algorytm k-NN. Algorytm ten pozwala na zaklasyfikowanie każdych danych przedstawionych w postaci wektorów w przestrzeni n-wymiarowej koncentrując się na wyliczaniu odległości pomiędzy nimi przy użyciu metryk lub miar podobieństwa. W celu wyliczenia odległości pomiędzy wektorami wykorzystano następujące metryki i miary podobieństwa:

- **metryka euklidesowa** - w celu obliczenia odległości d między dwoma wektorami należy obliczyć pierwiastek kwadratowy z sumy potęg różnic wartości współrzędnych o tych samych indeksach, zgodnie ze wzorem:

$$d(\mathbf{p}, \mathbf{q}) = d(\mathbf{q}, \mathbf{p}) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \cdots + (q_n - p_n)^2}$$

Wzór 1. Odległość euklidesowa.

- **metryka Manhattan** - w celu obliczenia odległości d między dwoma wektorami należy obliczyć sumę wartości bezwzględnych różnic współrzędnych punktów x oraz y, zgodnie ze wzorem:

$$d_m(x,y) = \sum_{k=1}^{n} |x_k - y_k|$$

Wzór 2. Odległość uliczna.

- **metryka Czebyszewa** - w celu obliczenia odległości d między dwoma wektorami należy obliczyć maksymalną wartość bezwzględnych różnic współrzędnych punktów x oraz y, zgodnie ze wzorem:

$$d_{ch}(x,y) = \max_i |x_i - y_i|$$

Wzór 3. Odległość Czebyszewa.

- **indeks Jaccarda** - w celu obliczenia odległości d między dwoma wektorami należy obliczyć iloraz mocy części wspólnej zbiorów i mocy sumy zbiorów, zgodnie ze wzorem:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

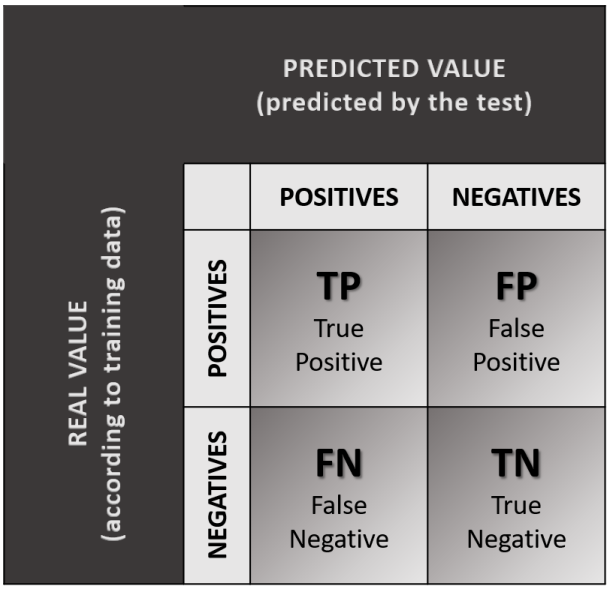
Wzór 4. Współczynnik podobieństwa Jaccarda.

- **podobieństwo kosinusowe** - w celu obliczenia odległości d między dwoma wektorami należy obliczyć iloraz wartości bezwzględnej sumy wartości wektora V1 i V2 przez pierwiastek z iloczynu sum wartości wektorów podniesionych do kwadratu, zgodnie ze wzorem:

$$r_{ca}(V_1,V_2) = \frac{\left|\sum_{i=1}^{n} v_{1i} \cdot v_{2i}\right|}{\sqrt{\sum_{i=1}^{n} v_{1i}^2 \cdot \sum_{i=1}^{n} v_{2i}^2}}$$

Wzór 5. Podobieństwo kosinusowe.

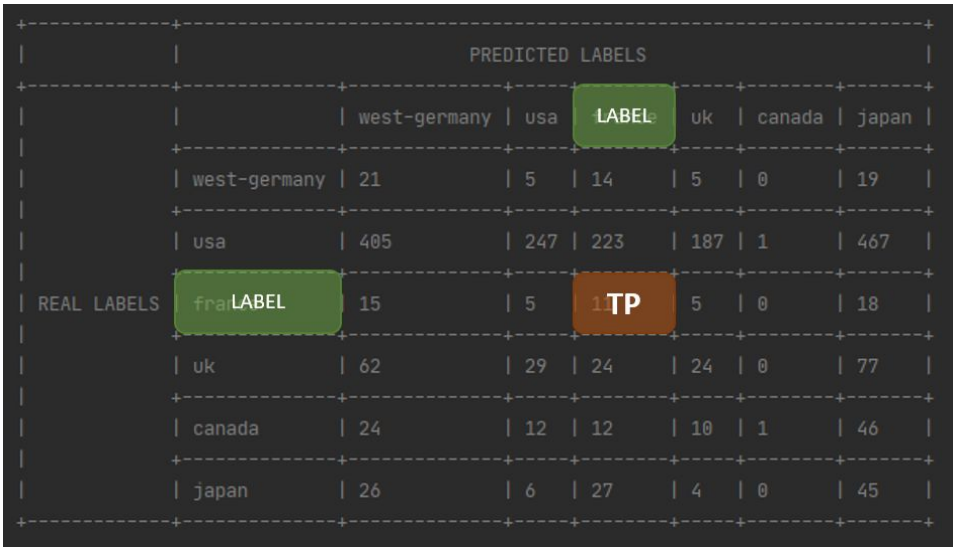
Po wykonaniu klasyfikacji, w celu zbadania sprawności klasyfikatora wyliczyliśmy współczynniki Accuracy, Precision oraz Recall używając do tego macierzy błędów. Opierając się na ocenie jakości klasyfikacji binarnej wyliczono wszystkie wartości dla każdej z klas, które po podsumowaniu względem wszystkich klas umożliwiły uzyskanie ogólnej oceny sprawności klasyfikatora.



Rysunek 1. Macierz pomyłek dla klasyfikacji binarnej.

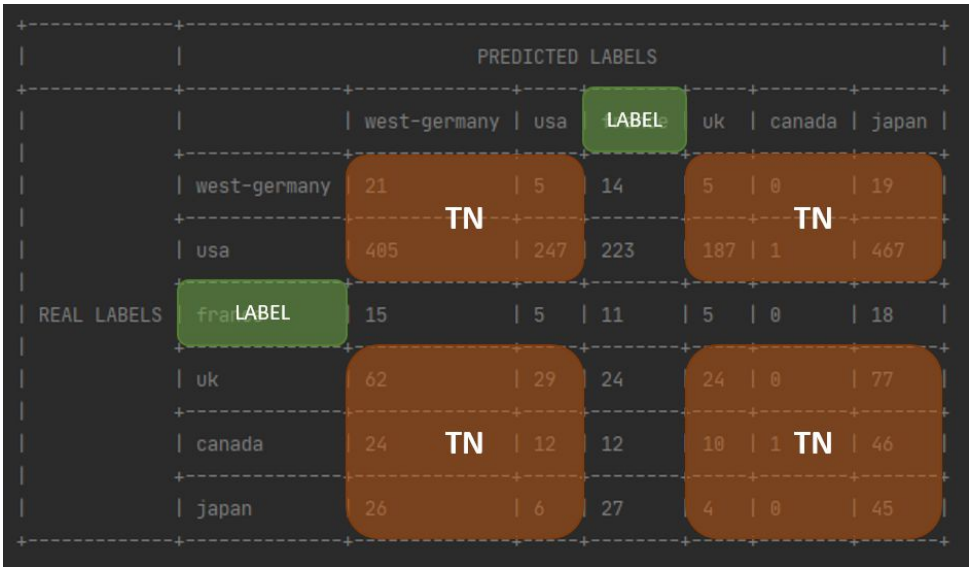
Obliczenia wykonano przy następujących założeniach:

- **TP – True Positive** – liczba obserwacji poprawnie zaklasyfikowanych do klasy pozytywnej.
    - *Przykład: Teksty prawidłowo przyporządkowane do klasy “france”. Pole macierzy oznaczone kolorem czerwonym (rys. 2).*



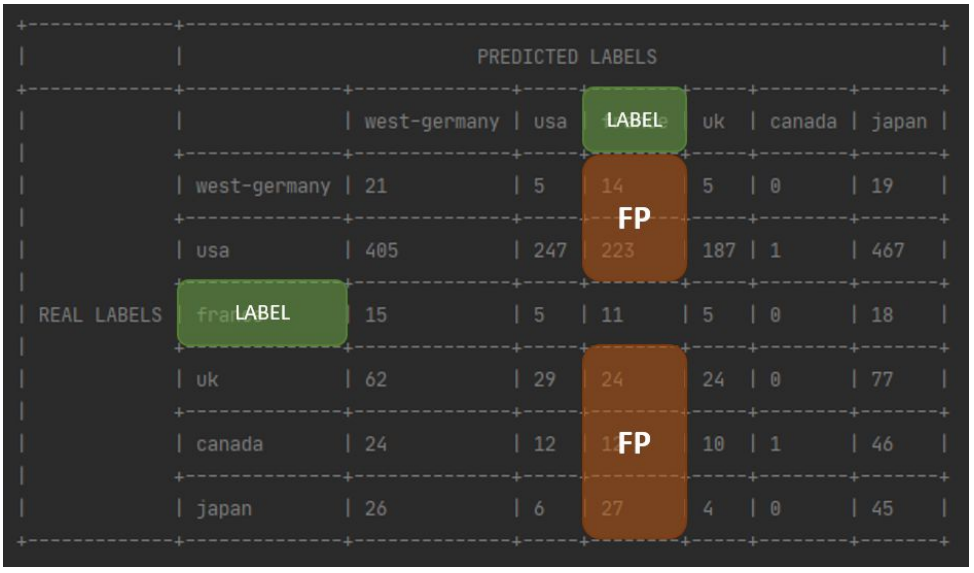
Rysunek 2. Przykładowa wartość TP w macierzy błędów dla pojedynczej etykiety.

- **TN – True Negative** – liczba obserwacji poprawnie zaklasyfikowanych do klasy negatywnej.
    - *Przykład: Teksty prawidłowo nie przyporządkowane do klasy “france”. Jest to suma pól macierzy oznaczonych kolorem czerwonym (rys. 3).*



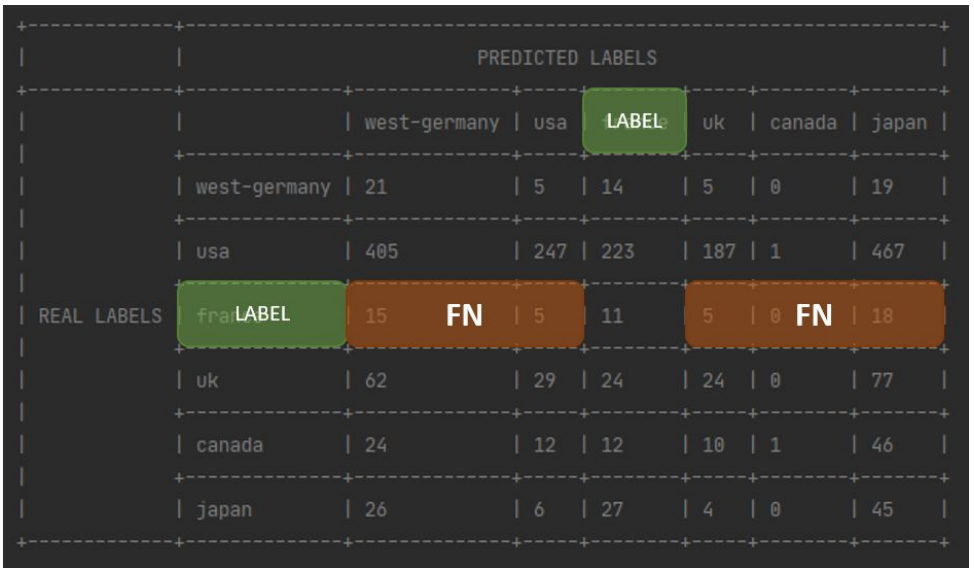
Rysunek 3. Przykładowa wartość TN w macierzy błędów dla pojedynczej etykiety.

- **FP – False Positive** – liczba obserwacji zaklasyfikowanych do klasy pozytywnej podczas, gdy w rzeczywistości pochodzą z klasy negatywnej.
    - *Przykład: Teksty błędnie przyporządkowane do klasy “france”, które w rzeczywistości powinny być przypisane do innej klasy. Jest to suma pól oznaczonych kolorem czerwonym (rys. 4).*



Rysunek 4. Przykładowa wartość FP w macierzy błędów dla pojedynczej etykiety.

- **FN – False Negative** – liczba obserwacji zaklasyfikowanych do klasy negatywnej podczas, gdy w rzeczywistości pochodzą z klasy pozytywnej.
    - *Przykład: Teksty błędnie przyporządkowane do klasy innej niż "france", które w rzeczywistości powinny zostać do niej przypisane. Jest to suma pól oznaczonych kolorem czerwonym (rys. 5).*



Rysunek 5. Przykładowa wartość FN w macierzy błędów dla pojedynczej etykiety.

- **ACC (Accuracy)** – sprawność klasyfikatora, określa prawdopodobieństwo poprawnej klasyfikacji, czyli stosunek poprawnych klasyfikacji do wszystkich klasyfikacji.

$$ACC = \frac{TP + TN}{TP + TN + FP + FN}$$

Wzór 6. Dokładność metody pomiaru dla klasyfikatora binarnego.

- **PPV (Precision)** - stosunek poprawnie sklasyfikowanych elementów z klasy A do wszystkich, które nasz klasyfikator oznaczył jako klasę A.

$$PPV = \frac{TP}{TP + FP}$$

Wzór 7. Wartość predykcyjna dodatnia dla klasyfikatora binarnego.

- **TPR (Recall)** - stosunek poprawnie rozpoznanych elementów z klasy A do wszystkich, które powinien rozpoznać, czyli do całej klasy A.

$$TPR = \frac{TP}{TP + FN}$$

Wzór 8. Wartość czułości dla klasyfikatora binarnego.

3. Opis implementacji

Aplikacja została napisana w języku programowania Java z wykorzystaniem frameworka “Spring” jako aplikacja webowa z wbudowaną bazą danych H2 w formie lokalnego pliku “mv.db”.

Aplikacja składa się z następujących modułów:

- **config** - konfiguracja aplikacji (pliki z rozszerzeniem .yaml)
- **controller** - moduł odpowiedzialny za komunikację aplikacji z użytkownikiem
- **model** - moduł zawierający modele danych wykorzystywane w aplikacji
- **repository** - moduł odpowiedzialny za operacje wykonywane na bazie danych
- **service** - najważniejszy moduł odpowiedzialny za logikę związaną z ekstrakcją jak i klasyfikacją cech
    - **classifier** - podmoduł odpowiedzialny za klasyfikację oraz ocenę klasyfikatora
    - **extractors** - podmoduł wykorzystywany do ekstrakcji cech z artykułów
    - **metrics** - podmoduł wykonujący obliczenia stosowane dla metody k-NN tj. odległości i podobieństwa wektorów cech
    - **utils** - podmoduł zajmujący się przetwarzaniem plików zawierających artykuły

4. Materiały i metody

4.1 Procesowanie tekstów

Przed samą ekstrakcją cechy każdy z artykułów został przeprocesowany przy użyciu popularnych metod eksploracji danych. Zastosowano kolejno:

- **tokenizację** - podzielono każdy tekst bezpośrednio na pojedyncze słowa
- **stemizację** - konwersja słów o tym samym znaczeniu do ich najprostszej formy
- **stopwords** - każde ze słów znajdujące się w stop liście zostało usunięte, aby pozbyć się słów popularnych o małym znaczeniu nie wpływającym na identyfikację tekstu

W celu zoptymalizowania działania aplikacji zbiór artykułów został przetworzony i oczyszczony z artykułów które mogą nie wpływać na wyniki klasyfikacji, a które mogą tylko wydłużać działanie programu (np. artykuły nie posiadające treści - puste pola “body”). Dodatkowo klasyfikacja została przeprowadzona tylko dla artykułów których pole “places” przyjmuje wartości: west-germany, usa, france, uk, canada, japan i są to ich jedyne etykiety.

4.2 Ekstrakcja cech

Kolejny krokiem była ekstrakcja cech z każdego tekstu. W celu optymalizacji działania klasyfikatora proces ten zapisuje się w bazie danych, dzięki czemu w każdej chwili można sprawdzić wszystkie wartości jakie zostały przypisane do danego tekstu. Zaimplementowano następujące cechy wykorzystując poniższe oznaczenia:

- **SUKDAW (SumUniqueKeywordsDivideAmountWords)** - suma ilości unikalnych słów kluczowych w całym przeprocesowanym artykule podzielona przez ilość wszystkich słów w artykule oryginalnym.
    - **UW (unique word)** - słowo w zbiorze słów kluczowych, które pojawia się tylko raz
    - **W (word)** - pojedyncze słowo w artykule przed procesowaniem
    - Dziedzina: $x \, \varepsilon \, [0.0761, 0.769]$

$$SUKDAW = \frac{\sum UW}{\sum W}$$

Wzór 9. Wartość cechy SUKDAW.

- **SCKDAW (SumCommonKeywordsDivideAmountWords)** - suma ilości pospolitych słów kluczowych w całym przeprocesowanym artykule podzielona przez ilość wszystkich słów w artykule oryginalnym.
    - **CW (common word)** - słowo kluczowe występujące więcej niż jeden raz w artykule
    - **W (word)** - pojedyncze słowo w artykule przed procesowaniem
    - Dziedzina: $x \, \varepsilon \, [0.0, 0.68]$

$$SCKDAW = \frac{\sum CW}{\sum W}$$

Wzór 10. Wartość cechy SCKDAW.

- **AOSUKDS (AvgOfSumUniqueKeywordsDivideSentence)** - w celu uzyskania wartości cechy segregujemy artykuł na zdania, wyliczamy dla każdego zdania ilość unikalnych słów kluczowych i dzielimy ją przez ilość wszystkich słów kluczowych w tym samym zdaniu. Dla każdego zdania otrzymamy iloraz, następnie liczymy średnią z wszystkich ilorazów.
    - **UW (unique word)** - słowo w zbiorze słów kluczowych, które pojawia się tylko raz (w tym przypadku w kontekście pojedynczego zdania)
    - **KW (keyword)** - słowo kluczowe (w tym przypadku w kontekście pojedynczego zdania)
    - **n** - ilość zdań w artykule
    - Dziedzina: $x \, \varepsilon \, [0.0, 1.0]$

$$AOSUKDS = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\sum UW}{\sum KW}\right)$$

Wzór 11. Wartość cechy AOSUKDS.

- **AOSCKDS (AvgOfSumCommonKeywordsDivideSentence)** - dla każdego ze zdań liczymy ilość pospolitych słów kluczowych występujących w danym zdaniu, następnie dzielimy każdą przez ogólną ilość słów kluczowych występujących w tym samym zdaniu. Ostatecznie liczymy średnią wartość z wcześniej otrzymanych ilorazów.
    - **CW (common word)** - słowo kluczowe występujące więcej niż jeden raz w artykule (w tym przypadku w kontekście pojedynczego zdania)
    - **KW (keyword)** - słowo kluczowe (w tym przypadku w kontekście pojedynczego zdania)
    - **n** - ilość zdań w artykule
    - Dziedzina: $x \in [0, 1]$

$$AOSCKDS = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\sum CW}{\sum KW}\right)$$

Wzór 12. Wartość cechy AOSCKDS.

- **AOSUKDP (AvgOfSumUniqueKeywordsDivideParagraph)** - dla każdego z akapitów liczymy ilość unikalnych słów kluczowych występujących w danym akapicie, następnie dzielimy każdą z liczb przez ogólną ilość słów kluczowy występujących w tym samym akapicie. Na koniec wyliczamy średnią wartość z wcześniej otrzymanych ilorazów.
    - **UW (unique word)** - słowo w zbiorze słów kluczowych, które pojawia się tylko raz (w tym przypadku w kontekście pojedynczego akapitu)
    - **KW (keyword)** - słowo kluczowe (w tym przypadku w kontekście pojedynczego akapitu)
    - **n** - ilość akapitów w artykule
    - Dziedzina: $x \in [0.05, 1.0]$

$$AOSUKDP = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\sum UW}{\sum KW}\right)$$

Wzór 13. Wartość cechy AOSUKDP.

- **AOSCKDP (AvgOfSumCommonKeywordsDivideParagraph)** - dla każdego z akapitów liczymy ilość pospolitych słów kluczowych występujących w danym akapicie, następnie dzielimy każdą z liczb przez ogólną ilość słów kluczowy występujących w tym samym akapicie. Na koniec wyliczamy średnią wartość z wcześniej otrzymanych ilorazów.
    - **CW (common word)** - słowo kluczowe występujące więcej niż jeden raz w artykule (w tym przypadku w kontekście pojedynczego akapitu)
    - **KW (keyword)** - słowo kluczowe (w tym przypadku w kontekście pojedynczego akapitu)
    - **n** - ilość akapitów w artykule
    - Dziedzina: $x \in [0.0, 0.64]$

$$AOSCKDP = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{\sum CW}{\sum KW} \right)$$

Wzór 14. Wartość cechy AOSCKDP.

- **PUACK (ProportionUniqueAndCommonKeywords)** - proporcja ilości unikalnych słów kluczowych do pospolitych słów kluczowych w tekście.
    - **UW (unique word)** - słowo w zbiorze słów kluczowych, które pojawia się tylko raz (w tym przypadku w kontekście całego artykułu)
    - **CW (common word)** - słowo kluczowe występujące więcej niż jeden raz w artykule (w tym przypadku w kontekście całego artykułu)
    - Dziedzina: $x \in [0.0, 23.0]$

$$PUACK = \frac{\sum UW}{\sum CW}$$

Wzór 15. Wartość cechy PUACK.

- **PUKIPOA (ProportionUniqueKeywordsInPartOfArticle)** - z artykułu wyciągamy pewien procent słów (pierwsze 20%), liczymy proporcję ilości unikalnych słów kluczowych występujących w danej części tekstu i ilości wszystkich słów kluczowych w danej części tekstu.
    - **UW (unique word)** - słowo w zbiorze słów kluczowych, które pojawia się tylko raz (w tym przypadku w kontekście pewnej części artykułu)
    - **KW (keyword)** - słowo kluczowe (w tym przypadku w kontekście pewnej części artykułu)
    - Dziedzina: $x \in [0.18, 1.0]$

$$PUKIPOA = \frac{\sum UW}{\sum KW}$$

Wzór 16. Wartość cechy PUKIPOA.

- **AKOP (AmountKeywordsOnParagraph)** - liczymy ilość słów kluczowych w każdym z akapitów i wyciągamy z tego średnią arytmetyczną.
    - **KW (keyword)** - słowo kluczowe (w tym przypadku w kontekście pojedynczego akapitu)
    - **n** - ilość akapitów w artykule
    - Dziedzina: $x \in [2.83, 72.0]$

$$AKOP = \frac{1}{n} \sum_{i=1}^{n} \left(\sum KW \right)$$

Wzór 17. Wartość cechy AKOP.

- **AWWTNOSDAK (AmountWordsWithTheNotOnStartDivideAmountKeywords)** - liczymy ilość słów kluczowych z przedimkiem "THE" i dzielimy przez ilość słów kluczowych
    - Uwaga: Tą cechę wyjątkowo liczymy w inny sposób, tzn. zaraz po tokenizacji i stemizacji wstrzymujemy się z wykorzystaniem stoplisty, aby zebrać słowa z przedimkiem “the”. Taką listę słów w następnym kroku dopiero filtrujemy stoplistą i przyrównujemy do ogólnej ilości słów kluczowych w tekście.
    - **AW (article word)** - słowo kluczowe, które występuje po przedimku “the”
    - **KW (keyword)** - słowo kluczowe (w tym przypadku w kontekście całego artykułu)
    - Dziedzina: $x \in [0.0, 0.3]$

$$AWWTNOSDAK = \frac{\sum AW}{\sum KW}$$

Wzór 18. Wartość cechy AWWTNOSDAK.

- **AL (AverageLevenshtein)** - liczymy odległość Lewensztajna dla każdego unikalnego słowa kluczowego w artykule względem innych wszystkich pospolitych słów kluczowych (zbiór CW nie będzie zawierać duplikatów) i wyciągamy z nich średnią arytmetyczną. Otrzymamy zbiór wartości o kardynalności równej ilości unikalnych słów kluczowych w artykule. Z otrzymanego zbioru wyliczamy średnią, która będzie naszą wartością końcową.
    - $lev(a, b)$ - odległość Levenshteina (edycyjna) pomiędzy słowami (a i b)
    - **UW (unique word)** - słowo w zbiorze słów kluczowych, które pojawia się tylko raz (w tym przypadku w kontekście całego artykułu)
    - **CW (common word)** - słowo kluczowe występujące więcej niż jeden raz w artykule (w tym przypadku w kontekście całego artykułu)
    - Dziedzina: $x \in [0.0, 12.25]$

Przykład zdania: *Alice has big cat and fish. This cat is big as Alice's dog.*

$$UW \in \{dog, fish\}$$
$$CW \in \{alice, big, cat\}$$

UW	CW	WARTOŚĆ ODLEGŁOŚCI	ŚREDNIA	WARTOŚĆ CECHY
DOG	ALICE	5	3,(3)	3,5
	BIG	2		
	CAT	3		
FISH	ALICE	4	3,(6)7	
	BIG	3		
	CAT	4		

Rysunek 6. Przykładowe obliczanie cechy AL.

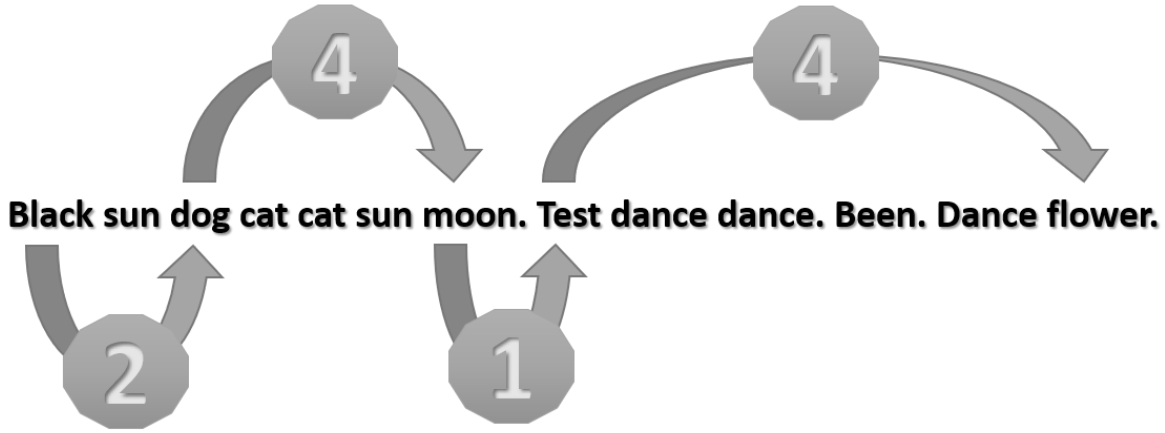
$$AL = \frac{1}{n} \sum_{i=1}^{n} lev_{a,b}(UW_{i_a}, CW_{i_b})$$

Wzór 19. Wartość cechy AL.

- **AWBUK (AmountWordsBetweenUniqueKeywords)** - Wyszukujemy unikalne słowa kluczowe w całym tekście, następnie liczymy odległości pomiędzy każdym następnym napotkanym słowem kluczowym. Mając w tekście zbiór słów kluczowych {A, B, C, D, E, F, G}, którego podzbiór unikalnych słów kluczowych to {B, D, F} wyliczamy kolejno odległość pomiędzy unikalnym słowami kluczowymi B i D oraz D i F. Z wszystkich odległości wyciągamy średnią - jest to ostateczna wartość cechy.
    - **UW (unique word)** - słowo w zbiorze słów kluczowych, które pojawia się tylko raz (w tym przypadku w kontekście całego artykułu)
    - **KW (keyword)** - słowo kluczowe (w tym przypadku w kontekście całego artykułu)
    - $dist_i$ - odległość pomiędzy słowami kluczowymi
    - Dziedzina: $x \, \varepsilon \, [1.0, 5.4]$

Przykład zdania: *Black sun dog cat cat sun moon. Test dance dance. Been. dance flower.*

$$UW \varepsilon \{black, \ dog, \ moon, \ test, \ flower\}$$



Rysunek 7. Przykład obliczania odległości pomiędzy unikalnymi słowami kluczowymi.

$$AWBUK = \frac{1}{n} \sum_{i=1}^{n} dist_i$$

Wzór 20. Wartość cechy AWBUK.

4.3 Eksperymenty

Eksperymenty, które przeprowadzaliśmy skupiały się na wpływie różnych parametrów na wyniki klasyfikacji. Przed samą ekstrakcją cech zbiór danych (tj. artykuły) losowo wymieszano i podzielono na dziesięć części (każdy zbiór zawierał prawie identyczną ilość artykułów). Taki podział umożliwił sprawne przeprowadzenie eksperymentu B oraz wszystkich innych pod kątem używania tych samych zbiorów danych dla każdej z klasyfikacji.

- **Eksperyment A** - wpływ parametru k na klasyfikację

W celu zbadania wpływu parametru k, który jest głównym współczynnikiem wykorzystywanym w metodzie k-NN (odpowiada za przynależność sąsiedztwa w zależności od używanej metryki lub miary podobieństwa) zdecydowano się na następujący zbiór wartości:

$$k \in \{0.1,\ 0.2,\ 0.5,\ 0.8,\ 1.0,\ 1.5,\ 2.0,\ 2.5,\ 3.0,\ 5.0,\ 10.0\}$$

Pozostałe wartości dla eksperymentu pozostały niezmienione:

PARAMETR	WARTOŚĆ		OPIS
MetricType	EUCLIDEAN		Wykorzystano metrykę – odległość euklidesowa.
DataBreakdown	L60T40		Proporcja podziału danych: 60 % zbiór uczący / 40 % zbiór testowy.
UsedFeatures	SUKDAW AOSUKDP AOSCKDP AL PUACK PUKIPOA	SCKDAW AOSUKDS AOSCKDS AWWTNOSDAK AKOP AWBUK	Użyto wszystkich możliwych cech przy klasyfikacji.

Tabela 1. Wartości stałe dla eksperymentu A.

- **Eksperyment B** - wpływ podziału zbioru uczącego na klasyfikację

Jako kolejny eksperyment obrano wyznaczenie wpływu wielkości proporcji podziału zbioru uczącego i testowego na klasyfikację. Ostatecznie eksperyment został przeprowadzony w następującej kombinacji:

Zbiór treningowy	Zbiór testowy
10 %	90 %
30 %	70 %
50 %	50 %
70 %	30 %
90 %	10 %

Tabela 2. Proporcje podziału danych dla eksperymentu B.

Pozostałe wartości dla eksperymentu pozostały niezmienione:

PARAMETR	WARTOŚĆ		OPIS
MetricType	EUCLIDEAN		Wykorzystano metrykę – odległość euklidesowa.
K	0.5		Parametr K będący wyznacznikiem sąsiedztwa dla artykułów.
UsedFeatures	SUKDAW AOSUKDP AOSCKDP AL PUACK PUKIPOA	SCKDAW AOSUKDS AOSCKDS AWWTNOSDAK AKOP AWBUK	Użyto wszystkich możliwych cech przy klasyfikacji.

Tabela 3. Wartości stałe dla eksperymentu B.

- **Eksperyment C** - wpływ wyboru metryki lub miary podobieństwa na klasyfikację

Zgodnie z założeniami projektu dla tego eksperymentu sprawdzono wpływ każdej z metryk lub miary podobieństwa na wyniki klasyfikacji tj. euklidesowa, manhattan, czebyszewa, indeks jaccarda, podobieństwo kosinusowe.

Pozostałe wartości dla eksperymentu pozostały niezmienione:

PARAMETR	WARTOŚĆ		OPIS
K	1.0		Parametr K będący wyznacznikiem sąsiedztwa dla artykułów.
DataBreakdown	L60T40		Proporcja podziału danych: 60 % zbiór uczący / 40 % zbiór testowy.
UsedFeatures	SUKDAW AOSUKDP AOSCKDP AL PUACK PUKIPOA	SCKDAW AOSUKDS AOSCKDS AWWTNOSDAK AKOP AWBUK	Użyto wszystkich możliwych cech przy klasyfikacji.

Tabela 4. Wartości stałe dla eksperymentu C.

- **Eksperyment D** - wpływ podzbiorów cech na wyniki klasyfikacji

Do przeprowadzenia tego eksperymentu podzielono zbiór 12 cech na 4 podzbiory w sposób następujący:

PODZBIÓR	CECHY
I	AOSCKDP AOSCKDS AOSUKDP AOSUKDS PUACK PUKIPOA SCKDAW SUKDAW
II	AKOP AWBUK AWWTNOSDAK AL PUACK PUKIPOA SCKDAW SUKDAW
III	AKOP AWBUK AWWTNOSDAK AL AOSCKDP AOSCKDS AOSUKDP AOSUKDS SCKDAW SUKDAW
IV	AKOP AWBUK AWWTNOSDAK AL AOSCKDP AOSCKDS AOSUKDP AOSUKDS PUACK PUKIPOA

Tabela 5. Podział podzbiorów cech dla eksperymentu D.

Pozostałe wartości dla eksperymentu pozostały niezmienione:

PARAMETR	WARTOŚĆ	OPIS
MetricType	EUCLIDEAN	Wykorzystano metrykę – odległość euklidesowa.
K	0.5	Parametr K będący wyznacznikiem sąsiedztwa dla artykułów.
DataBreakdown	L60T40	Proporcja podziału danych: 60 % zbiór uczący / 40 % zbiór testowy.

Tabela 6. Wartości stałe dla eksperymentu D.

4.4 Ocena klasyfikatora

Zgodnie z teorią dotyczącą macierzy błędów wyliczono współczynniki sprawności klasyfikatora dla każdej z klas/etykiet. Aby uzyskać pełne wyniki dotyczące całej klasyfikacji, aniżeli samych etykiet wartości wyliczono następująco:

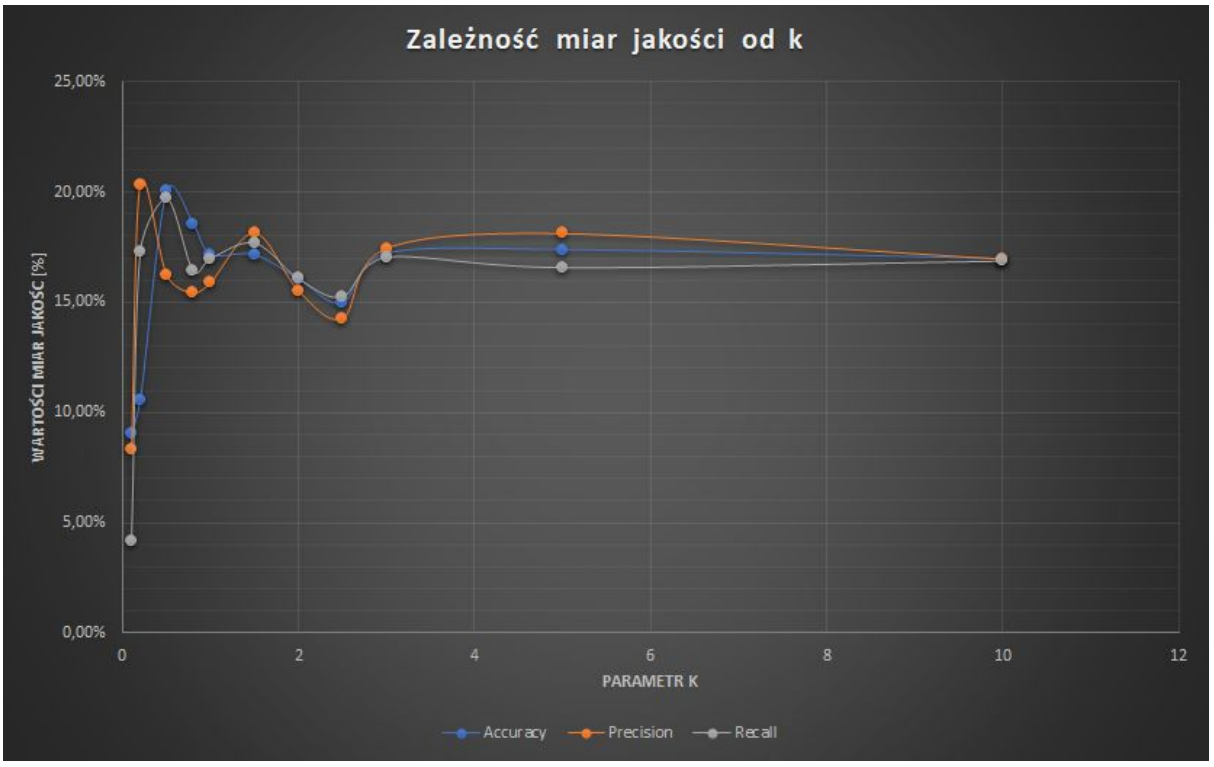
- **Total Accuracy** - całkowita sprawność klasyfikatora została wyliczona na podstawie wszystkich poprawnie zaklasyfikowanych tekstów względem wszystkich istniejących tekstów.
- **Total Precision** - całkowita wartość predykcji dodatniej dla klasyfikatora została wyliczona jako średnia arytmetyczna dla wszystkich wartości predykcji dodatniej otrzymanych dla każdej z etykiet.
- **Total Recall** - całkowita wartość czułości dla klasyfikacji została również wyliczono jako średnia arytmetyczna dla wszystkich wartości czułości otrzymanej dla każdej z etykiet.

5. Wyniki

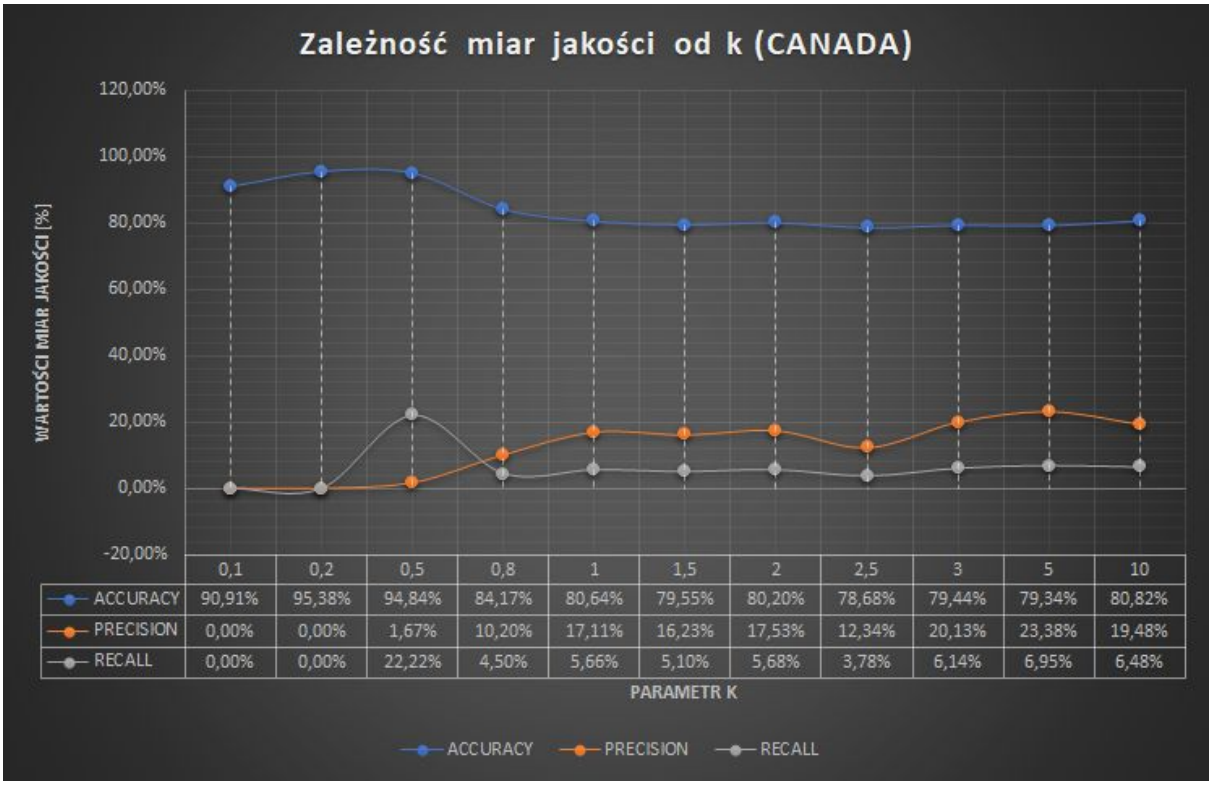
5.1 Eksperyment A

k	0,1	0,2	0,5	0,8	1	1,5	2	2,5	3	5	10
Accuracy	9,09%	10,58%	20,05%	18,57%	17,18%	17,15%	16,05%	14,98%	17,22%	17,39%	16,98%
Precision	8,33%	20,35%	16,25%	15,43%	15,91%	18,15%	15,53%	14,29%	17,46%	18,13%	16,97%
Recall	4,17%	17,29%	19,75%	16,44%	17,00%	17,70%	16,08%	15,26%	17,06%	16,59%	16,88%

Tabela 7. Porównanie poszczególnych współczynników sprawności klasyfikatora.

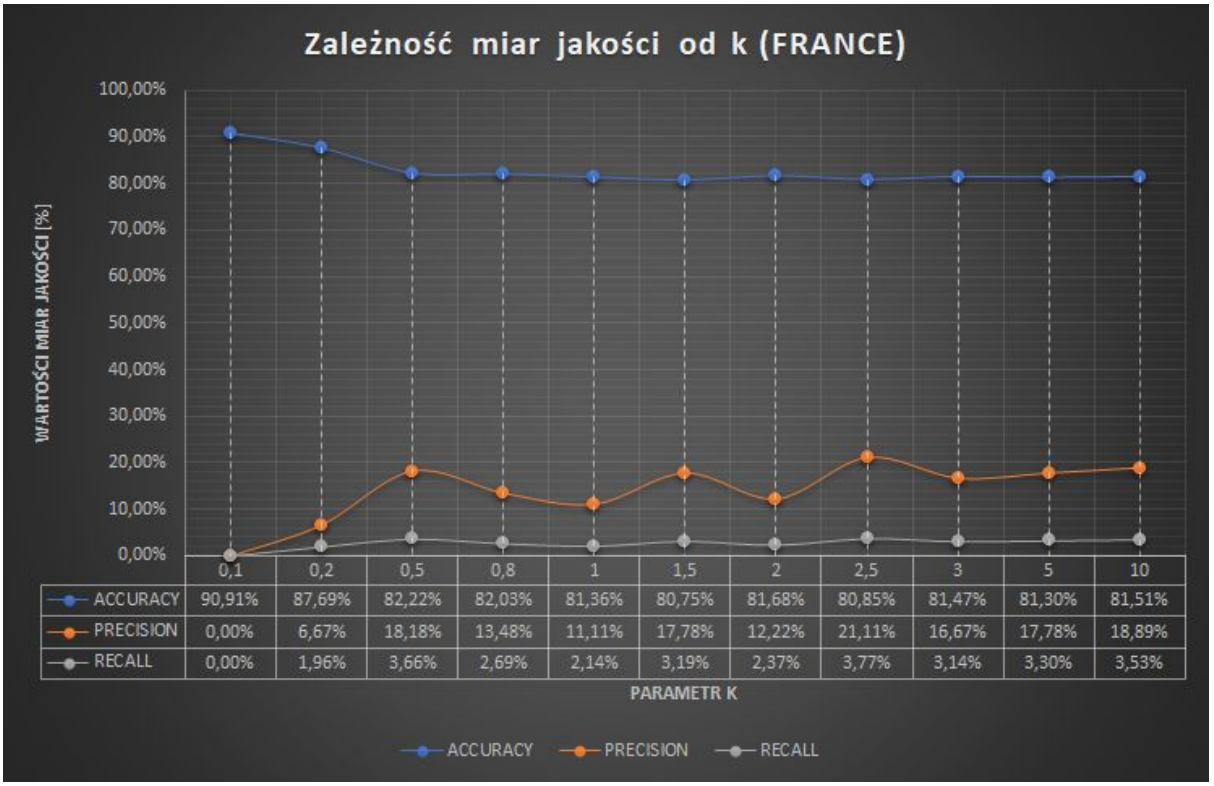


Wykres 1. Zależność miar jakości od k.



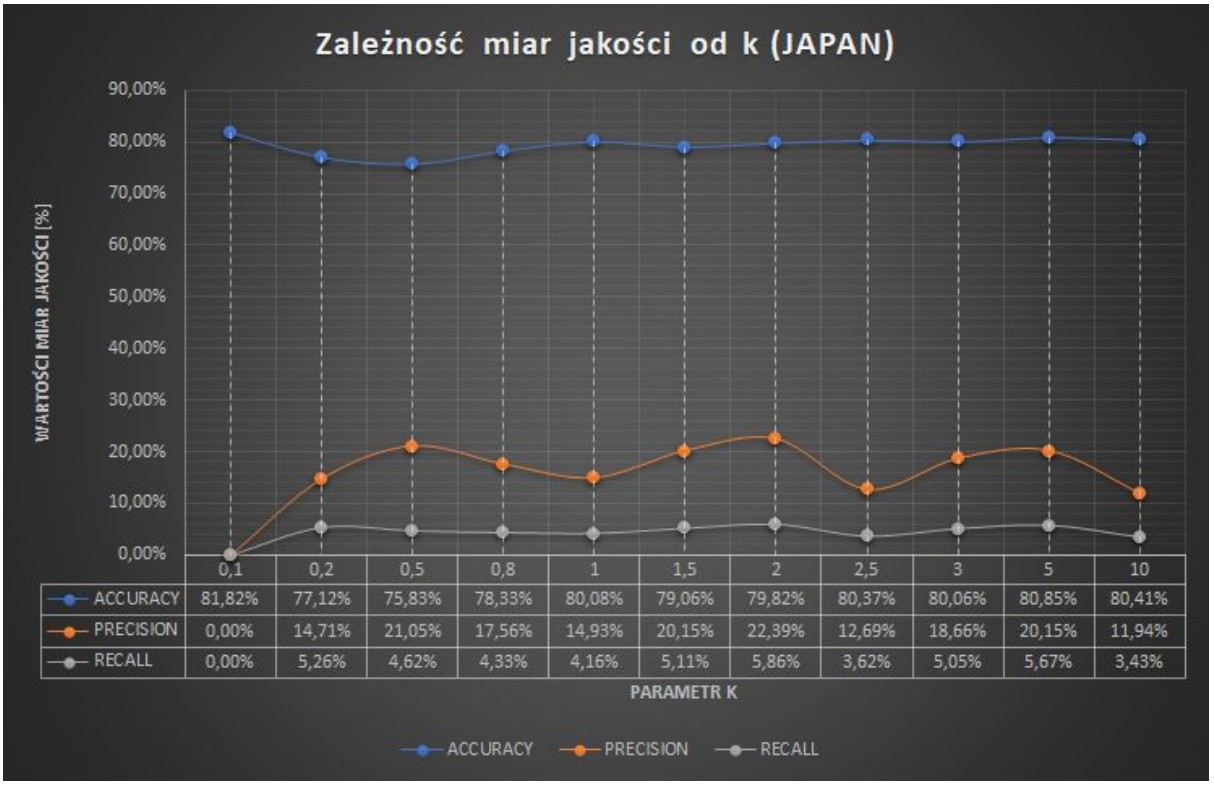
	0,1	0,2	0,5	0,8	1	1,5	2	2,5	3	5	10
ACCURACY	90,91%	95,38%	94,84%	84,17%	80,64%	79,55%	80,20%	78,68%	79,44%	79,34%	80,82%
PRECISION	0,00%	0,00%	1,67%	10,20%	17,11%	16,23%	17,53%	12,34%	20,13%	23,38%	19,48%
RECALL	0,00%	0,00%	22,22%	4,50%	5,66%	5,10%	5,68%	3,78%	6,14%	6,95%	6,48%

Wykres 2. Zależność miar jakości od k dla etykiety “CANADA”.



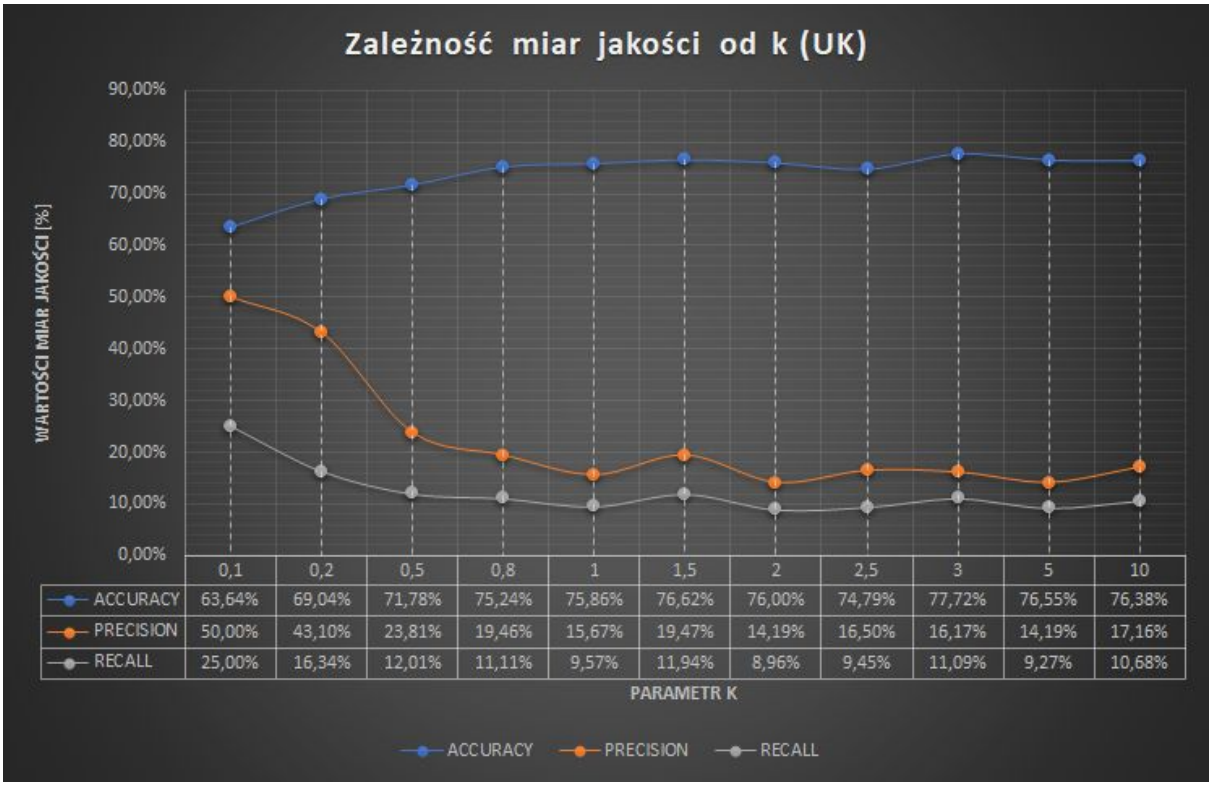
	0,1	0,2	0,5	0,8	1	1,5	2	2,5	3	5	10
ACCURACY	90,91%	87,69%	82,22%	82,03%	81,36%	80,75%	81,68%	80,85%	81,47%	81,30%	81,51%
PRECISION	0,00%	6,67%	18,18%	13,48%	11,11%	17,78%	12,22%	21,11%	16,67%	17,78%	18,89%
RECALL	0,00%	1,96%	3,66%	2,69%	2,14%	3,19%	2,37%	3,77%	3,14%	3,30%	3,53%

Wykres 3. Zależność miar jakości od k dla etykiety “FRANCE”.



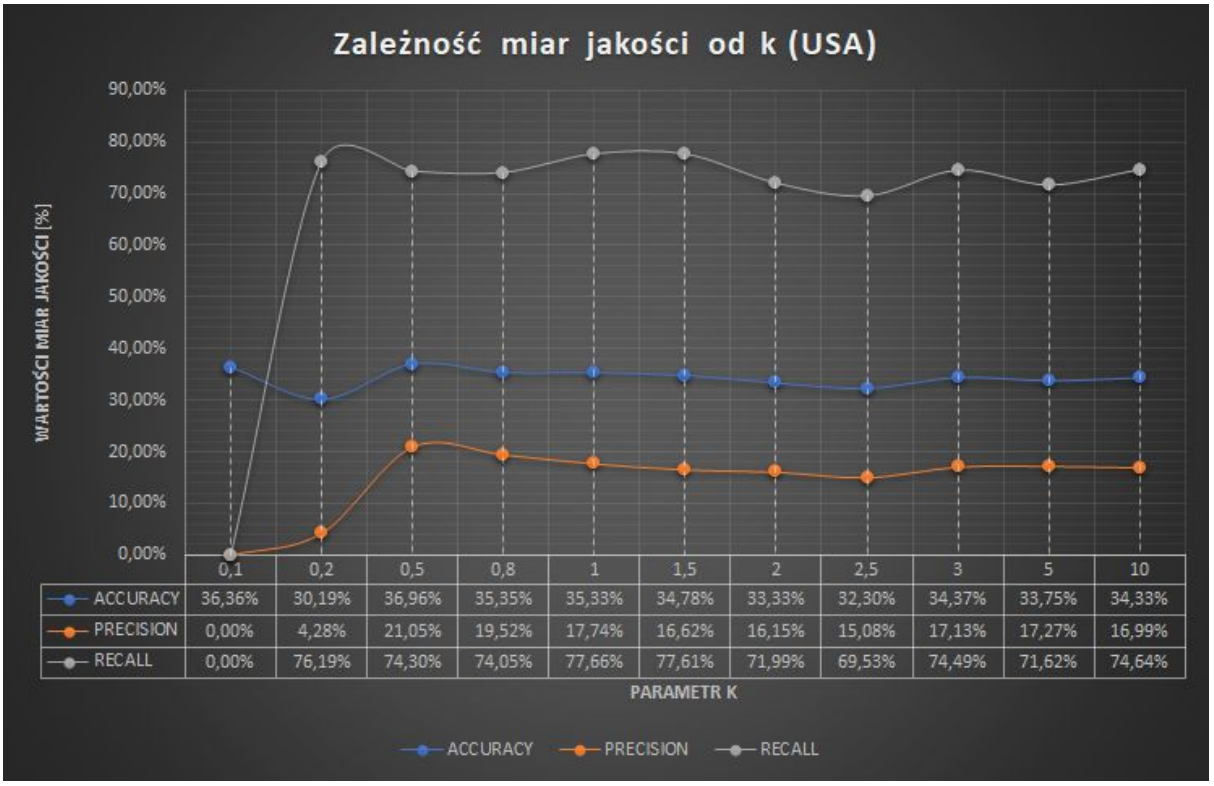
	0,1	0,2	0,5	0,8	1	1,5	2	2,5	3	5	10
ACCURACY	81,82%	77,12%	75,83%	78,33%	80,08%	79,06%	79,82%	80,37%	80,06%	80,85%	80,41%
PRECISION	0,00%	14,71%	21,05%	17,56%	14,93%	20,15%	22,39%	12,69%	18,66%	20,15%	11,94%
RECALL	0,00%	5,26%	4,62%	4,33%	4,16%	5,11%	5,86%	3,62%	5,05%	5,67%	3,43%

Wykres 4. Zależność miar jakości od k dla etykiety “JAPAN”.



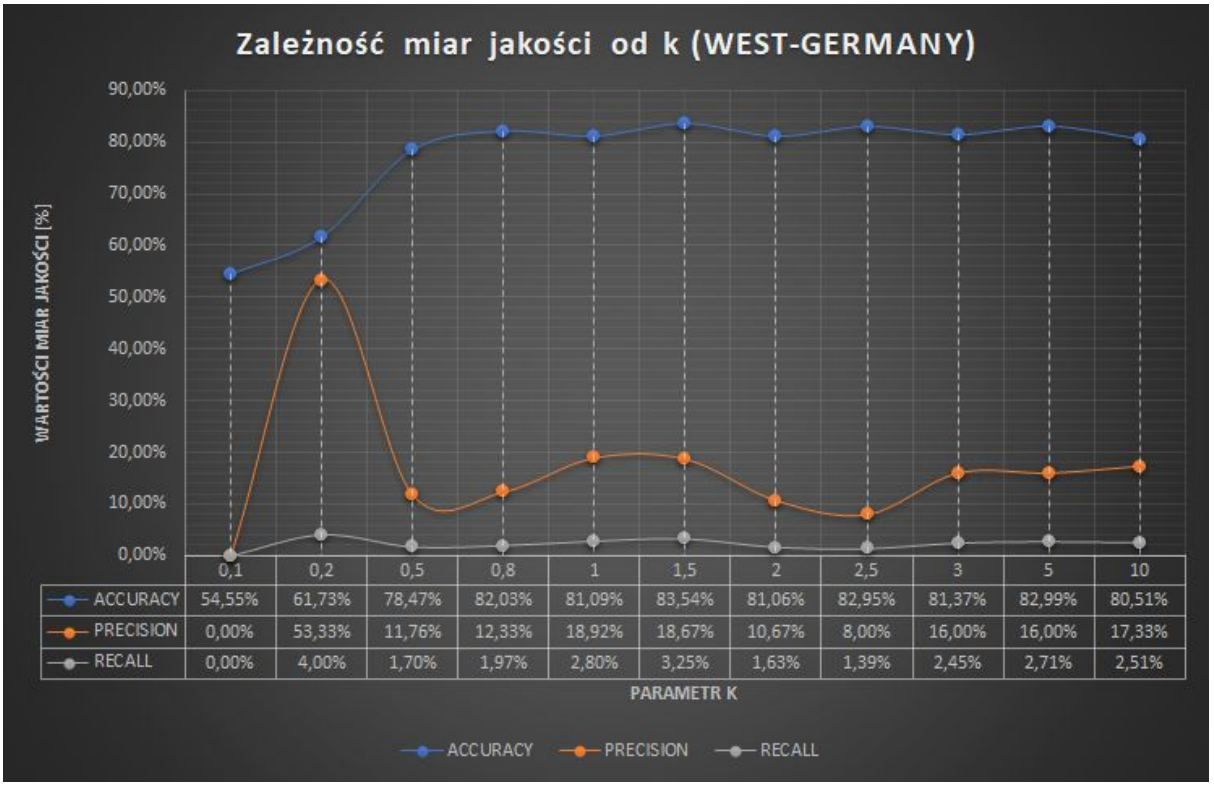
	0,1	0,2	0,5	0,8	1	1,5	2	2,5	3	5	10
ACCURACY	63,64%	69,04%	71,78%	75,24%	75,86%	76,62%	76,00%	74,79%	77,72%	76,55%	76,38%
PRECISION	50,00%	43,10%	23,81%	19,46%	15,67%	19,47%	14,19%	16,50%	16,17%	14,19%	17,16%
RECALL	25,00%	16,34%	12,01%	11,11%	9,57%	11,94%	8,96%	9,45%	11,09%	9,27%	10,68%

Wykres 5. Zależność miar jakości od k dla etykiety “UK”.



	0,1	0,2	0,5	0,8	1	1,5	2	2,5	3	5	10
ACCURACY	36,36%	30,19%	36,96%	35,35%	35,33%	34,78%	33,33%	32,30%	34,37%	33,75%	34,33%
PRECISION	0,00%	4,28%	21,05%	19,52%	17,74%	16,62%	16,15%	15,08%	17,13%	17,27%	16,99%
RECALL	0,00%	76,19%	74,30%	74,05%	77,66%	77,61%	71,99%	69,53%	74,49%	71,62%	74,64%

Wykres 6. Zależność miar jakości od k dla etykiety “USA”.



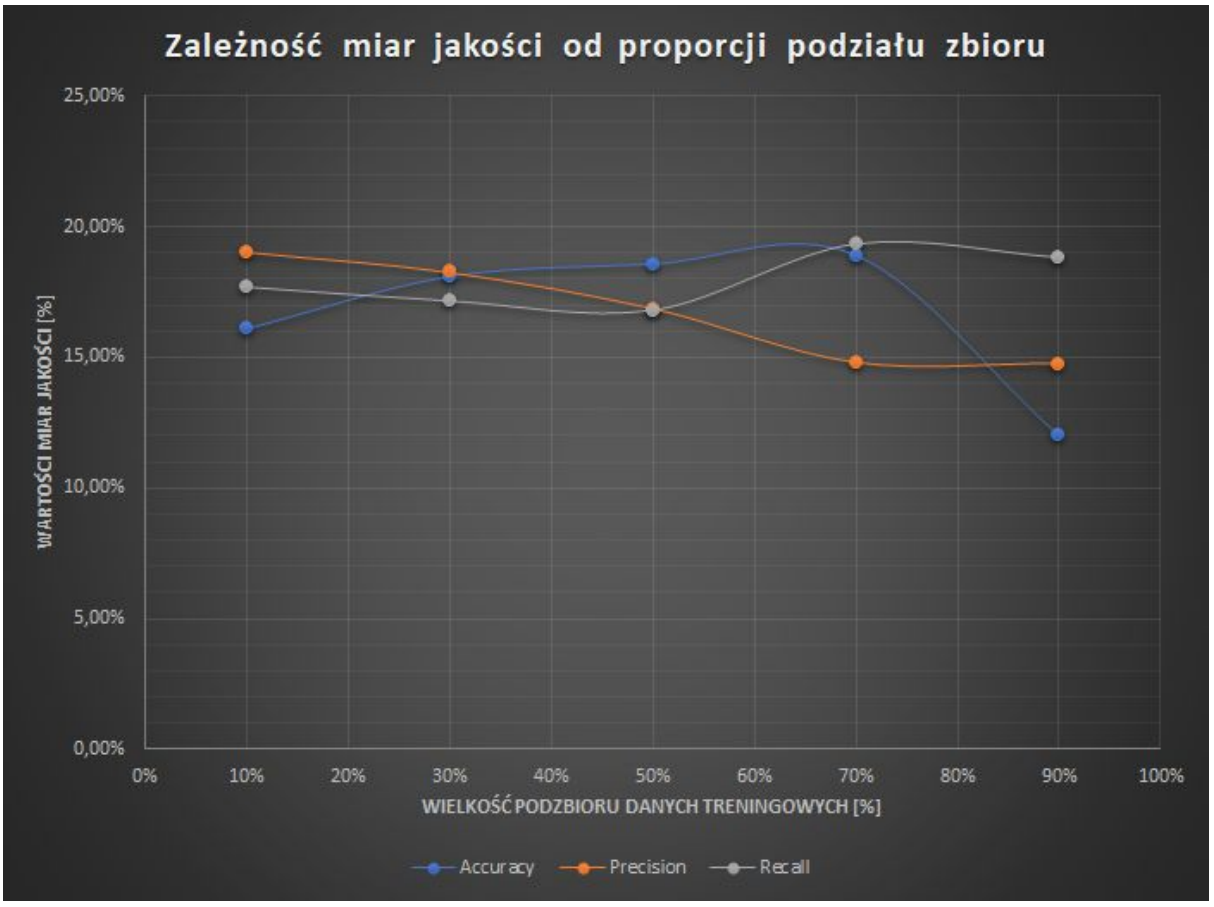
	0,1	0,2	0,5	0,8	1	1,5	2	2,5	3	5	10
ACCURACY	54,55%	61,73%	78,47%	82,03%	81,09%	83,54%	81,06%	82,95%	81,37%	82,99%	80,51%
PRECISION	0,00%	53,33%	11,76%	12,33%	18,92%	18,67%	10,67%	8,00%	16,00%	16,00%	17,33%
RECALL	0,00%	4,00%	1,70%	1,97%	2,80%	3,25%	1,63%	1,39%	2,45%	2,71%	2,51%

Wykres 7. Zależność miar jakości od k dla etykiety “WEST-GERMANY”.

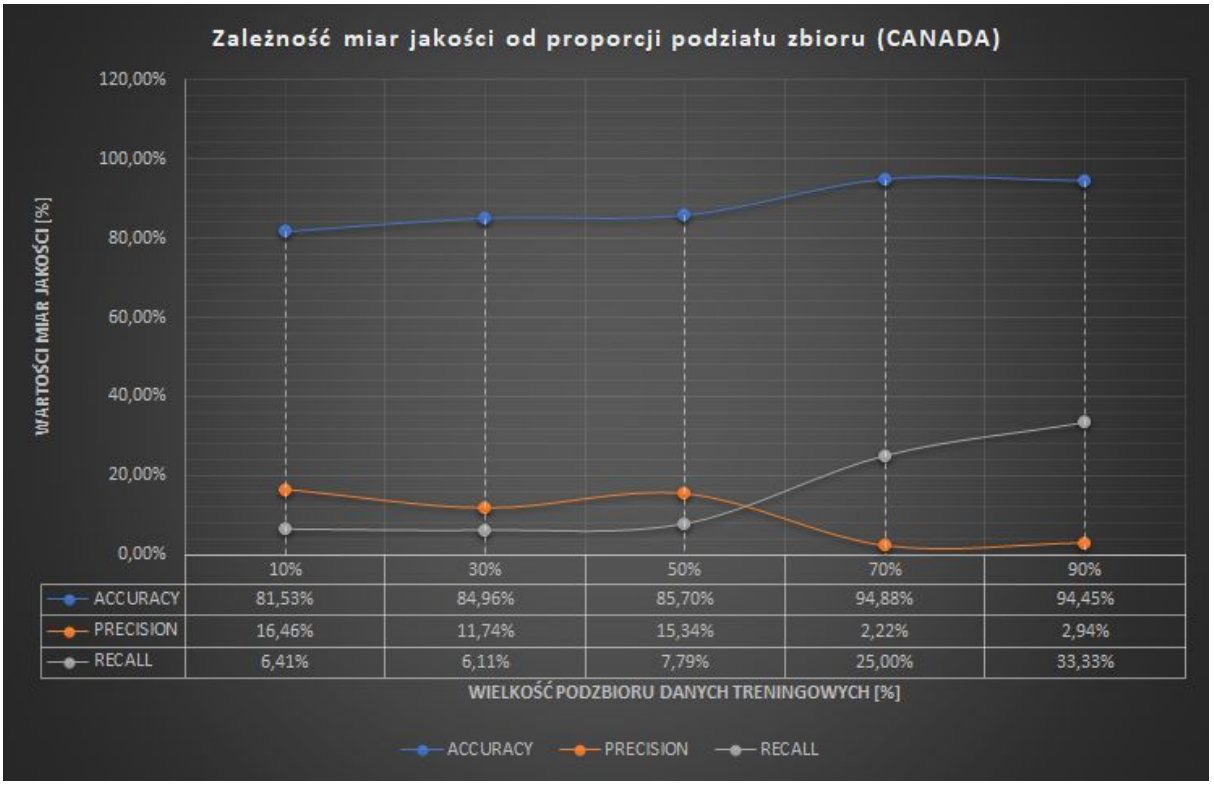
5.2 Eksperyment B

DataBreakdown	10%	30%	50%	70%	90%
Accuracy	16,08%	18,07%	18,55%	18,89%	12,04%
Precision	19,04%	18,26%	16,86%	14,79%	14,75%
Recall	17,69%	17,15%	16,80%	19,35%	18,84%

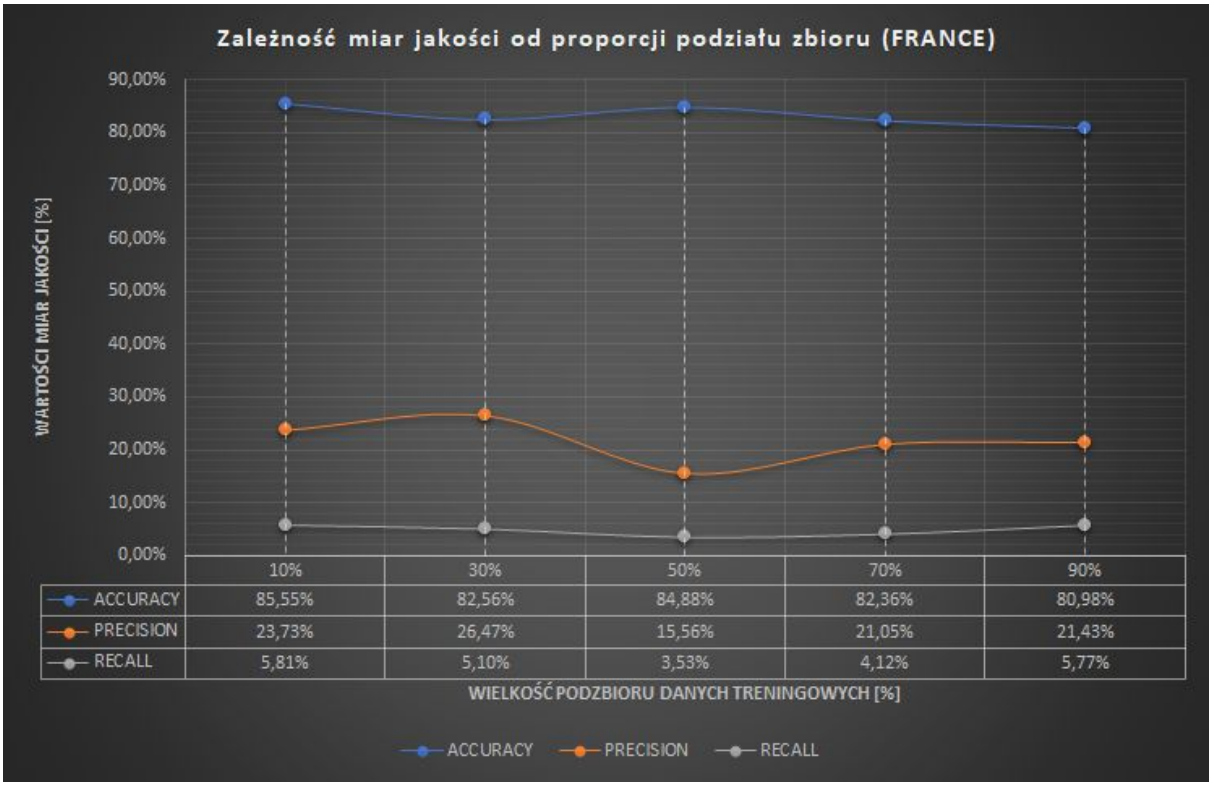
Tabela 8. Porównanie poszczególnych współczynników sprawności klasyfikatora.



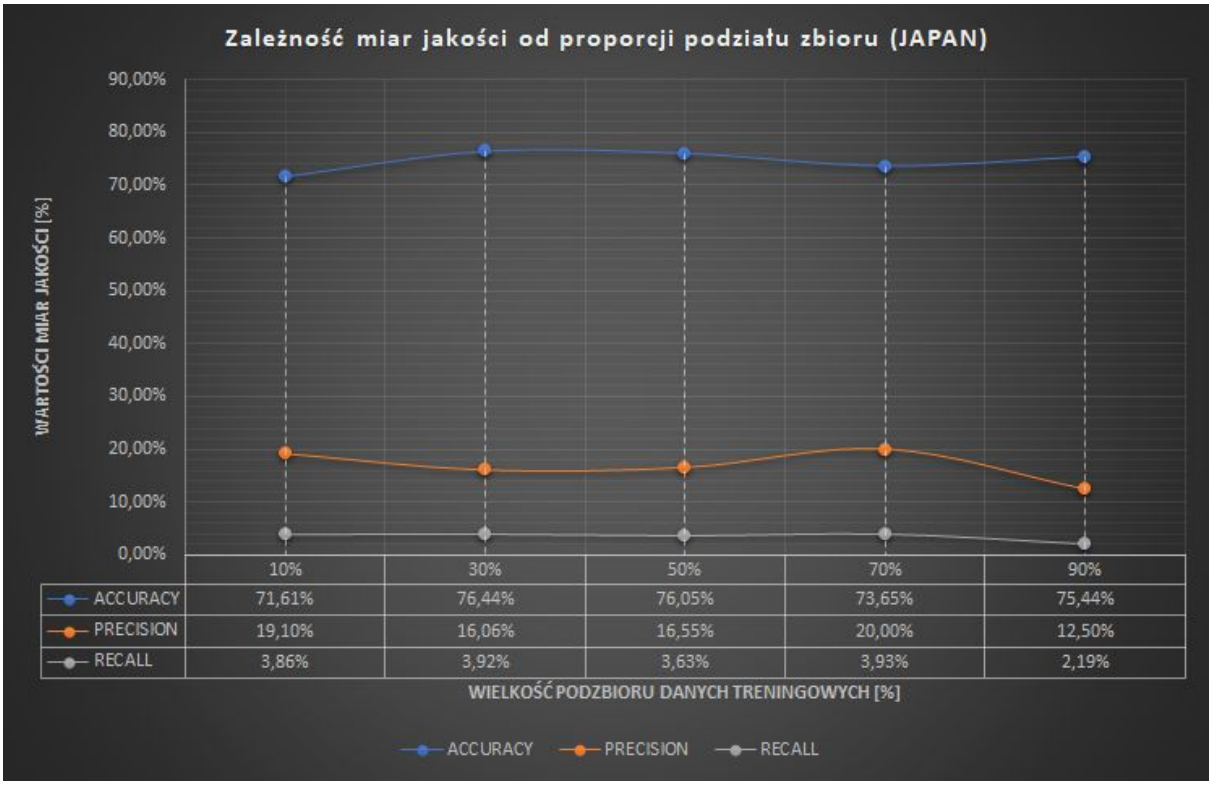
Wykres 8. Zależność miar jakości proporcji podziału zbioru.



Wykres 9. Zależność miar jakości od proporcji podziału zbioru dla etykiety “CANADA”.

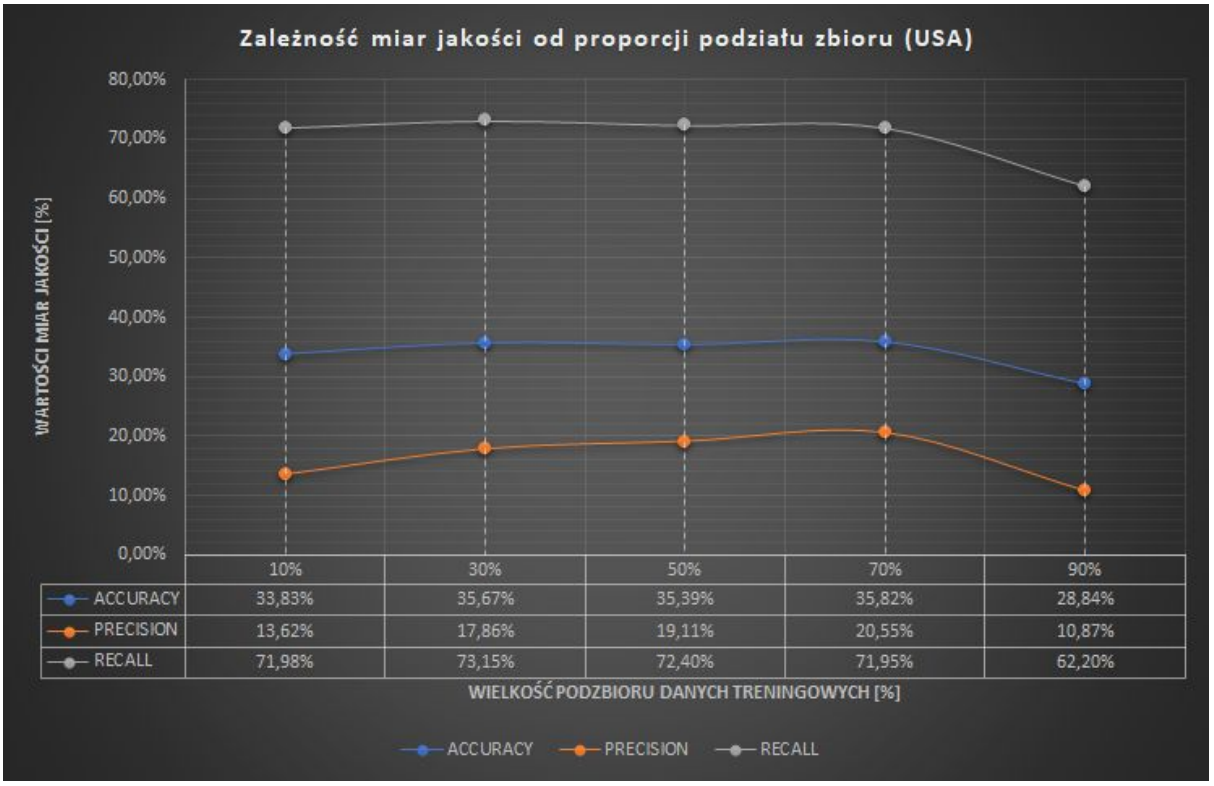


Wykres 10. Zależność miar jakości od proporcji podziału zbioru dla etykiety “FRANCE”.

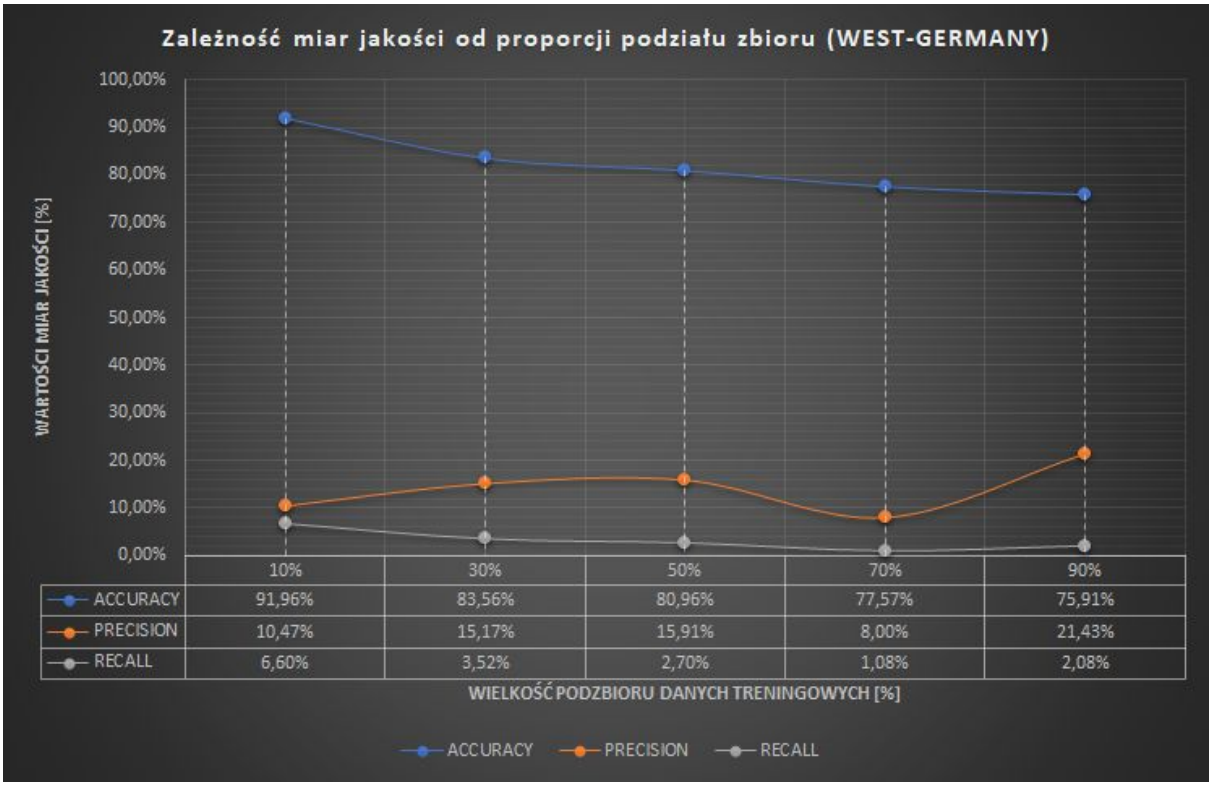


Wykres 11. Zależność miar jakości od proporcji podziału zbioru dla etykiety “JAPAN”.

Wykres 12. Zależność miar jakości od proporcji podziału zbioru dla etykiety “UK”.



Wykres 13. Zależność miar jakości od proporcji podziału zbioru dla etykiety “USA”.

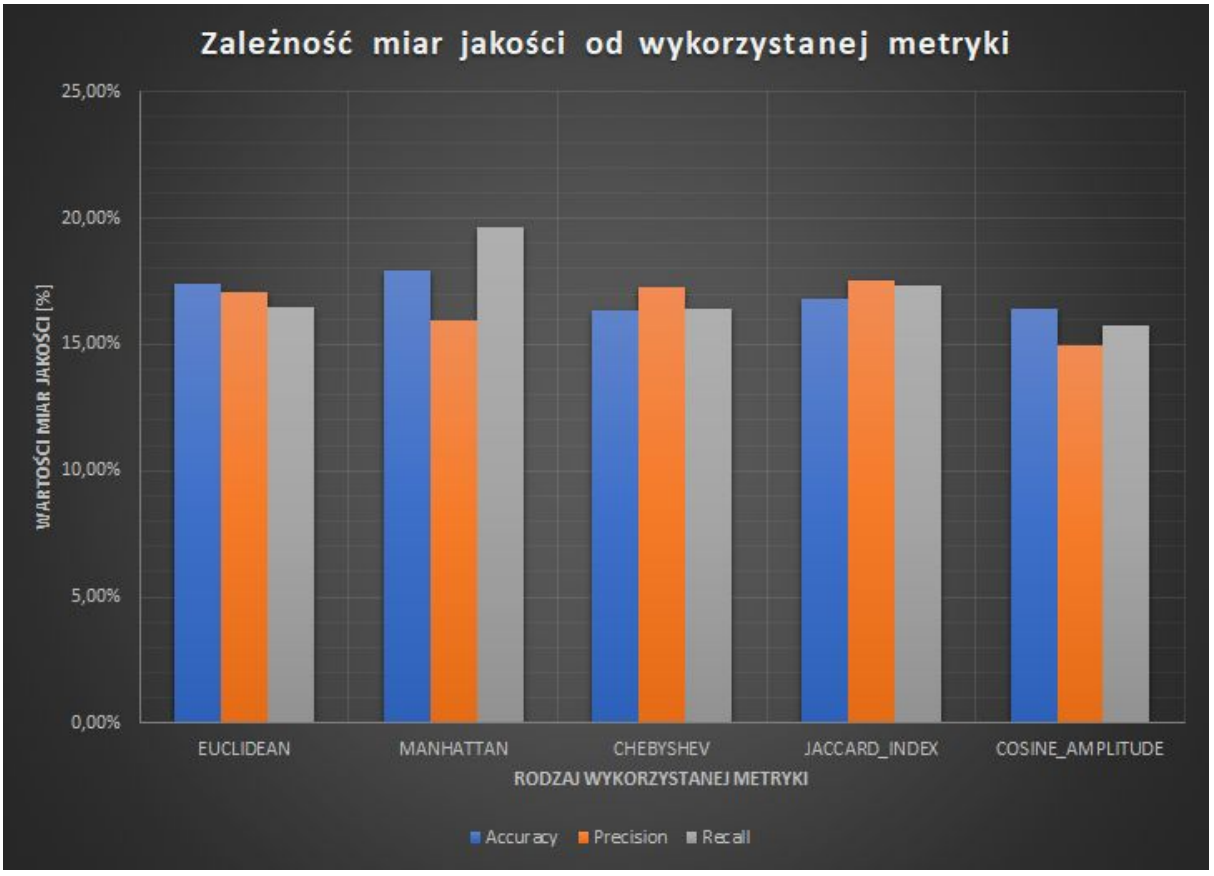


Wykres 14. Zależność miar jakości od proporcji podziału zbioru dla etykiety “WEST-GERMANY”.

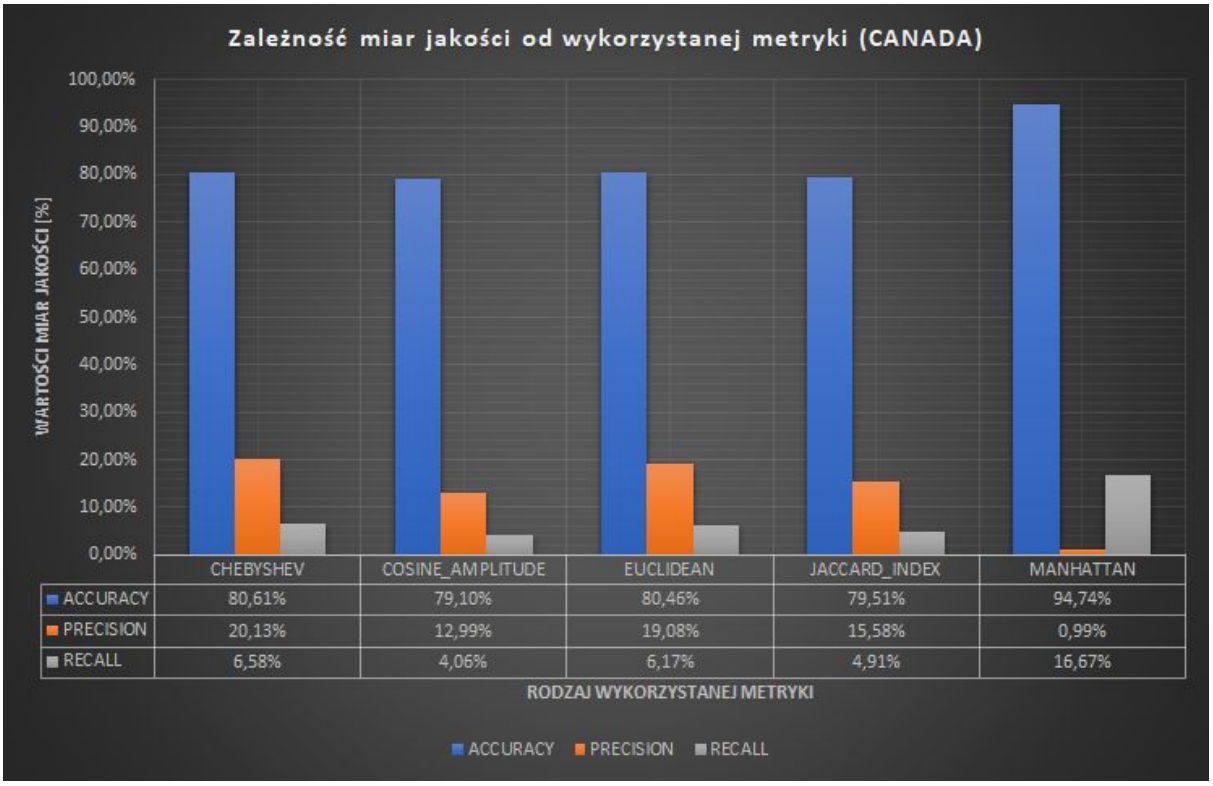
5.3 Eksperyment C

MetricType	EUCLIDEAN	MANHATTAN	CHEBYSHEV	JACCARD INDEX	COSINE AMPLITUDE
Accuracy	17,39%	17,93%	16,32%	16,80%	16,39%
Precision	17,09%	15,98%	17,26%	17,56%	14,93%
Recall	16,50%	19,61%	16,44%	17,32%	15,76%

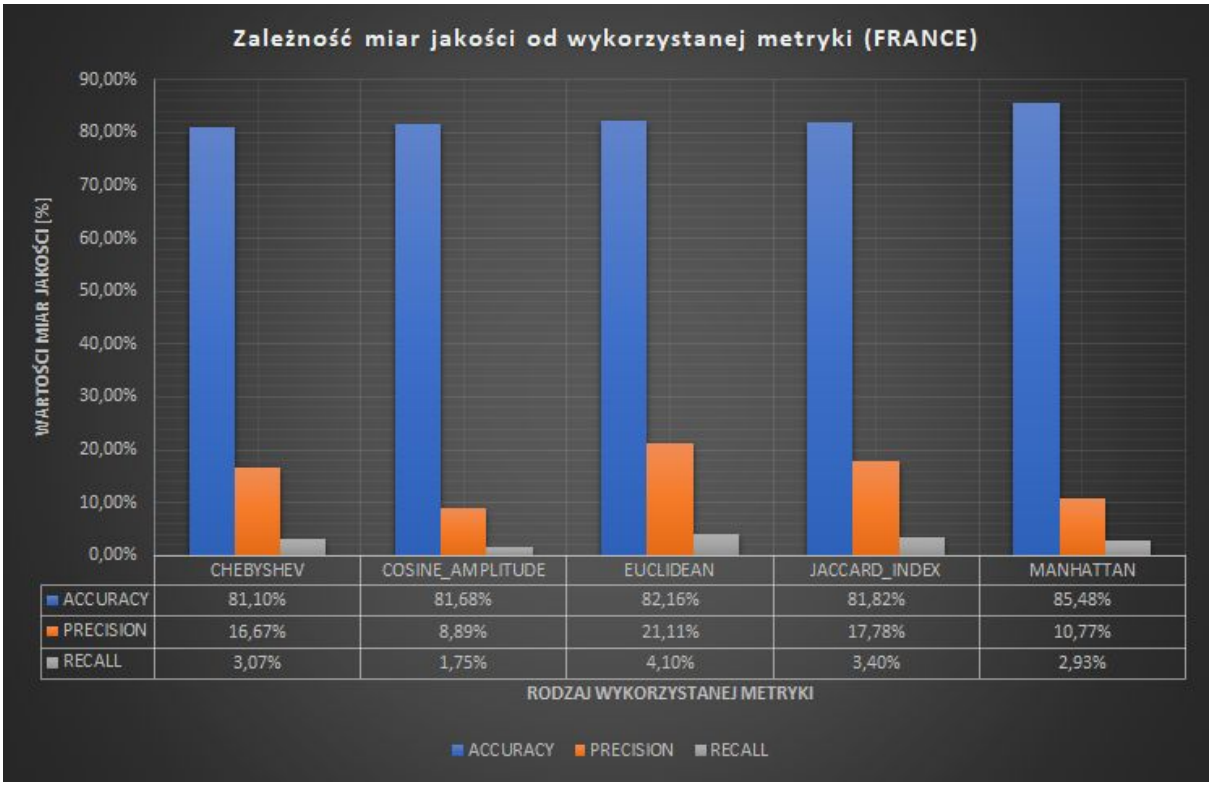
Tabela 9. Porównanie poszczególnych współczynników sprawności klasyfikatora.



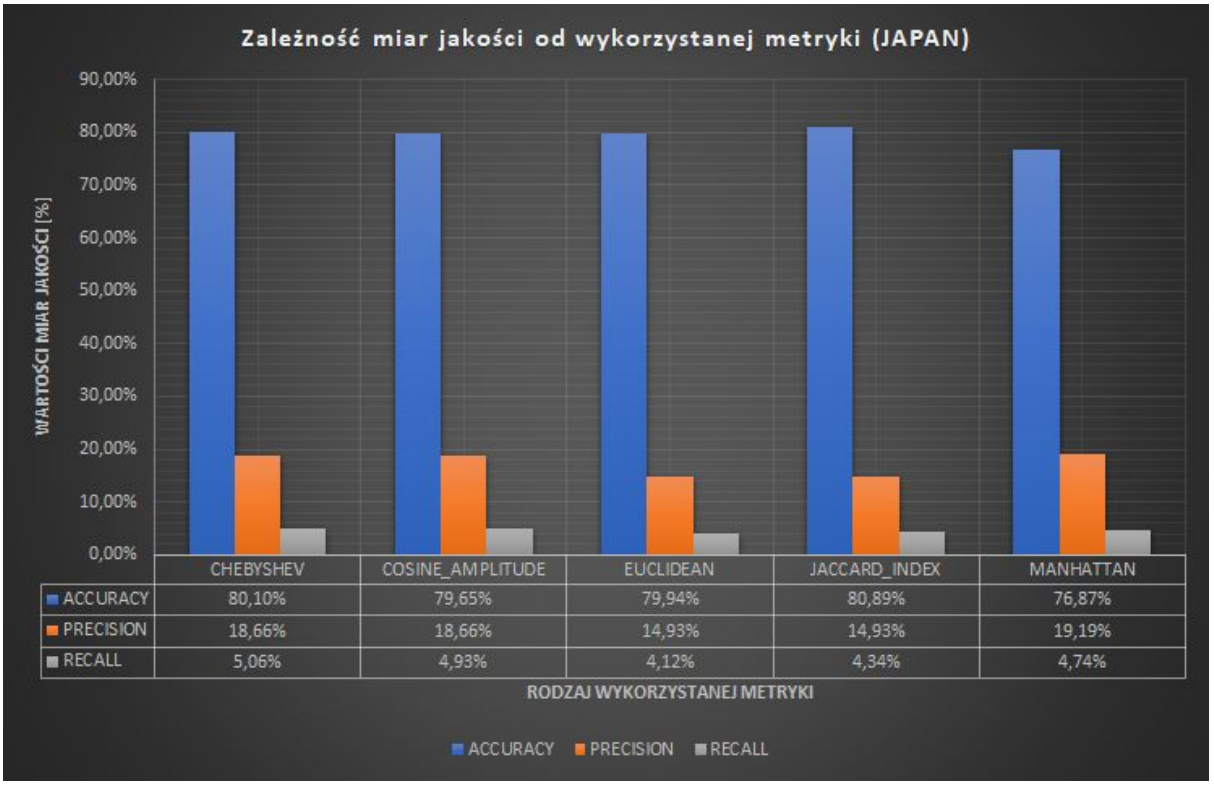
Wykres 15. Zależność miar jakości od wykorzystanej metryki/miary podobieństwa.



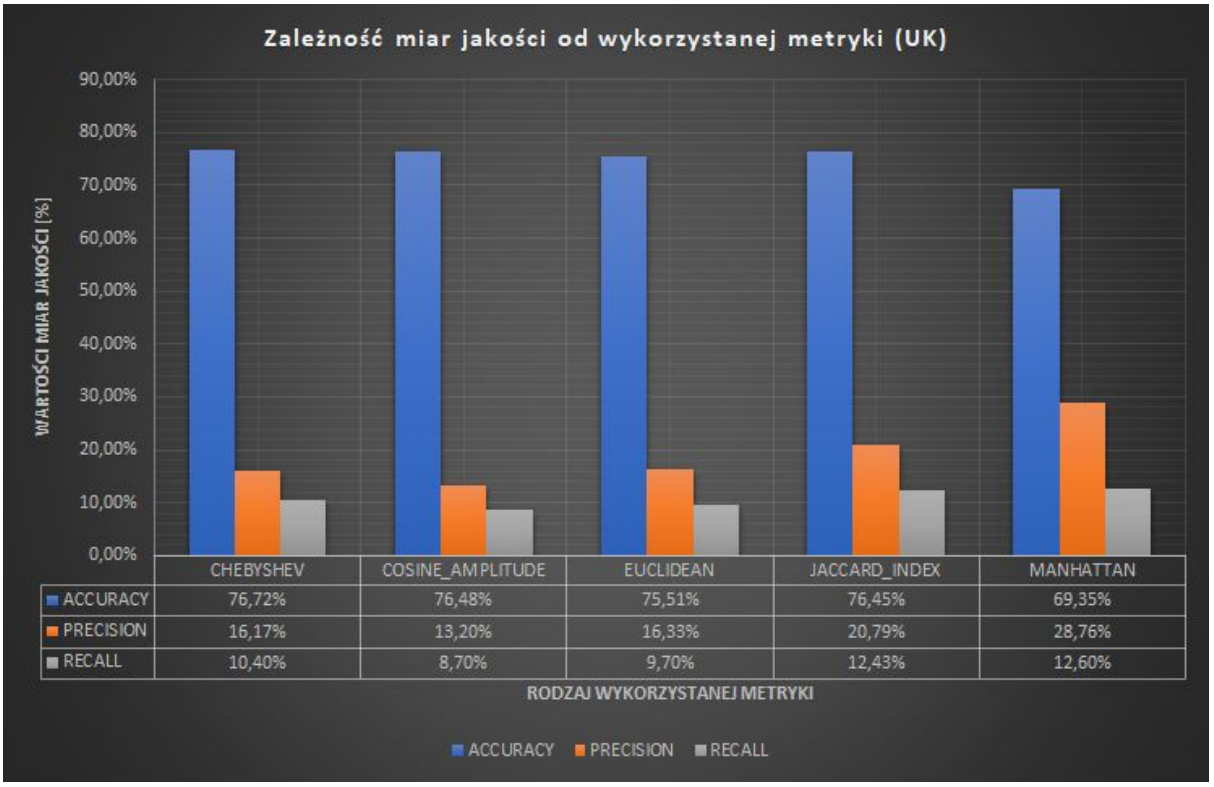
Wykres 16. Zależność miar jakości od wykorzystanej metryki/miary podobieństwa dla etykiety “CANADA”.



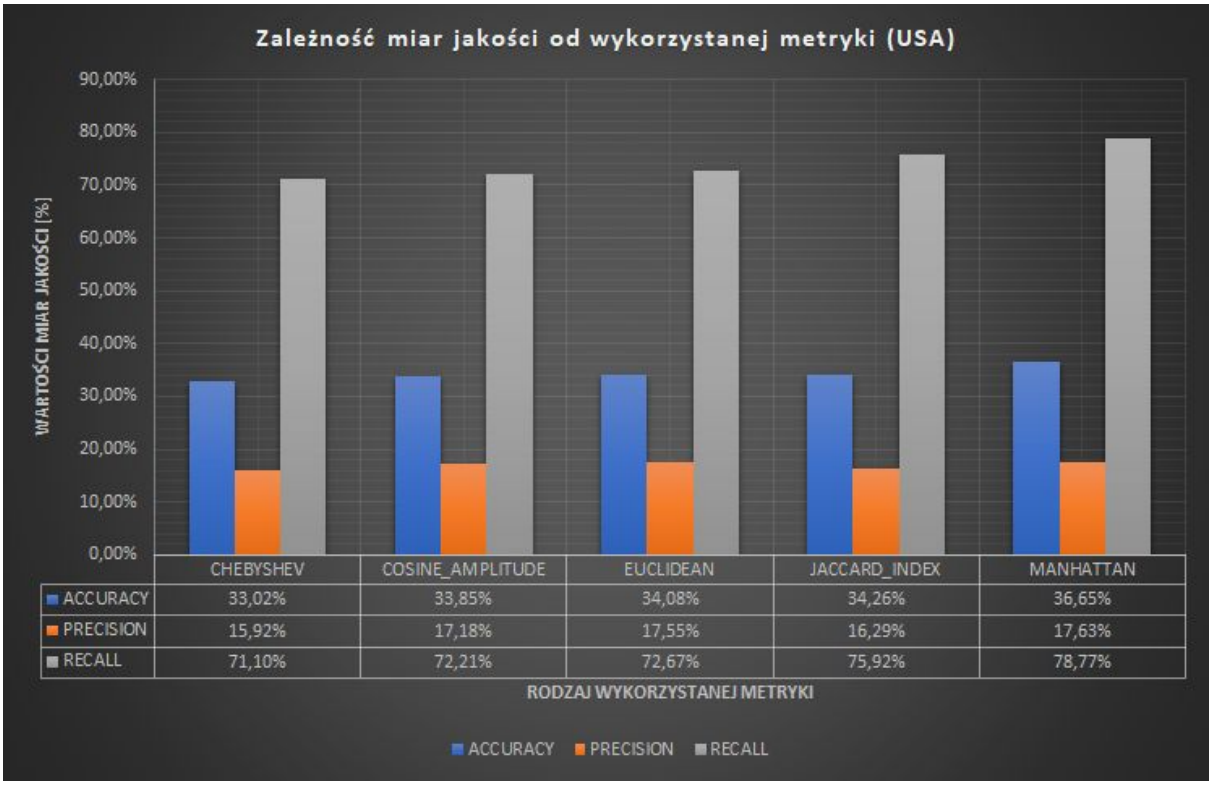
Wykres 17. Zależność miar jakości od wykorzystanej metryki/miary podobieństwa dla etykiety “FRANCE”.



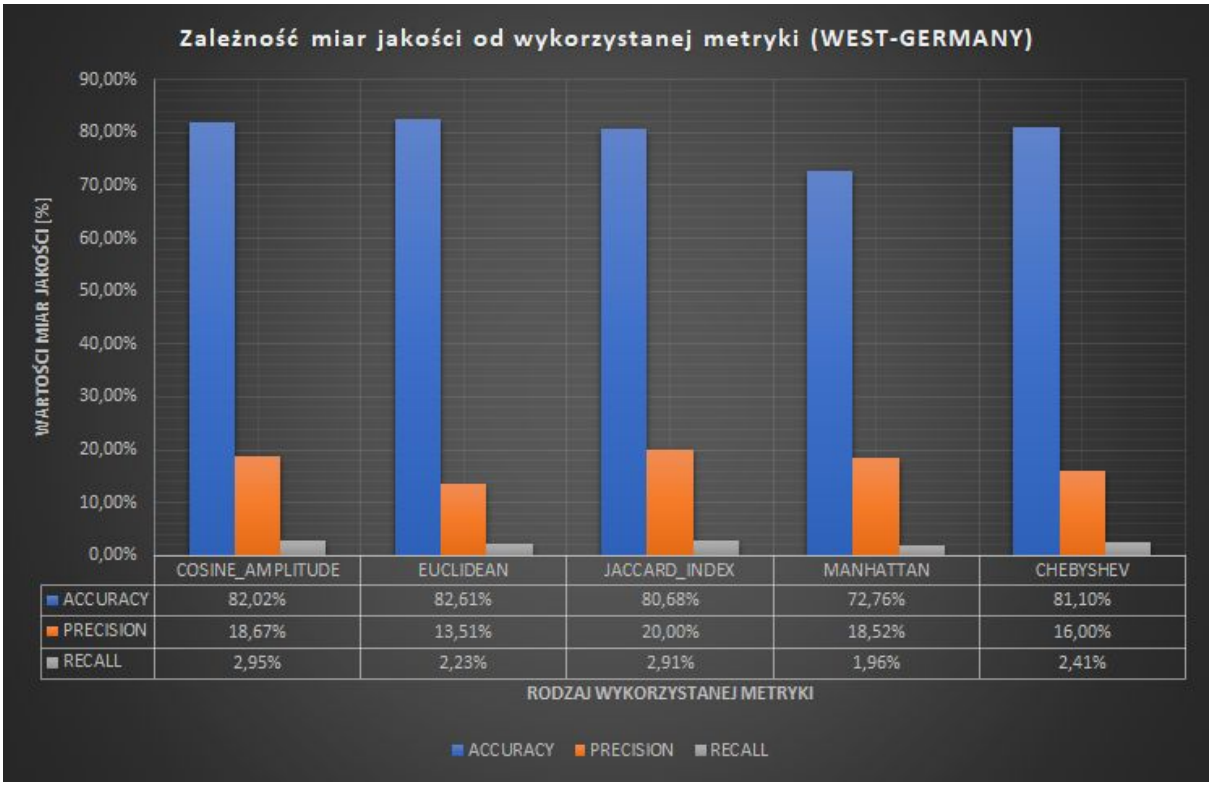
Wykres 17. Zależność miar jakości od wykorzystanej metryki/miary podobieństwa dla etykiety “JAPAN”.



Wykres 18. Zależność miar jakości od wykorzystanej metryki/miary podobieństwa dla etykiety “UK”.



Wykres 19. Zależność miar jakości od wykorzystanej metryki/miary podobieństwa dla etykiety “USA”.



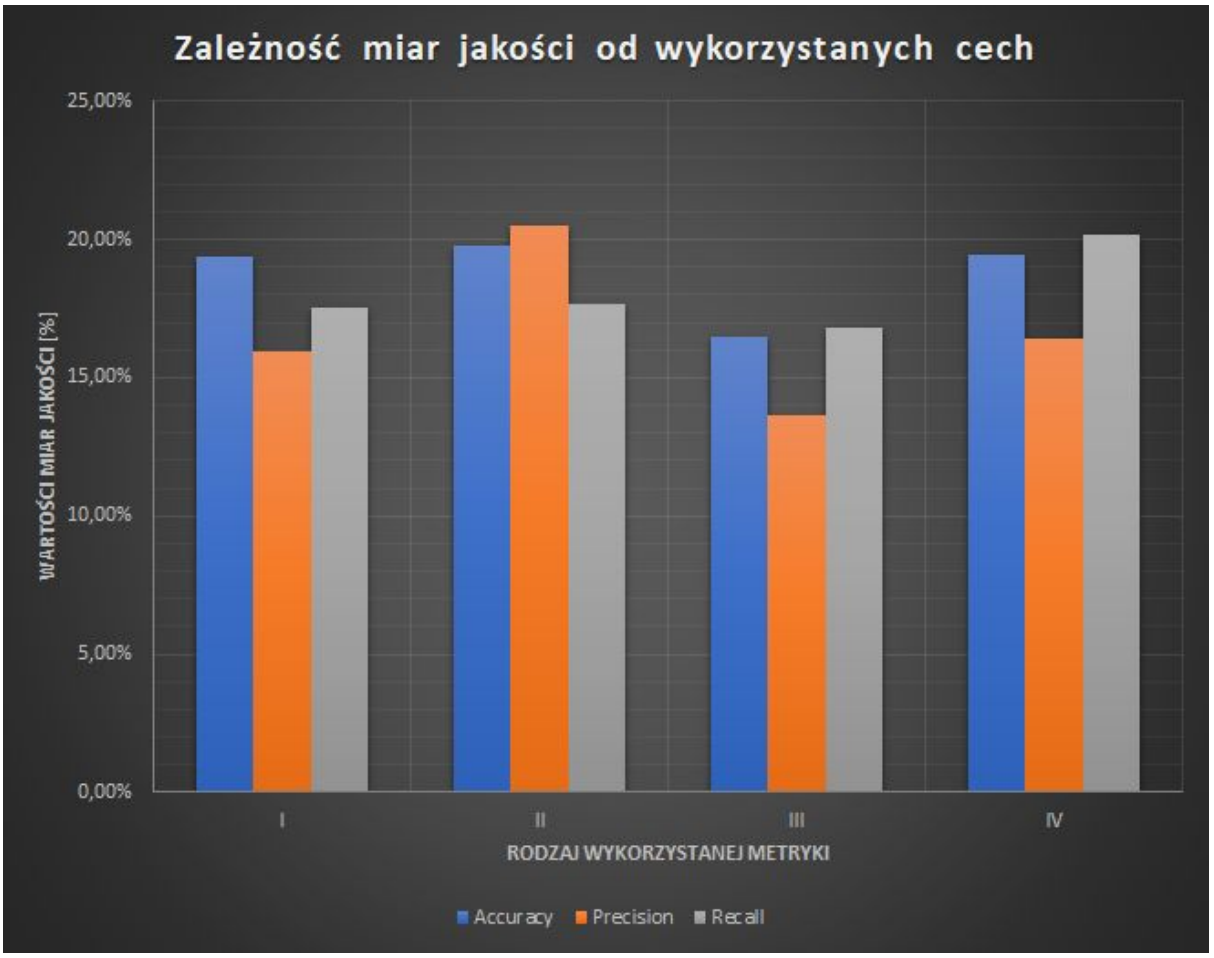
Wykres 20. Zależność miar jakości od wykorzystanej metryki/miary podobieństwa dla etykiety “WEST-GERMANY”.

5.4 Eksperyment D

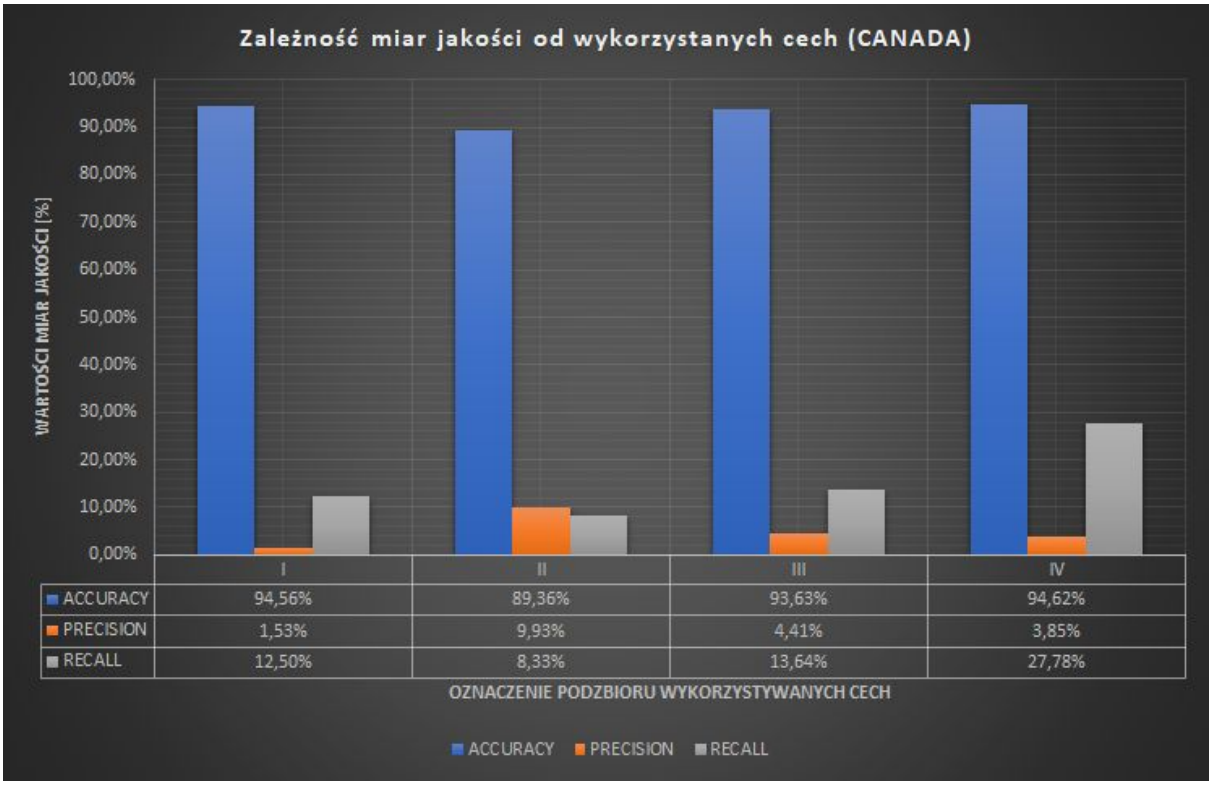
Zgodnie z tabelą podziału zbioru cech na podzbiory (tab. 5) w poniższej tabeli oraz wykresie zastosowano notację rzymską (tj. I, II, III, IV).

FeatureSet	I	II	III	IV
Accuracy	19,36%	19,80%	16,48%	19,42%
Precision	15,96%	20,50%	13,67%	16,40%
Recall	17,52%	17,65%	16,79%	20,16%

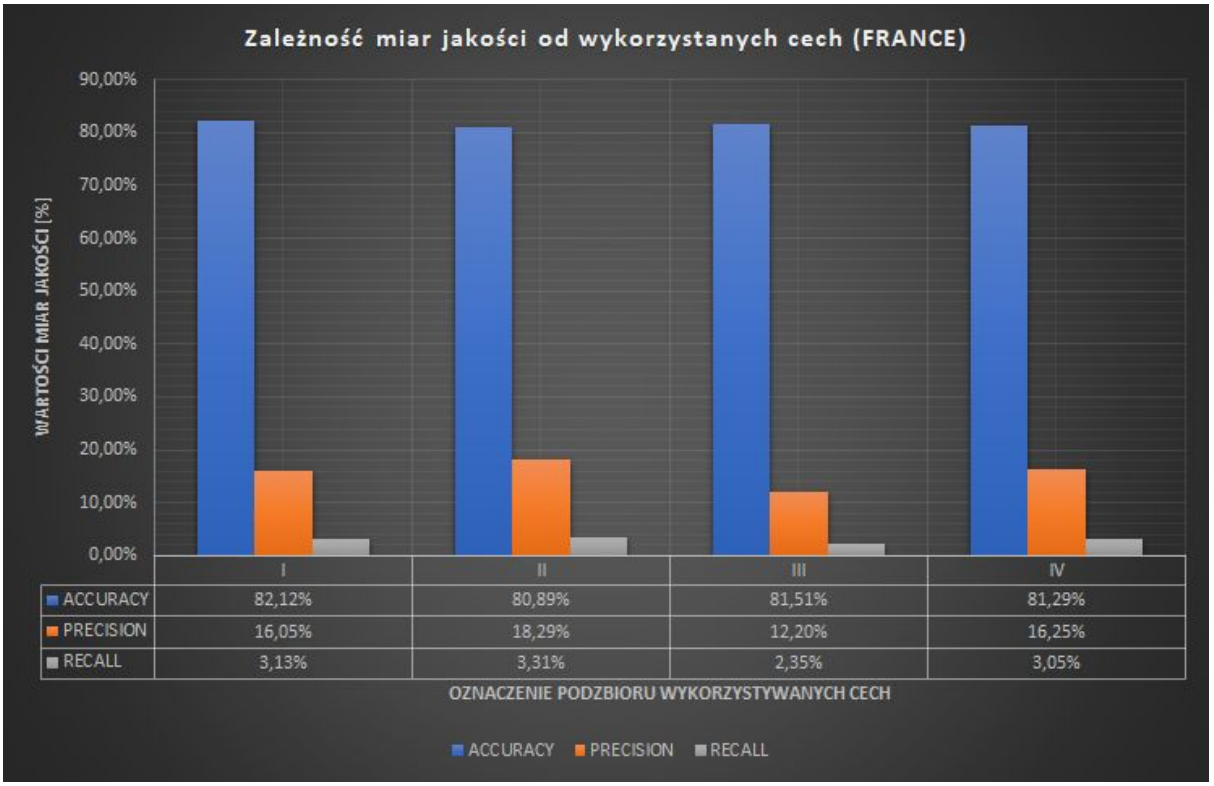
Tabela 10. Porównanie poszczególnych współczynników sprawności klasyfikatora.



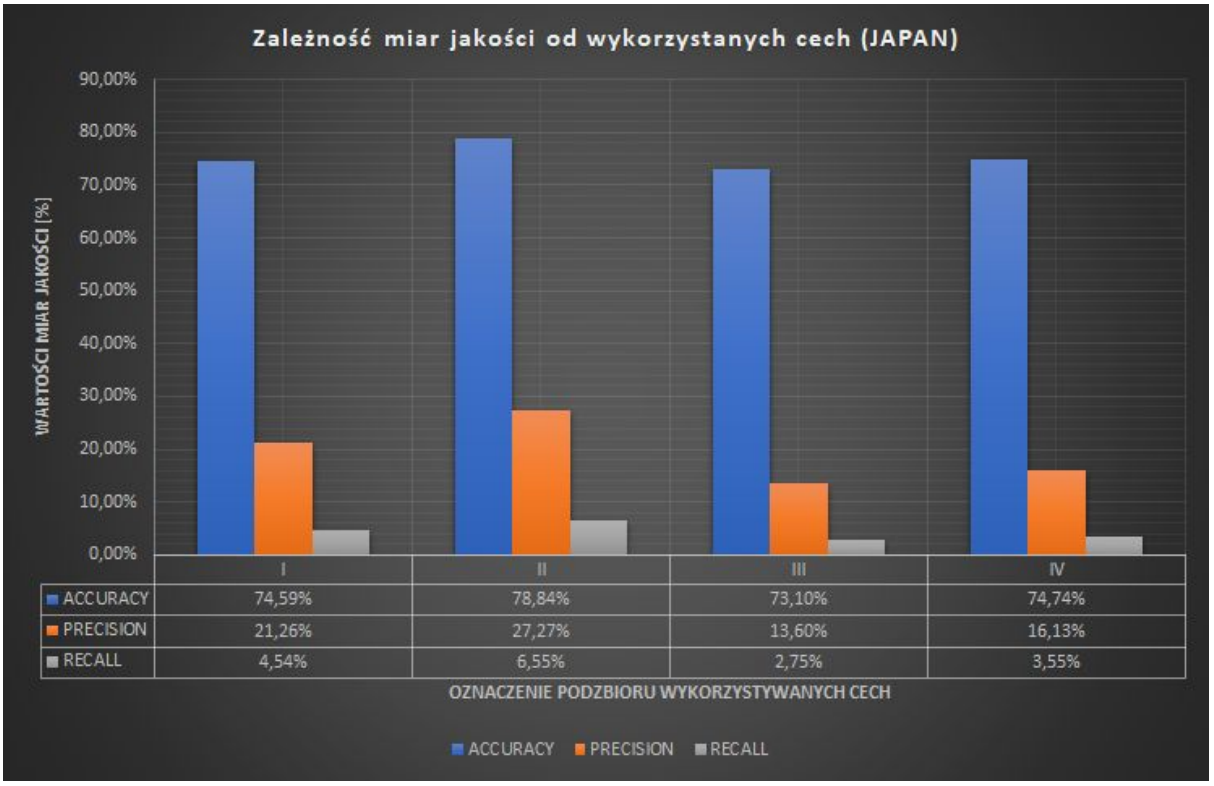
Wykres 21. Zależność miar jakości od wykorzystanych cech.



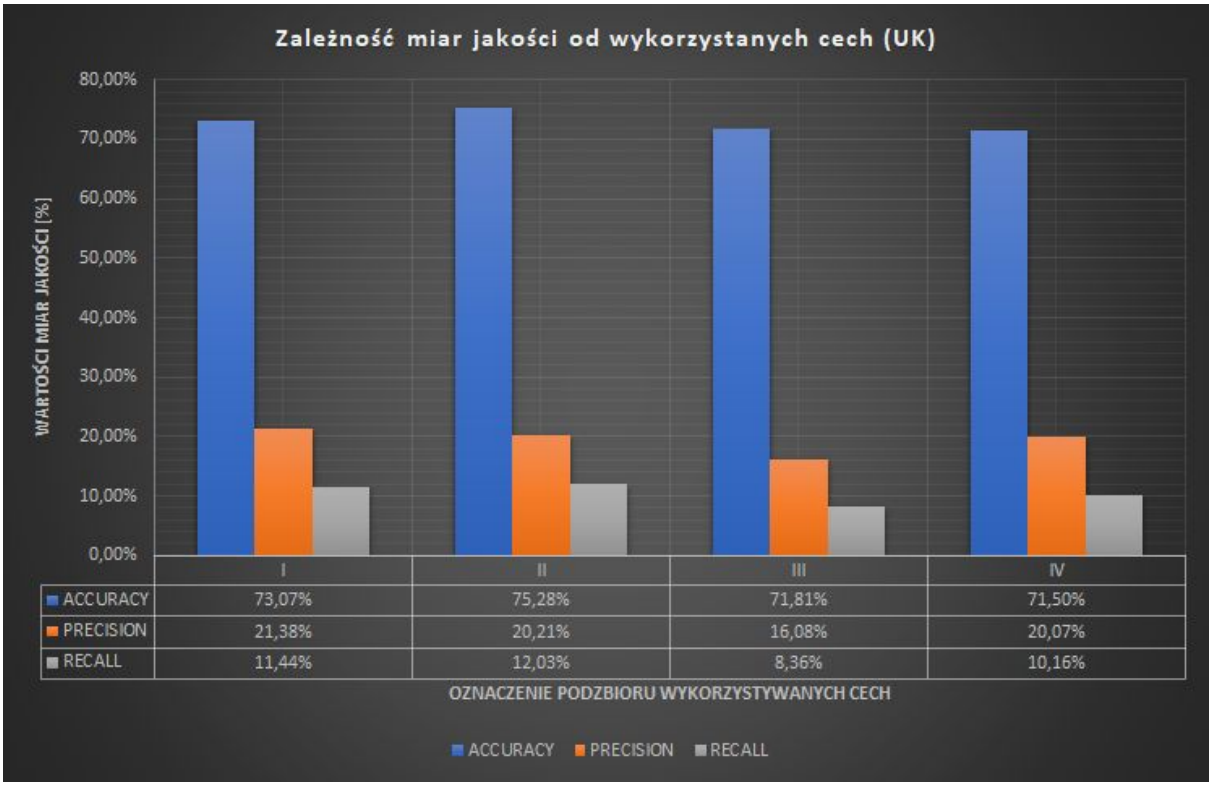
Wykres 22. Zależność miar jakości od wykorzystanych cech dla etykiety “CANADA”.



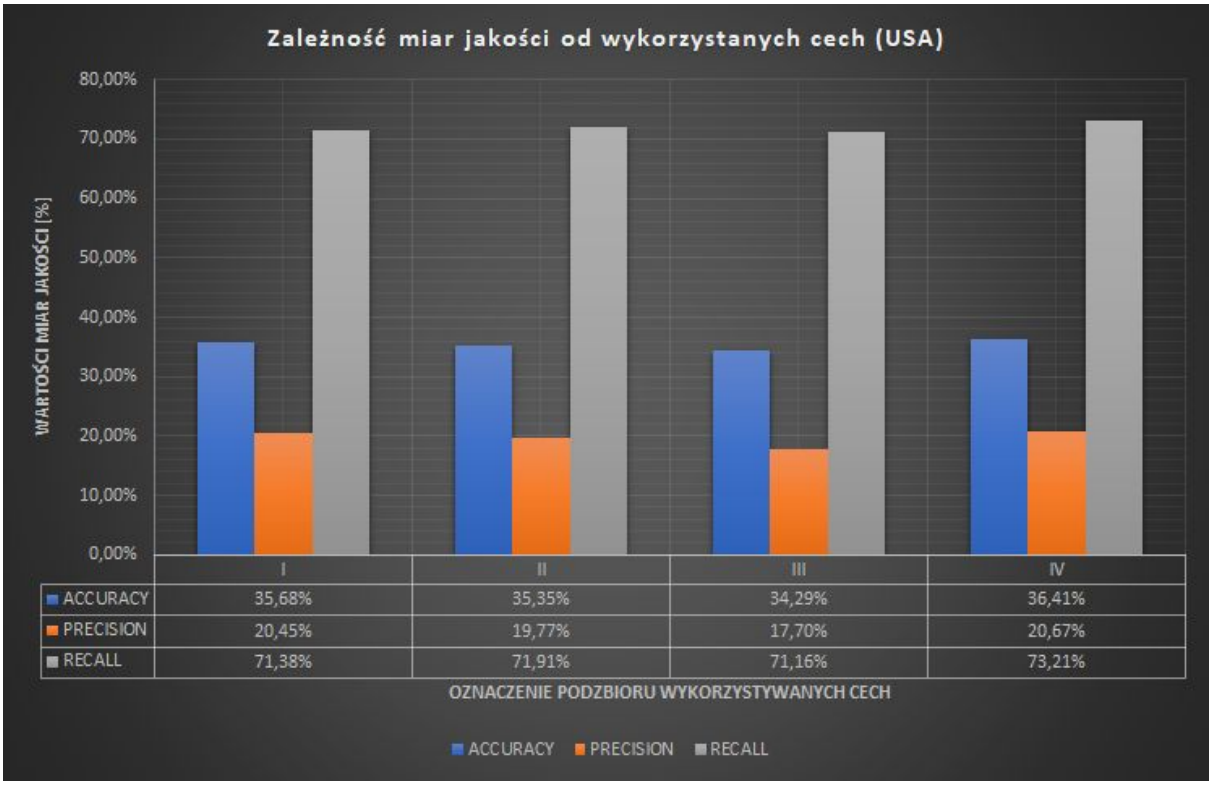
Wykres 23. Zależność miar jakości od wykorzystanych cech dla etykiety “FRANCE”.



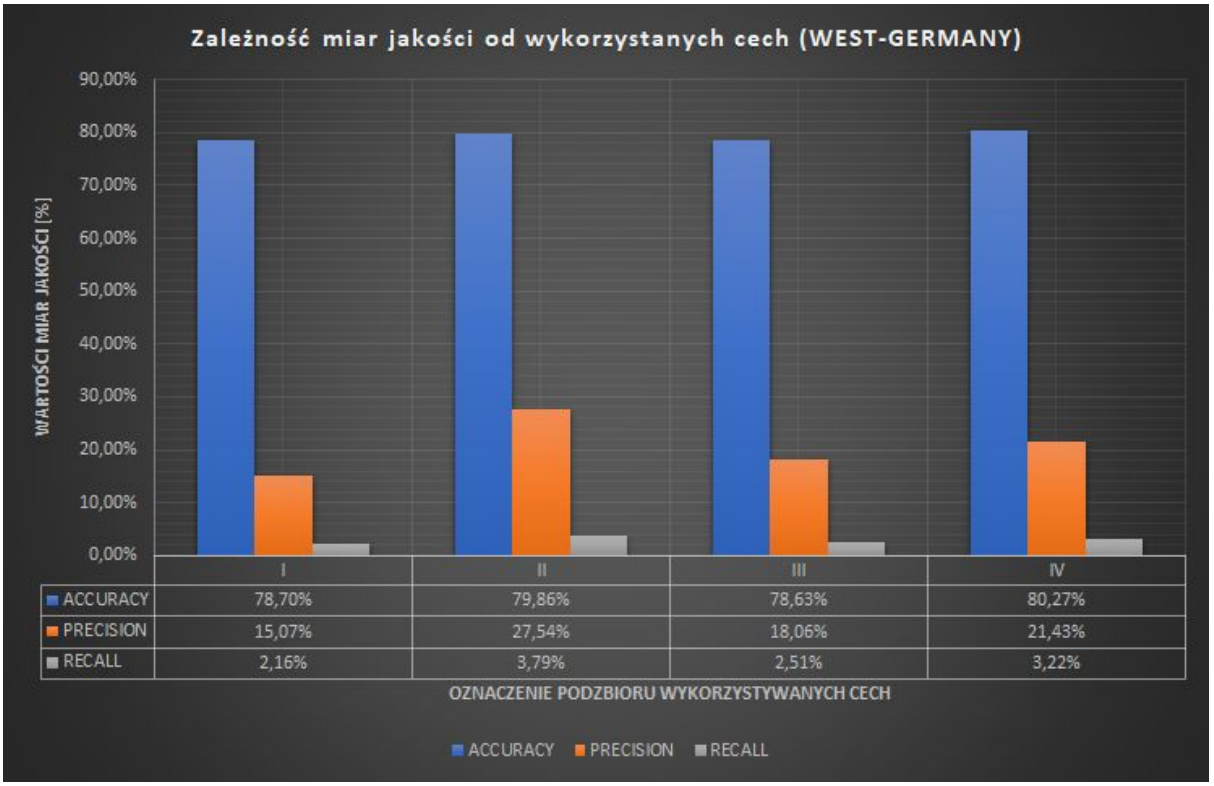
Wykres 24. Zależność miar jakości od wykorzystanych cech dla etykiety “JAPAN”.



Wykres 25. Zależność miar jakości od wykorzystanych cech dla etykiety “UK”.



Wykres 26. Zależność miar jakości od wykorzystanych cech dla etykiety “USA”.



Wykres 27. Zależność miar jakości od wykorzystanych cech dla etykiety “WEST-GERMANY”.

6. Dyskusja

Naszym głównym celem było zbadanie wpływu parametrów na wyniki otrzymywane podczas przeprowadzania eksperymentów. W związku z tym przeprowadzaliśmy 4 eksperymenty, w każdym sprawdzając inne warunki klasyfikatora tj:

- *wpływ parametru k na klasyfikację*
- *wpływ podziału zbioru uczącego na klasyfikację*
- *wpływ wyboru metryki lub miary podobieństwa na klasyfikację*
- *wpływ podzbiorów cech na wyniki klasyfikacji*

Eksperymenty te przeprowadzaliśmy dla całego zbioru artykułów jak i również dla poszczególnych zbiorów oznaczonych odpowiednimi etykietami państw.
W ramach każdego eksperymentu testowaliśmy różne wartości atrybutów aby ostatecznie ustalić wartość najbardziej optymalną lub taką która daje najlepsze wyniki.

Zbiór wszystkich artykułów

Wpływ parametru k na klasyfikację

Doświadczenie przeprowadzone w eksperymencie A, które miało na celu zbadanie wyników w zależności od parametru k przyjmującego wartości w zakresie od 0.1 do 10, pozwoliło nam ustalić najlepszy dobór parametru k do klasyfikacji. Z danych otrzymanych przez nas można odczytać, że najlepsza wartość parametru wynosi 0.5 dla większości metryk. Wartość Accuracy dla tego parametru jest największa wśród wszystkich otrzymanych wyników. Można również zauważyć zależność, iż wraz ze wzrostem parametru k dane statystyczne Accuracy, Precision i Recall przybierają coraz to bardziej zbliżone do siebie wartości.

Wpływ podziału zbioru uczącego na klasyfikacje

W przypadku eksperymentu B, badaliśmy wpływ proporcji danych uczących do danych testowych przyjmujących wartości: 10%, 30%, 50%, 70% oraz 90%. Otrzymane przez nas wyniki, przeważnie nie posiadały dużych rozbieżności. Po dokładnej weryfikacji wyników możemy przypuszczać, że w przypadku zbioru uczącego na poziomie 90% nastąpiło przeuczenie, w konsekwencji ocena klasyfikatora jest zauważalnie słabsza. Zaś najlepszą wartością parametru DataBreakdown jest wartość na poziomie 30%, w którym to wyniki danych statystycznych osiągają poniekąd największe wartości.

Wpływ wyboru metryki lub miary podobieństwa na klasyfikację

W kolejnym eksperymencie badaliśmy, która metryka pozwala nam na otrzymanie najlepszych wyników klasyfikacji. Z wyjątkiem metryki Manhattan, metryki przyjmują podobne wartości danych statystycznych Accuracy, Precision, i Recall. Podobieństwo kosinusowe przyjmuje nieznacznie mniejsze wartości przez co można powiedzieć, że jest to dla metody k-NN najgorszy z badanych sposób obliczania odległości w przestrzeni

n-wymiarowej, który wpływa bezpośrednio na klasyfikację. Metryką która się wyróżnia, oraz posiada największe wartości otrzymane danych statystycznych jest wyżej wspomniana metryka Manhattan. Możemy powiedzieć że dzięki najwyższym współczynnikom w całych badaniach Accuracy i Recall jest ona najlepszą metryką do przeprowadzania klasyfikacji. Wartości cech wyliczone dla tekstów są zazwyczaj bardzo zbliżone do siebie, więc jest to zapewne główny powód przewagi metryki ulicznej nad pozostałymi.

Wpływ podzbiorów cech na wyniki klasyfikacji

W celu przeanalizowania wpływu pewnych cech na wyniki klasyfikacji, podzieliliśmy wszystkie cechy na 4 podzbiory. Dokładny opis cech znajduje się w rozdziale 4.2, a podział cech na zbiory w rozdziale 4.3. Z otrzymanych przez nas danych możemy odczytać, że najlepszymi cechami jest zbiór drugi. Przyjmuje on największe wartości parametrów Accuracy oraz Precision. Jest to zbiór który przyjmuje poniekąd najmniej cech, jednak te cechy najbardziej różnią się od siebie. Zbiorem, który przyjmuje najsłabsze wyniki klasyfikacje jest zbiór trzeci. Posiada on najwięcej cech oraz cechy te w głównej mierze opierają się na liczbie słów kluczowych w pewnych fragmentach tekstu, gdzie bierzemy pod uwagę pewne części artykułu ostatecznie wyciągając średnią ze wszystkich wartości. Pozwala to nam przypuszczać, że cechy te są mało efektywne.

Zbiory artykułów według etykiet państw

Wpływ parametru k na klasyfikację

Wpływ parametru k na klasyfikację nie jest jednoznaczny w przypadku artykułów z różnymi etykietami państw. Dla niektórych państw rosnący współczynnik powoduje wyższy procent poprawnie zakwalifikowanych artykułów, dla innych powoduje spadek efektywności kwalifikacji a jeszcze dla innych procent utrzymuje się na tym samym poziomie. Jedną z kluczowych uwag jest fakt iż współczynnik Accuracy utrzymuje się przeważnie w okolicy 80%, wyjątkiem od tej reguły jest etykieta “USA” której granica to 40%. Analizując wykresy można powiedzieć z lekką nutą niepewności że najlepszą wartością do klasyfikacji jest wartość współczynnika 0.5.

Wpływ podziału zbioru uczącego na klasyfikacje

Podział zbiorów na treningowe i testowe, nie posiada bardzo istotnego wpływu na wyniki klasyfikacji. Różnice w zależności od podziału zbioru wahają się od 3% do maksymalnie 10%. Ponownie można zauważyć zależność iż wartości poprawnej klasyfikacji jest zdecydowanie mniejsza dla etykiety USA, niż dla innych państw. W tym przypadku klasyfikacja zwraca o połowę mniej poprawnych wartości. Analizując dane można powiedzieć że największy procent poprawnych klasyfikacji osiągnięto najczęściej przy 50% następnie przy 10% a raz przy 90%.

Wpływ wyboru metryki lub miary podobieństwa na klasyfikację

Analizując otrzymane wyniki można powiedzieć że najlepszą metryką w naszym przypadku jest metryka Manhatan. Osiąga ona niebywale dużą wartość poprawnych klasyfikacji na poziomie prawie 95% dla etykiety CANADA. Pozostałe metryki otrzymują standardowe wartości średnie jakie były otrzymywane podczas innych eksperymentów. Wartości wszystkich etykiet oscylują wokół wartości 80%, z wyjątkiem etykiety USA której wartości spadają do 30%.

Wpływ podzbiorów cech na wyniki klasyfikacji

Wpływ podzbiorów cech na wyniki nie jest dość istotnym czynnikiem, dlatego podczas tego eksperymentu wyniki otrzymane dla różnych etykiet są dość podobne do wyników otrzymywanych podczas poprzednich doświadczeń z poszczególnymi etykietami. Można z nich odczytać iż najlepiej klasyfikowane są artykuły z etykietą CANADA , a najgorzej tak jak w poprzednich eksperymentach artykuły z etykietą USA. Otrzymane wyniki są zbliżone do poprzednich eksperymentów, i nie posiadają istotnych różnic.

7. Wnioski

- Wartość parametru k wraz z wybraną metryką podobieństwa są najważniejszymi parametrami wpływającymi na wyniki klasyfikacji.
- Wartości parametru k po przekroczeniu wartości równej 3.0 mają niski wpływ na klasyfikację i mogą jedynie wydłużyć czas potrzebny na wykonanie obliczeń.
- Wektor cech nie może składać się z pojedynczych cech jak i również tych cech nie może być zbyt dużo dla optymalnej klasyfikacji.
- Cechy, które głównie ograniczają się do ilości słów w pewnym fragmencie tekstu są mało efektywne (zapewne jest to spowodowane tym, iż końcowa wartość cechy jest średnią arytmetyczną pewnych wartości).
- Zbiór uczący musi zostać odpowiednio dobrany aby nie doszło do niedouczenia czy też przeuczenia klasyfikatora.
- Skuteczność klasyfikatora oscyluje w granicy od 15 do 20 procent co może głównie świadczyć o źle dobranych cechach lub o źle dobranych innych parametrach podczas przeprowadzania eksperymentu. Jednakże na wszystkie wyniki ogromny wpływ miały przeważające w bazie danych artykuły z etykietą “USA”, które znacznie pogorszyły skuteczność klasyfikatora.
- Możliwe, iż niektóre cechy dają lepsze efekty przy klasyfikacji dla konkretnie dobranych pozostałych parametrów klasyfikacji. Uzyskanie optymalnego algorytmu i kombinacji “cecha” - “pozostałe parametry” jest ostatecznie bardzo czasochłonne oraz nieopłacalne. Lepszym wyjściem jest skupianie się na cechach uniwersalnych tj. dających podobne wyniki przy identycznych pozostałych parametrach.
- Ilość etykiet danej klasy znacząco wpływa na wyniki klasyfikacji zaburzając ogólną ocenę skuteczności klasyfikatora. Najlepiej klasyfikują się artykuły z etykietą CANADA, najgorzej zaś etykietą USA.

8. Bibliografia

- **A. Niewiadomski:** *ksr-wyklad.pdf,* [online]
    - Dostępny za pośrednictwem platformy WIKAMP
- **R. Tadeusiewicz:** *Rozpoznawanie obrazów*, PWN, Warszawa, 1991
- **Kai Ming Ting:** *Confusion Matrix*, Springer Nature, 2011, [online]
    - Dostępny w Internecie: https://link.springer.com/referenceworkentry/10.1007%2F978-0-387-30164-8_157
- **D. Chicco, G. Jurman:** *The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation*, 2020, [online]
    - Dostępny w Internecie: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6941312/