Informatyka, studia STACJONARNE, I st. semestr VI

Komputerowe systemy rozpoznawania **2020/2021**

Prowadzący: prof. dr hab. inż. Adam Niewiadomski poniedziałek, 12:00

Data oddania: _____ Ocena: _____

Julia Szymańska 224441
Przemysław Zdrzalik 224466

Projekt 1. Klasyfikacja dokumentów tekstowych

1. Cel projektu

Celem projektu jest stworzenie aplikacji klasyfikującej zadany zbiór danych tekstowych metodą K najbliższych sąsiadów (k-NN). Aplikacja ma za zadanie dokonać ekstrakcji cech na zbiorach tekstów[1] oraz następnie dokonać ich klasyfikacji.

2. Klasyfikacja nadzorowana metodą k-NN

Metoda K najbliższych sąsiadów, w skrócie metoda k-NN[1], jest to algorytm stosowany do klasyfikacji, który nie wymaga etapu uczenia. Polega na zaklasyfikowaniu rozpatrywanego elementu do grupy ze zbioru uczącego, gdzie spośród k najbliższych rozpatrywanemu elementowi sąsiadów najwięcej z nich należy do tej grupy. Klasyfikator przyjmuje cztery parametry wejściowe takie jak: warotść k - liczba rozpatrywanych sąsiadów, proporcje podziału zbiorów na zbior uczący i zbiór testowy, zbiór cech, a także metrykę i/lub miarę prawdopodobieństwa. Wynikiem klasyfikacji jest zaklasyfikowanie elementu do jednego ze zbiorów uczących.

2.1. Ekstrakcja cech, wektory cech

Na zbiorach danych tekstowych należy dokonać ekstrakcji cech, które będą wartościami rzeczywistymi oraz tekstowymi. Dane cechy będą repre-

zentowały tekst w postaci wektora cech podczas procesu klasyfikacji. Przed dokonaniem ekstrakcji cech, z tekstów usuwane są słowa znajdujące się na stop liście. Teksty ze zbioru danych tekstowych posiadają strukturę:

$$\begin{aligned} &< TEXT > \\ &\quad < TITLE/ > \\ &\quad < AUTHOR/ > \\ &\quad < DATELINE/ > \\ &\quad < BODY/ > \\ &< /TEXT > \end{aligned} \tag{1}$$

1. Liczba słów - cecha ta oznacza liczbę słów które składają się na pobrany tekst. Cecha ta będzie charakteryzowała długość dokumentu w postaci liczby całkowitej
$$c_1 = len \tag{2}$$
gdzie len - liczba słów w tekście.

2. Data z tagu <Dateline> - Każdy tekst w swoim body posiada tag <Dateline>, w którym znajduje się miasto oraz data podana w postaci miesiąca i dnia. Data będzie konwertowana na wartość liczbową, gdzie liczbą tą będzie numer podanego dnia w ciągu roku, licząc rok tak jakby rok był rokiem przestępnym, przykładowo data 1 marca będzie reprezentowana poprzez wartość 61. Cechę traktujemy jako cechę w postaci liczby całkowitej. Wartość będzie oznaczana poprzez symbol c_2.

3. Lokacja z tagu <Dateline>- jak wyżej. Lokację traktujemy jako cechę tekstową. Wartość będzie oznaczana poprzez symbol c_3.

4. Tytuł z tagu <Title>- Każdy tekst w swoim body posiada tag <Title>. Tytuł traktujemy jako cechę tekstową. Wartość będzie oznaczana poprzez symbol c_4.

5. Autor z tagu <Author>- Większość tekstów w swoim body posiada tag <Author>. Autora traktujemy jako cechę tekstową. Wartość będzie oznaczana poprzez symbol c_5.

6. Najczęściej występująca nazwa kraju - wybieramy najczęściej występującą w analizowanym tekście nazwę kraju. Nazwy krajów pobieramy z dołączonego pliku all-places-strings.lc, przykładowo krajem występującym w pliku jest 'albania'.Nazwę kraju traktujemy jako cechę tekstową.Wartość będzie oznaczana poprzez symbol c_6.

7. Zbiór występujących słów kluczowych. Za słowa kluczowe przyjmujemy słowa znajdujące się w dołączonych plikach o rozszerzeniach .lc.txt. Cechę traktujemy jako cechę tekstową.
$$c_7 : c_7 \in N \cap t \tag{3}$$

gdzie N - zbiór wszystkich słów kluczowych, t - zbiór słów należących do tekstu

8. Liczba wystąpień słów kluczowych - traktujemy jako cechę w postaci liczby całkowitej.

$$c_8 = |c_7| \tag{4}$$

gdzie c_8 - zbiór występujących słów kluczowych

9. Nasycenie tekstu ilością słów kluczowych - traktujemy jako cechę w postaci liczby zmienno przecinkowej.

$$c_9 = c_8/c_1 \tag{5}$$

gdzie c_9 - liczba wystąpień słów kluczowych w tekscie, c_1 - liczba słów w tekście

10. Najczęściej występujące słowo kluczowe - wybieramy najczęściej występujące w analizowanym tekście słowo kluczowe. Cechę traktujemy jako cechę tekstową. Wartość będzie oznaczana poprzez symbol c_{10}.

11. Liczba unikatowych słów - zliczamy liczbę unikatowych słów, to znaczy występujących dokładnie raz w analizowanym tekście. Cechę traktujemy jako cechę w postaci liczby całkowitej. Wartość będzie oznaczana poprzez symbol c_{11}.

Wektor cech będzie reprezentowany w postaci:

$$w = [c_1, c_2, c_3, c_4, c_5, c_6, c_7, c_8, c_9, c_{10}, c_{11}] \tag{6}$$

2.2. Miary jakości klasyfikacji

W celu określenia jakości wykonanej klasyfikacji korzystamy z czterech miar jakości klasyfikacji. Aby obliczyć każdą z miar tworzymy tablicę pomyłek, inaczej macierz błędu [4]. Tablica składa się z dwóch wierszy i dwóch kolumn, gdzie wiersze to klasy predykowane, a kolumny to klasy rzeczywiste. Dane oznaczone jako dane pozytywne i negatywne poddawane są klasyfikacji, która przypisuje im predykowaną klasę pozytywną bądź negatywną.

We wzorach zostały użyte oznaczenia:
— TP - sumaryczna liczba poprawnie zaklasyfikowanych tekstów rozpatrywanej klasy
— TN - sumaryczna liczba poprawnie zaklasyfikowanych tekstów pozostałych klas
— FP - sumaryczna liczba tekstów pozostałych klas zaklasyfikowanych do rozpatrywanej klasy
— FN - sumaryczna liczba tekstów rozpatrywanej klasy zaklasyfikowanych do pozostałych klas

		Klasa rzeczywista	
		Pozytywna	Negatywna
Klasa predykowana	Pozytywna	prawdziwie pozytywna (TP)	fałszywie pozytywna (FP)
	Negatywna	fałszywie negatywna (FN)	prawdziwie negatywna (TN)

Tabela 1. Wzór tablicy pomyłek[4].

Stosowane miary jakości klasyfikacji:

— Dokładność (ang. accuracy), ACC - jest to stosunek poprawnie zaklasyfikowanych tekstów do wszystkich klasyfikowanych tekstów.

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \tag{7}$$

— Precyzja (ang. precision), PPV - jest to stopień zgodności wyników uzyskanych w określonych warunkach z wielokrotnych pomiarów. Precyzja to stosunek liczby poprawnie zaklasyfikowanych tekstów rozpatrywanej klasy do liczby wszystkich tekstów zaklasyfikowanych do rozpatrywanej klasy.

$$PPV = \frac{TP}{TP + FP} \tag{8}$$

Dla całego zbioru dokumentów wartość miary jest liczona jako średnia ważona obliczonych precyzji dla pojedyńczych klas, gdzie wagą jest stosunek liczebności tej klasy do liczebności wszystkich klas.

$$PPV_{calk} = \sum_{n=1}^{m} (PPV_n * \frac{k_n}{k}) \tag{9}$$

Gdzie PPV_{calk} - precyzja obliczona dla wszystkich klas klasyfikowanych dokumentów, m - liczba rozpatrywanych klas, PPV_n - precyzja dla n-tej klasy, k_n - liczebność rzeczywista dokumentów klasy n, k - liczebność wszystkich klasyfikowanych dokumentów

— Czułość (ang. recall), TPR - jest to stosunek liczby poprawnie zaklasyfikowanych tekstów do rozpatrywanej klasy do liczby tekstów z rozpatrywanej klasy.

$$TPR = \frac{TP}{TP + FN} \tag{10}$$

Dla całego zbioru dokumentów wartość miary jest liczona jako średnia ważona obliczonych czułości dla pojedyńczych klas, gdzie wagą jest stosunek liczebności tej klasy do liczebności wszystkich klas.

$$TPR_{calk} = \sum_{n=1}^{m} (TPR_n * \frac{k_n}{k}) \tag{11}$$

Gdzie TPR_{calk} - czułość obliczona dla wszystkich klas klasyfikowanych dokumentów, m - liczba rozpatrywanych klas, TPR_n - czułość dla n-tej klasy, k_n - liczebność rzeczywista dokumentów klasy n, k - liczebność wszystkich klasyfikowanych dokumentów

— Miara F1 - średnia harmoniczna miar Precyzja i Czułość.

$$F1 = \frac{2}{\frac{1}{PPV} + \frac{1}{TPR}} \tag{12}$$

3. Klasyfikacja z użyciem metryk i miar podobieństwa tekstów

W procesie klasyfikacji możliwe jest wykorzystanie jednej z trzech metryk: metryka Euklidesowa, metryka Czebyszewa, metryka Uliczna. Metryki służą obliczeniu odległości pomiędzy dwoma wektora o dowolnym rozmiarze.

Metryka Eukliedesowa[1] jest opisana wzorem:

$$d(x, y) = \sqrt{(y_1 -^* x_1)^2 + ... + (y_n -^* x_n)^2} \tag{13}$$

gdzie: *d(x, y)* - odległość pomiędzy wektorem x i y; x, y - wektory o tym samym rozmiarze; n - rozmiar wektorów x i y; x_n, y_n - składowe wektora, $-^*$ - dla dwóch liczb rzeczywistych odejmowanie klasyczne, dla wartości tekstowych jest to wynik działania funkcji opisanej wzorem 17 przy użyciu miary podobieństwa 16.

Metryka Czebyszewa[1] jest opisana wzorem:

$$d(x, y) = max(|y_i -^* x_i|) \tag{14}$$

gdzie: *d(x, y)* - odległość pomiędzy wektorem x i y; x, y - wektory o tym samym rozmiarze; n - rozmiar wektorów x i y; x_i, y_i - i-ta składowa wektora, $-^*$ - dla dwóch liczb rzeczywistych odejmowanie klasyczne, dla wartości tekstowych jest to wynik działania funkcji opisanej wzorem 17 przy użyciu miary podobieństwa 16.

Metryka Uliczna[1] jest opisana wzorem:

$$d(x, y) = \sum_{i=1}^{n} |x_i -^* y_i| \tag{15}$$

gdzie: *d(x, y)* - odległość pomiędzy wektorem x i y; x, y - wektory o tym samym rozmiarze; n - rozmiar wektorów x i y; x_n, y_n - składowe wektora, $-^*$ - dla dwóch liczb rzeczywistych odejmowanie klasyczne, dla wartości tekstowych jest to wynik działania funkcji opisanej wzorem 17 przy użyciu miary podobieństwa 16.

By móc obliczyć odległość pomiędzy wektorami cech zadanych tekstów, należy wcześniej skorzystać z miary podobieństwa tekstu by zamienić cechy o wartościach tekstowych na liczby w wektorach. W programie została zastosowana metoda bigramów[3]. Korzystając z tej metody obliczamy współczynnik podobieństwa tej samej cechy tekstowej dla dwóch tekstów zgodnie ze wzorem:

$$s = \frac{1}{N-1} \sum_{i=0}^{N-1} h(i) \tag{16}$$

Gdzie s - wartość liczbowa będąca podobieństwem cechy tekstowej obu dokumentów zawierająca się w przedziale [0, 1], N - długość dłuższej cechy tekstowej z obu dokumentów, *h(i)* - przyjmuje wartość 1 gdy podciąg zaczynający się od i-tej pozycji w jednej cesze tekstowej dokumentu występuje w cesze tekstowej drugiego dokumentu, w przeciwnym wypadku przyjmuje wartość 0.

Obliczone w ten sposób podobieństwo cechy tekstowej dla dwóch tekstów wykorzystywane jest do obliczenia odległości pomiędzy nimi:

$$d = 1 - s \tag{17}$$

Gdzie d - odległość cechy tekstowej obu dokumentów, s - podobieństwo cechy tekstowej obu dokumentów.

Wstępna klasyfikacja na ograniczonym zbiorze tekstów została przeprowadzona dla trzech różnych zestawów parametrów wejściowych.
Parametry wejściowe dla pierwszej klasyfikacji wstępnej:

Tabela 2. Parametry wejściowe dla pierwszej wstępnej klasyfikacji.

K	**Metryka**	**Procent zbioru trenującego**	**Wybrane cechy**
5	Czebyszewa	80%	Wszystkie cechy

Wstępne wyniki miary Accuracy dla pierwszej klasyfikacji wstępnej:

Tabela 3. Wstępne wyniki miary Accuracy dla pierwszej klasyfikacji wstępnej.

Liczba tekstów	**Liczba poprawnie sklasyfikowanych tekstów**	**Accuracy**
394	238	0,60

Parametry wejściowe dla drugiej klasyfikacji wstępnej:

Tabela 4. Parametry wejściowe dla drugiej wstępnej klasyfikacji.

K	**Metryka**	**Procent zbioru trenującego**	**Wybrane cechy**
3	Euklidesowa	95%	c_3, c_4, c_6, c_7

Wstępne wyniki miary Accuracy dla drugiej klasyfikacji wstępnej:

Tabela 5. Wstępne wyniki miary Accuracy dla drugiej klasyfikacji wstępnej.

Liczba tekstów	**Liczba poprawnie sklasyfikowanych tekstów**	**Accuracy**
394	383	0,97

Parametry wejściowe dla trzeciej klasyfikacji wstępnej:

Tabela 6. Parametry wejściowe dla trzeciej wstępnej klasyfikacji.

K	**Metryka**	**Procent zbioru trenującego**	**Wybrane cechy**
9	Uliczna	73%	c_7, c_8, c_9, c_{10}

Wstępne wyniki miary Accuracy dla trzeciej klasyfikacji wstępnej:

Tabela 7. Wstępne wyniki miary Accuracy dla trzeciej klasyfikacji wstępnej.

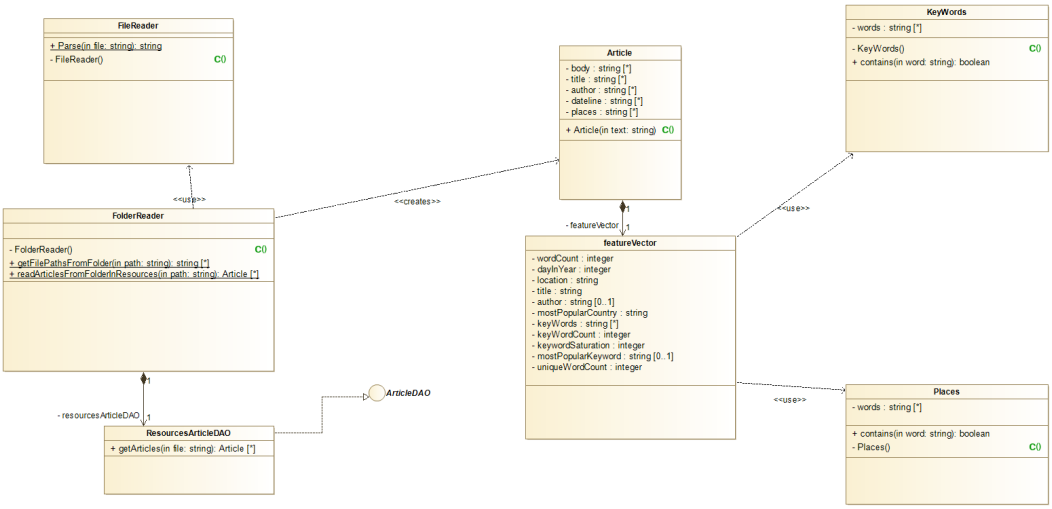
Liczba tekstów	**Liczba poprawnie sklasyfikowanych tekstów**	**Accuracy**
394	235	0,60

Najlepsze wyniki zostały uzyskane dla drugiej wstępnej klasyfikacji, w której został ograniczony zbiór cech, wybrane cechy to: Lokalizacja z tagu <Dateline>, Tytuł z tagu <Title>, Najczęściej występująca nazwa kraju, Zbiór występujących słów kluczowych.

4. Budowa aplikacji

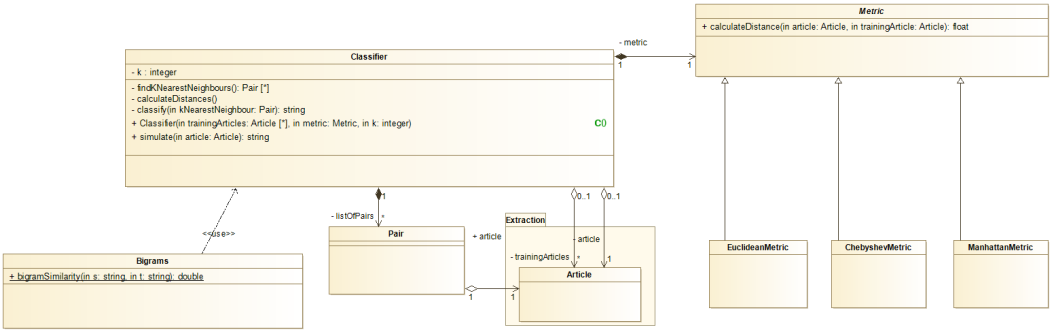
4.1. Diagramy UML

Aplikacja będzie składała się z dwóch modułów: z modułu ekstrakcji cech oraz z modułu klasyfikacji. Moduł ekstrakcji wczytuje pliki z treścią artykułów. Następnie tworzone są obiekty artykułów. Dla każdego obiektu usuwane są słowa ze stop listy oraz kolejno tworzone są wektory cech artykułów.



Rysunek 1. Diagram klas modułu ekstrakcji cech.

Moduł klasyfikacji oblicza odległości pomiędzy artykułem zadanym a każdym z artykułów ze zbioru trenującego za pomocą jednej z zadanych metryk [1] : metryki Euklidesowej, metryki Ulicznej, metryki Czebyszewa. Dla cech zapisanych w postaci tekstowej ich odległość jest obliczana za pomocą metody bigramów. W ten sposób tworzone są pary zawierające artykuł i odległość od zadanego artykułu. Następnie znajdowanych jest k najbliższych sąsiadów dla zadanego artykułu, gdzie poprzez słowo sąsiad rozumiemy artykuł ze zbioru trenującego. Ostatecznie artykuł jest klasyfikowany do klasy, której obiekty najczęściej wystąpiły wśród k najbliższych sąsiadów.



Rysunek 2. Diagram klas modułu klasyfikacji.

4.2. Prezentacja wyników, interfejs użytkownika

Po uruchomieniu programu użytkownik proszony jest o podanie poprzez konsolę kolejnych parametrów klasyfikacji. Na początku użytkownik podaje wartość parametru k, następnie wybiera jedną z trzech metryk, kolejno podawany jest procent zbioru treningowego w stosunku do zbioru wszystkich tekstów oraz użytkownik może podać cechy tekstów do klasyfikacji. Wybór parametrów w konsoli prezentuje się:

```
Podaj wartość k:
3
Wybierz metrykę:
1. Euklidesowa
2. Czebyszewa
3. Uliczna
2
Podaj procent artykułów treningowych:
95
Czy chcesz wybrać zestaw cech do klasyfikacji:
1. Tak
2. Nie
1
Cechy do wyboru:
1.  Liczba słów
2.  Autor z tagu <Author>
3.  Liczba unikatowych słów
4.  Data z tagu <Dateline>
5.  Lokalizacja z tagu <Dateline>
6.  Tytuł z tagu <Title>
7.  Najczęściej występująca nazwa kraju
8.  Zbiór występujących słów kluczowych
9.  Liczba wystąpień słów kluczowych
10. Nasycenie tekstu ilością słów kluczowych
11. Najczęściej występujące słowo kluczowe
1 2 3 4 5 6
```

Rysunek 3. Wybór parametrów klasyfikacji przez użytkownika.

Po wprowadzeniu przez użytkownika wszystkich parametrów klasyfikacji, rozpoczynane jest wczytywanie danych oraz wykonanie klasyfikacji.

```
Rozpoczęto wczytywanie danych.


Rozpoczęto klasyfikację.
```

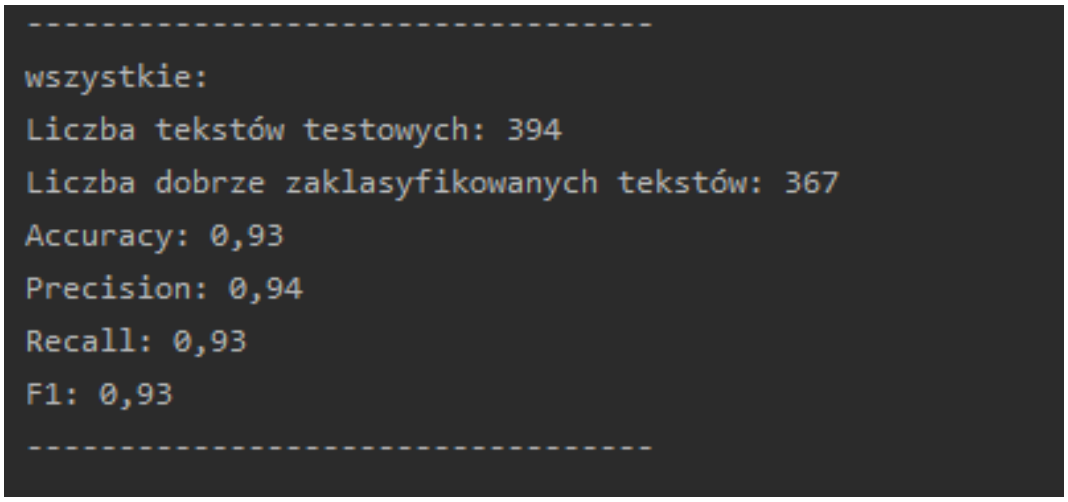
Rysunek 4. Wczytywanie danych i klasyfikacja.

Po wykonanej klasyfikacji na konsoli wyświetlane są obliczone parametry dla poszczególnych klas klasyfikacji oraz wyliczone parametry dla całego zbioru dokumentów. Dla poszczególnych klas klasyfikacji do obliczonych parametrów zaliczamy liczbę tekstów klasy, liczbę poprawnie zaklasyfikowanych tekstów do rozpatrywanej klasy, liczbę tekstów innych klas zaklasfyfikowanych do rozpatrywanej klasy oraz miary jakości: Precision, Recall, F1.

```
---------------------------------
west-germany:
Liczba tekstów klasy: 27
Liczba poprawnie zaklasyfikowanych tekstów: 20
Liczba tekstów innych klas zaklasyfikowanych do tej klasy: 0
Precision: 1,00
Recall: 0,74
F1: 0,85
---------------------------------
```

Rysunek 5. Wynik klasyfikacji dla pojedyńczej klasy klasyfikacji - klasa west-germany.

Dla całego zbioru dokumentów do obliczonych parametrów zaliczamy liczbę tekstów testowych, liczbę poprawnie zaklasyfikowanych tekstów oraz miary jakości: Accuracy, Precision, Recall, F1.



Rysunek 6. Wynik klasyfikacji dla całego zbioru dokumentów.

Do uruchomienia programu wymagana jest wersja Javy: 11.

5. Wyniki klasyfikacji dla różnych parametrów wejściowych

5.1. Eksperyment 1 - Wpływ wartości parametru k na miary jakości klasyfikacji

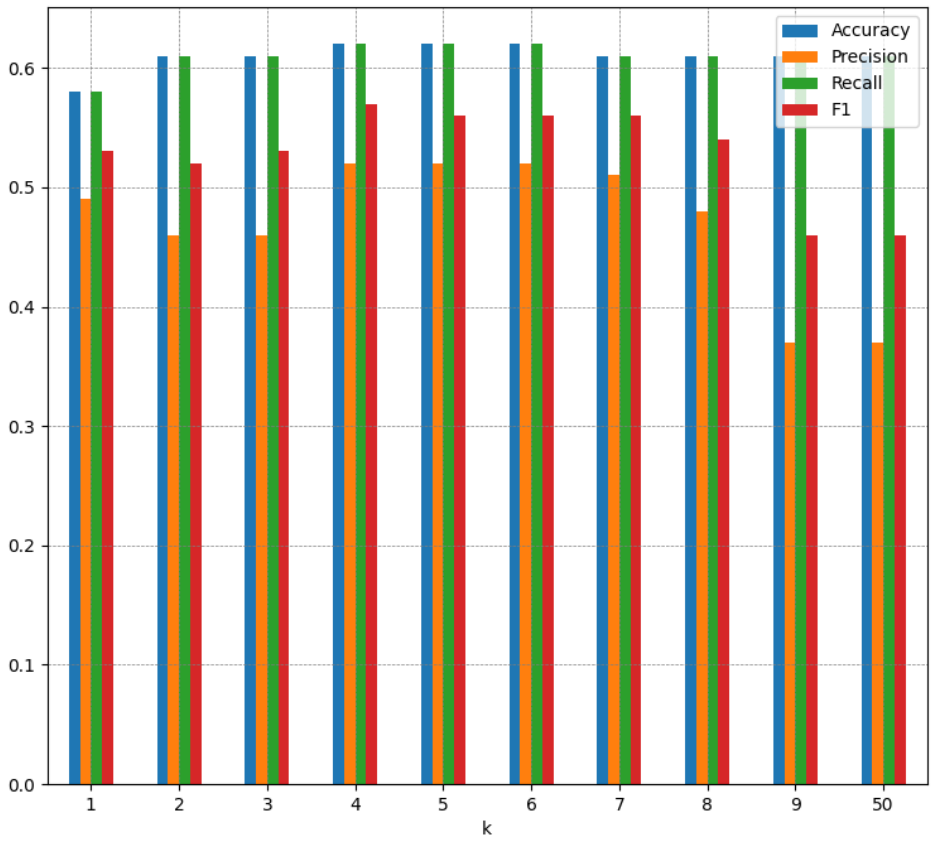
Eksperymenty zostały rozpoczęte od porównania wyników klasyfikacji dla różnych wartości parametru k. W tym celu przyjęliśmy nastepujące parametry:

Tabela 8. Parametry wejściowe dla eksperymentu porównującego różne wartości parametru k.

K	Metryka	Procent zbioru trenującego	Wybrane cechy
-	Euklidesowa	80%	Wszystkie cechy

Tabela 9. Wyniki klasyfikacji dla różnych wartości parametru k. Wartość Nan występująca w tabeli pojawia się, gdy następuje dzielenie przez 0 w Precision. Wartość Nan oznacza, że żaden tekst nie został zaklasyfikowany do tej klasy. Wartość Nan w F1 jest skutkiem wartości Nan w Precision.

K	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**50**
West-germany Precision	0,12	NaN	0,00	0,25	0,00	NaN	NaN	NaN	NaN	NaN
West-germany Recall	0,07	0,00	0,00	0,04	0,00	0,00	0,00	0,00	0,00	0,00
West-germany F1	0,09	NaN	0,00	0,06	0,00	NaN	NaN	NaN	NaN	NaN
Usa Precision	0,65	0,63	0,62	0,63	0,62	0,62	0,61	0,61	0,61	0,61
Usa Recall	0,89	0,97	0,99	1,00	1,00	1,00	0,99	1,00	1,00	1,00
Usa F1	0,75	0,76	0,76	0,77	0,77	0,76	0,76	0,76	0,76	0,76
France Precision	0,00	NaN	NaN	NaN	0,00	NaN	0,33	NaN	NaN	NaN
France Recall	0,00	0,00	0,00	0,00	0,00	0,00	0,03	0,00	0,00	0,00
France F1	0,00	NaN	NaN	NaN	0,00	NaN	0,06	NaN	NaN	NaN
Uk Precision	0,31	0,80	1,00	0,75	1,00	1,00	NaN	NaN	NaN	NaN
Uk Recall	0,15	0,12	0,06	0,09	0,06	0,03	0,00	0,00	0,00	0,00
Uk F1	0,20	0,21	0,11	0,16	0,11	0,06	NaN	NaN	NaN	NaN
Canada Precision	0,17	0,00	0,00	0,00	NaN	NaN	1,00	NaN	NaN	NaN
Canada Recall	0,05	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
Canada F1	0,08	0,00	0,00	0,00	NaN	NaN	NaN	NaN	NaN	NaN
Japan Precision	0,38	0,10	0,00	0,50	0,50	0,50	1,00	1,00	NaN	NaN
Japan Recall	0,07	0,02	0,00	0,02	0,02	0,02	0,02	0,02	0,00	0,00
Japan F1	0,12	0,04	0,00	0,04	0,04	0,04	0,05	0,05	NaN	NaN
Accuracy	0,58	0,61	0,61	0,62	0,62	0,62	0,61	0,61	0,61	0,61
Precision	0,49	0,46	0,46	0,52	0,52	0,52	0,51	0,48	0,37	0,37
Recall	0,58	0,61	0,61	0,62	0,62	0,62	0,61	0,61	0,61	0,61
F1	0,53	0,52	0,53	0,57	0,56	0,56	0,56	0,54	0,46	0,46



Rysunek 7. Wykres przedstawia zależność miar Accuracy, Precision, Recall, i F1 od wartości parametru k.

Zmiana parametru k miała największy wpływ na miarę jakości Precision. Najgorszy wynik był dla wartości parametru od 9 wzwyż, a najlepszy dla wartości parametru równego 4. Od wartości k równej 9 wszystkie obiekty zostały zaklasyfikowane do klasy USA, co możemy odczytać z wartości NaN przy obliczaniu precyzji. Wartość taka pojawia się przy dzieleniu przez 0, co oznacza że nie istniał taki obiekt, który zostałby sklasyfikowany do danej klasy.

5.2. Eksperyment 2 - Wpływ wyboru metryki na miary jakości klasyfikacji

Następnym eksperymentem jest wykonanie klasyfikacji dla różnych metryk. Pozostałe parametry pozostają niezmienne.

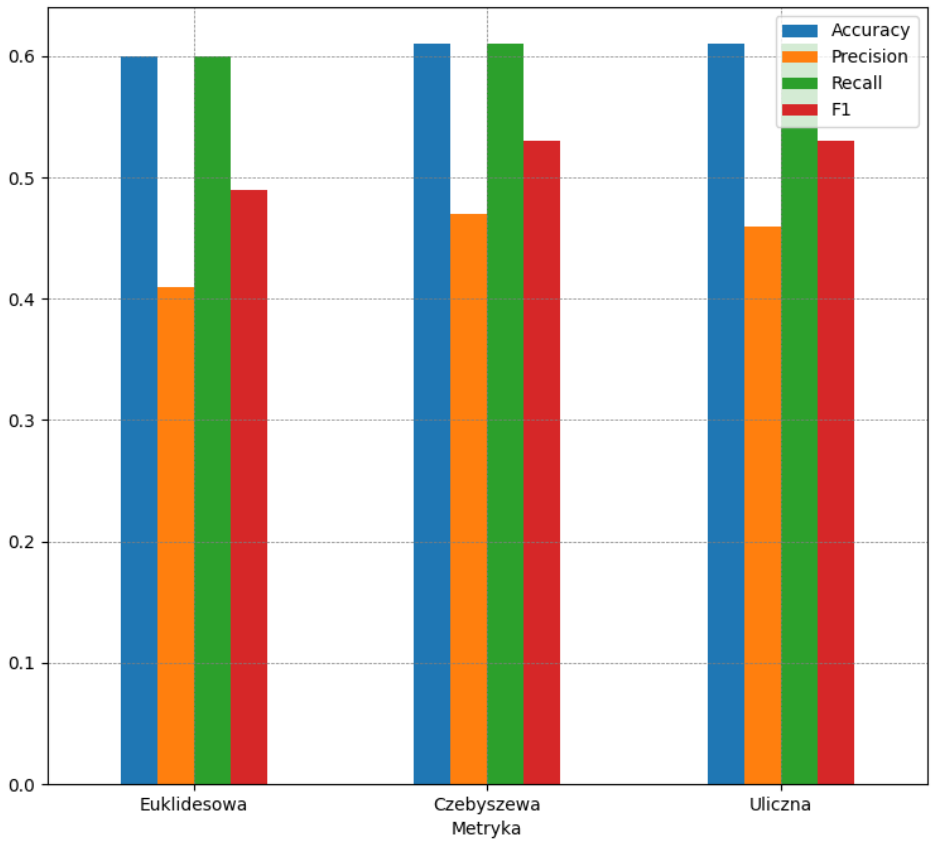
Wybrane parametry przedstawione są w poniższej tabeli:

Tabela 10. Parametry wejściowe dla eksperymentu porównującego różne metryki.

K	Metryka	Procent zbioru trenującego	Wybrane cechy
4	-	90%	Wszystkie cechy

Tabela 11. Wyniki klasyfikacji dla różnych metryk. Wartość Nan występująca w tabeli pojawia się, gdy następuje dzielenie przez 0 w Precision. Wartość Nan oznacza, że żaden tekst nie został zaklasyfikowany do tej klasy. Wartość Nan w F1 jest skutkiem wartości Nan w Precision.

Metryka	Euklidesowa	Czebyszewa	Uliczna
West-germany Precision	0,00	0,00	0,20
West-germany Recall	0,00	0,00	0,04
West-germany F1	0,00	0,00	0,06
Usa Precision	0,62	0,63	0,62
Usa Recall	0,98	0,98	0,98
Usa F1	0,76	0,76	0,76
France Precision	0,00	0,00	0,00
France Recall	0,00	0,00	0,00
France F1	0,00	0,00	0,00
Uk Precision	0,40	0,33	0,40
Uk Recall	0,06	0,09	0,06
Uk F1	0,11	0,14	0,11
Canada Precision	0,00	0,00	NaN
Canada Recall	0,00	0,00	0,00
Canada F1	0,00	0,00	NaN
Japan Precision	0,00	0,50	0,33
Japan Recall	0,00	0,02	0,02
Japan F1	0,00	0,04	0,04
Accuracy	0,60	0,61	0,61
Precision	0,41	0,47	0,46
Recall	0,60	0,61	0,61
F1	0,49	0,53	0,53



Rysunek 8. Wykres przedstawia zależność miar Accuracy, Precision, Recall, i F1 od metryki.

Metryki uzyskały zbliżone wyniki, spośród nich najlepszą okazała się metryka Czebyszewa, niewiele gorszą, bo z precyzją mniejszą o zaledwie 0.01 była metryka Uliczna.

5.3. Eksperyment 3 - Wpływ wyboru zestawu cech na miary jakości klasyfikacji

Następnym eksperymentem jest wykonanie klasyfikacji dla różnych podzbiorów cech. Pozostałe parametry pozostają niezmienne. W tym celu wybieramy sześć różnych zestawów cech.

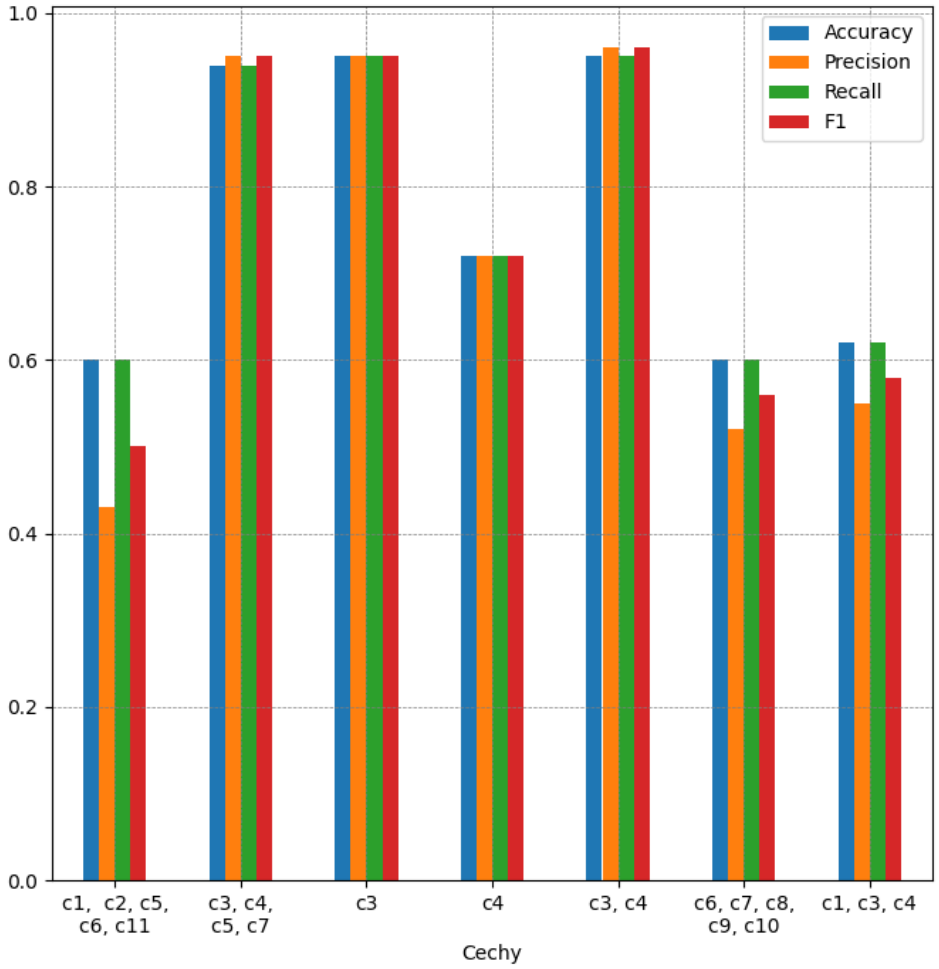
Parametry wejściowe dla eksperymentu porównującego różne podzbiory cech:

Tabela 12. Parametry wejściowe dla eksperymentu porównującego różne podzbiory cech.

K	Metryka	Procent zbioru trenującego	Wybrane cechy
4	Euklidesa	90	-

Tabela 13. Wyniki klasyfikacji dla różnych podzbiorów cech. Wartość Nan występująca w tabeli pojawia się, gdy następuje dzielenie przez 0 w Precision. Wartość Nan oznacza, że żaden tekst nie został zaklasyfikowany do tej klasy. Wartość Nan w F1 jest skutkiem wartości Nan w Precision.

Wybrany numer zestawu cech	**c_1, c_2, c_5, c_6, c_{11}**	**c_3, c_4, c_5, c_7**	**c_3**	**c_4**	**c_3, c_4**	**c_6, c_7, c_8, c_9, c_{10}**	**c_1, c_3, c_4**
West-germany Precision	0,00	1,00	1,00	0,71	1,00	0,22	NaN
West-germany Recall	0,00	0,85	0,93	0,37	0,96	0,07	0,00
West-germany F1	0,00	0,92	0,96	0,49	0,98	0,11	NaN
Usa Precision	0,62	0,92	0,93	0,73	0,93	0,64	0,63
Usa Recall	0,98	1,00	1,00	0,97	1,00	0,90	0,98
Usa F1	0,75	0,96	0,96	0,83	0,96	0,75	0,77
France Precision	0,00	1,00	1,00	0,67	1,00	0,00	1,00
France Recall	0,00	1,00	1,00	0,18	1,00	0,00	0,09
France F1	0,00	1,00	1,00	0,29	1,00	0,00	0,17
Uk Precision	0,67	1,00	1,00	0,55	1,00	0,08	0,43
Uk Recall	0,06	0,97	0,97	0,33	0,97	0,03	0,09
Uk F1	0,11	0,98	0,98	0,42	0,98	0,04	0,15
Canada Precision	0,00	1,00	1,00	0,75	1,00	0,31	0,29
Canada Recall	0,00	0,79	0,85	0,23	0,85	0,10	0,05
Canada F1	0,00	0,89	0,92	0,35	0,92	0,15	0,09
Japan Precision	0,00	0,97	0,97	0,76	0,97	0,67	0,67
Japan Recall	0,00	0,81	0,79	0,44	0,79	0,28	0,05
Japan F1	0,00	0,89	0,87	0,56	0,87	0,39	0,09
Accuracy	0,60	0,94	0,95	0,72	0,95	0,60	0,62
Precision	0,43	0,95	0,95	0,72	0,96	0,52	0,55
Recall	0,60	0,94	0,95	0,72	0,95	0,60	0,62
F1	0,50	0,95	0,95	0,72	0,96	0,56	0,58



Rysunek 9. Wykres przedstawia zależność miar Accuracy, Precision, Recall, i F1 od podzbioru cech.

Po porównaniu zestawów cech odkryliśmy, że najlepszą z cech jest Lokalizacja z tagu <Dateline>. Korzystając tylko z tej jednej cechy uzyskaliśmy prawie najlepsze wyniki, które nieznacznie poprawiło jedynie dodanie do tej cechy Tytułu z tagu <Title>. Dodanie do tych dwu cech autora oraz zbioru cech kluczowych nieznacznie pogorszyło wynik. Inne cechy uzyskały znacznie gorsze wyniki.

5.4. Eksperyment 4 - Wpływ wartości proporcji podziału zbioru na miary jakości klasyfikacji

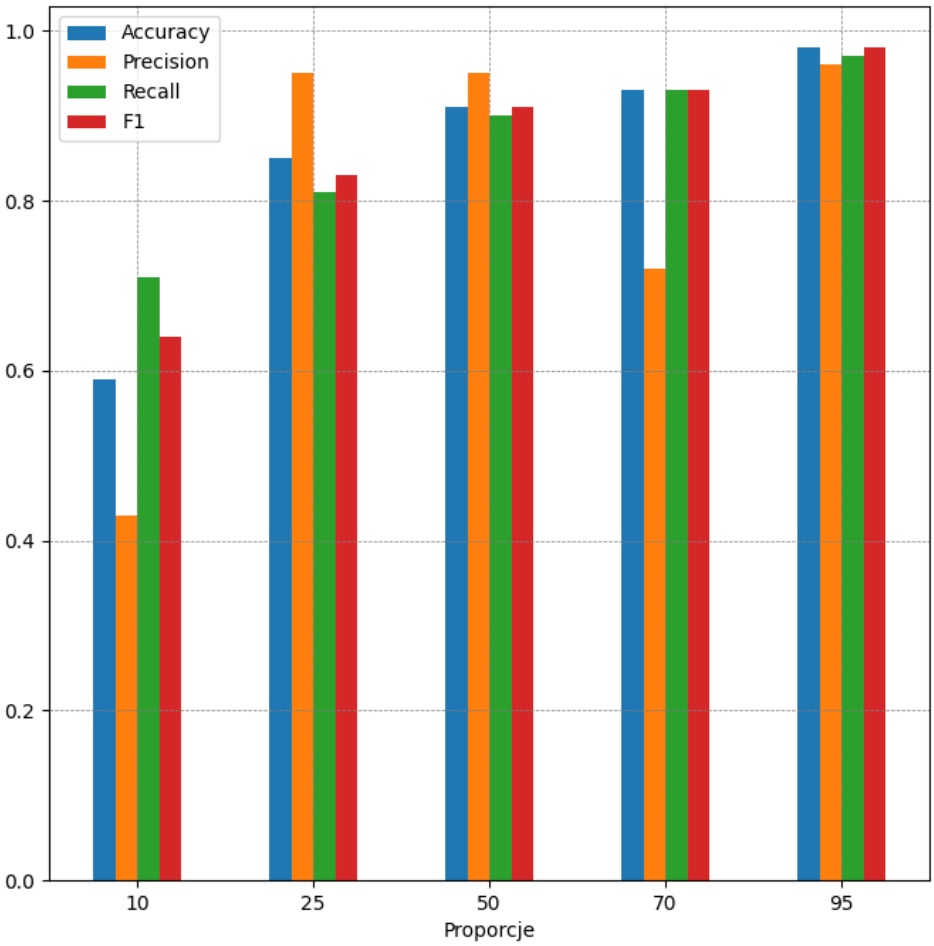
Ostatnim eksperymentem jest wykonanie klasyfikacji dla różnych wartości proporcji podziału zbioru. Pozostałe parametry pozostają niezmienne. Wybrane parametry przedstawione są w poniższej tabeli:

Tabela 14. Parametry wejściowe dla eksperymentu porównującego różne wartości proporcji podziału zbioru.

K	**Metryka**	**Procent zbioru trenującego**	**Wybrane cechy**
4	Euklidesa	-	c_4, c_5

Tabela 15. Wyniki klasyfikacji dla różnych wartości proporcji podziału zbioru. Wartość Nan występująca w tabeli pojawia się, gdy następuje dzielenie przez 0 w Precision. Wartość Nan oznacza, że żaden tekst nie został zaklasyfikowany do tej klasy. Wartość Nan w F1 jest skutkiem wartości Nan w Precision.

Proporcje	**10**	**25**	**50**	**70**	**95**
West-germany Precision	NaN	1,00	1,00	1,00	1,00
West-germany Recall	0,00	0,48	0,48	0,74	0,96
West-germany F1	NaN	0,65	0,65	0,85	0,98
Usa Precision	0,68	0,78	0,86	0,90	0,96
Usa Recall	1,00	1,00	1,00	1,00	1,00
Usa F1	0,81	0,88	0,92	0,94	0,98
France Precision	NaN	1,00	1,00	1,00	1,00
France Recall	0,00	1,00	1,00	1,00	1,00
France F1	NaN	1,00	1,00	1,00	1,00
Uk Precis=sion	0,94	0,86	1,00	1,00	1,00
Uk Recall	0,97	0,97	0,97	0,97	0,97
Uk F1	0,96	0,91	0,98	0,98	0,98
Canada Precis=sion	NaN	1,00	1,00	0,97	1,00
Canada Recall	0,00	0,41	0,59	0,74	0,87
Canada F1	NaN	0,58	0,74	0,84	0,93
Japan Precis=sion	0,83	1,00	1,00	1,00	1,00
Japan Recall	0,12	0,19	0,79	0,79	0,93
Japan F1	0,20	0,31	0,88	0,88	0,96
Accuracy	0,71	0,81	0,90	0,93	0,97
Precision	0,59	0,85	0,91	0,93	0,98
Recall	0,71	0,81	0,90	0,93	0,97
F1	0,64	0,83	0,91	0,93	0,98



Rysunek 10. Wykres przedstawia zależność miar Accuracy, Precision, Recall, i F1 od proporcji podziału zbioru.

Najlepszą jakość klasyfikacji otrzymano dla proporcji 95 - 5. Zmniejszanie ilości dokumentów w zbiorze treningowym pogarszało jakość klasyfikacji.

5.5. Eksperyment 5 - Wpływ wyboru metryki na miary jakości klasyfikacji przy ograniczonym zbiorze cech

Ze względu na zbliżone wyniki jakości klasyfikacji dla metryki Czebyszewa i Ulicznej, ponawiamy eksperyment wyboru metryk przy ograniczonym zbiorze cech.

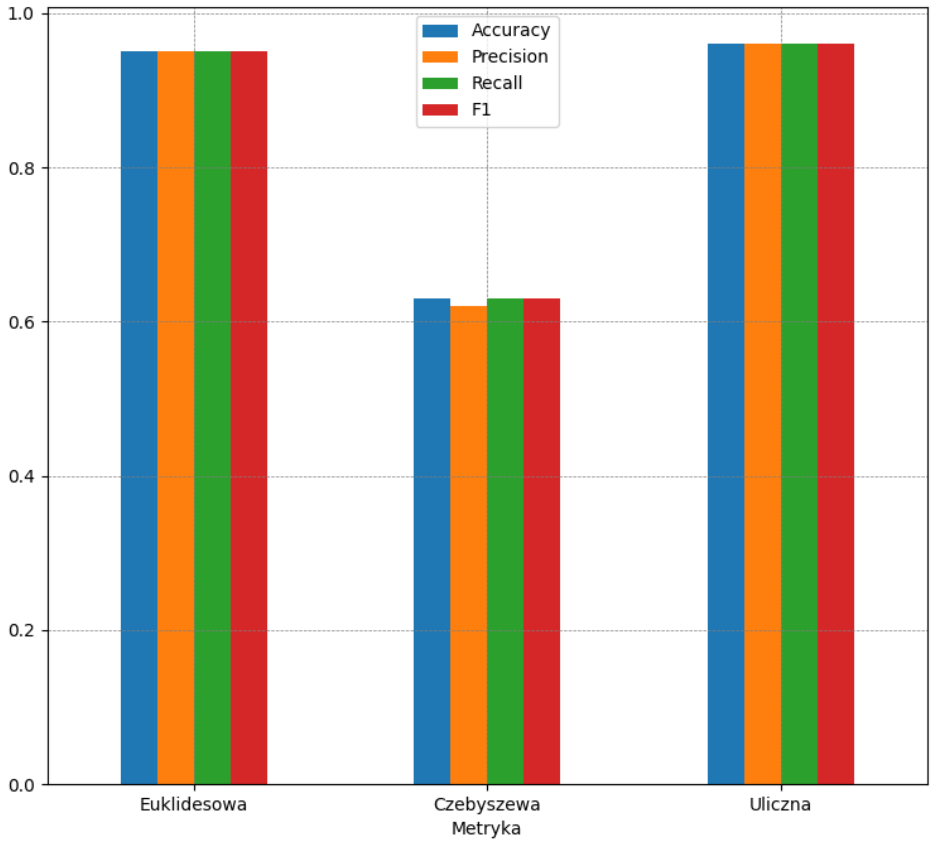
Wybrane parametry przedstawione są w poniższej tabeli:

Tabela 16. Parametry wejściowe dla eksperymentu porównującego różne metryk przy ograniczonym zbiorze cech.

K	**Metryka**	**Procent zbioru trenującego**	**Wybrane cechy**
4	-	90%	c_3, c_4, c_5, c_7,

Tabela 17. Wyniki klasyfikacji dla różnych metryk przy ograniczonym zbiorze cech. Wartość Nan występująca w tabeli pojawia się, gdy następuje dzielenie przez 0 w Precision. Wartość Nan oznacza, że żaden tekst nie został zaklasyfikowany do tej klasy. Wartość Nan w F1 jest skutkiem wartości Nan w Precision.

Metryka	**Euklidesowa**	**Czebyszewa**	**Uliczna**
West-germany Precision	1,00	1,00	1,00
West-germany Recall	0,85	0,04	0,93
West-germany F1	0,92	0,07	0,96
Usa Precision	0,92	0,63	0,94
Usa Recall	1,00	1,00	1,00
Usa F1	0,96	0,77	1,00
France Precision	1,00	0,00	1,00
France Recall	1,00	0,00	1,00
France F1	1,00	0,00	1,00
Uk Precision	1,00	0,75	1,00
Uk Recall	0,97	0,09	0,97
Uk F1	0,98	0,16	0,98
Canada Precision	1,00	NaN	1,00
Canada Recall	0,82	0,00	0,85
Canada F1	0,90	NaN	0,92
Japan Precision	0,97	1,00	0,97
Japan Recall	0,81	0,07	0,84
Japan F1	0,89	0,13	0,90
Accuracy	0,95	0,63	0,96
Precision	0,95	0,62	0,96
Recall	0,95	0,63	0,96
F1	0,95	0,63	0,96



Rysunek 11. Wykres przedstawia zależność miar Accuracy, Precision, Recall, i F1 od metryki przy ograniczonym zbiorze cech.

Metryka Euklidesowa i Uliczba uzyskały zbliżone wyniki, spośród nich najlepszą okazała się metryka Uliczna, niewiele gorszą, bo z precyzją mniejszą o zaledwie 0.01 była metryka Euklidesowa. Najgorzej wypadającą metryką w tym eksperymencie jest metryka Czebyszewa.

6. Dyskusja, wnioski

6.1. Wpływ wartości parametru k na miary jakość klasyfikacji

W sekcji 5.1 badającej wpływ wartości parametru k na miary jakość klasyfikacji na podstawie Wykresu 7 zauważamy, że wartość miary accuracy zmienia się wraz ze zmianą wartości paramatru k. Jakość klasyfikacji wzrasta wraz ze wzrostem wartości parametru k, aż do osiągnięcia wartości maksymalnej. Następnie jakość klasyfikacji zaczyna spadać, aż do momentu ustabilizowania się. Po ustabilizowaniu się, zwiększanie parametru k nie wpływa na jakość klasyfikacji. Po przeanalizowaniu wartości precision i recall dla pojedynczych klas w Tabeli 9 zauważamy, że zmniejszanie się jakości klasyfikacji

przy zwiększaniu wartości parametru k jest spowodowane coraz częstszym przyporządkowywaniem elementów do klasy USA. Stabilizowanie się jakości klasyfikacji po wartości progowej parametru k jest wynikiem przyporządkowywania wszystkich elementów do klasy USA. Dla każdej klasy, oprócz klasy USA, wartość recall maleje wraz ze wzrostem parametru k, dla klasy USA wartość miary recall wzrasta aż do osiągnięcia wartości maksymalnej (1.00). Dominacja klasy USA wynika z faktu, że teksty z klasy USA stanowią około 61% wszystkich klasyfikowanych tekstów. Z tego powodu przy największych wartościach parametru k wartość miary accuracy wynosi okolo 0.61, co wskazuje, że wszystkie teksty zostały zaklasyfikowane do klasy USA. Wartość parametru k ma zauważalny wpływ na jakość klasyfikacji. Dla badanej klasyfikacji oraz dla klasyfikacji o podobnym charakterze istotne jest dobranie odpowiedniej wartości parametru k.

6.2. Wpływ wyboru metryki na miary jakości klasyfikacji

W sekcji 5.2 zostały porównane trzy metryki: metryka Euklidesowa, metryka Czebyszewa, metryka Uliczna. W eksperymencie przyjeliśmy wartość parametru k równą 4, ponieważ w sekcji 5.1 ta wartość okazała się być wartością, dla której dla zadanego zbioru dokumentów jakość klasyfikacji była najlepsza. Na podstawie Wykresu 8 zauważamy, że najgorzej wypadającą metryką jest metryka Euklidesowa. Dla metryki Czebyszewa i Ulicznej na podstawie Wykresu 8 nie jesteśmy w stanie określić, która z metryk posiada lepsze wyniki miar jakości klasyfiakcji. Po porównaniu dokładnych wyników w Tabeli 11, zauważamy, że metryka Czebyszewa i Uliczna uzyskały niemalże takie same wyniki, jedyną różniącą się miarą była precision gdzie metryka Czebyszewa uzyskała wynik lepszy o mniej niż jeden procent.
Ze względy na uzyskane wyniki podczas opracowywania wyniosków postanowiliśmy przeprowadzić kolejny eksperyment. W tym eksperymencie jako parametr przekazaliśmy ograniczony zbiór cech, który w sekcji 5.3 należał do jednego z trzech zbiorów cech, które uzyskały najlepsze wyniki klasyfikacji i równocześnie podany zbiór cech zawierał najwięcej cech, natomiast pozostałe parametry wejściowe są identycznej jak w sekcji 5.5. Na podstawie Wykresu 11 zauważamy, że najgorzej wypadającą metryką jest metryka Czebyszewa. Dla metryki Euklidesowej i Ulicznej na podstawie Wykresu 11 nie jesteśmy w stanie określić, która z metryk posiada lepsze wyniki miar jakości klasyfiakcji. Po porównaniu dokładnych wyników w Tabeli 17, zauważamy, że metryka Euklidesowa i Uliczna uzyskały niemalże takie same wyniki, gdzie róznica wyniosła mniej niż jeden procent i wystąpiła we wszystkich miarach jakości.
Po podsumowaniu wyników z obu sekcji stwierdzamy, że najlepszą z metryk jest metryka Uliczna. Na podstawie wyników wnioskujemy, że metryka Czebyszewa może zostać zastosowana w celu znalezienia cechy negatywnie wpływającej na jakość klasyfikacji. Metryka Czebyszewa jest bardziej wrażliwa na cechy mające negatywny wpływ na jakość klasyfikacji niż metryka Uliczna i Euklidesowa, ponieważ metryka ta jako odleglość wybiera pare cech, dla których odległość jest największa.

6.3. Wpływ wyboru zestwu cech na miary jakości klasyfikacji

W sekcji 5.3 porównaliśmy sześć zestawów cech. Do porównania przyjęliśmy stałe parametry: wartość k równą 4, metrykę euklidesową i proporcje zbioru trenującego do testowego 90:10. W zależności od wyboru zestawu cech wyniki klasyfikacji były diametralnie rózne. Najlepsze wyniki uzyskaliśmy dla zestawów cech zawierających cechę c_3 - lokacja z tagu <Dateline>. Dodanie do tej cechy cechy c_4 - tytuł z tagu <Dateline>nieznacznie poprawiło wynik klasyfikacji, jednak cecha ta samodzielnie uzyskała wynik dużo gorszy niż cecha c_3. Najlepsze wyniki klasyfikacji uzyskaliśmy używając cech tekstowych, bez użycia cech liczbowych. Cechy liczbowe bardzo pogarszały jakość klasyfikacji, prawdopodobnie wynika to z faktu, że odległość pomiędzy cechami tekstowymi przy wykorzystanych przez nas metodach zawsze znajduje się w przedziale <0,1>, podczas gry odległość dla cech liczbowych może być znacznie większa, więc cechy te dominują cechy tekstowe. Nie znaleźliśmy dobrego sposobu na poradzenie sobie z tym problemem. Jednym z mozliwych rozwiązań mogło by być zastosowanie wag dla odpowiednich cech.

7. Braki w realizacji projektu 1.

Wszystkie obowiązkowe elementy projektu zostały zrealizowane.

Literatura

[1] R. Tadeusiewicz: Rozpoznawanie obrazów, PWN, Warszawa, 1991.

[2] A. Niewiadomski, Methods for the Linguistic Summarization of Data: Applications of Fuzzy Sets and Their Extensions, Akademicka Oficyna Wydawnicza EXIT, Warszawa, 2008.

[3] A. Niewiadomski, ksr-wyklad-2009.pdf, 2009.

[4] Internet forum. Wikipedia: The Free Encyclopedia, Dostępny w: `https://pl.wikipedia.org/wiki/Tablica_pomy%C5%82ek?fbclid=IwAR1yFbhG8HoSicSBnyA43YhpyU0tJiaIpI6ghUdNZvzDhPtMPwAWHtrdPUQ`

[5] Machine Learning Repository. UCI:, Dostępny w: `http://archive.ics.uci.edu/ml/datasets/Reuters-21578+Text+Categorization+Collection`