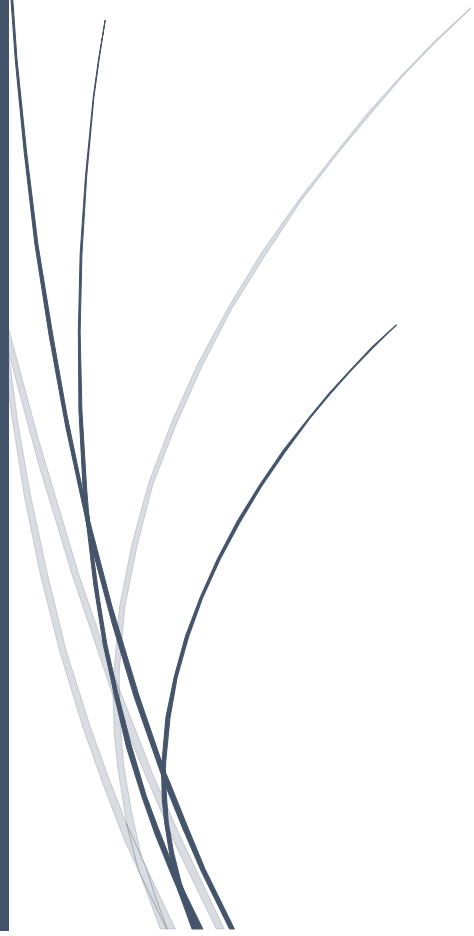
Klasyfikacja tekstu

Sprawozdanie z ćwiczenia 4.

Sztuczna Inteligencja i Inżynieria Wiedzy - laboratorium



Spis treści

Klasyfikacja tekstu 3

Badania 3

Selekcja i ekstrakcja cech 3

Badanie 1 Wpływ rozmiaru słownika na czas budowy modelu oraz jego skuteczność dla Naiwnego Bayesa oraz drzewa decyzyjnego bez selekcji cech 3

Badanie 2: Wpływ liczby wyselekcjonowanych cech na skuteczność algorytmów oraz czas budowy modeli 5

Badanie 3 Wpływ rozmiaru słownika na czas budowy modelu oraz jego skuteczność dla Naiwnego Bayesa oraz drzewa decyzyjnego z selekcją cech 6

Badanie 4: Wpływ liczby wyselekcjonowanych cech na skuteczność algorytmów oraz czas budowy modeli 7

Badanie 5: Porównanie skuteczności różnych metod selekcji cech 9

Badanie 6: Porównanie skuteczności różnych metod lematyzacji 10

Badanie 7: Porównanie skuteczności różnych stop list 11

Badanie 8: Porównanie skuteczności różnych metod tokenizacji 12

Naive Bayes 13

Badanie 9: Porównanie skuteczności Naiwnego Bayesa z rozkładem dwumianowym i wielomianowym oraz zliczaniem liczby słów lub badania tylko ich wystąpienia 13

Parametry drzewa decyzyjnego 14

Badanie 10: Porównanie skuteczności drzewa decyzyjnego z włączonym i wyłączonym pruningiem 14

Badanie 11: Porównanie skuteczności drzewa decyzyjnego dla różnych wartości ConfidenceFactor 15

Badanie 12: Porównanie skuteczności drzewa decyzyjnego z subTreeRaising i subTreeReplacement dla różnych wartości ConfidenceFactor 17

Badanie 13: Porównanie skuteczności drzewa decyzyjnego z reducedErrorPruning dla różnej części wykorzystywanych do pruningu 18

Badanie 14: Porównanie skuteczności drzewa decyzyjnego z reducedErrorPruning dla różnej minimalnej liczby instancji 19

Pozostałe badania 20

Badanie 15: Działanie klasyfikatorów dla różnych podzbiorów klas decyzyjnych – po usunięciu klas z najwyższym FP Rate 20

Badanie 16: Działanie klasyfikatorów dla różnych podzbiorów klas decyzyjnych – po usunięciu klas z najwyższym TP Rate 21

Badanie 17: Określenie zestawów cech istotnych dla poszczególnych klas 22

Podsumowanie 31

Klasyfikacja tekstu

Celem tego zadania było porównanie algorytmów służących do klasyfikacji tekstu. Dane, na jakich bazowano, to artykuły z polskiej Wikipedii podzielone na 32 klasy. Do budowy i analizy modeli wykorzystano pakiet Weka, jego wersję z GUI. Porównywano algorytmy *Naiwnego* Bayesa oraz *drzewa decyzyjnego C4.5*.

Dane były ładowane z odpowiedniego pliku, następnie ekstraktowano z tekstów słowa oraz dokonywano ich selekcji. W następnym kroku dane zostały podzielone na dane treningowe (90%) oraz testowe (10%). Na danych treningowych z użyciem walidacji krzyżowej były budowane odpowiednie modele klasyfikatorów, a następnie były one oceniane na danych testowych i wybierany najlepszy.

Badania

Selekcja i ekstrakcja cech

Badanie 1 Wpływ rozmiaru słownika na czas budowy modelu oraz jego skuteczność dla Naiwnego Bayesa oraz drzewa decyzyjnego bez selekcji cech

Cel badania: Porównanie czasu budowy modelu oraz jego skuteczność dla Naiwnego Bayesa oraz drzewa decyzyjnego w zależności od rozmiaru słownika.

Stałe w badaniu:

selekcja cech: brak
zliczanie wystąpień: tak
stemmer: *IteratedLovinsStemmer*
stopWordsHandler: *MultiStopWords*
tokenizer: *WordTokenizer*parametry drzewa: z pruningiem, confidence factor: 0.25, subTreeRaising

Zmienne w badaniu:

algorytmy: *NaiveBayesMultinomial* oraz drzewo decyzyjne C4.5
rozmiar słownika: od 100 do 500 słów na klasę

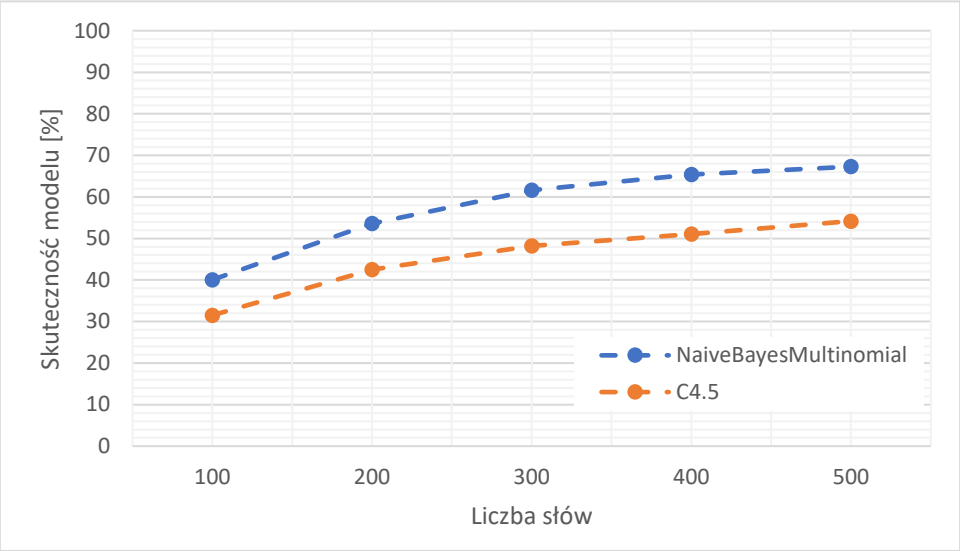
Przebieg badania: Załadowanie danych z pliku, ekstrakcja cech z tekstu, podział zbioru na testowy i treningowy, zapisanie zbiorów do plików, wczytanie danych treningowych, podział walidacją krzyżową oraz zbudowanie odpowiednich modeli, zapis modeli do pliku, wczytanie danych testowych oraz modeli z pliku, walidacja modeli, zapisanie najlepszej osiągniętej skuteczności.

Wyniki:

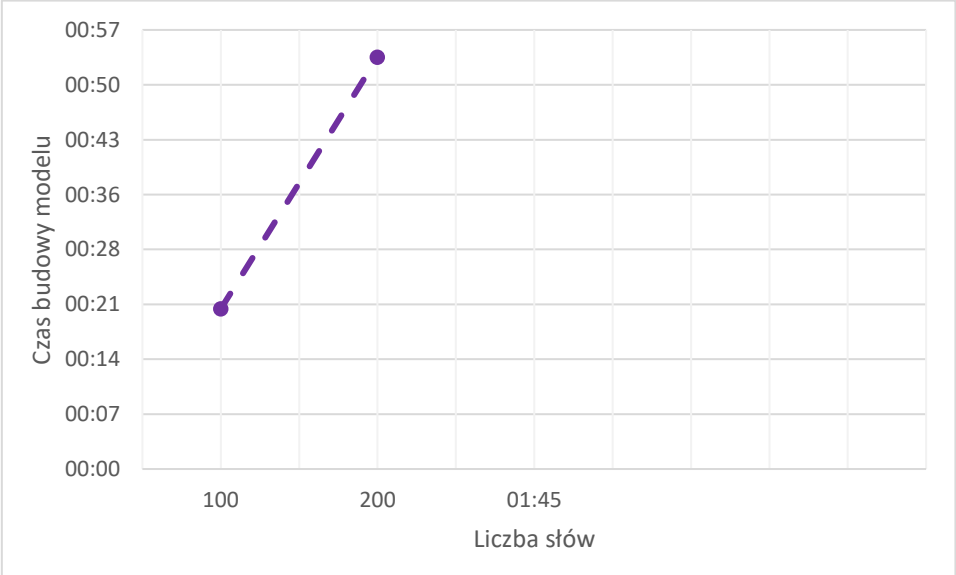
Tabela 1 Czasy budowy modelu oraz skuteczność w zależności od rozmiaru słownika - wyniki badania 1

	NaiveBayesMultinomial		C4.5	
wordsToKeep	Time (model build)	Correct[%]	Time (model build)	Correct[%]
100	-	40.0407	00:21	31.5041
200	-	53.5569	00:54	42.4797
300	-	61.5854	01:45	48.1707

400	-	65.3455	02:52	51.0163
500	00:01	67.2764	04:11	54.1667



Wykres 1 Skuteczność algorytmów w zależności od wielkości słownika



Wykres 2 Czas budowy drzewa decyzyjnego w zależności od wielkości słownika

Wnioski:

Zwiększenie rozmiaru słownika wpływa na poprawę skuteczności algorytmów, jednak w przypadku drzewa decyzyjnego znacząco zwiększa także czas budowy modelu.

Badanie 2: Wpływ liczby wyselekcjonowanych cech na skuteczność algorytmów oraz czas budowy modeli

Cel badania: Porównanie czasu budowy modelu oraz jego skuteczność dla Naiwnego Bayesa oraz drzewa decyzyjnego w zależności od liczby wyselekcjonowanych słów.

Stałe w badaniu:

selekcja cech: *ChiSquare*
zliczanie wystąpień: tak
stemmer: *IteratedLovinsStemmer*
stopWordsHandler: *MultiStopWords*
tokenizer: *WordTokenizer*
parametry drzewa: z pruningiem, confidence factor: 0.25, subTreeRaising
rozmiar słownika: 500 słów na klasę

Zmienne w badaniu:

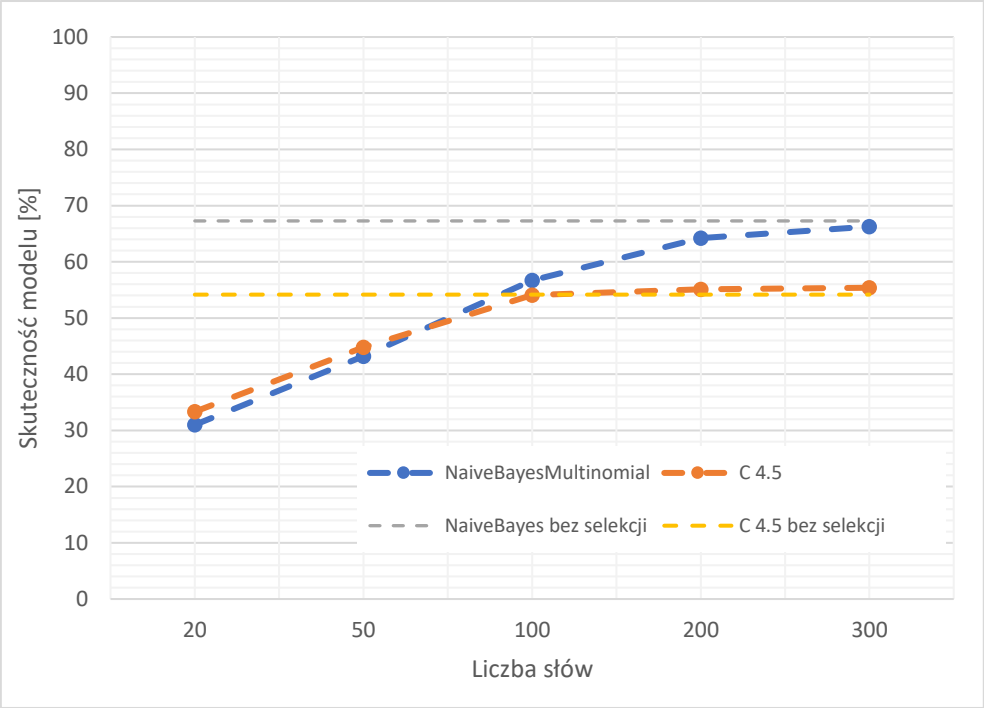
algorytmy: *NaiveBayesMultinomial* oraz drzewo decyzyjne C4.5
liczba wyselekcjonowanych cech: od 20 do 300

Przebieg badania: Załadowanie danych z pliku, ekstrakcja cech z tekstu, selekcja cech, podział zbioru na testowy i treningowy, zapisanie zbiorów do plików, wczytanie danych treningowych, podział walidacją krzyżową oraz zbudowanie odpowiednich modeli, zapis modeli do pliku, wczytanie danych testowych oraz modeli z pliku, walidacja modeli, zapisanie najlepszej osiągniętej skuteczności.

Wyniki:

Tabela 2 Czasy budowy modelu oraz skuteczność w zależności od liczby cech

	NaiveBayesMultinomial		C 4.5	
numToSelect	Time(buildmodel)	Correct%	Time(buildmodel)	Correct%
20	-	30.9959	00:01	33.3333
50	-	43.1911	00:07	44.8171
100	-	56.7073	00:22	54.0650
200	-	64.2276	01:00	55.0813
300	-	66.2602	01:50	55.3862
500 (bez selekcji)	00:01	67.2764	04:11	54.1667



Wykres 3 Skuteczność algorytmów w zależności od selekcji cech z 500 w słowniku

Wnioski:

Nadmierna selekcja cech zmniejsza skuteczność algorytmu.

Badanie 3 Wpływ rozmiaru słownika na czas budowy modelu oraz jego skuteczność dla Naiwnego Bayesa oraz drzewa decyzyjnego z selekcją cech

Cel badania: Porównanie czasu budowy modelu oraz jego skuteczność dla Naiwnego Bayesa oraz drzewa decyzyjnego w zależności od rozmiaru słownika.

Stałe w badaniu:

selekcja cech: *ChiSquare*, wybór 300 słów
zliczanie wystąpień: tak
stemmer: *IteratedLovinsStemmer*
stopWordsHandler: *MultiStopWords*
tokenizer: *WordTokenizer*parametry drzewa: z pruningiem, confidence factor: 0.25, subTreeRaising

Zmienne w badaniu:

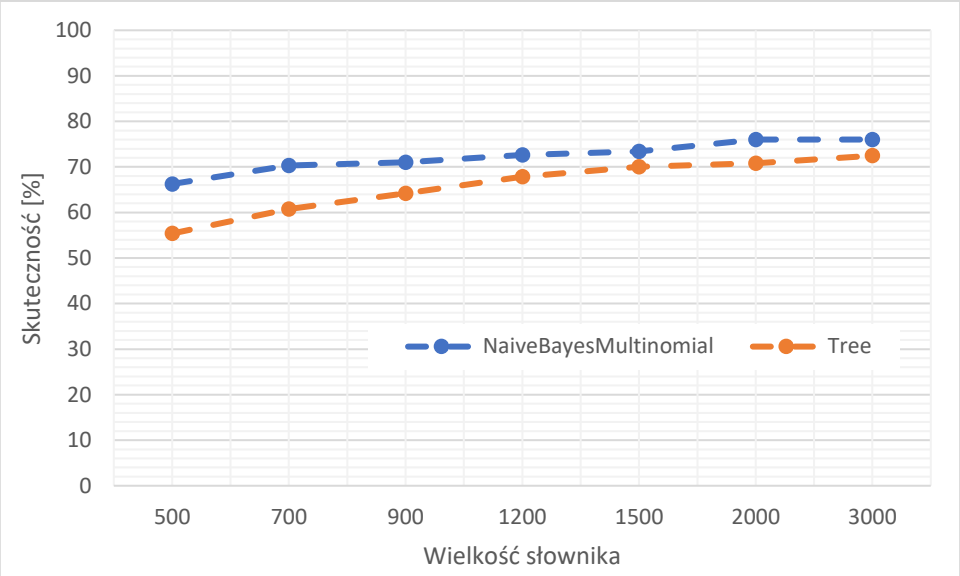
algorytmy: *NaiveBayesMultinomial* oraz drzewo decyzyjne C4.5
rozmiar słownika: od 500 do 3000 słów na klasę

Przebieg badania: Załadowanie danych z pliku, ekstrakcja cech z tekstu, selekcja cech, podział zbioru na testowy i treningowy, zapisanie zbiorów do plików, wczytanie danych treningowych, podział walidacją krzyżową oraz zbudowanie odpowiednich modeli, zapis modeli do pliku, wczytanie danych testowych oraz modeli z pliku, walidacja modeli, zapisanie najlepszej osiągniętej skuteczności.

Wyniki:

Tabela 3 Skuteczność algorytmów w zależności od wielkości słownika

	NaiveBayesMultinomial		Tree	
wordsToKeep	Time(buildmodel)	Correct%	Time(buildmodel)	Correct%
500		66.2602	01:50	55.3862
700		70.3252	01:59	60.7724
900		71.0366	01:58	64.2276
1200		72.6626	02:03	67.8862
1500		73.374	01:36	70.0203
2000		76.0163	01:44	70.8333
3000		76.0163	01:41	72.4593



Wykres 4 Skuteczność algorytmów z selekcją w zależności od wielkości słownika

Wnioski: Zwiększenie liczby słów w słowniku poprawia wyniki algorytmów.

Badanie 4: Wpływ liczby wyselekcjonowanych cech na skuteczność algorytmów oraz czas budowy modeli

Cel badania: Porównanie czasu budowy modelu oraz jego skuteczność dla Naiwnego Bayesa oraz drzewa decyzyjnego w zależności od liczby wyselekcjonowanych słów **dla większej liczby słów w słowniku**.

Stałe w badaniu:

selekcja cech: *ChiSquare*
zliczanie wystąpień: tak
stemmer: *IteratedLovinsStemmer*
stopWordsHandler: *MultiStopWords*
tokenizer: *WordTokenizer*
parametry drzewa: z pruningiem, confidence factor: 0.25, subTreeRaising
rozmiar słownika: 2000 słów na klasę

Zmienne w badaniu:

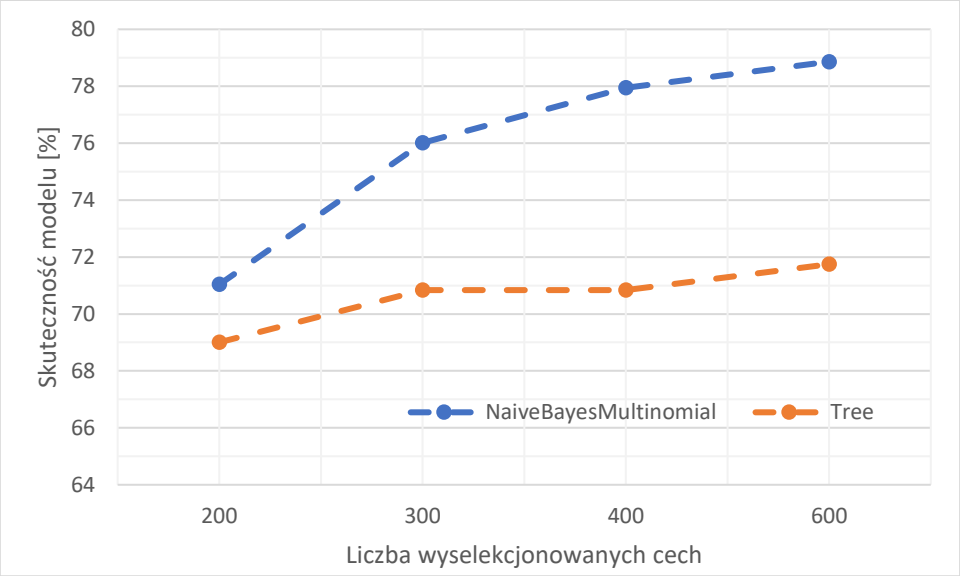
algorytmy: *NaiveBayesMultinomial* oraz drzewo decyzyjne C4.5
liczba wyselekcjonowanych cech: od 20 do 300

Przebieg badania: Załadowanie danych z pliku, ekstrakcja cech z tekstu, selekcja cech, podział zbioru na testowy i treningowy, zapisanie zbiorów do plików, wczytanie danych treningowych, podział walidacją krzyżową oraz zbudowanie odpowiednich modeli, zapis modeli do pliku, wczytanie danych testowych oraz modeli z pliku, walidacja modeli, zapisanie najlepszej osiągniętej skuteczności.

Wyniki:

Tabela 4 Wpływ liczby wyselekcjonowanych cech na skuteczność algorytmów

	NaiveBayesMultinomial		Tree	
numToSelect	Time(buildmodel)	Correct%	Time(buildmodel)	Correct%
200		71.0366	01:11	69.0041
300		76.0163	01:44	70.8333
400		77.9472	03:02	70.8333
600		78.8618	06:18	71.748



Wykres 5 Wpływ liczby wyselekcjonowanych cech na skuteczność algorytmów

Wnioski: Dalsze zwiększanie liczby wyselekcjonowanych słów poprawia skuteczność algorytmów. Dla drzewa decyzyjnego jednak, od pewnego momentu przyrost skuteczności jest niewielki, ale czasu budowy znaczący.

Badanie 5: Porównanie skuteczności różnych metod selekcji cech

Cel badania: Porównanie skuteczności NB i drzewa przy selekcji używając statystyki Chi Kwadrat, współczynników korelacji Pearsona oraz przyrostu informacyjnego

Stałe w badaniu:

zliczanie wystąpień: tak
stemmer: *IteratedLovinsStemmer*
stopWordsHandler: *MultiStopWords*
tokenizer: *WordTokenizer*
parametry drzewa: z pruningiem, confidence factor: 0.25, subTreeRaising
rozmiar słownika: 2000 słów na klasę

Zmienne w badaniu:

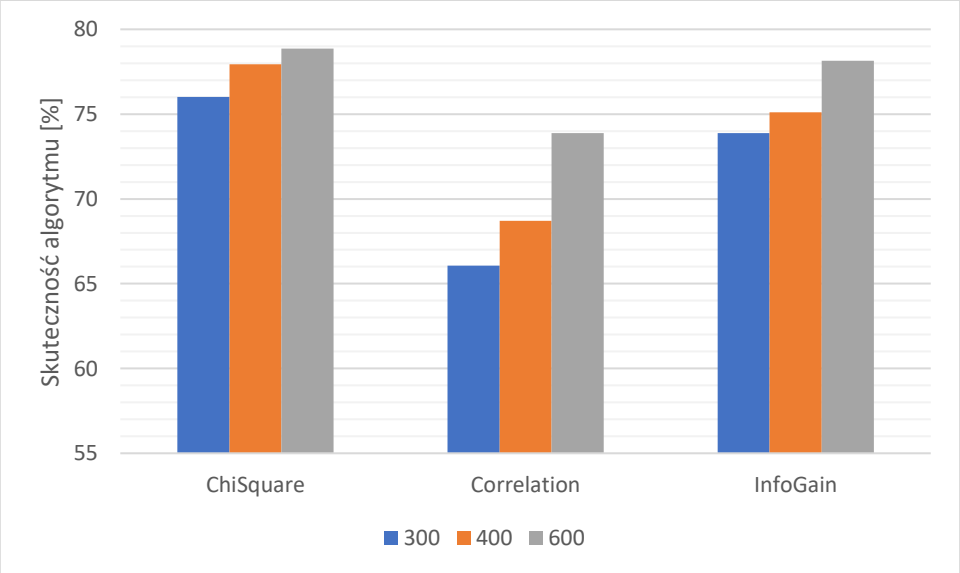
algorytmy: *NaiveBayesMultinomial* oraz drzewo decyzyjne C4.5
liczba wyselekcjonowanych cech: 300, 400, 600
selekcja cech: *ChiSquare*

Przebieg badania: Załadowanie danych z pliku, ekstrakcja cech z tekstu, selekcja cech, podział zbioru na testowy i treningowy, zapisanie zbiorów do plików, wczytanie danych treningowych, podział walidacją krzyżową oraz zbudowanie odpowiednich modeli, zapis modeli do pliku, wczytanie danych testowych oraz modeli z pliku, walidacja modeli, zapisanie najlepszej osiągniętej skuteczności.

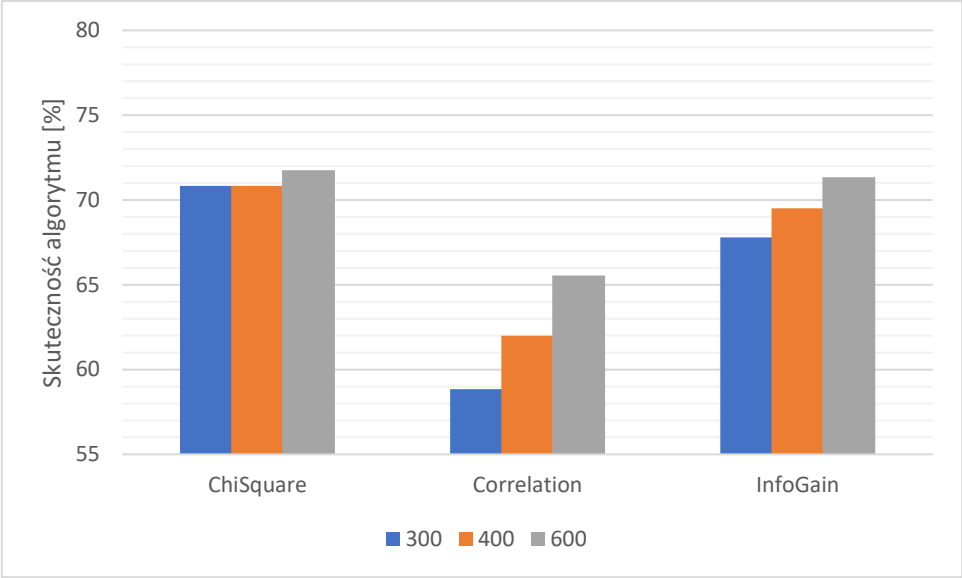
Wyniki:

Tabela 5 Skuteczność algorytmów w zależności od metody selekcji cech

liczba wyselekcjo-nowanych słów	**ChiSquare**			**Correlation**			**InfoGain**		
	NaiveBayes Multinomial	Tree		NaiveBayes Multinomial	Tree		NaiveBayes Multinomial	Tree	
	Correct %	Model build time	Correct %	Correct %	Model build time	Correct %	Correct %	Model build time	Correct %
300	76.0163	01:44	70.8333	66.0569	02:15	58.8415	73.8821	01:07	67.7846
400	77.9472	03:02	70.8333	68.6992	03:39	61.9919	75.1016	03:45	69.5122
600	78.8618	06:18	71.748	73.8821	07:42	65.5488	78.1504	07:58	71.3415



Wykres 6 Skuteczność Naiwnego Bayesa w zależności od metody selekcji cech oraz liczby wyekstraktowanych cech



Wykres 7 Skuteczność drzewa decyzyjnego w zależności od metody selekcji cech oraz liczby wyekstraktowanych cech

Wnioski: Najlepszą metodą do selekcji cech spośród przetestowanych jest metoda wykorzystująca charakterystykę Chi Kwadrat. Lepszą skutecznością charakteryzuje się jednak metoda *Naiwnego Bayesa* od drzewa decyzyjnego.

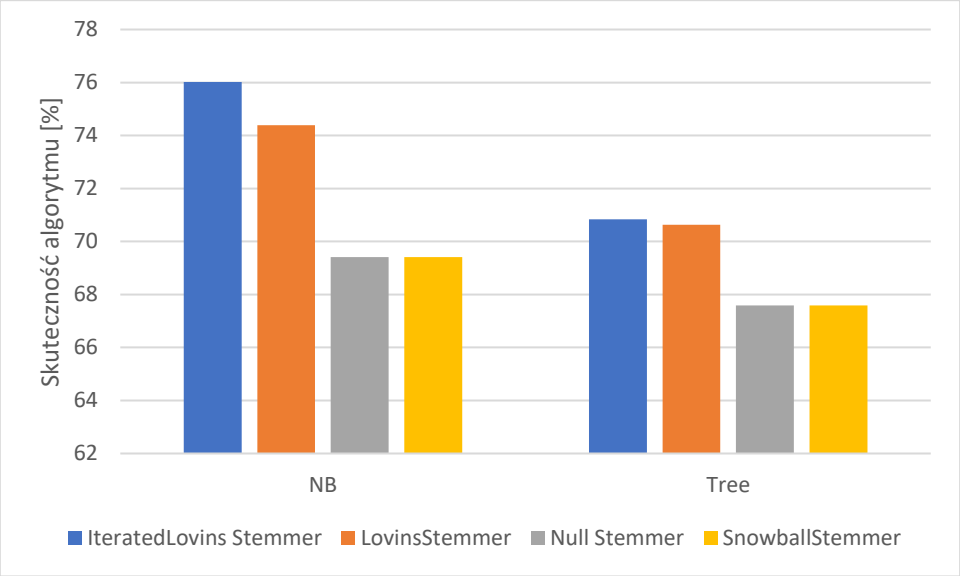
Badanie 6: Porównanie skuteczności różnych metod lematyzacji

Cel badania: Wybór najlepszego stemmera
Stałe w badaniu:
zliczanie wystąpień: tak
stopWordsHandler: *MultiStopWords*
tokenizer: *WordTokenizer*
parametry drzewa: z pruningiem, confidence factor: 0.25, subTreeRaising
rozmiar słownika: 2000 słów na klasę
liczba wyselekcjonowanych cech: 300
selekcja cech: *ChiSquare*
Zmienne w badaniu:
algorytmy: *NaiveBayesMultinomial* oraz drzewo decyzyjne C4.5
stemmer: *IteratedLovinsStemmer*, *LovinsStemmer*, *SnowballStemmer*, brak
Przebieg badania: Załadowanie danych z pliku, ekstrakcja cech z tekstu, selekcja cech, podział zbioru na testowy i treningowy, zapisanie zbiorów do plików, wczytanie danych treningowych, podział walidacją krzyżową oraz zbudowanie odpowiednich modeli, zapis modeli do pliku, wczytanie danych testowych oraz modeli z pliku, walidacja modeli, zapisanie najlepszej osiągniętej skuteczności.

Wyniki:

Tabela 6 Skuteczność różnych stemmerów

stemmer	NB	Tree
IteratedLovins Stemmer	76.0163	70.8333
LovinsStemmer	74.3902	70.6301
Null Stemmer	69.4106	67.5813
SnowballStemmer	69.4106	67.5813



Wykres 8 Skuteczność różnych stemmerów

Wnioski: *IteratedLovinsStemmer* ma najlepszą skuteczność zarówno dla naiwnego Bayesa, jak i drzewa decyzyjnego.

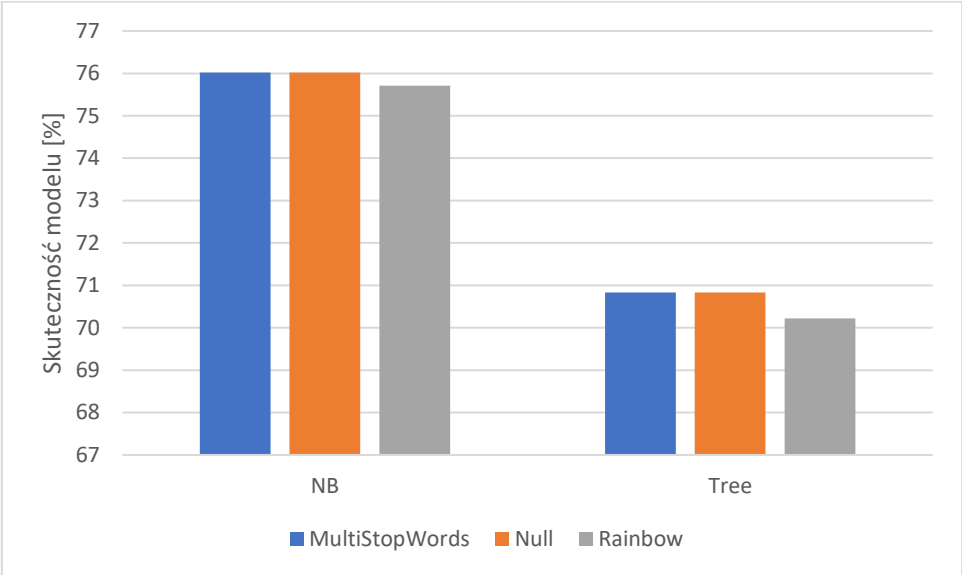
Badanie 7: Porównanie skuteczności różnych stop list

Cel badania: Wybór najlepszej stop listy
Stałe w badaniu:
zliczanie wystąpień: tak
stemmer: *IteratedLovinsStemmer*
tokenizer: *WordTokenizer*
parametry drzewa: z pruningiem, confidence factor: 0.25, subTreeRaising
rozmiar słownika: 2000 słów na klasę
liczba wyselekcjonowanych cech: 300
selekcja cech: *ChiSquare*
Zmienne w badaniu:
algorytmy: *NaiveBayesMultinomial* oraz drzewo decyzyjne C4.5
stopWordsHandler: *MultiStopWords, Rainbow,* brak
Przebieg badania: Załadowanie danych z pliku, ekstrakcja cech z tekstu, selekcja cech, podział zbioru na testowy i treningowy, zapisanie zbiorów do plików, wczytanie danych treningowych, podział walidacją krzyżową oraz zbudowanie odpowiednich modeli, zapis modeli do pliku, wczytanie danych testowych oraz modeli z pliku, walidacja modeli, zapisanie najlepszej osiągniętej skuteczności.

Wyniki:

Tabela 7 Skuteczność różnych stop list

stop lista	NB	Tree
MultiStopWords	76.0163	70.8333
Null	76.0163	70.8333
Rainbow	75.7114	70.2236



Wykres 9 Skuteczność różnych stop list

Wnioski: Zastosowanie stop listy *Rainbow* obniża skuteczność algorytmów.

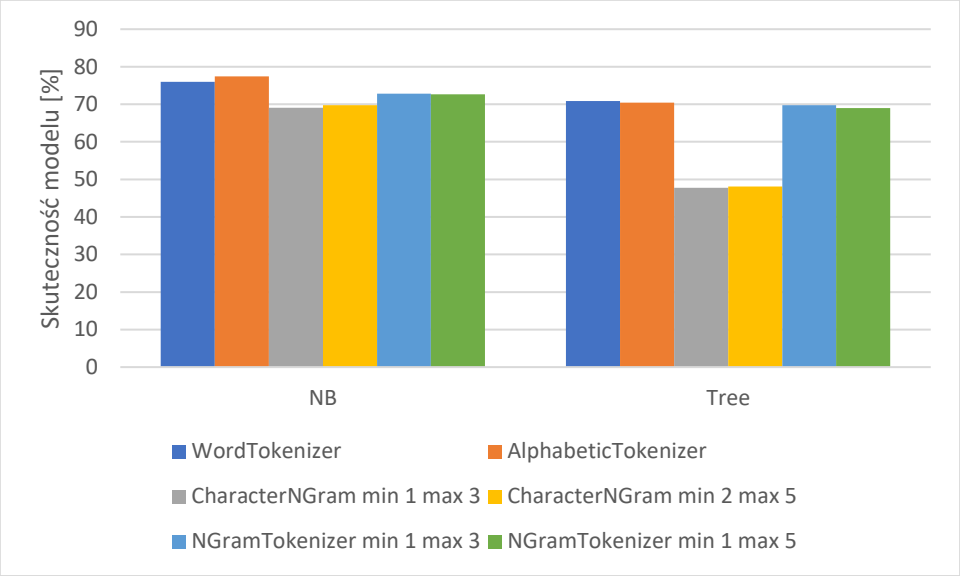
Badanie 8: Porównanie skuteczności różnych metod tokenizacji

Cel badania: Wybór najlepszego tokenizera
Stałe w badaniu:
zliczanie wystąpień: tak
stemmer: *IteratedLovinsStemmer*
stopWordsHandler: *MultiStopWords*
parametry drzewa: z pruningiem, confidence factor: 0.25, subTreeRaising
rozmiar słownika: 2000 słów na klasę
liczba wyselekcjonowanych cech: 300
selekcja cech: *ChiSquare*
Zmienne w badaniu:
algorytmy: *NaiveBayesMultinomial* oraz drzewo decyzyjne C4.5
tokenizer: *WordTokenizer, AlphabeticTokenizer, CharacterNGram, NGramTokenizer*
Przebieg badania: Załadowanie danych z pliku, ekstrakcja cech z tekstu, selekcja cech, podział zbioru na testowy i treningowy, zapisanie zbiorów do plików, wczytanie danych treningowych, podział walidacją krzyżową oraz zbudowanie odpowiednich modeli, zapis modeli do pliku, wczytanie danych testowych oraz modeli z pliku, walidacja modeli, zapisanie najlepszej osiągniętej skuteczności.

Wyniki:

Tabela 8 Wyniki dla różnych tokenizerów

Tokenizer	NB	Tree
WordTokenizer	76.0163	70.8333
AlphabeticTokenizer	77.439	70.4268
CharacterNGram min 1 max 3	69.1057	47.7642
CharacterNGram min 2 max 5	69.7154	48.0691
NGramTokenizer min 1 max 3	72.8659	69.7154
NGramTokenizer min 1 max 5	72.6626	69.0041



Wykres 10 Wyniki dla różnych tokenizerów

Wnioski: Dla Bayesa najlepiej sprawdza się *AlphabeticTokenizer*, a dla drzewa standardowy *WordTokenizer*. *CharacterNGramTokenizer* sprawdza się dużo słabiej dla drzewa.

Naive Bayes

Badanie 9: Porównanie skuteczności Naiwnego Bayesa z rozkładem dwumianowym i wielomianowym oraz zliczaniem liczby słów lub badania tylko ich wystąpienia

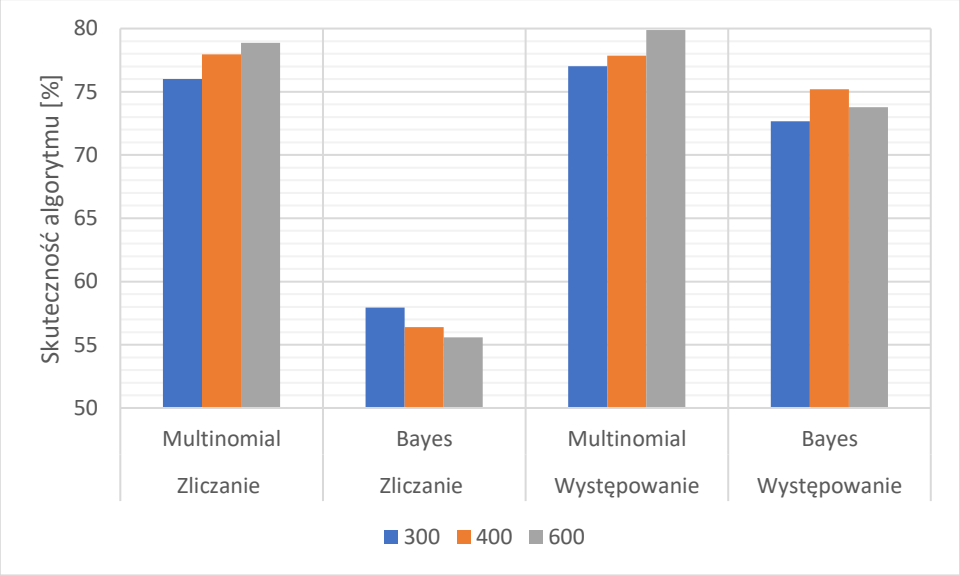
Cel badania: Porównanie Naiwnego Bayesa z rozkładem dwu- i wielomianowym
Stałe w badaniu:
stemmer: *IteratedLovinsStemmer*
stopWordsHandler: *MultiStopWords*
tokenizer: *WordTokenizer*
parametry drzewa: z pruningiem, confidence factor: 0.25, subTreeRaising
rozmiar słownika: 2000 słów na klasę
liczba wyselekcjonowanych cech: 300
selekcja cech: *ChiSquare*
Zmienne w badaniu:
algorytmy: *NaiveBayesMultinomial* oraz *NaiveBayes*
zliczanie wystąpień

Przebieg badania: Załadowanie danych z pliku, ekstrakcja cech z tekstu, selekcja cech, podział zbioru na testowy i treningowy, zapisanie zbiorów do plików, wczytanie danych treningowych, podział walidacją krzyżową oraz zbudowanie odpowiednich modeli, zapis modeli do pliku, wczytanie danych testowych oraz modeli z pliku, walidacja modeli, zapisanie najlepszej osiągniętej skuteczności.

Wyniki:

Tabela 9 Porównanie Naiwnego Bayesa z rozkładem dwu- i wielomianowym ze zliczaniem wystąpień słów i bez

liczba cech	Zliczanie		Występowanie	
	Multinomial	Binomial	Multinomial	Bayes
300	76.0163	57.9268	77.0325	72.6626
400	77.9472	56.4024	77.8455	75.2033
600	78.8618	55.5894	79.878	73.7805



Wykres 11 Porównanie Naiwnego Bayesa z rozkładem dwu- i wielomianowym ze zliczaniem wystąpień słów i bez

Wnioski: Naiwny Bayes działa lepiej z rozkładem wielomianowym. Wszystkie te algorytmy działają lepiej sprawdzając tylko wystąpienia cech, bez ich zliczania, a Bayes z rozkładem dwumianowym radzi sobie dużo gorzej ze zliczaniem wystąpień cech.

Sprawdzenie różnych rodzajów wygładzania dla Naiwnego Bayesa w Wece w GUI Knowledge Flow jest niemożliwe.

Parametry drzewa decyzyjnego

Badanie 10: Porównanie skuteczności drzewa decyzyjnego z włączonym i wyłączonym pruningiem

Cel badania: Sprawdzenie wpływu pruningu
Stałe w badaniu:
stemmer: *IteratedLovinsStemmer*

stopWordsHandler: *MultiStopWords*
tokenizer: *WordTokenizer*
rozmiar słownika: 2000 słów na klasę
liczba wyselekcjonowanych cech: 200
selekcja cech: *ChiSquare*
zliczanie wystąpień: tak
algorytm: drzewo C 4.5 (J48)

Zmienne w badaniu:

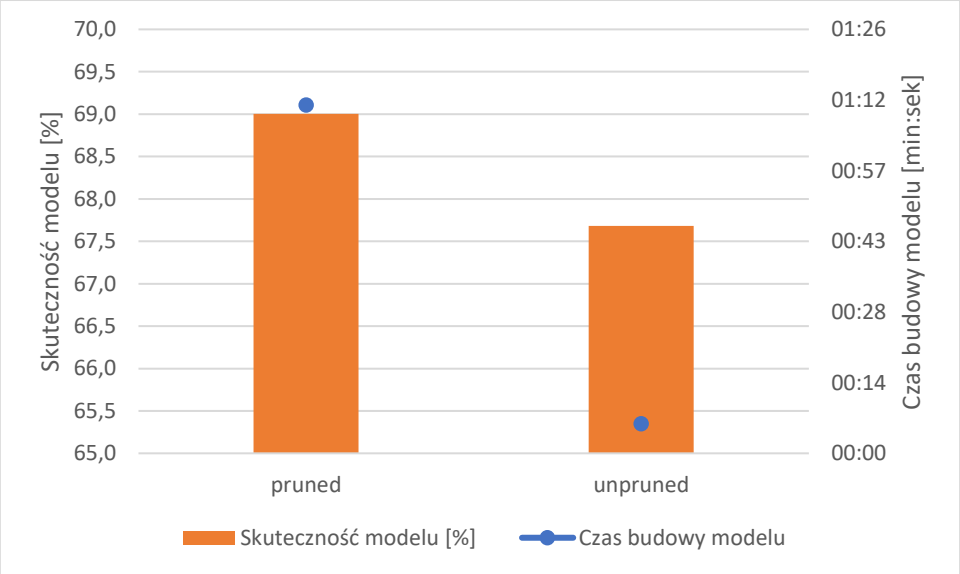
pruning: włączony/wyłączony

Przebieg badania: Załadowanie danych z pliku, ekstrakcja cech z tekstu, selekcja cech, podział zbioru na testowy i treningowy, zapisanie zbiorów do plików, wczytanie danych treningowych, podział walidacją krzyżową oraz zbudowanie odpowiednich modeli, zapis modeli do pliku, wczytanie danych testowych oraz modeli z pliku, walidacja modeli, zapisanie najlepszej osiągniętej skuteczności.

Wyniki:

Tabela 10 Porównanie drzewa decyzyjnego z pruningiem i bez

	Czas budowy modelu	Skuteczność modelu [%]
pruned	01:11	69.0041
unpruned	00:06	67.6829



Wykres 12 Porównanie drzewa z pruningiem i bez

Wnioski: Wyłączenie pruningu znacząco skraca czas budowy drzewa. Powoduje jednak także spadek skuteczności przez zwiększony overfitting.

Badanie 11: Porównanie skuteczności drzewa decyzyjnego dla różnych wartości ConfidenceFactor

Cel badania: Sprawdzenie wpływu *confidence factor* i znalezienie jego optymalnej wartości.

Stałe w badaniu:

stemmer: *IteratedLovinsStemmer*
stopWordsHandler: *MultiStopWords*
tokenizer: *WordTokenizer*
rozmiar słownika: 2000 słów na klasę
liczba wyselekcjonowanych cech: 200
selekcja cech: *ChiSquare*
zliczanie wystąpień: tak
algorytm: drzewo C 4.5 (J48)
parametry drzewa: z pruningiem, subTreeRaising

Zmienne w badaniu:

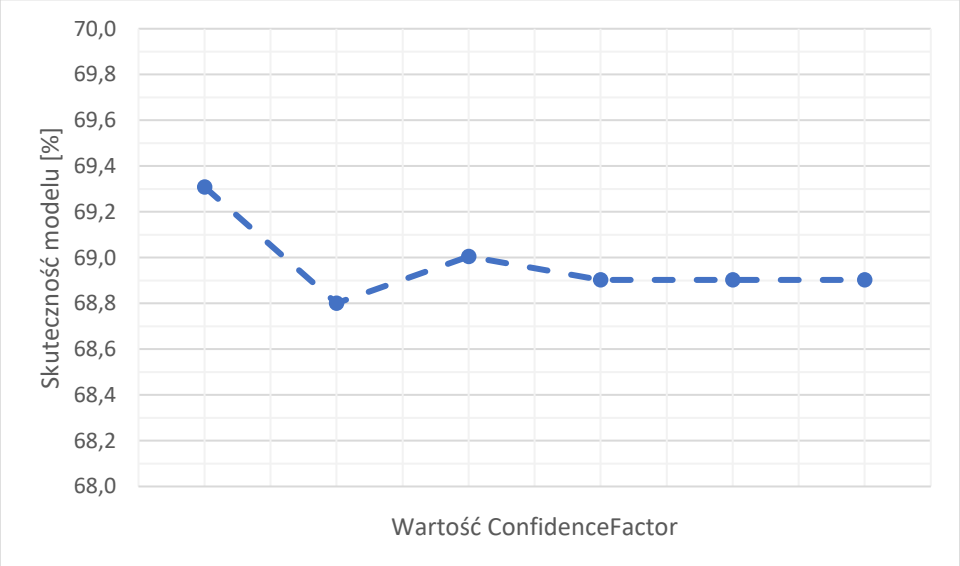
confidenceFactor

Przebieg badania: Załadowanie danych z pliku, ekstrakcja cech z tekstu, selekcja cech, podział zbioru na testowy i treningowy, zapisanie zbiorów do plików, wczytanie danych treningowych, podział walidacją krzyżową oraz zbudowanie odpowiednich modeli, zapis modeli do pliku, wczytanie danych testowych oraz modeli z pliku, walidacja modeli, zapisanie najlepszej osiągniętej skuteczności.

Wyniki:

Tabela 11 Wpływ confidenceFactor na skuteczność modelu

C	Czas budowy	Skuteczność [%]
0.10	00:01:35	69.3089
0.20	00:00:58	68.8008
0.25	00:00:54	69.0041
0.30	00:00:57	68.9024
0.40	00:01:35	68.9024
0.50	00:01:00	68.9024
0.55	00:24:43	68.1911



Wykres 13 Wpływ confidenceFactor na skuteczność modelu

Przyjmuje się, że współczynnik *confidenceFactor* powinien dla J48 w Wece przyjmować wartości (0, 0.5]. Może przyjmować wartości wyższe, wtedy jednak pruning nie jest wykonywany (ale czas budowy drzewa znacząco rośnie).

Wnioski: Lepiej jest określić *confidenceFactor* na poziomie 0.1, aby zwiększyć pruning i zapobiec overfittingowi.

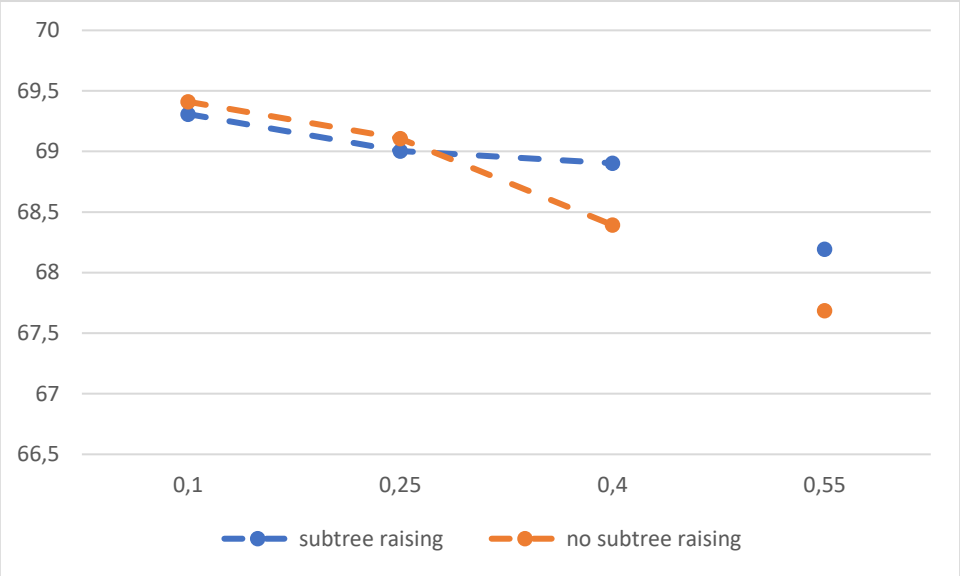
Badanie 12: Porównanie skuteczności drzewa decyzyjnego z subTreeRaising i subTreeReplacement dla różnych wartości ConfidenceFactor

Cel badania: Sprawdzenie wpływu *subtreeRaising* oraz *confidence factor*
Stałe w badaniu:
stemmer: *IteratedLovinsStemmer*
stopWordsHandler: *MultiStopWords*
tokenizer: *WordTokenizer*
rozmiar słownika: 2000 słów na klasę
liczba wyselekcjonowanych cech: 200
selekcja cech: *ChiSquare*
zliczanie wystąpień: tak
algorytm: drzewo C 4.5 (J48)
parametry drzewa: z pruningiem
Zmienne w badaniu:
confidenceFactor
subTreeRaising: on/off
Przebieg badania: Załadowanie danych z pliku, ekstrakcja cech z tekstu, selekcja cech, podział zbioru na testowy i treningowy, zapisanie zbiorów do plików, wczytanie danych treningowych, podział walidacją krzyżową oraz zbudowanie odpowiednich modeli, zapis modeli do pliku, wczytanie danych testowych oraz modeli z pliku, walidacja modeli, zapisanie najlepszej osiągniętej skuteczności.

Wyniki:

Tabela 12 Porównanie włączonego i wyłączonego subtree raisingu dla różnych confidence factor

	subtree raising		no subtree raising	
C	czas	skuteczność	czas	skuteczność
0.1	00:01:35	69.3089	01:02	69.4106
0.25	00:00:54	69.0041	00:43	69.1057
0.4	00:01:35	68.9024	00:55	68.3943
0.55	00:24:43	68.1911	00:17:46	67.6829



Wykres 14 Porównanie włączonego i wyłączonego subtree raisingu dla różnych confidence factor

Wnioski: Ciężko jest jednoznacznie określić, co jest lepsze.

Badanie 13: Porównanie skuteczności drzewa decyzyjnego z reducedErrorPruning dla różnej części wykorzystywanych do pruningu

Cel badania: Sprawdzenie reducedErrorPruningu oraz znalezienie optymalnej wartości numFolds

Stałe w badaniu:

stemmer: *IteratedLovinsStemmer*
stopWordsHandler: *MultiStopWords*
tokenizer: *WordTokenizer*
rozmiar słownika: 2000 słów na klasę
liczba wyselekcjonowanych cech: 200
selekcja cech: *ChiSquare*
zliczanie wystąpień: tak
algorytm: drzewo C 4.5 (J48)
parametry drzewa: z pruningiem reducedErrorPruning

Zmienne w badaniu:

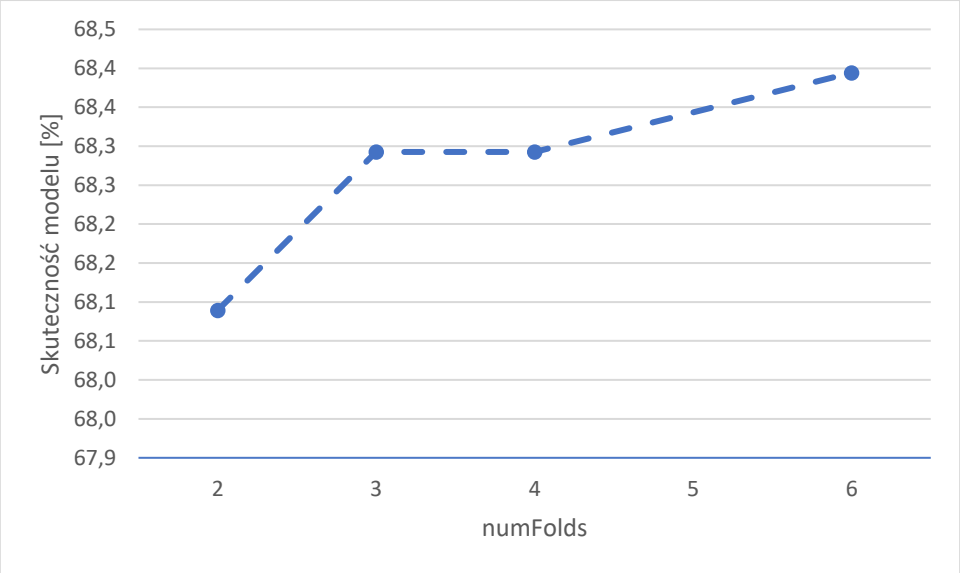
numFolds

Przebieg badania: Załadowanie danych z pliku, ekstrakcja cech z tekstu, selekcja cech, podział zbioru na testowy i treningowy, zapisanie zbiorów do plików, wczytanie danych treningowych, podział walidacją krzyżową oraz zbudowanie odpowiednich modeli, zapis modeli do pliku, wczytanie danych testowych oraz modeli z pliku, walidacja modeli, zapisanie najlepszej osiągniętej skuteczności.

Wyniki:

Tabela 13 Wyniki dla różnych numFolds

numFolds	time	correct %
2	00:27	68.0894
3	00:33	68.2927
4	00:42	68.2927
6	00:38	68.3943
8	00:41	67.6829
10	00:43	68.2927



Wykres 15 Wyniki dla różnych numFolds

Wnioski: Lepiej jest wybrać ½ danych do pruningu, aby stworzyć bardziej ogólny model.

Badanie 14: Porównanie skuteczności drzewa decyzyjnego z reducedErrorPruning dla różnej minimalnej liczby instancji

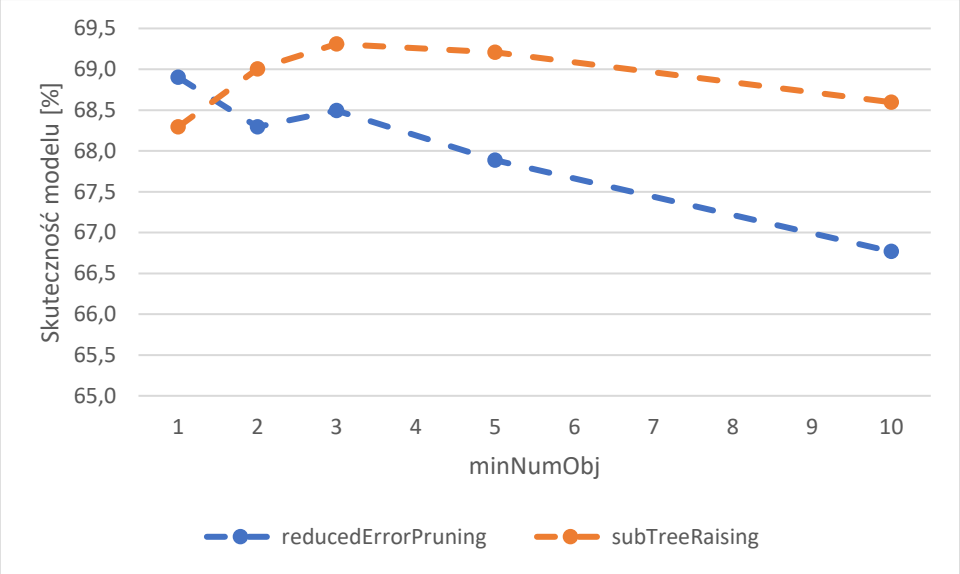
Cel badania: Znalezienie optymalnej wartości minNumObj
Stałe w badaniu:
stemmer: *IteratedLovinsStemmer*
stopWordsHandler: *MultiStopWords*
tokenizer: *WordTokenizer*
rozmiar słownika: 2000 słów na klasę
liczba wyselekcjonowanych cech: 200
selekcja cech: *ChiSquare*
zliczanie wystąpień: tak
algorytm: drzewo C 4.5 (J48)
parametry drzewa: z pruningiem
Zmienne w badaniu:
minNumObj
pruning: reducedErrorPruning (numFolds: 3), subtree raising (confidence factor: 0.25)
Przebieg badania: Załadowanie danych z pliku, ekstrakcja cech z tekstu, selekcja cech, podział zbioru na testowy i treningowy, zapisanie zbiorów do plików, wczytanie danych treningowych, podział walidacją krzyżową oraz zbudowanie odpowiednich modeli, zapis

modeli do pliku, wczytanie danych testowych oraz modeli z pliku, walidacja modeli, zapisanie najlepszej osiągniętej skuteczności.

Wyniki:

Tabela 14 Wyniki w zależności od minimalnej liczby instancji

	reducedErrorPruning -n3		subTreeRaising C 0.25	
minNumObj	time	correct %	time	correct %
1	00:35	68.9024	01:08	68.2927
2	00:33	68.2927	00:53	69.0041
3	00:29	68.4959	01:00	69.3089
5	00:27	67.8862	01:04	69.2073
10	00:24	66.7683	00:45	68.5976



Wykres 16 Wyniki w zależności od minimalnej liczby instancji

Pozostałe badania

Badanie 15: Działanie klasyfikatorów dla różnych podzbiorów klas decyzyjnych – po usunięciu klas z najwyższym FP Rate

Cel badania: Zbadanie skuteczności modelu po usunięciu części klas
Stałe w badaniu:
stemmer: *IteratedLovinsStemmer*
stopWordsHandler: *MultiStopWords*
tokenizer: *WordTokenizer*
rozmiar słownika: 2000 słów na klasę
liczba wyselekcjonowanych cech: 400
selekcja cech: *ChiSquare*
zliczanie wystąpień: tak
algorytm: NaiveBayesMultiNomial , drzewo

Przebieg badania: Załadowanie danych z pliku, ekstrakcja cech z tekstu, selekcja cech, podział zbioru na testowy i treningowy, zapisanie zbiorów do plików, wczytanie danych treningowych, podział walidacją krzyżową oraz zbudowanie odpowiednich modeli, zapis modeli do pliku, wczytanie danych testowych oraz modeli z pliku, walidacja modelu, zapis informacji o prawdopodobieństwie słowa pod warunkiem klasy.

Wyniki:
usunięto z puli danych 3 klasy o najwyższym wskaźniku *False Positive Rate*:

Klasa	FP Rate
propaganda-polityczna	0,022
niemieccy wojskowi	0,020
komiksy	0,016

	NB	Tree
Wszystkie klasy	77.9472	70.8333
Usunięta część klas	83.1096	77.2931

Wnioski: Usunięcie klas, które algorytm najsłabiej rozpoznaje wpływa na poprawę jego skuteczności.

Badanie 16: Działanie klasyfikatorów dla różnych podzbiorów klas decyzyjnych – po usunięciu klas z najwyższym TP Rate

Cel badania: Zbadanie skuteczności modelu po usunięciu części klas
Stałe w badaniu:
stemmer: *IteratedLovinsStemmer*
stopWordsHandler: *MultiStopWords*
tokenizer: *WordTokenizer*
rozmiar słownika: 2000 słów na klasę
liczba wyselekcjonowanych cech: 400
selekcja cech: *ChiSquare*
zliczanie wystąpień: tak
algorytm: NaiveBayesMultiNomial , drzewo
Przebieg badania: Załadowanie danych z pliku, ekstrakcja cech z tekstu, selekcja cech, podział zbioru na testowy i treningowy, zapisanie zbiorów do plików, wczytanie danych treningowych, podział walidacją krzyżową oraz zbudowanie odpowiednich modeli, zapis modeli do pliku, wczytanie danych testowych oraz modeli z pliku, walidacja modelu, zapis informacji o prawdopodobieństwie słowa pod warunkiem klasy.

Wyniki:
usunięto z puli danych 3 klasy o najwyższym wskaźniku *True Positive Rate*: Karkonosze, Amerykanscy-prozaicy, Wojska-pancerne.

	NB	Tree
Wszystkie klasy	77.9472	70.8333
Usunięta część klas	78.4358	71.6201

Wnioski: Usunięcie części klas, nawet jeżeli klasy te były dobrze rozpoznawane przez algorytm, zwiększa jego skuteczność.

Badanie 17: Określenie zestawów cech istotnych dla poszczególnych klas

Cel badania: Znalezienie cech istotnych dla poszczególnych klas oraz co one mówią nam o poszczególnych klasach w danych

Stałe w badaniu:

stemmer: *IteratedLovinsStemmer*
stopWordsHandler: *Null*
tokenizer: *WordTokenizer*
rozmiar słownika: 3000 słów na klasę
liczba wyselekcjonowanych cech: 100
selekcja cech: *ChiSquare*
zliczanie wystąpień: tak
algorytm: NaiveBayesMultiNomial

Przebieg badania: Załadowanie danych z pliku, ekstrakcja cech z tekstu, selekcja cech, podział zbioru na testowy i treningowy, zapisanie zbiorów do plików, wczytanie danych treningowych, podział walidacją krzyżową oraz zbudowanie odpowiednich modeli, zapis modeli do pliku, wczytanie danych testowych oraz modeli z pliku, walidacja modelu, zapis informacji o prawdopodobieństwie słowa pod warunkiem klasy.

Wyniki:

Tabela 15 Klasa: Karkonosze - najczęstsze słowa

słowa	prawdop.
karkonosz	0.11
turystyczn	0.06
gór	0.06
wydawnictw	0.06
karkonoskieg	0.05
czesk	0.04
karkonoszach	0.04
25nbsp	0.04
grzbietu	0.04
jel	0.03
pttk	0.03
sudetów	0.03
turystycznej	0.03
staff	0.03
geograf	0.03
szlak	0.03
kraków	0.03
marek	0.03
8370051685map	0.02
bibliografiasłownik	0.02
sudetach	0.02
granitu	0.02

Tabela 16 Klasa: Gry-komputerowe - najczęstsze słowa

słowa	P
gry	0.31
gracz	0.21
gier	0.11
th	0.09
of	0.09
punktów	0.03
silnik	0.02
czołgów	0.01
czołg	0.01
mistrzostw	0.01
samochód	0.01
amerykańsk	0.01
fant	0.01

Tabela 17 Katolicyzm - najczęstsze słowa i ich prawdopodobieństwo

słowa	Katolicyzm
ur	0.26
kraków	0.1
of	0.07
th	0.06
wydawnictw	0.06
amerykańsk	0.04
gór	0.02
marek	0.02
monet	0.02
kar	0.02
pisarz	0.02
samochód	0.01
turystyczn	0.01
czesk	0.01
geograf	0.01
kot	0.01
karierz	0.01
roślin	0.01
wynikiem	0.01
powieśc	0.01
najwyższ	0.01
kadłub	0.01
egiptu	0.01
dotychczasowej	0.01
awer	0.01

Tabela 18 Narkomania - najczęstsze słowa i ich prawdopodobieństwo

słowa	Narkomania
lek	0.25
chemiczn	0.18
organiczn	0.16
roślin	0.13
th	0.06
of	0.03
kraków	0.01
wydawnictw	0.01
amerykańsk	0.01

Tabela 19 Monety - najczęstsze słowa i ich prawdopodobieństwo

słowa	Monety
monet	0.61
bit	0.14
awer	0.08
th	0.01
of	0.01
kraków	0.01
wydawnictw	0.01
amerykańsk	0.01
gór	0.01
marek	0.01
czesk	0.01

Tabela 20 Galezie-prawa - najczęstsze słowa i ich prawdopodobieństwo

słowa	Galezie-prawa
ur	0.2
najwyższ	0.1
of	0.09
kar	0.09
kraków	0.06
wydawnictw	0.04
th	0.03
amerykańsk	0.02
pisarz	0.02
gór	0.01
marek	0.01
organiczn	0.01
pancernych	0.01
medycznej	0.01
podatku	0.01

Tabela 21 Amerykanscy-prozaicy - najczęstsze słowa i ich prawdopodobieństwo

słowa	Amerykanscy-prozaicy
th	0.5
of	0.24
powieśc	0.05
ur	0.03
amerykańsk	0.03
pisarz	0.03
fant	0.03
fict	0.03
wydawnictw	0.02

Tabela 22 Narciarstwo - najczęstsze słowa i ich prawdopodobieństwo

słowa	Narciarstwo
mistrzostw	0.2
mistrzostwach	0.18
ur	0.13
narciarsk	0.13
punktów	0.06
turnieju	0.06
karierz	0.05
wynikiem	0.03
amerykańsk	0.01
najwyższ	0.01
kar	0.01
gór	0.01
drużynowych	0.01

Tabela 23 Wojska-pancerne - najczęstsze słowa i ich prawdopodobieństwo

słowa	Wojska-pancerne
czołgów	0.2
czołg	0.18
pancern	0.14
czołgu	0.1
pancernych	0.09
pancernej	0.09
kadłub	0.04
silnik	0.04
prototyp	0.02
amerykańsk	0.01
gór	0.01
wydawnictw	0.01
samochód	0.01
KM	0.01
podwoz	0.01

Tabela 24 Filmy-animowane- najczęstsze słowa i ich prawdopodobieństwo

słowa	Filmy-animowane
th	0.25
animowan	0.18
amerykańsk	0.08
of	0.07
marek	0.05
kot	0.04
samochód	0.03
gry	0.03
wydawnictw	0.02
powieśc	0.02
komik	0.02
czołg	0.01
gór	0.01
wynikiem	0.01
najwyższ	0.01
kar	0.01
fant	0.01
egiptu	0.01
gracz	0.01
gier	0.01
samolot	0.01
samolotu	0.01
rank	0.01
okręt	0.01
samolotów	0.01
chińsk	0.01

Tabela 25 Pilka-nożna - najczęstsze słowa i ich prawdopodobieństwo

słowa	Pilka-nozna
kar	0.21
piłkarz	0.19
mistrzostw	0.16
ur	0.14
turnieju	0.05
karierz	0.03
gry	0.02
gracz	0.02
mistrzostwach	0.02
th	0.01
of	0.01
wynikiem	0.01
najwyższ	0.01
egiptu	0.01
punktów	0.01
kraków	0.01
turniejach	0.01

Tabela 26 Choroby - najczęstsze słowa i ich prawdopodobieństwo

słowa	Choroby
lek	0.23
th	0.08
of	0.08
wydawnictw	0.05
roślin	0.04
organiczn	0.03
chemiczn	0.03
kotów	0.03
kar	0.02
wynikiem	0.02
amerykańsk	0.02
medycznej	0.02
ur	0.01
gry	0.01
najwyższ	0.01
kraków	0.01
marek	0.01
kot	0.01
powieśc	0.01
gór	0.01
KM	0.01
pisarz	0.01
jel	0.01
szlak	0.01
podzielił	0.01

Tabela 27 Samoloty - najczęstsze słowa i ich prawdopodobieństwo

słowa	Samoloty
samolot	0.29
samolotu	0.17
samolotów	0.11
silnik	0.09
kadłub	0.08
prototyp	0.06
podwoz	0.04
KM	0.03
oblatan	0.03
samolotem	0.03
amerykańsk	0.02
th	0.01
of	0.01

Tabela 28 Egipt - najczęstsze słowa i ich prawdopodobieństwo

słowa	Egipt
egiptu	0.18
egipc	0.13
ur	0.05
of	0.04
th	0.03
gór	0.02
mistrzostw	0.02
samolot	0.01
wydawnictw	0.01
roślin	0.01
kar	0.01
wynikiem	0.01
najwyższ	0.01
kot	0.01
szlak	0.01
okręt	0.01
czołgów	0.01
turystyczn	0.01
tiran	0.01

Tabela 29 Kultura Chin-- najczęstsze słowa i ich prawdopodobieństwo

słowa	Kultura-Chin
chińsk	0.16
th	0.11
of	0.1
ur	0.09
gracz	0.09
gór	0.07
punktów	0.03
wydawnictw	0.02
roślin	0.02
najwyższ	0.02
gry	0.02
powieśc	0.02
pisarz	0.02
kar	0.01
amerykańsk	0.01

Tabela 30 Pierwiastki-chemiczne - najczęstsze słowa i ich prawdopodobieństwo

słowa	Pierwiastki-chemiczne
chemiczn	0.26
pierwiastek	0.23
roślin	0.04
th	0.03
monet	0.03
of	0.02
wydawnictw	0.02
kar	0.02
silnik	0.02
organiczn	0.02
chińsk	0.01
najwyższ	0.01
amerykańsk	0.01
wynikiem	0.01
samolotu	0.01
samolotów	0.01
kadłub	0.01
lek	0.01

Tabela 31 Komputery - najczęstsze słowa i ich prawdopodobieństwo

słowa	Komputery
bit	0.07
jel	0.06
th	0.05
of	0.05
punktów	0.05
gry	0.05
prototyp	0.05
amerykańsk	0.03
gracz	0.03
gier	0.03
rank	0.03
wydawnictw	0.02
chińsk	0.02
wynikiem	0.02
25nbsp	0.02
ur	0.02
pierwiastek	0.01
kar	0.01
silnik	0.01
najwyższ	0.01
lek	0.01
mistrzostw	0.01
KM	0.01
medycznej	0.01
marek	0.01
animowan	0.01
szachowych	0.01
szachow	0.01

Tabela 32 Rachunkowosc - najczęstsze słowa i ich prawdopodobieństwo

słowa	Rachunkowosc
podatku	0.66
of	0.06
wydawnictw	0.03
th	0.02
amerykańsk	0.01
gier	0.01
wynikiem	0.01
ur	0.01
kar	0.01
silnik	0.01
najwyższ	0.01
gór	0.01
kraków	0.01
samochód	0.01

Tabela 33 Propaganda-polityczna- najczęstsze słowa i ich prawdopodobieństwo

słowa	Propaganda-polityczna
ur	0.11
of	0.08
wydawnictw	0.07
pisarz	0.07
th	0.04
powieśc	0.04
czołg	0.04
czołgu	0.04
kar	0.03
marek	0.03
albańsk	0.03
amerykańsk	0.02
gór	0.02
kraków	0.02
czołgów	0.02
pancern	0.02

alban	0.02
kompoz	0.02
wynikiem	0.01
najwyższ	0.01
samochód	0.01
bit	0.01
gry	0.01
chińsk	0.01
monet	0.01
samolotu	0.01
samolotów	0.01
kadłub	0.01
podzielił	0.01
samolot	0.01
tiran	0.01
kotów	0.01
pancernych	0.01
pancernej	0.01
czesk	0.01
geograf	0.01
granitu	0.01

Tabela 34 Niemieccy-wojskowi - najczęstsze słowa i ich prawdopodobieństwo

słowa	Niemieccy-wojskowi
ur	0.37
wydawnictw	0.08
of	0.07
kraków	0.05
th	0.04
okręt	0.04
kar	0.03
samolotów	0.02
samolot	0.02
pancernych	0.02
pisarz	0.01
marek	0.01
amerykańsk	0.01
czołgów	0.01
pancern	0.01
najwyższ	0.01
samolotu	0.01
pancernej	0.01
geograf	0.01
silnik	0.01
samolotem	0.01

Tabela 35 Arabowie - najczęstsze słowa i ich prawdopodobieństwo

słowa	Arabowie
ur	0.27
mistrzostw	0.08
th	0.07
of	0.05
wydawnictw	0.03
pisarz	0.03
egiptu	0.03
mistrzostwach	0.03
kraków	0.02
samolotu	0.02
powieśc	0.02
egipc	0.02
kar	0.01
samolot	0.01
pancernych	0.01
amerykańsk	0.01
geograf	0.01
gór	0.01
kompoz	0.01
wynikiem	0.01
samochód	0.01
podzielił	0.01
kot	0.01
piłkarz	0.01
turnieju	0.01
fict	0.01
bibliografiasłownik	0.01

Tabela 36 Astronautyka - najczęstsze słowa i ich prawdopodobieństwo

słowa	Astronautyka
silnik	0.18
amerykańsk	0.11
of	0.09
kadłub	0.07
ur	0.05
samolot	0.04
podwoz	0.04
th	0.03
kar	0.03
samolotów	0.03
samolotu	0.02

okręt	0.02
chińsk	0.02
wydawnictw	0.01
gór	0.01
wynikiem	0.01
czołgu	0.01
punktów	0.01
prototyp	0.01
chemiczn	0.01
roślin	0.01

Tabela 37 Kotowate - najczęstsze słowa i ich prawdopodobieństwo

słowa	Kotowate
kot	0.46
kotów	0.15
ogon	0.09
kotowatych	0.08
of	0.04
th	0.02
amerykańsk	0.01
chińsk	0.01
gór	0.01
powieśc	0.01
egipc	0.01
jel	0.01

Tabela 38 Albania - najczęstsze słowa i ich prawdopodobieństwo

słowa	Albania
alban	0.31
tiran	0.16
albańsk	0.15
ur	0.08
of	0.02
th	0.02
powieśc	0.02
wydawnictw	0.02
pisarz	0.02
gór	0.01
kar	0.01
punktów	0.01
mistrzostw	0.01
mistrzostwach	0.01
kompoz	0.01
lek	0.01

Tabela 39 Ekologia-roslin - najczęstsze słowa i ich prawdopodobieństwo

słowa	Ekologia-roslin
roślin	0.64
gór	0.03
wydawnictw	0.02
organiczn	0.02
of	0.01
th	0.01
chińsk	0.01
chemiczn	0.01
kraków	0.01
marek	0.01
najwyższ	0.01
szlak	0.01
sudetach	0.01

Tabela 40 Optyka - najczęstsze słowa i ich prawdopodobieństwo

słowa	Optyka
chemiczn	0.11
of	0.1
organiczn	0.08
th	0.07
roślin	0.04
ur	0.04
wydawnictw	0.03
silnik	0.03
gór	0.02
punktów	0.02
amerykańsk	0.02
wynikiem	0.02
marek	0.01
najwyższ	0.01
pisarz	0.01
kar	0.01
kompoz	0.01
ogon	0.01
egipc	0.01
samolot	0.01
okręt	0.01
prototyp	0.01
podzielił	0.01
bibliografiasłownik	0.01
czołgów	0.01
gry	0.01
monet	0.01

Tabela 41 System-opieki-zdrowotnej-w-Polsce - najczęstsze słowa i ich prawdopodobieństwo

słowa	System-opieki-zdrowotnej -w-Polsce
medycznej	0.27
ur	0.19
wydawnictw	0.08
of	0.07
lek	0.05
roślin	0.03
th	0.02
marek	0.02
chemiczn	0.01
gór	0.01
wynikiem	0.01
najwyższ	0.01
kar	0.01
kraków	0.01
rank	0.01

Tabela 42 Zegluga - najczęstsze słowa i ich prawdopodobieństwo

słowa	Zegluga
okręt	0.46
th	0.08
kadłub	0.07
of	0.05
amerykańsk	0.04
ur	0.03
samolotów	0.03
silnik	0.02
samolot	0.02
wydawnictw	0.01
gór	0.01
wynikiem	0.01
monet	0.01
chińsk	0.01
szlak	0.01
powieśc	0.01
mistrzostw	0.01
samolotu	0.01
pancern	0.01
KM	0.01

Tabela 43 Muzyka-powazna - najczęstsze słowa i ich prawdopodobieństwo

słowa	Muzyka-powazna
kompoz	0.3
ur	0.13
th	0.11
of	0.07
kraków	0.06
wydawnictw	0.05
gry	0.03
amerykańsk	0.01
marek	0.01
kar	0.01
punktów	0.01
pisarz	0.01
turnieju	0.01
karierz	0.01
fant	0.01

Tabela 44 Samochody - najczęstsze słowa i ich prawdopodobieństwo

słowa	Samochody
silnik	0.34
KM	0.24
samochód	0.21
prototyp	0.03
of	0.02
podwoz	0.02
th	0.01
amerykańsk	0.01
punktów	0.01
kadłub	0.01
mistrzostw	0.01
mistrzostwach	0.01
pancernych	0.01

Tabela 45 Zydzi - najczęstsze słowa i ich prawdopodobieństwo

słowa	Zydzi
ur	0.22
th	0.12
of	0.1
wydawnictw	0.06
amerykańsk	0.03
marek	0.03
kar	0.03

kraków	0.02
pisarz	0.02
pancernych	0.01
kompoz	0.01
karierz	0.01
gór	0.01
powieśc	0.01
pancern	0.01
medycznej	0.01
kot	0.01
egiptu	0.01
pancernej	0.01
czołg	0.01

Tabela 46 Szachy - najczęstsze słowa i ich prawdopodobieństwo

słowa	Szachy
mistrzostw	0.09
turnieju	0.08
mistrzostwach	0.07
szach	0.06
ur	0.05
punktów	0.05
podzielił	0.05
szachistów	0.05
kar	0.04
karierz	0.04
wynikiem	0.04
najwyższ	0.04
rank	0.04
turniejach	0.04
arcymistrz	0.04
szachow	0.03
drużynowych	0.03
megab	0.03
bibliografiachessb	0.03
szachowych	0.02
dotychczasowej	0.02
th	0.01
gry	0.01

Tabela 47 Sporty-silowe - najczęstsze słowa i ich prawdopodobieństwo

słowa	Sporty-silowe
strongman	0.36
mistrzostw	0.32
ur	0.1
mistrzostwach	0.09
wynikiem	0.02
kar	0.01
drużynowych	0.01
th	0.01
of	0.01
amerykańsk	0.01

Tabela 48 Komiksy - najczęstsze słowa i ich prawdopodobieństwo

słowa	Komiksy
th	0.25
komik	0.18
of	0.12
wydawnictw	0.08
ur	0.03
amerykańsk	0.03
gry	0.03
kar	0.02
powieśc	0.02
gier	0.02
turnieju	0.01
marek	0.01
kraków	0.01
pisarz	0.01
gór	0.01
kot	0.01
czołg	0.01
samochód	0.01
fant	0.01
samolot	0.01
ogon	0.01
fict	0.01
animowan	0.01

Wnioski: Najczęściej występujące słowa często należą do tej samej rodziny wyrazów, co trzon klasy, ale też są powiązane tematycznie z klasą. Istnieją jednak klasy, z którymi powiązanych jest dużo słów, które nie identyfikują ich bardzo jednoznacznie.

Podsumowanie

SI ma wiele zastosowań. Do rozpoznawania tekstu w tym przypadku lepiej sprawdza się model NaiveBayesMultinomial – ma on wyższe wyniki i krótszy czas budowy niż drzewo decyzyjne, jest bardziej odporny na szumy i overfitting.