Jarosław Butajło
Paweł Wąsowski

Metody odkrywania wiedzy - projekt
Temat: Filtrowanie poczty elektronicznej

Temat projektu

Tematem projektu jest filtrowanie wiadomości poczty elektronicznej. Zadaniem jest próba stworzenia klasyfikatora wiadomości poczty elektronicznej, filtrującego pożądaną korespondencję od niechcianych wiadomości (spamu). Klasyfikacja wiadomości ma odbywać się na podstawie tekstu ich treści.
Celem projektu jest sprawdzenie, z jaką skutecznością, można na podstawie treści, odfiltrować niepożądaną korespondencję, i który z przetestowanych algorytmów nadaje się najlepiej do tego zadania. Przy ocenie użyteczności algorytmów wzięto pod uwagę to, jaka część spamu zostanie poprawnie zaklasyfikowana.

Zbiór danych

Klasyfikatory trenowano i testowano na wiadomościach ze zbioru http://spamassassin.apache.org/old/publiccorpus.
Zbiór ten zawiera 6049 wiadomości w języku angielskim, z czego 31 % stanowi spam. Wykorzystany zbiór pożądanych wiadomości został podzielony przez udostępniającego na dwa podzbiory wiadomości, opisane jako łatwo i trudno odróżnialne od spamu.
Wiadomości w zbiorze danych, oprócz treści wiadomości, zawierają metadane, m.in. adresy serwerów pocztowych, nadawców i odbiorców wiadomości, tematy wiadomości.
Treść części wiadomości zawiera znaczniki *HTML* i słowa w języku angielskim.

Przygotowanie zbioru danych do analizy

Przed przystąpieniem do analizy danych, z wiadomości wyekstrahowano ich treść, odrzucając metadane. W tekście pozostawiono znaczniki *HTML*, ale usunięto ich nawiasy i domknięcia – np. para *<td> </td>* została przekształcona na *td*.
Wszystkie litery tekstów zostały zamienione na małe. Ponadto usunięto tzw. *stopwords*, słowa krótsze niż 3-literowe, słowa występujące rzadziej, niż w 1% dokumentów i znaki interpunkcyjne – te termy prawdopodobnie nie miałyby dużego wpływu na jakość klasyfikacji.

Wiadomości o różnych poziomach trudności klasyfikacji wymieszano ze sobą, tworząc jeden zbiór, zawierający 6049 elementów.

Wektorowe reprezentacje tekstu

Na potrzeby analizy, teksty przkształcono do postaci wektorów liczb. Wykorzystano 3 popularne reprezentacje wektorowe tekstu:

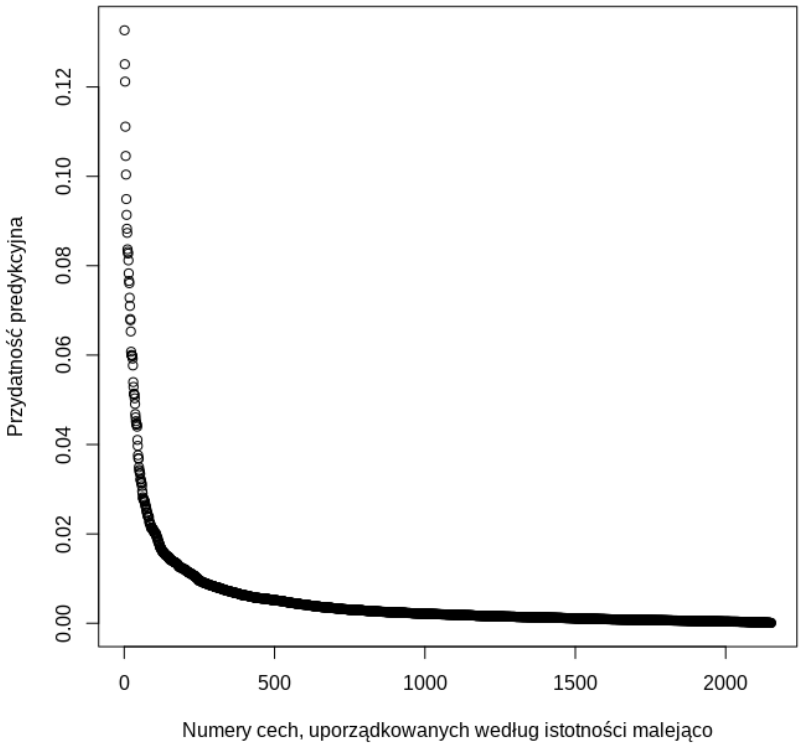
- tf (ang. *term-frequency*) – każdemu słowu przypisana jest liczba wystąpień w danym tekście
- tf-idf (ang. *term frequency – inverse document frequency*) z normalizacją tf – każdemu słowu przypisana jest liczba wystąpień w danym tekście, podzielona przez liczbę wszystkich słów w tym tekście i pomnożona przez czynnik normalizujący idf:

idf = log (liczba dokumentów w zbiorze danych / liczba dokumentów, w których dane słowo występuje przynajmniej raz)

- binarną - przypisująca wagę „1” słowom występującym w danym tekście i „0” pozostałym słowom

Selekcja atrybutów

Selekcji atrybutów dodkonano na podstawie analizy ich predykcyjnej przydatności. Miarą przydatności atrybutu była wartość indeksu Giniego.



Przydatność predykcyjna cech dla reprezentacji tf-idf

Do dalszej analizy wybrano po 200 cech o największych przydatnościach, wyznaczonych osobno dla każdej z reprezentacji wektorowych. Przydatność predykcyjna pozostałych cech jest stosunkowo mała, co widać na powyższym wykresie.

Testy klasyfikatorów

Opis procedury testej – część wspólna dla wszystkich eksperymentów

Testów jakości wszystkich klasyfikatorów dokonywano metodą 5-krotnej walidacji krzyżowej. Zbiór testowy w każdej iteracji walidacji krzyżowej miał 1210 elementów, spośród których spam stanowił średnio 31%. Badano średnie wartości błędu klasyfikacji oraz komórek macierzy pomyłek:

- TP (ang. true-positive) – liczba poprawnie zaklasyfikowanych pożądanych wiadomości
- FP (ang. false-positive) – liczba wiadomości niechcianych, zaklasyfikowanych jako pożądana korespondencja
- TN (ang. true-negative) – liczba poprawnie klasyfikowanych niechcianych wiadomości
- FN (ang. false-negative) – liczba wiadomości pożądanych, zaklasyfikowanych jako spam

Dla każdego z klasyfikatorów przeprowadzono testy klasyfikacji z każdą z reprezentacji wektorowych tekstów.

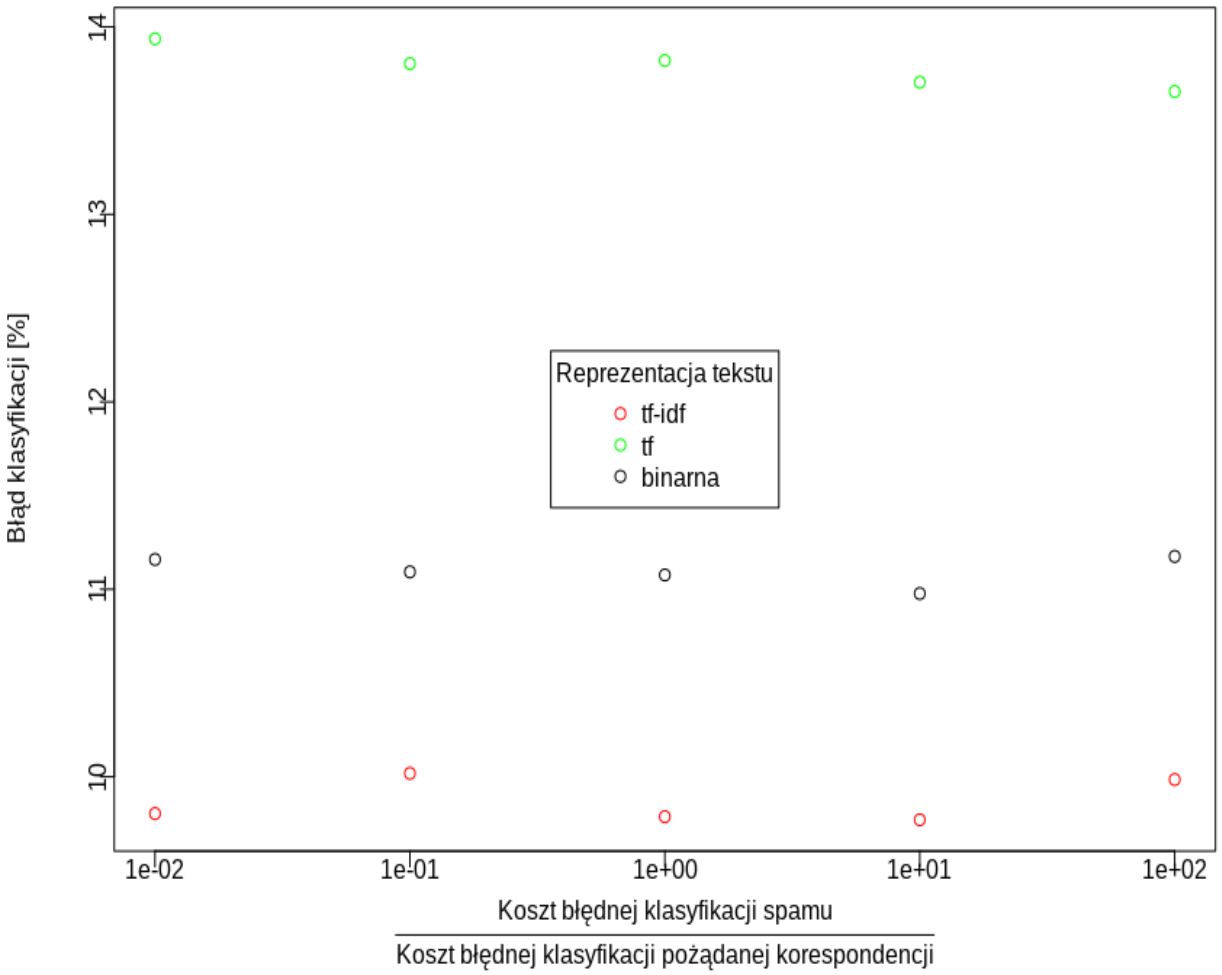
Naiwny klasyfikator Bayesa

Klasyfikator Bayesa może zwracać wyniki predykcji w postaci etykiety przewidywanej klasy, albo prawdopodobieństw poszczególnych klas. Wykorzystując drugi rodzaj wyniku sprawdzono, jaki wpływ na błąd klasyfikacji i wartości w macierzy pomyłek ma ustalenie różnych kosztów dla błędnej klasyfikacji spamu i pożądanej korespondencji.

Sprawdzono 5 różnych stosunków kosztu błędnej klasyfikacji spamu do kosztu błędnej klasyfikacji pożądanych wiadomości: 0,01; 0,1; 1; 10; 100. Koszt błędnej klasyfikacji pożądanej korespondencji zawsze wynosił 1.

W eksperymentach wykorzystano implementację klasyfikatora *naiveBayes* z pakietu *e1071*. Uwzględnianie kosztów pomyłek zostało zaimplementowane przez wykonujących projekt.

Wyniki eksperymentów są przedstawione na poniższym wykresie i w tabeli.



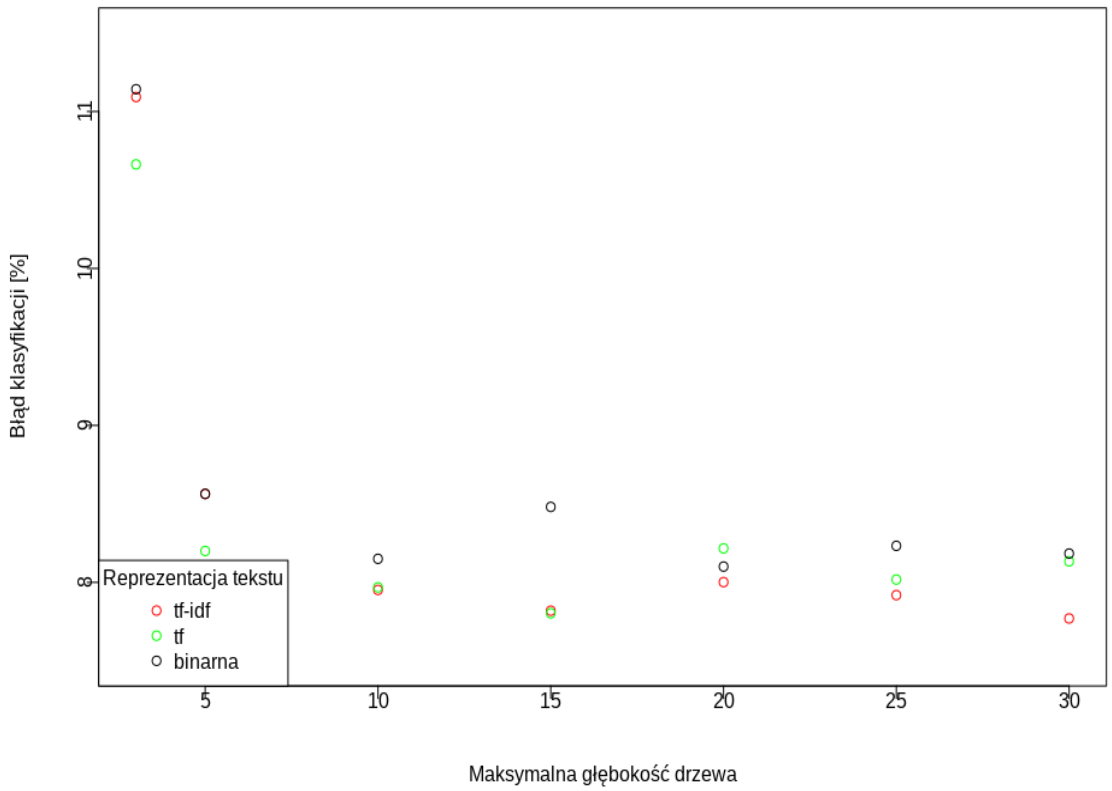
Reprezentacja tekstu	Koszt błędnej klasyfikacji spamu	Średni błąd klasyfikacji [%]	Średni TP	Średni FP	Średni FN	Średni TN	Średni koszt pomyłki
tf-idf	0.01	9.80	773.6	61,8	56,8	317,6	0.05
	0.1	10.02	773.4	64,2	57,0	315,2	0.06
	1	9.79	773,2	61,2	57,2	318,2	0.10
	10	9.77	772,2	60,0	58,2	319,4	0.53
	100	9.99	771,8	62,2	58,6	317,2	4.90
tf	0.01	13,94	796,6	134,8	33,8	244,6	0,11
	0.1	13,80	796,4	133,0	34,0	246,4	0,11
	1	13,82	795,8	132,6	34,6	246,8	0,14
	10	13,70	795,8	131,2	34,6	248,2	0,39
	100	13,66	795,0	129,8	35,4	249,6	3,03
binarna	0.01	11,16	774,2	78,8	56,2	300,6	0,07

	0.1	11,09	773,2	77,0	57,2	302,4	0,07
	1	11,08	773,2	76,8	57,2	302,6	0,11
	10	10,98	772,8	75,2	57,6	304,2	0,54
	100	11,18	771,6	76,4	58,8	303,0	4,92

Obserwacje i wnioski
W wynikach testów widać duże rozbieżności pomiędzy reprezentacją tf, a pozostałymi. Różnica widoczna jest przede wszystkim w liczbie wiadomości klasyfikowanych, jako pożądana korespondencja (TP, FP) – klasyfikator przewidujący w oparciu o reprezentację tf przydziela znacznie więcej wiadomości do tej kategorii, niż dwa pozostałe. Może to wynikać z większej podatności klasyfikatora na prawdopodobieństwa apriori klas – klasyfikator częściej przypisuje etykiety pożądanej korespondencji, bo ma ona większe prawdopodobieństwo apriori.
W przypadku pozostałych reprezentacji, prawdopodobieństwo apriori wydaje się mieć mniejszy wpływ na wynik predykcji.
Pomimo tego, że reprezentacja tf dała największe błędy klasyfikacji, w rzeczywistym systemie filtracji spamu, wykorzystującym klasyfikator Bayesa, to ona może sprawdzić się lepiej niż pozostałe - przy założeniu, że błędna klasyfikacja pożądanej wiadomości, skutkująca nieotrzymaniem jej przez adresata, jest o wiele większym błędem, niż błędna klasyfikacja spamu.

Drzewa decyzyjne
Podczas eksperymentów sprawdzono wpływ maksymalnej głębokości drzewa i minimalnego rozmiaru podziału na jakość klasyfikacji. Wyniki eksperymentów są przedstawione na poniższych wykresach i w tabeli.
W eksperymentach wykorzystano implementację drzew decyzyjnych z pakietu *rpart*.

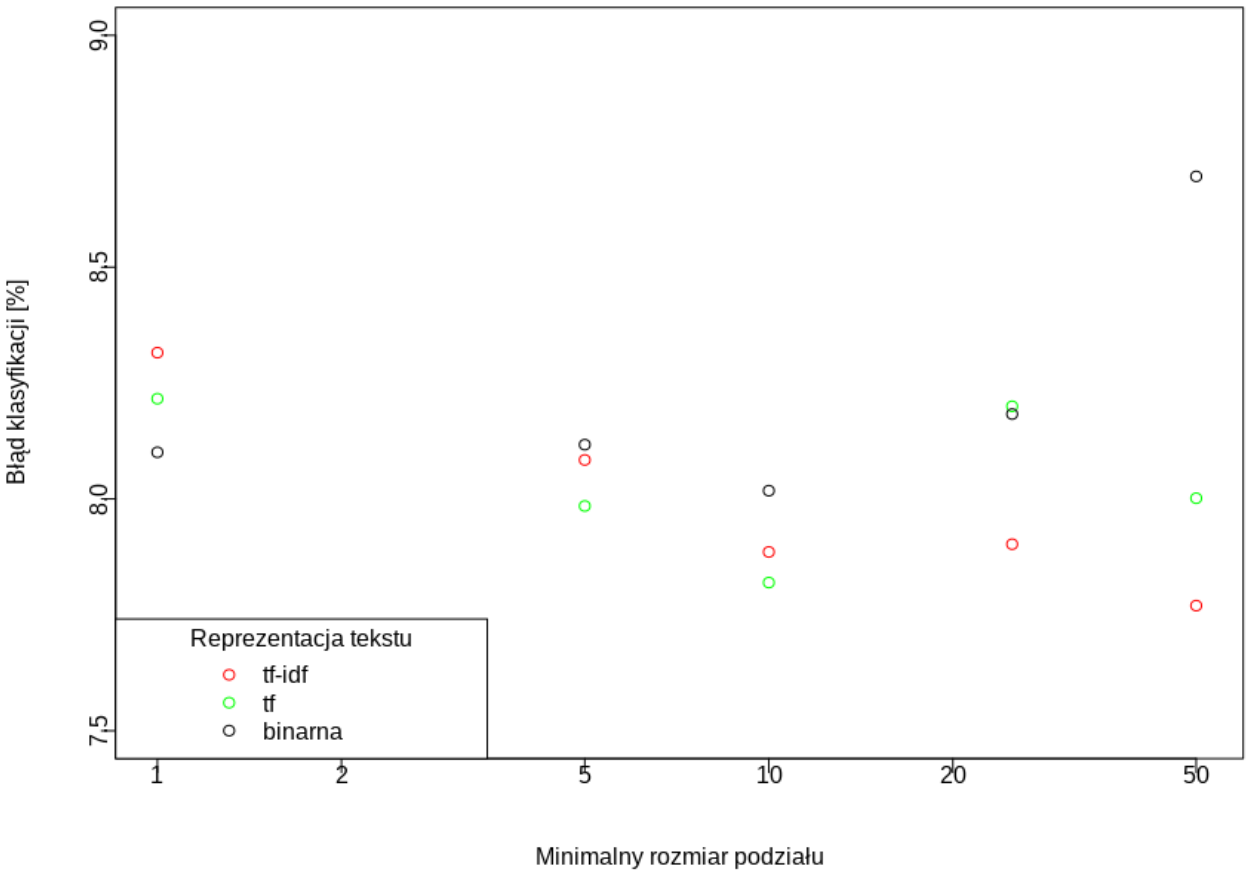
Badanie wpływu maksymalnej głębokości drzewa na błąd klasyifkacji przeprowadzono z minimalnym rozmiarem podziału ustalonym na 5.



Wykres nie uwzględnia wyników dla maksymalnej głębokości drzewa równej 1 – wartości błędów były znacznie większe, niż pozostałe i dodanie ich do wykresu pogorszyłoby jego czytelność, ze względu na zwiększenie granicy osi Y.

Reprezentacja tekstu	Maksymalna głębokość drzewa	Średni błąd klasyfikacji [%]	Średni TP	Średni FP	Średni FN	Średni TN
tf-idf	1	18,23	791,2	181,4	39,2	198,0
	3	11,09	791,2	95,0	39,2	284,4
	5	8,56	793,6	66,8	36,8	312,6
	10	7,95	795,2	61,0	35,2	318,4
	15	7,82	793,4	57,6	37,0	321,8
	20	8,00	794,0	60,4	36,4	319,0
	25	7,92	795,2	60,6	35,2	318,8
	30	7,77	798,0	61,6	32,4	317,8
tf	1	18,22	791,2	181,2	39,2	198,2
	3	10,66	787,8	86,4	42,6	293,0
	5	8,20	795,0	63,8	35,4	315,6
	10	7,97	790,8	56,8	39,6	322,6
	15	7,80	790,4	54,4	40,0	325,0
	20	8,22	789,2	58,2	41,2	321,2
	25	8,02	787,4	54,0	43,0	325,4
	30	8,13	785,2	53,2	45,2	326,2
binarna	1	18,22	791,2	181,2	39,2	198,2
	3	11,14	781,2	85,6	49,2	293,8
	5	8,56	786,8	60,0	43,6	319,4
	10	8,15	785,8	54,0	44,6	325,4
	15	8,48	784,0	56,2	46,4	323,2
	20	8,10	786,4	54,0	44,0	325,4
	25	8,23	784,8	54,0	45,6	325,4
	30	8,18	786,0	54,6	44,4	324,8

Badanie wpływu minimalnego podziału drzewa na błąd klasyifkacji przeprowadzono z maksymalną głębokością ustaloną na 10.



Reprezentacja tekstu	Minimalny rozmiar podziału	Średni błąd klasyfikacji [%]	Średni TP	Średni FP	Średni FN	Średni TN
tf-idf	1	8,32	793,2	63,4	37,2	316,0
	5	8,08	794,0	61,4	36,4	318,0
	10	7,89	795,6	60,6	34,8	318,8
	25	7,90	793,2	58,4	37,2	321,0
	50	7,77	797,8	61,4	32,6	318,0
tf	1	8,22	792,6	61,6	37,8	317,8
	5	7,98	787,6	53,8	42,8	325,6
	10	7,82	788,0	52,2	42,4	327,2
	25	8,20	788,6	57,4	41,8	322,0
	50	8,00	788,2	52,6	44,2	326,8
binarna	1	8,10	787,0	54,6	43,4	324,8
	5	8,12	788,0	55,8	42,4	323,6
	10	8,02	791,0	57,6	39,4	321,8

	25	8,18	788,6	57,2	41,8	322,2
	50	8,70	781,8	56,6	48,6	322,8

Obserwacje i wnioski
W teście wpływu maksymalnej głębokości drzew na błąd klasyfikacji zaobserwowano znaczne różnice w jakości klasyfikacji tylko pomiędzy małymi wartościami parametru – 1, 3, 5. Dla większych wartości, błędy klasyfikacji zmieniały się nieznacznie.
Nie zaobserwowano zwiększenia błędu klasyfikacji dla drzew o dużej maksymalnej głębokości – nawet dla drzew o maksymalnej głębokości 30, czyli maksymalnej wartości, jaką można zadać w pakiecie rpart. Oznacza to, że nie wystąpiło nadmierne dopasowanie do zbioru trenującego, nawet dla najbardziej złożonych drzew.
Przeprowadzono dodatkowy test, mający na celu sprawdzenie, czy zmniejszenie wartości minimalnego podziału do 1, przy maksymalnej głębokości drzewa 30 będzie skutkował nadmiernym dopasowaniem do danych trenujących i zwiększy błąd. Dla reprezentacji tf-idf błąd klasyfikacji wyniósł 8.03%, czyli więcej, niż dla większego rozmiaru podziału, ale wciąż stosunkowo mało.
Ciekawym przypadkiem są drzewa, dokonujące tylko jednego podziału, klasyfikujące wiadomości zawierające hiperłącza (atrybut *href HTML*) jako spam. Pomimo prostoty, mylą się jedynie w około 18.2% przypadków – to znacznie lepszy rezultat, niż przydzielanie wiadomości losowej etykiety, albo zawsze etykiety pożądanej korespondencji, która występuje w zbiorze danych w 69% przypadków.

Rozmiar minimalnego podziału okazał się mieć małe znaczenie dla jakości klasyfikacji. Niewielki wzrost błędu klasyfikacji zaobserwowano dla małych i dużych wartości tego parametru.

W odróżnieniu od klasyfikatora Bayesa, wpływ wyboru reprezentacji tekstu na błąd klasyfikacji i wartości w macierzy pomyłek był niewielki – zaobserwowane maksymalne różnice błędów klasyfikacji pomiędzy reprezentacjami były mniejsze niż 1 punkt procentowy. Najmniejszy błąd klasyfikacji, najczęściej dawało wykorzystanie reprezentacji tf-idf.

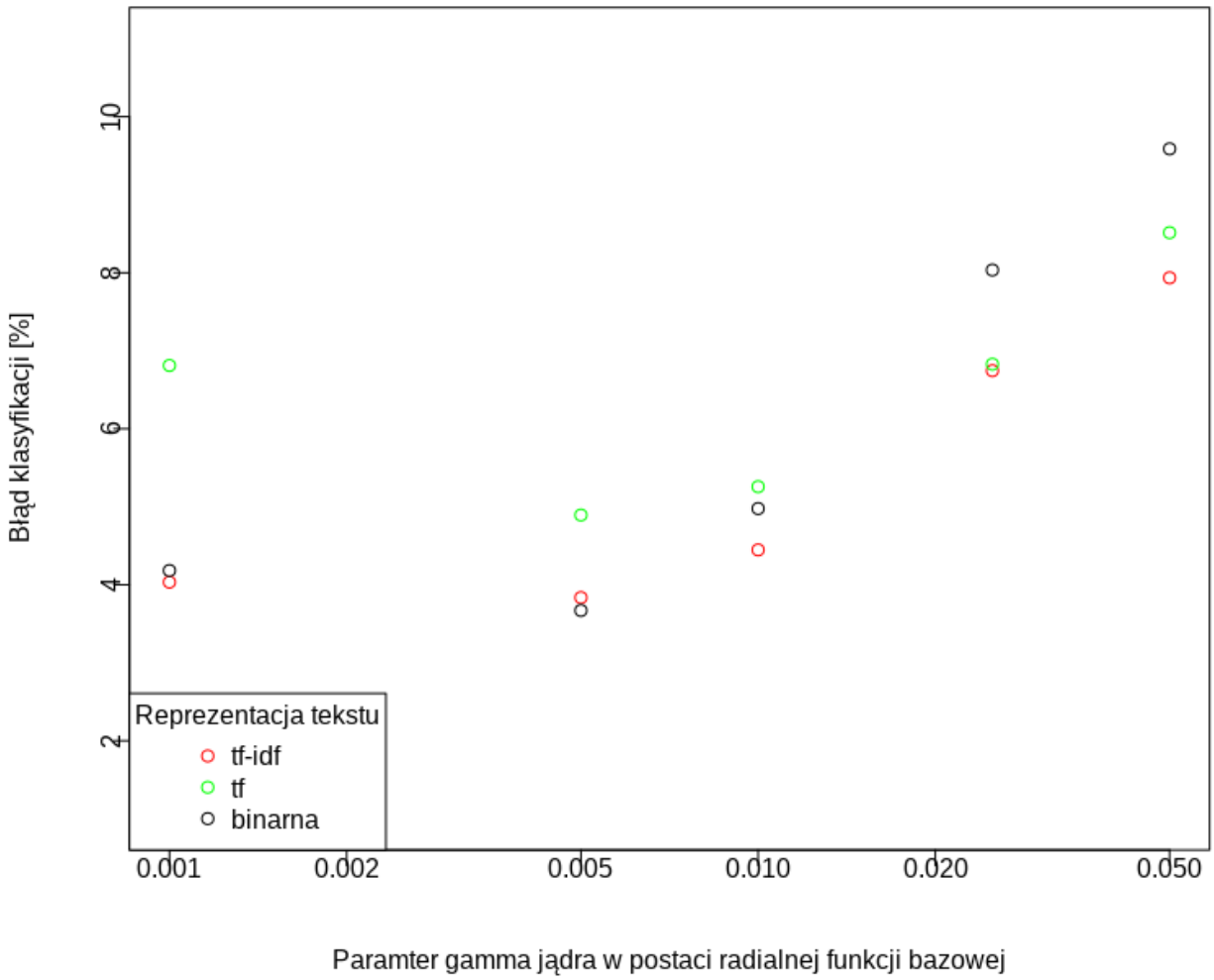
SVM
Podczas eksperymentów sprawdzono wpływ stosowania różnych funkcji jądrowych i ich parametrów oraz parametru kosztu naruszenia marginesu funkcji decyzyjnej C na błąd klasyfikacji.
Testy parametrów różnych jąder przeprowadzono przy domyślnie ustaionym parametrze kosztu C = 1.
W eksperymentach wykorzystano implementację SVM z pakietu *e1071*.

Wyniki eksperymentów są przedstawione na poniższych wykresach i w tabeli.

Jako pierwsze jądro, przetestowano radialną funkcję bazową (RBF):

$$K(x,y)=e^{-\gamma\|x-y\|^2}$$

Zbadano wpływ parametru γ na jakość klasyfikacji.



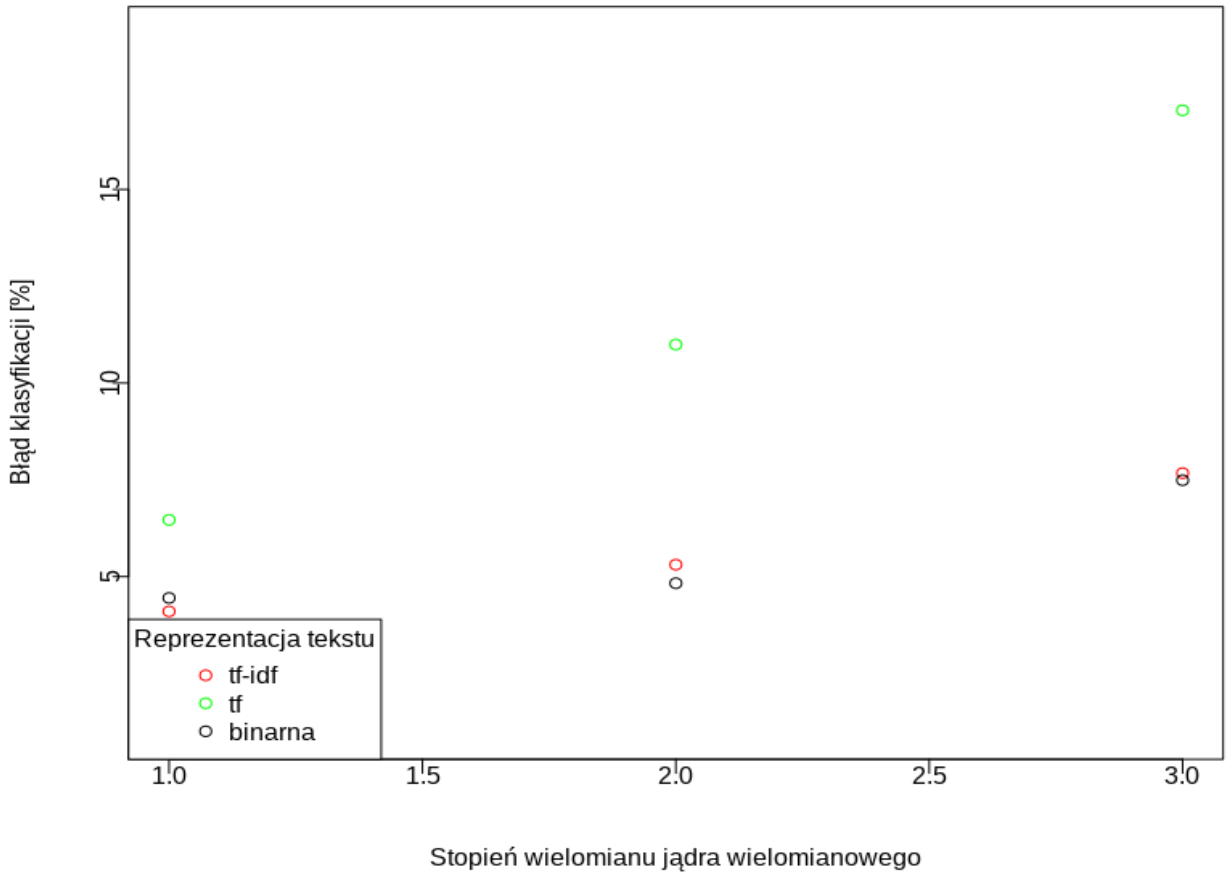
Reprezentacja tekstu	γ	Średni błąd klasyfikacji [%]	Średni TP	Średni FP	Średni FN	Średni TN
tf-idf	0.001	4,03	818,0	36,4	12,4	343,0
	0.005	3,84	807,4	23,4	23,0	356,0
	0.01	4,45	797,6	21,0	32,8	358,4
	0.025	6,74	766,2	17,4	64,2	362,0
	0.05	7,94	750,4	16,0	80,0	363,4
tf	0.001	6,81	811,2	63,2	19,2	316,2
	0.005	4,89	801,4	30,2	29,0	349,2
	0.01	5,26	794,0	27,2	36,4	352,2
	0.025	6,83	769,2	21,4	61,2	358,0
	0.05	8,51	747,0	19,6	83,4	359,8
binarna	0.001	4,18	811,8	32,0	18,6	347,4
	0.005	3,67	807,8	21,8	22,6	357,6
	0.01	4,98	790,2	20,0	40,2	359,4

	0.025	8,03	748,2	15,0	82,2	364,4
	0.05	9,59	726,4	12,0	104,0	367,4

Kolejnym przetestowanym jądrem był wielomian:

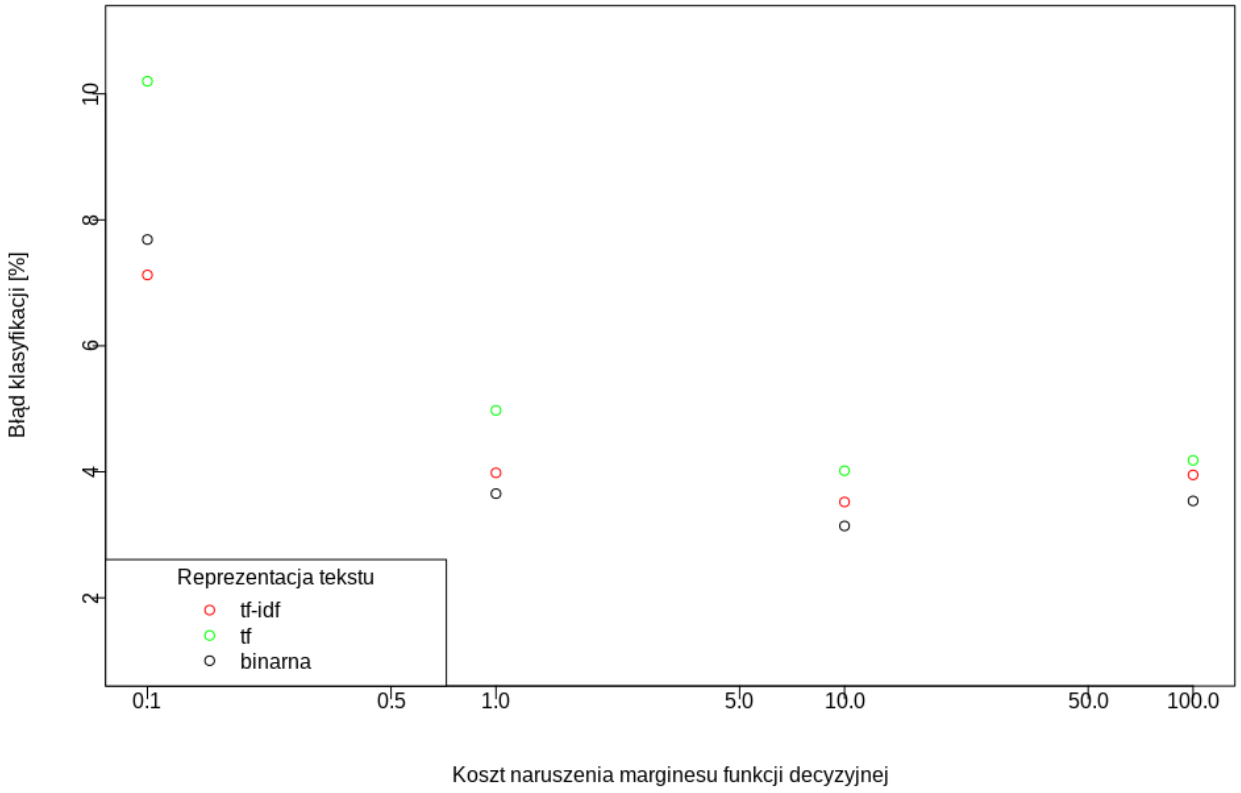
$$K(\boldsymbol{x},\boldsymbol{y})=(\boldsymbol{x}^T\boldsymbol{y})^{stopień\ wielomianu}$$

Zbadano wpływ stopnia wielomianu na jakość klasyfikacji.



Reprezentacja tekstu	Stopień wielomianu	Średni błąd klasyfikacji [%]	Średni TP	Średni FP	Średni FN	Średni TN
tf-idf	1	4,10	816,4	35,6	14,0	343,8
	2	5,31	821,8	55,6	8,6	323,8
	3	7,67	820,0	82,4	10,4	297,0
tf	1	6,46	811,2	59,0	19,2	320,4
	2	10,99	815,6	118,2	14,8	261,2
	3	17,04	816,2	192,0	14,2	187,4
binarna	1	4,45	812,2	35,6	18,2	343,8
	2	4,83	818,4	46,4	12,0	333,0
	3	7,49	820,2	80,4	10,2	299,0

Wpływ zmian parametru kosztu C na jakość klasyfikacji zbadano, wykorzystując jądro w postaci radialnej funkcji bazowej, z γ = 0.005, dla którego uzyskano najmniejszy błąd klasyfikacji w poprzednich testach.



Reprezentacja tekstu	C	Średni błąd klasyfikacji [%]	Średni TP	Średni FP	Średni FN	Średni TN
tf-idf	0.1	7,13	779,2	35,0	51,2	344,4
	1	3,98	807,4	25,2	23,0	354,2
	10	3,52	809,4	21,6	21,0	357,8
	100	3,95	803,8	21,2	26,6	358,2
tf	0.1	10,20	769,8	62,8	60,6	316,6
	1	4,98	802,4	32,2	28,0	347,2
	10	4,02	805,6	23,8	24,8	355,6
	100	4,18	803,8	24,0	26,6	355,4
binarna	0.1	7,69	768,8	31,4	61,6	348,0
	1	3,65	807,6	21,4	22,8	358,0
	10	3,14	809,2	16,8	21,2	362,6
	100	3,54	804,4	16,8	26,0	362,6

Obserwacje i wnioski

Zaobserwowano duży wpływ wartości parametru γ radialnej funkcji bazowej na błąd klasyfikacji – na przykład dla binarnej reprezentacji tekstów, zaobserwowano ponad dwukrotny wzrost błędu przy 10-krotnym zwiększeniu γ z 0,005 do 0,05.

Wartość radialnej funkcji bazowej maleje ze wzrostem wzajemnej odległości próbek w przestrzeni, co przekłada się na mały wpływ próbek odległych od klasyfikowanej na wynik klasyfikacji. Zmieniając wartość γ, można regulować maksymalną odległość w której znajdują się próbki, mające wpływ na wynik klasyfikacji. Wzrost γ sprawia, że tylko najbliższe próbki mają wpływ na wynik predykcji, a ich znaczenie również maleje ze wzrostem γ. To może wyjaśniać szybki wzrost błędu klasyfikacji, ze wzrostem γ.
Dla skrajnych wartości γ zaobserwowano stosunkowo dużą, sięgającą do 2.81 punktów procentowych różnicę błędów klasyfikacji pomiędzy różnymi reprezentacjami tekstu.

W przypadku jąder wielomianowych, dla wszystkich reprezentacji tekstu, zaobserwowano znaczne wzrosty błędów klasyfikacji, ze wzrostem stopnia wielomianu. Co zaskakujące, najlepsze rezultaty dał brak nieliniowej transformacji przestrzeni cech i zastosowanie najprostszego jądra liniowego.
Najlepsza, dla każdego z jąder wielomianowych okazała się reprezentacja binarna, a najgorsza tf. Różnice pomiędzy błędami klasyfikacji różnych reprezentacji tekstów sięgały nawet ponad 9 punktów procentowych.

Zaobserwowano również wpływ wartości kosztu C na jakość klasyfikacji – dla małych kar za przekroczenie marginesu funkcji decyzyjnej błąd klasyfikacji był znaczący i malał ze wzrostem kary.

W teście z jądrem RBF, z parametrem $\gamma = 0.005$. parametrem kosztu $C = 10$ i binarną reprezentacją tekstu, klasyfikator SVM osiągnął średni błąd klasyfikacji na poziomie jedynie 3,14%, co jest najlepszym wynikiem, ze wszystkich testowanych w projekcie klasyfikatorów.

Podsumowanie

Wszystkie przetestowane klasyfikatory, z zadanymi odpowiednimi parametrami, były w stanie odróżnić spam od pożądanej korespondencji w przynajmniej 90% przypadków. Predykcja, wykorzystująca jedynia znajomość prawdopodobieństw apriori klas, dałaby 69% skuteczność – sprawdzone metody sprawdziły się znacznie lepiej.
Najlepszy okazal się klasyfikator SVM z jądrem w postaci radialnej funkcji bazowej, poprawnie klasyfikujący 96,86% wiadomości.
Przeważnie najlepszą reprezentacją tekstu była tf-idf, chociaż czasami dawała gorszą jakość klasyfikacji, w tym dla najlepszego znalezionego klasyfikatora, niż reprezentacja binarna. Zwykle najgorsze wyniki klasyfikacji dawała reprezentacja tf. Znaleziono jedno, mało przekonujące, uzasadnienie jej użycia do klasyfikacji wiadomości – można jej użyć w przypadku klasyfikatora Bayesa, aby prewencyjnie zarówno poprawnie, jak i błędnie, klasyfikować więcej wiadomości jako pożądaną korespondencję.