Statystyczne metody klasyfikacji tekstów

WYDAWNICTWO
UNIWERSYTETU
ŁÓDZKIEGO

Adam Idczak
Jerzy Korzeniewski

Statystyczne metody klasyfikacji tekstów

WYDAWNICTWO
UNIWERSYTETU
ŁÓDZKIEGO
Łódź 2022

Adam Idczak – Uniwersytet Łódzki, Wydział Ekonomiczno-Socjologiczny
Instytut Statystyki i Demografii, Katedra Metod Statystycznych
90-214 Łódź, ul. Rewolucji 1905 r. nr 41/43

Jerzy Korzeniewski – Uniwersytet Łódzki, Wydział Ekonomiczno-Socjologiczny
Instytut Statystyki i Demografii, Katedra Demografii
90-214 Łódź, ul. Rewolucji 1905 r. nr 41/43

RECENZENCI
Paweł Lula, *Grażyna Trzpiot*

REDAKTOR INICJUJĄCY
Beata Koźniewska

REDAKTOR WYDAWNICTWA UŁ
Dorota Stępień

SKŁAD I ŁAMANIE
Munda – Maciej Torz

KOREKTA TECHNICZNA
Wojciech Grzegorczyk

PROJEKT OKŁADKI
Andrzej Pilichowski-Ragno

Zdjęcie wykorzystane na okładce autorstwa Andrzeja Pilichowskiego-Ragno

https://doi.org/10.18778/8220-786-6

Wydane przez Wydawnictwo Uniwersytetu Łódzkiego
Wydanie I. W.10497.21.0.K

Ark. wyd. 6,0; ark. druk. 8,875

ISBN 978-83-8220-786-6
e-ISBN 978-83-8220-787-3

Wydawnictwo Uniwersytetu Łódzkiego
90-237 Łódź, ul. Matejki 34A
www.wydawnictwo.uni.lodz.pl
e-mail: ksiegarnia@uni.lodz.pl
tel. 42 635 55 77

Spis treści

Wstęp

W ostatnich latach, wraz z szybkim rozwojem technologii komputerowych i internetowych, coraz większego znaczenia nabierają komputerowe metody badania tekstu (*text mining*). Badanie tekstu jest elementem interdyscyplinarnej dziedziny wiedzy, jaką jest przetwarzanie języka naturalnego (*natural language processing, NLP*), łączącej zagadnienia sztucznej inteligencji, informatyki i językoznawstwa, zajmującej się automatyzacją analizy, tłumaczenia i rozumienia języka naturalnego przez komputer. System maszynowy starający się zrozumieć język naturalny przekształca ten język na formalne symbole, łatwiejsze w użyciu dla programów komputerowych. Możliwości systemów komputerowych mogą być później wykorzystywane w takich zagadnieniach, jak: streszczanie tekstu, wyszukiwanie informacji z tekstu, sprawdzanie poprawności tekstu, maszynowe tłumaczenie tekstu, tworzenie słowników, określanie tematyki tekstu, określanie emocjonalnego charakteru tekstu itp. Termin „analiza sentymentu" (*sentiment analysis*) został użyty po raz pierwszy przez Nasukawę i Yi (2003) w sensie „ustalania subiektywnej polaryzacji (pozytywnej lub negatywnej) oraz siły polaryzacji (silnie pozytywna, średnio pozytywna, słabo pozytywna itp.) danego tekstu recenzującego, innymi słowy, ustalania opinii autora tekstu". W późniejszych latach rozszerzono trochę tę definicję i wyróżnione zostały różne rodzaje sentymentu (por. podrozdział 1.1), na przykład: sentyment dokumentu, sentyment zdania, sentyment zwrotu czy sentyment opisywanego podmiotu (*entity*). W niniejszej monografii skoncentrujemy się na analizie sentymentu w najpopularniejszym sensie tego terminu, to znaczy w odniesieniu do sentymentu całego dokumentu. Inne określenie pełniące funkcję synonimu analizy sentymentu w literaturze anglojęzycznej to badanie opinii (*opinion mining*). Obszar zastosowań analizy sentymentu jest dość szeroki – od badania sentymentu krótkich komentarzy-wypowiedzi zamieszczanych w komunikatorach internetowych lub mediach społecznościowych, lub odpowiedzi ankietowych na pojedyncze pytania, poprzez opinie klientów o produktach i usługach oraz recenzje efektów pracy twórczej, do podsumowywania tekstów (*text summarization*) w celu ich skróconej archiwizacji. Można też spotkać już dość popularne zastosowania badania opinii

wchodzące w obszar problemów naukowych z dziedziny ekonomii i finansów, jak na przykład klasyfikowanie opinii znalezionych w mediach społecznościowych w celu opracowania strategii inwestycyjnej dla rynków kapitałowych. W tej publikacji nacisk jest położony na badanie sentymentu krótkich dokumentów tekstowych. Analizę sentymentu można ująć w postaci problemu klasyfikacyjnego zbioru dokumentów (*text classification*, *text categorization*). Wówczas możemy wyróżnić różne rodzaje klasyfikacji w zależności od liczby klas, na jaką należy podzielić zbiór dokumentów. Gdy klasyfikujemy dokumenty do jednej z dwóch klas, to wówczas mówimy o problemie dwuklasowym lub binarnym (*binary classification*), zaś klasy dokumentów określamy mianami klasy pozytywnej i negatywnej. Gdy klas jest więcej, mamy do czynienia z klasyfikacją wieloklasową (*multiclass classification*), zazwyczaj są to trzy klasy: pozytywna, neutralna i negatywna. W niniejszej monografii nacisk zostanie położony na problemy klasyfikacji binarnej. Istnieje wiele różnych podejść do problemu badania opinii, niektóre z nich wykorzystują źródła zewnętrzne, tj. słowniki, tezaurusy i inne opracowania leksykalne, w innych są stosowane tylko metody maszynowego uczenia się, jeszcze inne posługują się głównie ustalonymi regułami (por. podrozdział 1.1). W przypadku języka polskiego nie są jeszcze wystarczająco rozwinięte jego słownikowe zasoby, tezaurusy czy słowosieci elektroniczne, które mogłyby zostać wykorzystane w badaniu opinii, dlatego metody odwołujące się do źródeł leksykalnych zostały w tej publikacji pominięte. Należy również wspomnieć o dwóch zasadniczych rodzajach badania opinii w zależności od dostępności informacji o oczekiwanej wartości wyjściowej klasyfikacji dokumentów. Jeżeli dysponujemy taką informacją, to mówimy wówczas o klasyfikacji z nadzorem (*supervised classification*), jeżeli takiej informacji nie mamy, to będzie to klasyfikacja bez nadzoru (*unsupervised classification*). W prezentowanej monografii został położony nacisk na metody z nadzorem, ale ze względu na aspekty praktyczne, z jak najmniejszym rozmiarem zbioru uczącego. Naszym dalszym celem badawczym jest opracowywanie metod klasyfikacji dokumentów bez nadzoru, ale pozytywnym początkiem może być praca nad metodami ze zbiorem uczącym o małej liczebności. Inną istotną cechą metod analizy sentymentu jest to, czy przy selekcji terminów, na podstawie których będzie dokonywana klasyfikacja dokumentów, jest konieczne korzystanie z ustalonej metody klasyfikacji. Jeśli tak, to taką metodę zaliczamy do grupy metod zwanych *wrapper*. Jeśli metoda selekcji terminów jest niezależna od metody klasyfikacji dokumentów i jej zasadą jest filtrowanie terminów, to zaliczamy ją do grupy metod zwanych *filtering methods*. W publikacji został również położony nacisk na badanie pierwszego etapu metod, tj. selekcji terminów. Uważamy, że ten etap klasyfikacji dokumentów ma zasadnicze znaczenie dla jakości całej klasyfikacji. Spośród wszystkich standardowych klasyfikatorów najlepsze osiągają bardzo podobne wyniki, które są silnie uzależnione od tego, jakie zbiory terminów są wykorzystywane oraz jak są one wykorzystywane.

Cele niniejszej monografii są następujące:

- dokonanie przeglądu porównawczego wybranych metod klasyfikacji tekstów anglojęzycznych ze względu na ich sentyment, ze szczególnym uwzględnieniem selekcji terminów;
- zbadanie jakości wybranych metod klasyfikacji tekstów ze względu na ich sentyment w zastosowaniu do dokumentów w języku polskim;
- zaproponowanie nowych metod, które poprawiałyby jakość klasyfikacji lub posiadały inne atuty badawcze.

Rozdział 1

Wprowadzenie w problematykę klasyfikacji tekstów

1.1. Podstawowe pojęcia

Jeśli chcemy wykorzystać komputery i narzędzia statystyczne do klasyfikacji dokumentów tekstowych, to konieczne jest przyjęcie jakiejś formy opisu matematycznego tekstu. Najogólniejszą klasyfikację modeli, które mogą być wykorzystane w tym celu, można znaleźć w pracy Luli (2018). Jak zauważają w innej publikacji Lula i Wójcik (2011), w analizie sentymentu dokumentów tekstowych można wykorzystywać cztery różne podejścia bazujące na:

- analizie terminów (*word-based approach*);
- analizie wzorców (*pattern-based approach*);
- badaniu ontologii (*ontology-based approach*);
- komputerowych metodach uczenia się.

Przedmiotem niniejszej monografii będą metody z pierwszej i ostatniej grupy. Słowa, szerzej, terminy to cechy, czyli zmienne opisujące jednostkę statystyczną, którą jest dokument tekstowy. Komputerowe metody uczenia się mają kilka zasadniczych zalet (por. Lula i Wójcik, 2011), m.in. wysoką efektywność oraz możliwość wielokrotnego użycia niezależnie od kontekstu, środowiska itp. Tej ostatniej możliwości nie mają podejścia oparte na ontologiach, jak również, do pewnego stopnia, te bazujące na analizie wzorców. Inna klasyfikacja metod analizy sentymentu jest pokazana na rysunku 1.

Rahate i Emmanuel (2013) wyróżniają cztery rodzaje cech, którymi można opisać dokumenty: składniowe (*syntactic*), semantyczne (*semantic*), oparte na odnośnikach (*link based*) oraz stylistyczne (*stylistic*). Cechy składniowe to słowa, zwroty, części mowy. Cechy semantyczne odwołują się do badania znaczenia opartego na zależnościach pomiędzy słowami, zwrotami, znakami i symbolami. Zazwyczaj w tym celu używana jest metoda punktowa, odzwierciedlająca pozytywną bądź negatywną wymowę zwrotu. Cechy oparte na odnośnikach śledzą bazy odnośników podawane w dokumentach lub wypowiedziach w celu ustalenia sentymentu wypowiedzi. Takie podejście jest oparte na założeniu, że silnie połączone

linkami strony internetowe często prezentują takie same opinie na temat poruszanych kwestii. Cechy stylistyczne śledzą styl wypowiedzi, rozkład długości używanych słów, bogactwo słów, cechy leksykalne słów. Za cechy najistotniejsze, które będą podstawą do charakterystyki dokumentu w naszej monografii, wybraliśmy cechy składniowe.

Klasyfikacja dokumentów oraz selekcja terminów są od siebie nawzajem zależne, wobec czego istotne jest badanie tego, jak jeden obszar wpływa na drugi. W tym kontekście należy wymienić dwie kwestie. Po pierwsze, czy zależności, charakterystyki cech ustalone w trakcie klasyfikacji wykonanej przy użyciu jakiegoś klasyfikatora mogą być stosowane ogólniej, tzn. dla innych klasyfikatorów? Po drugie, czy różne metody selekcji cech spisują się tak samo dla różnych typów klasyfikatorów? Do tych kwestii odnosi się pośrednio lub bezpośrednio każdy artykuł naukowy dotyczący tematu badania opinii, zaś pewne uogólnienia starali się sformułować Mladenic i inni (2004).

W odniesieniu od tego, czy selekcja cech zależy od klasyfikatora czy też nie, metody selekcji można podzielić na dwie grupy: filtrujące (*filtering*) oraz zbudowane na klasyfikatorze (*wrapper*). W metodach typu *wrapper* ustalony jest klasyfikator i wybierany jest zbiór cech, które dają najlepszą jakość klasyfikacji tylko dla tego klasyfikatora. W metodach filtrujących badana jest tylko charakterystyka cech lub ich podzbiorów w oparciu o ustalone kryteria. Pokazano, że metody typu *wrapper* (por. Guyon i Elisseeff 2003; Yu i Liu 2004) mogą być lepszej jakości, ale są bardzo czasochłonne, więc nieprzydatne dla zbiorów danych wielowymiarowych. W kontekście klasyfikacji tekstów należy zauważyć, że na ogół mamy do czynienia z bardzo dużą liczbą cech/terminów, dlatego względy natury praktycznej przemawiają za metodami filtrującymi. Ponadto metody typu *wrapper* często tracą efektywność z powodu często obecnej w klasyfikacji wady nadmiernego dopasowania modelu.

Od samego początku klasyfikacji tekstów za pomocą komputerowych metod uczenia się, czyli od lat 90. ubiegłego wieku, dominowały dwie formy reprezentacji dokumentów, tj. binarna – informująca tylko o tym, czy dany termin występuje w danym dokumencie czy też nie występuje (np. Fragoudis i inni 2005), oraz częstościowa – podająca liczbę wystąpień danego terminu w dokumencie (McCallum i Nigam 1998; Yang i Liu 1999). Trudno jednoznacznie rozstrzygnąć, która z tych form jest efektywniejsza. Wraz z każdą nową metodą, której autor twierdzi, że uzyskał trochę lepsze wyniki, taki osąd należałoby zmieniać. Ponadto należy zaznaczyć, że istnieją bardziej zaawansowane formy opisu częstości występowania terminów (por. podrozdział 1.2), dla których autorzy je stosujący uzyskują również dobrą efektywność.

Jak zauważa Genkin i inni (2007), metody heurystyczne są wydajne pod względem czasowym i pamięciowym, ale wraz z nimi pojawiają się nowe problemy. Podstawy statystyczne większości metod heurystycznych są niejasne. Konsekwencją jest na przykład to, że niemożliwy jest wybór określonej liczby cech dla danego

problemu wedle jakiejś matematycznej reguły. Dodatkowo większość efektywnych metod selekcji cech rozpatruje je w izolacji od wielu charakterystyk dokumentów, co może powodować wybór cech redundantnych lub nieefektywnych.

W badaniu tekstu występują pojęcia: ekstrakcji cech (*feature extraction*) oraz selekcji cech (*feature selection*). Ekstrakcja cech polega na wydobywaniu z tekstu obiektów spójnych logicznie pod względem leksykalnym i przedstawieniu ich w formie modelu matematycznego. Selekcja to natomiast wybieranie podzbiorów cech spośród zbioru wszystkich cech wyjściowych.

Selekcja zmiennych nie jest zadaniem łatwym. Jak zauważa Joachims (1998), na ogół prawie wszystkie cechy są ważne. Jako przykład podaje najbardziej znany zbiór benchmarkowy *Reuters*, kategorię „acq" oraz uporządkowanie terminów metodą IG (*information gain*), a następnie użycie naiwnego estymatora Bayesa według następujących grup uporządkowanych terminów: 1–200, 201–500, 501–1000, 1001–2000, 2001–4000, 4001–9962. Okazuje się, że nawet terminy końcowe zawierają ważne w klasyfikacji dokumentów informacje. Jakość klasyfikacji na końcowych terminach jest dużo wyższa od tej zastosowanej do wybranych losowo. Wniosek z tego jest taki, że selekcja zmiennych nie może być zbyt agresywna.

Termin sentyment, będący podstawą klasyfikacji, można rozumieć na różne sposoby, może on odnosić się do sentymentu dokumentu, sentymentu zdania lub zwrotu oraz do sentymentu materii (*entity*). W kontekście sentymentu dokumentu zakłada się, że każdy pojedynczy dokument wyraża jakąś opinię o konkretnym podmiocie (który jest znany). Zadaniem statystyka jest ustalenie tego, czy dokument wyraża opinię pozytywną czy negatywną. Sentyment dokumentu można podzielić na mniejsze obszary odnoszące się do zdania, zwrotu, a nawet słowa. Można następnie klasyfikować te mniejsze jednostki i wyniki klasyfikacji agregować tak, by otrzymać odpowiedź na pytanie o opinię na poziomie dokumentu. Wtedy problem klasyfikacji opinii na poziomie zdania staje się problemem samym w sobie i napotyka również trudności. Przykładem mogą być wypowiedzi, które stwierdzają oczywiste fakty i wydawałoby się, że nie są obarczone żadną opinią, ale tak nie jest, bo samo wspomnienie o tych faktach może być odczytane jako opinia nawet skrajnie pozytywna lub negatywna. Wtedy problemem priorytetowym staje się odróżnienie wypowiedzi obiektywnych od subiektywnych. Z punktu widzenia analizy sentymentu całego dokumentu o wiele łatwiejsze do oceny stopnia spolaryzowania opinii są zdania subiektywne, na ogół nacechowane emocjonalnym słownictwem. Metody poświęcone ustalaniu obiektywności lub subiektywności zdania często korzystają ze słowników. Sentyment materii, niekiedy zwany też sentymentem aspektu, odnosi się do tego, że opinia zawarta w wypowiedzi powinna odnosić się do ściśle zdefiniowanej materii, innymi słowy – opinia powinna mieć ściśle określony aspekt. Na przykład w recenzji produktów, zwłaszcza we wstępnym etapie, można znaleźć zdania skrajnie opinionośne, wprowadzające w temat recenzji informację o tym, że

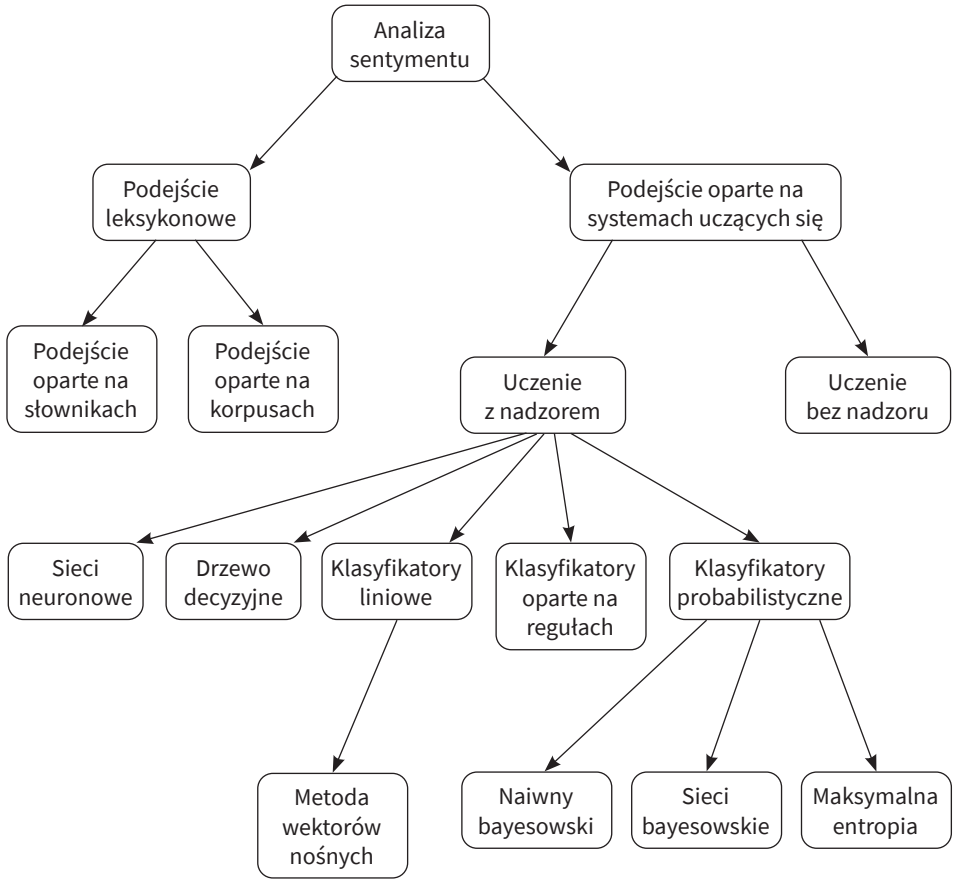
dany produkt jest nam absolutnie niezbędny do życia, natomiast ta opinia nie ma absolutnie nic wspólnego z oceną produktu. Analizowane wypowiedzi powinny charakteryzować się ścisłym przestrzeganiem sentymentu materii. Lula (2018) wymienia następujące czynniki, które utrudniają poprawną klasyfikację tekstów ze względu na ich sentyment:

- niejasne przedstawienie sensu wypowiedzi;
- sarkastyczny lub ironiczny ton wypowiedzi;
- błędy ortograficzne i stylistyczne;
- konieczność przeprowadzenia analizy nawiązań do obiektów badanych;
- badanie znaczenia powtórzeń zwrotów lub słów;
- konieczność właściwej interpretacji wyrażeń negujących;
- problem identyfikacji nazw własnych;
- problem rozpoznawania znaczenia porównań;
- problem wieloznaczności słów i dłuższych wypowiedzi.

Jak zauważają Saad i Saberi (2017), wśród wielu rodzajów metod służących badaniu opinii dominują dwie grupy (por. rysunek 1): metody maszynowego uczenia się oraz metody oparte na leksykonach. Każda z grup metod ukazana na rysunku 1 ma swoje wady i zalety. Diagram ten nie jest kompletny, co kilka miesięcy należałoby go aktualizować, gdyż pojawiają się nowe grupy metod. Należy pamiętać o tym, że wszelkie badania efektywności metod analizy sentymentu są obarczone różnymi czynnikami utrudniającymi porównywanie metod i ich ocenianie. Różne są metody wstępnej obróbki tekstów (i różne oprogramowanie), różne są proporcje liczebności zbioru testowego w stosunku do zbioru uczącego, jak również inne są formy statystycznej prezentacji tekstu (por. podrozdział 1.3). Jak zauważają Yazdani i inni (2017), dwie podstawowe wady metody *bag of words*, którą będziemy posługiwać się w naszych badaniach, to konieczność zmierzenia się z bardzo dużą liczbą cech/terminów, co prowadzi prostą drogą do szeroko znanego w statystyce zjawiska *curse of dimensionality* oraz braku możliwości wychwytywania zależności semantycznych pomiędzy terminami.

Metoda *bag of words* nie notuje kolejności występowania słów w dokumencie, a jedynie ich liczbę. Pomimo tych dwóch wad metoda jest bardzo popularna wśród badaczy, ponieważ ma zalety, których nie posiadają inne podejścia. Nie korzystamy z różnych słowników i leksykonów, co powodowałoby dalsze problemy związane, dla odmiany, z doborem tych źródeł, koniecznością ich ciągłego aktualizowania itp. Ustalenie liczebności zbioru uczącego też nie jest oczywiste przy badaniu opinii dokumentów. Standardowo w większości zadań związanych z klasyfikacją nadzorowaną przyjmuje się, że zbiór uczący powinien stanowić około 2/3 rozmiaru całego zbioru danych, pozostała 1/3 przeznaczana jest na zbiór testowy. Wśród badaczy efektywności metod analizy sentymentu nie ma zgody co do standardowego rozmiaru zbioru uczącego – waha się on od 50% do nawet 90%. W tej monografii nacisk został położony na badanie efektywności metod przy jak najmniejszym rozmiarze zbioru, zaczynamy nawet od 3% całego zbio-

ru danych. Takie podejście uzasadniamy względami praktycznymi. We współczesnym świecie, nacechowanym wysoką dynamiką zjawisk, bardzo ważna jest szybkość reakcji oraz, oczywiście, jej koszt. Klasycznym przykładem zastosowań badania sentymentu jest badanie dokumentów zbieranych on-line, często z wymaganą szybką reakcją również niemalże on-line. Jak w takiej sytuacji zatrudniać badaczy do określenia opinii dokumentów? Skąd brać tych badaczy? To generuje dodatkowe koszty oraz problemy natury technicznej. Co zrobić, jeśli za jakiś czas będziemy mogli zebrać następną porcję dokumentów, o której to możliwości mogliśmy nawet wcześniej nie wiedzieć? Na wszystkie tego typu pytania jest, naszym zdaniem, tylko jedna odpowiedź – trzeba opracowywać metody, które będą spisywać się dobrze nawet przy kilkuprocentowym rozmiarze zbioru uczącego, a najlepiej bez zbioru uczącego.



Rysunek 1. Wybrane metody klasyfikacji tekstów ze względu na ich sentyment.

Źródło: opracowanie własne.

Metoda wektorów nośnych (*Support Vector Machine* – SVM), wspomniana na rysunku 1, będzie wykorzystywana w badaniach z uwagi na dobrą opinię w literaturze przedmiotu. Jak podkreślają Wu i inni (2015), klasyfikator SVM dobrze radzi sobie z problemem wysokiej wymiarowości, nieliniowością danych, małymi rozmiarami prób oraz charakteryzuje się wysoką precyzją i raczej nie powoduje efektu nadmiernego dopasowania. Sieci neuronowe wymienione na rysunku 1 doczekały się wielu różnych wariacji (*recurrent neural networks*, *long short term memory networks*, *convolutional networks*), ale nie pozbyto się wrodzonych wad tego podejścia, takich jak konieczność uciążliwego dostrajania parametrów, dużej liczby parametrów, dużej niestabilności stosowanych miar regulacji działania sieci itp.

1.2. Uwagi terminologiczne i oznaczenia

W literaturze przedmiotu nie ma ujednoliconego systemu oznaczeń. Nawet podstawowe pojęcie, którym wszyscy się posługują, jest nazywane terminem (*term*), cechą (*feature*), słowem (*word*), zwrotem (*phrase*) lub jeszcze inaczej. Biorąc pod uwagę to, że nie wszystkie wczytane terminy, najogólniej należałoby powiedzieć: zestawy liter, są słowami, bo mogą być skrótami, dziwnymi nazwami własnymi, wyrazami zdumienia, zachwytu itp. trudnymi do zakwalifikowania pod hasłem słowo, zdecydowaliśmy się na używanie określenia „termin". Czasem zamiennie będziemy stosować określenie „cecha", bo terminy występujące w dokumentach są traktowane jako zmienne opisujące dokumenty. W odniesieniu do numerowania dokumentów będziemy używać litery n, natomiast w odniesieniu do numerowania terminów będziemy stosować literę m. Podstawowe oznaczenia są zatem następujące:

- liczba wszystkich dokumentów: N;
- liczba wszystkich cech lub, zazwyczaj, terminów: M;
- liczba dokumentów, w których występuje termin f: n_f;
- liczba dokumentów, w których nie występuje termin f: $n_{\bar{f}}$;
- liczba dokumentów, w których występuje termin f i które należą do klasy C_k: $n_{f,k}$;
- liczba dokumentów, w których nie występuje termin f i które należą do klasy C_k: $n_{k,\bar{f}}$;
- liczba dokumentów, w których nie występuje termin f i które nie należą do klasy C_k: $n_{\bar{k},\bar{f}}$;
- liczba dokumentów, w których występuje termin f i które nie należą do klasy C_k: $n_{\bar{k},f}$;
- liczba dokumentów, które należą do klasy C_k: n_{C_k}.

Ma miejsce zależność:

$$n_f = \sum_k n_{f,k}. \tag{1.1}$$

Prawdopodobieństwa przypisane cechom są rozumiane w odniesieniu do ich występowania w dokumentach. I tak, $P(f)$ to częstość dokumentów zawierających cechę f, czyli: $P(f) = \frac{n_f}{N}$.

Częstość dokumentów niezawierających cechy f to: $P(\overline{f}) = \frac{n_{\overline{f}}}{N}$.

Częstość klasy C_k też jest rozumiana w odniesieniu do liczby dokumentów:

$$P(C_k) = \frac{n_{C_k}}{N}. \tag{1.2}$$

Podobnie, to znaczy w odniesieniu do liczb dokumentów, rozumiane są prawdopodobieństwa warunkowe, na przykład prawdopodobieństwo terminu pod warunkiem założenia o tym, że rozważane są tylko dokumenty z ustalonej klasy. Te prawdopodobieństwa określamy następująco:

$$P(f|C_k) = \frac{n_{f,k}}{n_{C_k}} \qquad P(C_k|f) = \frac{n_{f,k}}{n_f} \qquad P(C_k|\overline{f}) = \frac{n_{k,\overline{f}}}{n_{\overline{f}}}. \tag{1.3}$$

Liczba wystąpień terminu f we wszystkich dokumentach: m_f.
Liczba wystąpień terminu f w klasie C_k: $m_{f,k}$.
Ma miejsce zależność:

$$m_f = \sum_k m_{f,k}. \tag{1.4}$$

Liczba terminów występujących w dokumencie d: m_d.

W związku z istnieniem zbiorów uczącego i testowego będziemy posługiwać się następującymi oznaczeniami:

- liczba dokumentów w zbiorze uczącym: n_{train};
- liczba dokumentów pozytywnych w zbiorze uczącym: $n_{1,\,train}$;
- liczba dokumentów negatywnych w zbiorze uczącym: $n_{0,\,train}$;
- liczba dokumentów w zbiorze uczącym zawierających termin f: $n_{f,\,train}$;
- liczba dokumentów pozytywnych w zbiorze uczącym zawierających termin f: $n_{f,\,1,\,train}$;
- liczba dokumentów negatywnych w zbiorze uczącym zawierających termin f: $n_{f,\,0,\,train}$.

W związku z propozycją nowej metody prezentowaną w ostatnim rozdziale, w której występują dwie listy słów: pozytywnych (SP) i negatywnych (SN), będziemy posługiwać się następującymi oznaczeniami:

- liczba terminów z listy SP mających niezerowe skorelowanie z terminem *f* występującym w zbiorze *d*: $m_{SP,f,d}$;
- liczba terminów z listy SN mających niezerowe skorelowanie z terminem *f* występującym w zbiorze *d*: $m_{SN,f,d}$.

1.3. Etapy wstępnej obróbki tekstu

Zanim dokumenty tekstowe będą mogły zostać poddane klasyfikacji ze względu na ich wydźwięk, należy poddać je wstępnej obróbce (*text preprocessing*). Tekst musi zostać przekształcony z postaci nieustrukturyzowanej do postaci ustrukturyzowanej, umożliwiającej dalsze wykorzystanie go przez klasyfikator. Korpus (*corpus*), czyli kolekcja dokumentów, zostaje poddany wstępnym przekształceniom mającym na celu usunięcie zbędnych lub nieistotnych, z punktu widzenia zadania text miningowego, elementów tekstu. Najczęściej stosowane modyfikacje tekstu opisano poniżej.

We wstępnej obróbce tekstu dąży się do tego, aby wyeliminować te elementy, które na ogół nie niosą za sobą informacji pomocnych w określeniu wydźwięku dokumentu oraz aby taki tekst doprowadzić do postaci ustrukturyzowanej. Eliminacja zbędnych wyrazów realizowana jest poprzez zastosowanie tzw. *stoplisty* (*stop word list*). Zawiera ona wyrazy, które zazwyczaj nie mają silnego związku z wyrażaniem opinii. Typowa stoplista składa się z przyimków, zaimków oraz spójników, a jej elementy mogą być specyficzne dla danego języka.

Kolejnym typowym zabiegiem zaliczającym się do wstępnej obróbki tekstu jest usunięcie liczb. Takie podejście często bywa słuszne, np.

> Młody człowiek poniósł tragiczną śmierć w wieku 27 lat.

W tym zdaniu pominięcie liczby 27 nie zmieni negatywnego nacechowania tekstu. Nie zawsze jednak przyniesie pożądany rezultat, szczególnie gdy autor wypowiedzi posiłkuje się oceną liczbową:

> Pobyt w tym hotelu oceniam na 5.

Taka sytuacja, choć jest możliwa, występuje rzadko. Bardzo często poza oceną liczbową w ślad za zadowoleniem/niezadowoleniem autora podążają określenia, które w znaczny sposób mogą posłużyć do zidentyfikowania sentymentu wypowiedzi, np.:

Pobyt w tym wspaniałym hotelu oceniam na 5.

Standardową czynnością poprzedzającą kolejne kroki analizy jest usunięcie znaków interpunkcyjnych. W wyniku usuwania słów znajdujących się na stopliście, liczb lub znaków interpunkcyjnych, w zależności od użytego oprogramowania, usuwane elementy mogą zostać zamienione na puste pola (*double spacing*). Nadmiarowe puste pola należy również usunąć z dokumentów tekstowych. Ponadto wszystkie wielkie litery zamieniane są na małe.

Stemmatyzacja (*stemming*) polega na sprowadzaniu do tej samej morfologicznej postaci wyrazów, które mają podobne znaczenie oraz wspólny rdzeń (*root*), czyli pochodzą od tego samego wyrazu podstawowego. Poszczególne wyrazy w języku polskim mogą być zapisane na wiele sposobów, np. rzeczowniki odmienia się przez liczby (pojedyncza lub mnoga), przypadki; czasowniki odmienia się przez rodzaje, osoby, liczby, czasy, tryby, strony; przymiotniki odmienia się przez przypadki, liczby i rodzaje etc. W celu uproszczenia struktury danych tekstowych, intuicyjnie wydaje się zasadne zastąpienie jednym słowem wielu form tego samego słowa, nie tracąc przy tym na semantycznej wartości samego tekstu. Stemmatyzację cechuje wysoki poziom automatyzacji, wymaga niewielkiej liczby reguł, które algorytmicznie wyszukują rdzeń w poszczególnych słowach. Dokonując stemmatyzacji na tekście, przykładowo zamienimy wyrazy *samolotu*, *samolotowi*, *samolotem*, *samolocie* na *samolot*. Proces ten nie jest pozbawiony wad, co ilustruje poniższy przykład:

Kasia wypiła wczoraj dwa kieliszki prosecco.

Tomek wyciął piłą cały las.

Wyrazy *wypiła* oraz *piłą* mogą zostać uznane za ten sam wyraz i zamienione na jeden wyraz podstawowy, np. *pić* albo *piła*.

Tej wady pozbawiona jest *lemmatyzacja* (*lemmatization*) – precyzyjniejsze podejście sprowadzania słów do ich słownikowej formy podstawowej, lemmy (*lemma*), uwzględniające części mowy (*part of speech* – POS). W podanym przykładzie w pierwszym zdaniu wyraz *wypiła* rozpoznany zostanie jako czasownik i sprowadzony do wyrazu *wypić*, natomiast w drugim zdaniu wyraz *piłą* rozpoznany zostanie jako rzeczownik i zamieniony na wyraz *piła*. Proces lemmatyzacji wymaga utworzenia pokaźnych słowników wyrazów oraz lemm, a także bogatej wiedzy domenowej, specyficznej dla danego języka.

Na ten moment tekst zawarty w dokumentach tekstowych oczyszczono z niepotrzebnych elementów oraz uproszczono nieco jego zawartość. Dane nadal pozostają wysoce nieustrukturyzowane. W kolejnych akapitach czytelnik pozna koncepty umiejscawiające nieustrukturyzowane dane w bardziej ustrukturyzowanych ramach.

Bardzo popularnym sposobem na przekształcenie tekstu w ustrukturyzowaną formę jest przedstawienie dokumentów tekstowych jako zbioru poszczególnych

wyrazów, zwanego *bag of words* (BOW). Tekst dzielony jest na osobne wyrazy (*unigramy*), np. BOW zdania:

> Witek spędził wakacje na Majorce.

jest taki:

> Witek, spędził, wakacje, na, Majorce.

Bag of words jest szczególnym przypadkiem szerszego konceptu – *bag of n-grams*, w którym treść dokumentu reprezentowana jest jako zbiór *n*-wyrazowych podciągów tekstu, np. dwuwyrazowe podciągi noszą nazwę *bigramów*, trójwyrazowe – *trigramów* etc. Bazując na wcześniejszym przykładzie, utworzono następującą reprezentację bigramową:

> Witek spędził, spędził wakacje, wakacje na, na Majorce.

Niektórzy badacze sądzą, że BOW prowadzi donikąd, podając następujący przykład:

> Piękny hotel w okropnym mieście.
> Okropny hotel w pięknym mieście.

Oba zdania mają identyczne BOW, a treści zupełnie inne. Przykład jest tyleż trafny, co i nie. Niuans polega na tym, że w obu zdaniach oceniane są dwa obiekty: hotel i miasto. Taka sytuacja występuje rzadko. W większości tekstów tematem jest jeden obiekt, a większość użytych przymiotników i określeń odnosi się do niego. Okazuje się, że nawet w przypadku recenzji filmowych (gdzie jest wiele postaci, postaw, obiektów dobrych i złych) reprezentacja w formie BOW może być przydatna. Intuicyjnie wydaje się, że im dłuższe teksty, tym prezentacja za pomocą BOW będzie tracić na znaczeniu.

Podejściu BOW można zarzucić, iż pomija ono istotne informacje zawarte w dokumencie, np. układ akapitów, kolejność wyrazów, struktury syntaktyczne, semantyczne powiązania pomiędzy wyrazami. Badacze poświęcili dużo uwagi uwzględnieniu tych braków. Pang i inni (2002), Dave i inni (2003), Joshi i Penstein-Rosé (2009) w swych pracach wykorzystali *n*-gramy wyższego rzędu, Hatzivassiloglou, Wiebe (2000) zastosowali części mowy (*part of speech* – POS), Na i inni (2005) *użyli* połączenia przeczeń z wyrazami sąsiadującymi, tworząc tzw. *negation phases*. Dave i inni (2003), Gamon (2004), Subrahmanian i Reforgiato (2008), Joshi i Penstein-Rosé (2009) zaproponowali sposoby na uchwycenie zależności językowych.

W kolejnym kroku tworzona jest reprezentacja dokumentów tekstowych za pomocą *modelu* VSM (*Vector Space Model*; Salton i inni 1975), w którym dokument tekstowy przedstawiony jest jako *j*-elementowy wektor. Poszczególne elementy

wektora odpowiadają liście *n*-gramów (zwanych *features* lub *terms*) utworzonych z całego korpusu. Kolekcję dokumentów można przedstawić w *macierzy* DTM (*document-term matrix*), w której poszczególne dokumenty reprezentowane są przez wiersze, natomiast poszczególne *n*-gramy reprezentowane są przez kolumny:

$$x = \left[x_{ij} \right], \tag{1.5}$$

gdzie:
x – macierz DTM,
x_{ij} – liczba wystąpień j-tego n-gramu w i-tym dokumencie,
$i = 1, \ldots, N$ (N – liczba dokumentów),
$j = 1, \ldots, M$ (M – liczba n-gramów).

Oznaczenie x_{ij} zostało przyjęte w bieżącym rozdziale dla wygody zapisu wzorów oraz dlatego, że zapisy mają odniesienie do obiektów o szerszym znaczeniu niż słowo lub termin, tzn. do n-gramów. W dalszej części książki będziemy posługiwać się symbolami $m_{f,k}$ w odniesieniu do częstości terminu i klasy, gdzie termin/cecha oznaczany jest literą f oraz $n_{f,k}$ w odniesieniu do częstości terminu, dokumentu i klasy. Zauważmy, że żaden z tych dwóch symboli nie oznacza częstości wystąpienia terminu f w jakimś dokumencie. Elementy macierzy DTM przyjmują wartość zero wówczas, gdy dany n-gram nie wystąpił w rozpatrywanym dokumencie. W przeciwnym przypadku macierz przyjmuje wartości różne od zera. Ostateczna wartość zależeć będzie od wybranego *schematu ważenia zmiennych* (*feature weighting scheme*), który pozwala badaczowi na różnorodne zaakcentowanie pewnych cech DTM. Do kanonu technik ważenia zmiennych w analizie sentymentu należą (omówione w pracach: Pang i inni 2002; Na i inni 2005):

1) postać binarna (*term presence*, *binary*):

$$x_{ij} = \begin{cases} 0, & \text{kiedy} \quad TF_{ij} = 0 \\ 1, & \text{kiedy} \quad TF_{ij} > 0 \end{cases}, \tag{1.6}$$

2) postać częstościowa (*term frequency* – TF):

$$x_{ij} = TF_{ij}, \tag{1.7}$$

3) postać TFIDF (*term frequency inverse document frequency* – TFIDF):

$$x_{ij} = \begin{cases} 0, & \text{kiedy} \quad TF_{ij} = 0 \\ TF_{ij} * \log\left(\dfrac{N}{nf_j}\right), & \text{kiedy} \quad TF_{ij} > 0 \end{cases}, \tag{1.8}$$

gdzie:
TF_{ij} – liczba wystąpień j-tego n-gramu w i-tym dokumencie,
N – liczba wszystkich dokumentów,
nf_j – liczba dokumentów, w których j-ty n-gram wystąpił.

Postać binarną cechuje prostota oraz łatwość interpretacji. W tym wariancie DTM przechowuje informacje o fakcie wystąpienia poszczególnego *n*-gramu w *i*-tym dokumencie. Minusem może być strata informacji w stosunku do wariantu częstościowego, ponieważ macierz składa się z samych zer oraz jedynek, pomijając przy tym wielokrotne wystąpienia tego samego *n*-gramu w *i*-tym dokumencie. Postać binarna przydaje się jednak wówczas, gdy metoda klasyfikacji wymaga danych wejściowych w formacie zero-jedynkowym (np. naiwny klasyfikator Bayesa w wariancie z rozkładem zero-jedynkowym).

Wariant częstościowy macierzy DTM zawiera liczności wystąpień poszczególnych *n*-gramów w poszczególnych dokumentach. Zaletą tego podejścia jest niewątpliwie większy wektor informacji wykorzystywany do dyskryminacji dokumentów ze względu na ich wydźwięk. W takiej reprezentacji częstość wystąpienia danego *n*-gramu ma znaczenie, stąd *n*-gramy o dużym ładunku emocjonalnym występujące często będą silniej separować zbiór dokumentów ze względu na wydźwięk. Niepożądanym efektem wykorzystania częstości do opisu dokumentów tekstowych jest podwyższenie znaczenia słów występujących często, zazwyczaj niemających związku z sentymentem, a bardziej wynikających ze specyfiki danego języka. W celu zniwelowania tego efektu można wykorzystać funkcję mitygującą (*damping function*) postaci $\sqrt{TF_{ij}}$ lub $\log(1 + TF_{ij})$.

Ostatni wyżej wymieniony wariant, TFIDF, waży liczności TF_{ij} *odwrotną liczebnością dokumentów* (*inverse document frequency* – IDF) – $\log(N / nf_j)$. Taki zabieg ma w zamyśle osłabienie wpływu *n*-gramów pojawiających się w dużej liczbie dokumentów, przypuszczalnie mających słabsze zdolności dyskryminacyjne zbioru dokumentów. TFIDF pomaga osłabić znaczenie słów m.in. znajdujących się na stopliście, kiedy badacz nie zdecyduje się na jej zastosowanie lub uzupełnić ją, w przypadku gdy stoplista nie jest kompletna.

Powyższa lista prezentuje jedynie podstawowe i sprawdzone techniki ważenia zmiennych, które przeniknęły do analizy sentymentu z obszaru klasyfikacji tekstu (*text classification*), a swoje korzenie mają w dyscyplinie *wyszukiwania informacji* (*information retrieval*). Bardziej wyszukane propozycje ważenia zmiennych zaproponowali w badaniach m.in. Paltoglou i Thelwall (2010) oraz Carvalho i Guedes (2020).

Uważny czytelnik zwróci uwagę na fakt, iż większość elementów w macierzy DTM jest równa zero. Wszak język dozwala przekazać tę samą informację przy pomocy różnych słów. Mnogość informacji do przekazania, szeroki wachlarz słownictwa, błędy ortograficzne, możliwe literówki etc. powodują, że macierz DTM jest rzadka oraz ma bardzo duży wymiar. Te właściwości omawianej macierzy nastręczają kłopotów natury numerycznej, wydłużają czas obliczeń oraz negatywnie oddziałują na klasyfikatory wrażliwe na duży wymiar zmiennych objaśniających (szczególnie, gdy liczba obserwacji znacznie przewyższa liczbę zmiennych objaśniających). Saif i inni (2012) zaproponowali dwie skuteczne metody rozwiązania problemu rzadkości macierzy DTM. Redukcję wymiaru badano wielokrotnie w publikacjach omawianych w dalszej części monografii.

1.4. Klasyfikatory wykorzystywane w badaniach[1]

1.4.1. Naiwny klasyfikator Bayesa

Reguła Bayesa (Domański, Pruska 2000) definiuje prawdopodobieństwo *a posteriori* wystąpienia dokumentu D należącego do klasy C_k jako:

$$P(C_k|D) = \frac{P(C_k)P(D|C_k)}{P(D)}, \tag{1.9}$$

gdzie:
$P(C_k)$ – prawdopodobieństwo wystąpienia dokumentu należącego do klasy C_k (prawdopodobieństwo *a priori*),
$P(D|C_k)$ – prawdopodobieństwo zaobserwowania dokumentu D pod warunkiem, że należy on do klasy C_k,
$P(D)$ – prawdopodobieństwo zaobserwowania dokumentu D.

Reguła klasyfikacyjna polega na przydzieleniu dokumentu D do klasy C_k, dla której zachodzi równość:

$$P(C_k|D) = \max_k \mathrm{P}(C_k|D), \tag{1.10}$$

co jest równoważne z:

$$P(C_k|D) = \max_k [P(C_k)P(D|C_k)]. \tag{1.11}$$

Wygodnie jest założyć w powyższej formule, że n-gramy x_j mają niezależne rozkłady w k-tej klasie:

$$P(D|C_k) = \prod_{j=1}^{M} P(x_j|C_k). \tag{1.12}$$

W rezultacie otrzymujemy *naiwny klasyfikator Bayesa* (*naive Bayes classifier* – NB), a jego naiwność wynika z przyjętego założenia o niezależności n-gramów, które w praktyce rzadko jest spełnione. W tym miejscu czytelnikowi należy się jeszcze kilka słów wyjaśnień na temat przyjętego uproszczenia. Po pierwsze, założenie o niezależności upraszcza sposób wyznaczania prawdopodobieństw warunkowych $P(x_j|C_k)$ z problemu wielowymiarowego do przypadku jednowymiarowego. Po drugie, naruszenie tego założenia niekoniecznie przekłada się na niską jakość klasyfikacji i nie musi mieć na nią większego wpływu, ilekroć relacja opisana równaniem (1.11) dla k-tej klasy nie zostaje zaburzona.

Zastosowanie naiwnego klasyfikatora Bayesa sprowadza się zatem do obliczenia prawdopodobieństw *a posteriori* $P(C_k|D)$ dla każdej z k klas oraz zaklasyfikowania

1 Tam, gdzie to możliwe, pominięto notację dotyczącą numeru dokumentu (indeks *i*).

dokumentu do klasy, dla której wartość obliczonego prawdopodobieństwa jest największa. W celu obliczenia $P(C_k|D)$ konieczna jest znajomość prawdopodobieństwa *a priori* $P(C_k)$ oraz prawdopodobieństwa warunkowego $P(D|C_k)$. Prawdopodobieństwo $P(C_k)$ wyznaczane jest na podstawie zbioru treningowego przy pomocy estymatora największej wiarygodności:

$$P(C_k) = \frac{n_k}{N}, \tag{1.13}$$

gdzie n_k oznacza liczbę dokumentów, które należą do *k*-tej klasy.

Prawdopodobieństwa warunkowe $P(D|C_k)$ zależą od przyjętego modelu generującego dokument *D* w dowolnej klasie. Model ten ma ścisły związek z zastosowanym schematem ważenia *n*-gramów. Poniżej przedstawiono kilka wybranych modeli prawdopodobieństwa.

1.4.1.1. Model zero-jedynkowy

Dla macierzy DTM w postaci binarnej poszczególnym *n*-gramom przypisywane są 0 i 1 w zależności od tego, czy dany *n*-gram pojawił się w dokumencie tekstowym czy też nie. Model zero-jedynkowy, odpowiadający takiej postaci DTM, zakłada, że *j*-ty *n*-gram może pojawić się w dokumencie *D* w klasie *k* z prawdopodobieństwem równym p_{jk} (oraz może się nie pojawić z prawdopodobieństwem $1 - p_{jk}$). Łączne prawdopodobieństwo wystąpienia dokumentu *D* w klasie *k*, $P(D|C_k)$, wyrażone jest jako iloraz poszczególnych prawdopodobieństw wystąpienia lub niewystąpienia danego *n*-gramu:

$$P(D|C_k) = \prod_{x_j \in D} p_{jk} \prod_{x_j \notin D} (1 - p_{jk}). \tag{1.14}$$

Do oszacowania parametru modelu zero-jedynkowego, p_{jk}, wykorzystuje się estymator największej wiarygodności danej wzorem:

$$p_{jk} = \frac{m_{jk}}{n_k} \tag{1.15}$$

gdzie n_{jk} oznacza liczbę dokumentów, które należą do *k*-tej klasy, zawierających *j*-ty *n*-gram.

1.4.1.2. Model wielomianowy

Model wielomianowy opisuje prawdopodobieństwo liczby sukcesów *j* w dokonanej liczbie *l* niezależnych prób, w której w każdych z tych prób istnieje stałe prawdopodobieństwo sukcesu p_j dla każdego z *j* zdarzeń (przy czym $\sum_{j=1}^{M} p_j = 1$). Przekładając to na nomenklaturę analizy sentymentu, można powiedzieć, że model wielomianowy opisuje prawdopodobieństwo liczby wystąpień poszczególnych *n*-gramów w zaobserwowanym dokumencie *D*. Taki model wpasowuje się w postać częstościową macierzy DTM, w której dokument tekstowy wyrażony jest jako

liczba wystąpień poszczególnych *n*-gramów, tj. $D = (x_1, x_2, ..., x_j)$. Wówczas prawdopodobieństwo pojawienia się określonych liczebności *n*-gramów w dokumencie należącym do klasy *k* wynosi:

$$P(D|C_k) = \frac{l!}{x_1!x_{2!}\ldots x_M!}\prod_{j=1}^{M} p_{jk}^{x_j}, \tag{1.16}$$

gdzie $l = \sum_{j=1}^{M} x_j$.

Brakującym elementem jest p_{jk}, oznaczającym prawdopodobieństwo pojawienia się *j*-tego *n*-gramu w dokumencie *D*, pod warunkiem, że należy on do *k*-tej klasy. Estymator największej wiarygodności dla tego parametru przedstawia się następująco:

$$p_{jk} = \frac{n_{jk}}{\sum_{j=1}^{M} n_{jk}}, \tag{1.17}$$

gdzie n_{jk} oznacza liczbę wystąpień *j*-tego *n*-gramu w dokumentach należących do klasy *k*.

1.4.1.3. Model Gaussa

Postać TFIDF transformuje liczebności poszczególnych *n*-gramów, a więc zmiennych dyskretnych, do postaci zmiennych ciągłych. Wówczas konieczne jest posiłkowanie się odpowiednim rozkładem prawdopodobieństwa. Odpowiednia funkcja gęstości może przyjąć postać rozkładu normalnego, tj.:

$$P(D|C_k) = (\sigma_{jk}\sqrt{2\pi})^{-1} \cdot \exp\left(-\frac{(X_j - \mu_{jk})^2}{2\sigma_{jk}^2}\right), \tag{1.18}$$

gdzie:
μ_{jk} – średnia dla *j*-tego *n*-gramu w *k*-tej klasie,
σ_{jk} – odchylenie standardowe dla *j*-tego *n*-gramu w *k*-tej klasie.

1.4.2. Regresja logistyczna

Załóżmy, że zmienna losowa *C* ma rozkład zero-jedynkowy z parametrem *p*:

$$C \sim ZJ(p) \tag{1.19}$$

oraz może przyjąć dwie wartości:

$$C = \begin{cases} 0, & \text{kiedy sentyment dokumentu tekstowego jest negatywny,} \\ 1, & \text{kiedy sentyment dokumentu tekstowego jest pozytywny,} \end{cases} \tag{1.20}$$

wówczas prawdopodobieństwo tego, że sentyment dokumentu D jest negatywny, p, można opisać przy pomocy regresji logistycznej (Hosmer, Lemeshow, Sturdivant 2013), tj.:

$$p = p(C = 0|D) = \frac{e^{\beta_0 + \boldsymbol{\beta}^t \boldsymbol{D}}}{1 + e^{\beta_0 + \boldsymbol{\beta}^t \boldsymbol{D}}}, \tag{1.21}$$

gdzie:
β_0 – wyraz wolny,
$\boldsymbol{\beta}$ – wektor szacowanych parametrów.

Ze względów praktycznych dobrze jest zastosować *przekształcenie logitowe* (*logit transformation*) na równaniu (1.21), otrzymując pożądane własności modelu liniowego:

$$ln\left(\frac{p}{1-p}\right) = \beta_0 + \boldsymbol{\beta}^t \boldsymbol{D}, \tag{1.22}$$

W rezultacie powyższe równanie jest liniowe względem parametrów. Pozytywną konsekwencją tego faktu jest możliwość interpretacji parametrów, w podobny sposób jak w modelu liniowym, w kontekście *ilorazu szans* $\left(\frac{e^{\beta_0 + \boldsymbol{\beta}^t \boldsymbol{D}'}}{e^{\beta_0 + \boldsymbol{\beta}^t \boldsymbol{D}}}\right)$[2].

Przykładowo, jeżeli jeden z elementów wektora D, x_j, wzrośnie o jednostkę (*ceteris paribus*), wówczas iloraz szans wzrośnie o e^{β_j}. Oznacza to, że szansa na to, że dokument ma pozytywny wydźwięk (zakładając zmianę x_j o jednostkę), wzrosła o $(e^{\beta_j} - 1) \cdot 100\%$.

Prawdopodobieństwo tego, że dokument D ma pozytywny sentyment, można wyznaczyć z poniższego równania:

$$p(C = 1|D) = 1 - p(C = 0|D). \tag{1.23}$$

Reguła klasyfikacyjna polega na przypisaniu dokumentowi D negatywnego wydźwięku, jeżeli poniższe równanie jest spełnione:

$$P(C = 0|D) = \max[p(C = 0|D), p(C = 1|D)], \tag{1.24}$$

w przeciwnym wypadku dokumentowi przypisuje się pozytywny wydźwięk. Innymi słowy reguła nadaje wydźwięk dokumentowi D w zależności od tego, które z prawdopodobieństw $P(C = 0|D)$, $P(C = 1|D)$ jest większe.

Parametry równania (1.22) znajdowane są przy pomocy metody największej wiarygodności, maksymalizując poniższą funkcję:

2 $e^{(\beta_0 + \boldsymbol{\beta}^t \boldsymbol{D}')}$ oznacza szansę obliczoną dla x_j podwyższonego o jedną jednostkę.

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{I} p(C_i|\boldsymbol{D}_i), \tag{1.25}$$

ze względu na parametry β_0 oraz $\boldsymbol{\beta}$:

$$\hat{\boldsymbol{\beta}} = \underset{\beta}{\operatorname{argmax}}\, L(\boldsymbol{\beta}). \tag{1.26}$$

1.4.3. Metoda SVM

Metoda wektorów nośnych (*Support Vector Machine* – SVM) znajduje obecnie szerokie zastosowanie w obszarze zwanym *machine learning*. Została zaproponowana przez Vladimira Vapnika oraz Alexeya Chervonenkisa w latach 60. ubiegłego wieku. Współczesną postać metody wektorów nośnych wyklarowały publikacje z lat 90. (Boser i inni 1992; Cortes i Vapnik 1995).

Załóżmy, że dokument tekstowy D o wydźwięku $y \in \{-1;+1\}$ (odpowiednio: negatywny i pozytywny) może być przedstawiony przy pomocy M n-gramów, tj. $D = (X_1, \ldots, X_M)$, wówczas liniowy klasyfikator opisują dwa poniższe równania:

$$\begin{aligned} F(D) &= znak(f(D)) \\ f(D) &= a_0 + \sum_{j=1}^{M} a_j X_j, \end{aligned} \tag{1.27}$$

gdzie:
$a_0, \ldots, a_M$ – parametry modelu,
funkcja *znak* przyjmuje wartość 1, gdy $f(D) > 0$, -1, gdy $f(D) < 0$ oraz 0, gdy $f(D) = 0$.

Tak zdefiniowany liniowy separator wyznacza granicę decyzyjną $f(D) = 0$ pomiędzy dokumentami o wydźwięku pozytywnym a dokumentami o wydźwięku negatywnym.

W przypadku, gdy dane są liniowo separowane, istnieje hiperpłaszczyzna $f(D)$, taka że dla każdego dokumentu D ze zbioru danych spełniony jest warunek:

$$y_i f(D_i) \geq 1. \tag{1.28}$$

Takich hiperpłaszczyzn można wyznaczyć nieskończenie wiele, jednak idea metody SVM polega na znalezieniu takiej hiperpłaszczyzny, która maksymalizuje jej odległość od najbliższego punktu (dokumentu) w układzie. Taki zamysł wydaje się zasadny, ponieważ dobra granica decyzyjna powinna separować dane, zachowując możliwie dużą odległość od klasyfikowanych obiektów. Odległość hiperpłaszczyzny od najbliższego punktu nosi nazwę *marginesu* (*margin*), natomiast punkty leżące na granicy marginesu noszą nazwę *wektorów nośnych*

(*support vectors*). Maksymalizacja marginesu jest realizowana poprzez minimalizację wyrażenia[3]:

$$\frac{1}{2}\sum_{j=1}^{M} a_j^2. \tag{1.29}$$

Uwzględniając powyższe postulaty, w celu wyznaczenia parametrów $f(D)$ minimalizowana jest funkcja celu dana poniższym wzorem:

$$\frac{1}{2}\sum_{j=1}^{M} a_j^2 + C \cdot \sum_{i=1}^{N} \xi_i \tag{1.30}$$

pod warunkami:

$$\xi_i \geq 1 - y_i f(D_i) \text{ oraz } \xi_i \geq 0,\ i = 1, \ldots, N. \tag{1.31}$$

gdzie:
W – wektor j parametrów odpowiadających n-gramom,
C – stała zwana też jako *slack penalty*,
ξ_i – zmienne dopełniające (*slack variables*), dozwalające na naruszenie marginesu przez i-tą obserwację.

Minimalizacja funkcji celu może być przedstawiona jako zadanie programowania kwadratowego, w którym maksymalizowana jest funkcja (Aggarwal 2018):

$$\sum_{i=1}^{N} \alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{k=1}^{N} \alpha_i \alpha_k y_i y_k \langle D_i, D_k \rangle \tag{1.32}$$

ze względu na α_i ($i = 1, \ldots, N$) pod warunkiem:

$$0 \leq \alpha_i \leq C, \tag{1.33}$$

gdzie $\langle D_i, D_k \rangle$ oznacza iloczyn skalarny D_i oraz D_k.

Parametry funkcji $f(D)$ wyliczane są zgodnie ze wzorem:

$$a_j = \sum_{i=1}^{N} \alpha_i y_i X_j. \tag{1.34}$$

Metoda SVM zyskuje na uniwersalności poprzez zastosowanie tzw. *kernel trick*. Zamysł polega na zastosowaniu liniowego klasyfikatora dla zmiennych $X_1, \ldots, X_M$ przekształconych z przestrzeni R^M do przestrzeni o większym wymiarze R^Z ($Z > M$). Takie przekształcenie umożliwia przeniesienie skomplikowanej (nieliniowej) granicy decyzyjnej z przestrzeni R^M do przestrzeni, w której dane będą separowalne

3 Odległość między hiperpłaszczyzną $a_0 + \sum_{j=1}^{M} a_j X_j = 1$ a $a_0 + \sum_{j=1}^{M} a_j X_j = -1$ wynosi $2/\sum_{j=1}^{M} a_j^2$, zatem minimalizowanie $1/2(\sum_{j=1}^{M} a_j^2)$ oznacza maksymalizację marginesu.

liniowo. Całość skutkuje większą elastycznością metody SVM, zachowując jednocześnie prostotę liniowego modelu.

Potencjał, jaki drzemie w równaniu (1.32), ukryty jest w powiązaniu funkcji straty z iloczynem skalarnym pomiędzy dwoma wektorami, tj. $\langle D_i, D_k \rangle$. Zastosowanie funkcji jądrowej zdecydowanie upraszcza problemy natury obliczeniowej, ponieważ umożliwia ono obliczenie iloczynu skalarnego dla wielowymiarowych przestrzeni (nawet nieskończonych, np. dla jądra gaussowskiego) na oryginalnym zbiorze danych, bez potrzeby znajomości samego przekształcenia $\varphi(\varphi: R^M \rightarrow R^Z)$, a zatem i bez potrzeby wyznaczenia projekcji zmiennych do wyższych wymiarów.

Do podstawowych jąder stosowanych w metodzie SVM należą jądra:

liniowe: $K(D_i, D_k) = D_i^T D_k$,
wielomianowe: $K(D_i, D_k) = (\gamma D_i^T D_k + r)^d$, $\gamma > 0$,
gaussowskie (RBF): $K(D_i, D_k) = exp(-\gamma||D_i - D_k||^2)$,
sigmoidalne: $K(D_i, D_k) = tanh(\gamma D_i^T D_k + r)$,

gdzie γ, r, d są parametrami jądra.

1.5. Miary jakości klasyfikacji

Naturalną, intuicyjnie oczywistą miarą jakości klasyfikacji wydaje się dokładność (trafność) klasyfikacji (*accuracy*), czyli liczba poprawnie sklasyfikowanych dokumentów podzielona przez liczbę wszystkich dokumentów. W przypadku dwóch klas dokładność będzie dana wzorem:

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN}, \tag{1.35}$$

gdzie:
TP (*true positives*) – liczba poprawnie sklasyfikowanych dokumentów pozytywnych,
FP (*false positives*) – liczba niepoprawnie sklasyfikowanych dokumentów negatywnych,
TN (*true negatives*) – liczba poprawnie sklasyfikowanych dokumentów negatywnych,
FN (*false negatives*) – liczba niepoprawnie sklasyfikowanych dokumentów pozytywnych.

Trafność to bardzo dobra miara, niestety tylko wtedy, gdy liczebności klas są mniej więcej jednakowe. Jeżeli tak nie jest, to należy odwołać się do innych miar, które zadbają o to, żeby klasyfikacja, która pominęła np. jedną nieliczną klasę, nie okazała się dobra. W literaturze przedmiotu dominują jeszcze dwie miary: precyzja (*precision*) oraz czułość (*recall*). W przypadku dwóch klas są one dane wzorami:

$$precision = \frac{TP}{TP + FP}, \tag{1.36}$$

$$recall = \frac{TP}{TP + FN}. \tag{1.37}$$

W celu zmniejszenia liczby miar, których nadmiar mógłby zniekształcać poprawną ocenę, bardzo często stosuje się syntezę precyzji oraz czułości w postaci miary F1, która jest średnią harmoniczną precyzji i czułości, a jest dana wzorem:

$$F1 = \frac{2 \cdot precision \cdot recall}{precision + recall}. \tag{1.38}$$

W niniejszej monografii będą stosowane dwie miary: dokładność klasyfikacji oraz miara F1.

1.6. Testowe zbiory danych

W klasycznej analizie statystycznej badacz może uogólniać przydatność nowych metod przy pomocy symulacji, w której zakłada określone rozkłady zmiennych. Ze względu na nieustrukturyzowany charakter danych tekstowych, klasyczne metody symulacji nie znajdują zastosowania w statystycznej analizie tekstu. Badania skupiające swoją uwagę na tekście na ogół mają charakter studium przypadku – uczeni w swych rozważaniach posługują się zazwyczaj ogólnodostępnymi *zbiorami benchmarkowymi* (*benchmark dataset*), które, w pewnym stopniu, zapewniają porównywalność uzyskanych rezultatów. W literaturze przedmiotu zdecydowanie w roli zbioru benchmarkowego dominuje zbiór anglojęzyczny *polarity dataset*, wchodzący w skład kompletu danych pod tytułem *Movie Review Data*. W roli dużego zbioru dokumentów występuje najczęściej zbiór *OHSUMED*. Do tych dwóch zbiorów będziemy odwoływać się często, podając wyniki różnych badań, w związku z czym poniżej przedstawiamy ich krótką charakterystykę. Niniejsze opracowanie jest poświęcone głównie badaniu sentymentu tekstów polskich, dlatego przedstawiamy siedem zbiorów polskojęzycznych dość różnorodnych pod względem rodzaju, zakresu dziedzinowego oraz języka wypowiedzi. Wadą tych zbiorów było to, że były one bardzo niezbilansowane pod względem kilkukrotnej przewagi dokumentów pozytywnych nad negatywnymi.

W takich warunkach wszelkie badania efektywności klasyfikacji były bardzo obciążone łatwością klasyfikacji (pomimo różnych miar stosowanych do oceny jej efektywności). Wobec tego, z sześciu spośród tych zbiorów utworzono sześć kolejnych zbiorów zbilansowanych pod względem odsetka dokumentów pozytywnych i negatywnych poprzez losowe usunięcie części zbiorów pozytywnych. Na zbiorach zbilansowanych wszystkie metody badania sentymentu mają, na ogół, trudniejsze zdanie.

1.6.1. Zbiór *Reuters-21578*

Zbiór *Reuters-21578* składa się z 21 578 krótkich artykułów na różne tematy. Każdy z artykułów został przypisany do jednej z dwóch kategorii: pozytywnej lub negatywnej. Autorzy zalecają korzystanie z tego zbioru według podziału na zbiór uczący (13 625 dokumentów) i testowy (6188 dokumentów) oraz pominięcie 1765 dokumentów.

1.6.2. Zbiór *Polarity*

Zbiór recenzji filmowych *Movie Review Data*[4] to dokumenty wygenerowane na podstawie danych zgromadzonych w popularnych internetowych filmowych bazach danych. Całość składa się z 3 plików opisanych poniżej.

Pool of HTML files

Plik składa się z 27 886 plików w formacie HTML. Dane zawierają surowe recenzje, które nie zostały poddane obróbce wstępnej tekstu oraz nie posiadają nadanych etykiet wydźwięku. Recenzje filmowe zostały pobrane z internetowej filmowej bazy danych o nazwie *Internet Movie Database* (IMDB)[5].

Polarity dataset

Najnowszy plik w wersji 2.0, wykorzystany w publikacji Pang i Lee (2004) i wielu innych, zawiera 2000 recenzji (po 1000 negatywnych i pozytywnych) wstępnie przetworzonych na podstawie wyżej opisanego pliku HTML. Ze względu na różne skale ocen surowych recenzji nadawanie etykiet sentymentu zależne było od tegoż formatu. Brano pod uwagę tylko te dokumenty tekstowe, w których autor wprost wyraził swoją ocenę. Oceny mierzone na skali pięciogwiazdkowej oznaczono jako pozytywne, kiedy ocena wyniosła 3,5 gwiazdki lub więcej, natomiast oceny 2 i mniej oceniono jako negatywne. Brano pod uwagę tylko te dokumenty

4 *Movie Review Data*, https://www.cs.cornell.edu/people/pabo/movie-review-data/ (dostęp: 15.10.2021).

5 *Internet Movie Database*, www.imdb.com (dostęp: 15.10.2021).

tekstowe, w których autor wprost wyraził swoją ocenę. Oceny mierzone na skali czterogwiazdkowej oznaczono jako pozytywne, kiedy ocena wyniosła 3 gwiazdki lub więcej, natomiast oceny 1,5 i mniej oceniono jako negatywne. Oceny wyrażone przy pomocy liter oznaczono jako pozytywne, kiedy ocena wynosiła B lub więcej, z kolei oceny oznaczone literą C lub mniej oznaczone zostały jako negatywne.

Sentence polarity dataset

Zbiór zawiera 10 662 zdań i skrawków wypowiedzi (*snippets*) pobranych z internetowej bazy danych *rottentomatoes*. Każde zdanie pochodzące z bazy oznaczono jako *fresh* lub jako *rotten* oraz nadano im odpowiednio etykiety wydźwięku: pozytywny lub negatywny. Sentyment rozkłada się po 5331 zdań pozytywnych oraz 5331 zdań negatywnych. Zbiór został po raz pierwszy wykorzystany w pracy Pang i Lee (2005).

1.6.3. Zbiór *OHSUMED*

Zbiór *OHSUMED* został po raz pierwszy wprowadzony do badań przez Hersha (Hersh i inni 1994). Składa się on z 348 566 dokumentów medycznych z lat 1987–1991. Dokumenty są krótkimi opisami pacjentów i ich leczenia. Każdy dokument jest przypisany do jednej z dwóch klas: pozytywnej albo negatywnej.

1.6.4. Zbiór *bank*

W skład polskojęzycznych recenzji dotyczących jednego z czołowych polskich banków wchodzą trzy zbiory. Dokumenty tekstowe zawierają teksty napisane w języku polskim przez klientów banku na tematy szeroko związane z bankowością (m.in. sprawy dotyczące aplikacji mobilnej, serwisu transakcyjnego, bankomatów, produktów finansowych, bezpieczeństwa, obsługi, windykacji, marketingu & PR, infolinii, reklamacji) oraz doświadczeniami na linii klient–bank. Piszący recenzję klient, oprócz wiadomości tekstowej, określa jej pozytywny lub negatywny wydźwięk przy pomocy wyboru odpowiedniej ikony, tj. „uśmiechniętej twarzy" lub „smutnej twarzy". Omawiane zbiory są dobrze zbilansowane, czyli liczba tekstów o pozytywnym sentymencie jest bliska liczbie tekstów o negatywnym sentymencie. Zbiór liczy 7859 dokumentów, z czego 3837 to dokumenty negatywne, natomiast 4022 – pozytywne.

1.6.5. Zbiór *perfumy*

Zbiór zawiera 2591 recenzji damskich i męskich perfum zakupionych przez klientów sklepów internetowych. Spośród wszystkich dokumentów 268 recenzji nacechowanych jest negatywnie, natomiast pozostałe 2323 dokumenty nace-

chowane są pozytywnie. Wydźwięk dokumentu tekstowego określono w oparciu o pięciogwiazdkowy system oceny. Teksty z oceną poniżej 3 gwiazdek oznaczono jako komentarze negatywne, z kolei teksty ocenione powyżej 3 gwiazdek zaliczono do komentarzy pozytywnych. Recenzje z oceną 3 wykluczono z analizy z uwagi na trudność związaną z jednoznacznym przyporządkowaniem odpowiedniego wydźwięku wypowiedzi.

1.6.6. Zbiór *perfumyzbil*

Zbiór *perfumyzbil* powstał w ten sposób, że spośród 2323 dokumentów pozytywnych ze zbioru *perfumy* wybrano losowo 2000 dokumentów. W efekcie otrzymano zbiór 591 dokumentów, z których 323 to dokumenty pozytywne, zaś 268 – negatywne.

1.6.7. Zbiór *ksiazki*

Zbiór składa się z 8933 recenzji książek. W całym zbiorze dokumentów 3415 recenzji nacechowanych jest negatywnie, natomiast pozostałe 5518 dokumentów nacechowanych jest pozytywnie. Wydźwięk dokumentu tekstowego określono w oparciu o dziesięciogwiazdkowy system oceny. Teksty z oceną poniżej 5 gwiazdek oznaczono jako komentarze negatywne, z kolei teksty ocenione powyżej 5 gwiazdek zaliczono do komentarzy pozytywnych. Recenzje z oceną 5 wykluczono z analizy z uwagi na trudność związaną z jednoznacznym przyporządkowaniem odpowiedniego wydźwięku wypowiedzi.

1.6.8. Zbiór *ksiazkizbil*

Zbiór powstał w ten sposób, że spośród 5518 dokumentów pozytywnych ze zbioru *ksiazki* wybrano losowo 1100 dokumentów. W efekcie otrzymano zbiór złożony z 7833 dokumentów, z których 4418 to dokumenty pozytywne, zaś 3415 – negatywne.

1.6.9. Zbiór *apteki*

Zbiór 5934 recenzji tworzą opinie klientów aptek internetowych. Komentarze dotyczą oceny produktów, obsługi, szybkości realizacji zamówienia etc. Zbiór zawiera 1007 dokumentów nacechowanych negatywnie oraz 4927 dokumentów nacechowanych pozytywnie. Wydźwięk dokumentu tekstowego określono w oparciu

o pięciogwiazdkowy system oceny. Teksty z oceną poniżej 3 gwiazdek oznaczono jako komentarze negatywne, z kolei teksty ocenione powyżej 3 gwiazdek zaliczono do komentarzy pozytywnych. Recenzje z oceną 3 wykluczono z analizy z uwagi na trudność związaną z jednoznacznym przyporządkowaniem odpowiedniego wydźwięku wypowiedzi.

1.6.10. Zbiór *aptekizbil*

Zbiór powstał w ten sposób, że spośród 4927 dokumentów pozytywnych ze zbioru *apteki* wybrano losowo 3900 dokumentów. W efekcie otrzymano zbiór 2034 dokumentów, z których 1027 to dokumenty pozytywne, zaś 1007 – negatywne.

1.6.11. Zbiór *esklepy*

W zbiorze zawarto 21 760 komentarzy klientów na temat popularnych sklepów internetowych ze sprzętem AGD. Komentarze dotyczą oceny produktów, obsługi, szybkości realizacji zamówienia etc. Zbiór zawiera 4366 dokumentów nacechowanych negatywnie oraz 17 394 dokumenty nacechowane pozytywnie. Wydźwięk dokumentu tekstowego określono w oparciu o pięciogwiazdkowy system oceny. Teksty z oceną poniżej 3 gwiazdek oznaczono jako komentarze negatywne, z kolei teksty ocenione powyżej 3 gwiazdek zaliczono do komentarzy pozytywnych. Recenzje z oceną 3 wykluczono z analizy z uwagi na trudność związaną z jednoznacznym przyporządkowaniem odpowiedniego wydźwięku wypowiedzi.

1.6.12. Zbiór *esklepyzbil*

Zbiór powstał w ten sposób, że spośród 17 394 dokumentów pozytywnych ze zbioru *esklepy* wybrano losowo 13 000 dokumentów. W efekcie otrzymano zbiór 8760 dokumentów, z których 4394 to dokumenty pozytywne, zaś 4366 – negatywne.

1.6.13. Zbiór *kurier*

Zbiór recenzji firm kurierskich składa się z 4137 komentarzy klientów na temat jakości świadczonych usług przez popularnych polskich przewoźników. Zbiór zawiera 539 dokumentów nacechowanych negatywnie oraz 3598 dokumentów nacechowanych pozytywnie. Wydźwięk dokumentu tekstowego określono w oparciu

o pięciogwiazdkowy system oceny. Teksty z oceną poniżej 3 gwiazdek oznaczono jako komentarze negatywne, z kolei teksty ocenione powyżej 3 gwiazdek zaliczono do komentarzy pozytywnych. Recenzje z oceną 3 wykluczono z analizy z uwagi na trudność związaną z jednoznacznym przyporządkowaniem odpowiedniego wydźwięku wypowiedzi.

1.6.14. Zbiór *kurierzbil*

Zbiór powstał w ten sposób, że spośród 3598 dokumentów pozytywnych ze zbioru *kurier* wybrano losowo 3000 dokumentów. W efekcie otrzymano zbiór 1137 dokumentów, z których 598 to dokumenty pozytywne, zaś 539 – negatywne.

1.6.15. Zbiór *hotele*

W zbiorze zawarto 3366 opinii klientów na temat pobytu w zagranicznych hotelach. Zbiór zawiera 1031 dokumentów nacechowanych negatywnie oraz 2335 dokumentów nacechowanych pozytywnie. Wydźwięk dokumentu tekstowego określono w oparciu o dziesięciostopniowy system oceny. Teksty z oceną poniżej 5 oznaczono jako komentarze negatywne, z kolei teksty ocenione powyżej 5 zaliczono do komentarzy pozytywnych. Recenzje z oceną 5 wykluczono z analizy z uwagi na trudność związaną z jednoznacznym przyporządkowaniem odpowiedniego wydźwięku wypowiedzi.

1.6.16. Zbiór *hotelezbil*

Zbiór powstał w ten sposób, że spośród 2335 dokumentów pozytywnych ze zbioru *hotele* wybrano losowo 1400 dokumentów. W efekcie otrzymano zbiór 1966 dokumentów, z których 935 to dokumenty pozytywne, zaś 1031 – negatywne.

1.7. Oprogramowanie używane w badaniach

Etapy wstępnej obróbki tekstu, trenowania klasyfikatorów, oceny klasyfikacji zostały przeprowadzone w całości przy użyciu języka i środowiska do obliczeń statystycznych *R*. Jest to darmowe narzędzie bardzo cenione przez statystyków

ze względu na mnogość zaimplementowanych metod statystycznych (udostępnianych w tzw. *pakietach*), umożliwiających realizację obliczeń oraz graficzną prezentację przy stosunkowo niewielkim nakładzie pracy. Język *R* może poszczycić się akademickim rodowodem, ponieważ prace nad nim zostały zapoczątkowane przez Rossa Ihakę oraz Roberta Gentlemana na Uniwersytecie w Auckland w Nowej Zelandii w 1991 roku. Pierwsza stabilna wersja beta 1.0 ujrzała światło dzienne 29 lutego 2000 roku.

1.7.1. Wstępna obróbka tekstu

Do wstępnego przetwarzania dokumentów wykorzystano pakiet *tm*[6], który jest frameworkiem dla zastosowań text miningowych, umożliwiającym wczytywanie dokumentów tekstowych z różnych źródeł, oczyszczanie oraz transformacje tekstu, tworzenie macierzy DTM wraz z podstawowymi schematami ważenia zmiennych. Dla zaznajomienia Czytelnika z możliwościami pakietu poniżej zamieszczono opis kluczowych funkcji oraz fragmenty kodów programów, które przyczyniły się do wykonania obliczeń w badaniach przedstawionych w niniejszej publikacji.

Pakiet *tm* oferuje funkcje ułatwiające modyfikację dokumentów tekstowych. Jedną z nich jest funkcja *tm_map*, która jako pierwszy argument przyjmuje korpus, w drugim można wyspecyfikować, o jaką operację (funkcję) modyfikującą tekst chodzi, a w kolejnych podać ewentualne argumenty potrzebne funkcji modyfikującej (por. przykład dla funkcji *stripWhitespace*). Do wyboru są m.in.:

- *tolower* – zmiana tekstu na małe litery;
- *removeNumbers* – usuwanie liczb;
- *removePunctuation* – usuwanie znaków interpunkcyjnych;
- *removeWords* – usuwanie słów, np. znajdujących się na stopliście – wówczas podajemy kolejny argument dodatkowy *words*, w którym przekazujemy słowa (jako wektor znakowy) do usunięcia (np. *words = nasza_stoplista*). Należy zwrócić uwagę, iż pakiet *tm* ma wbudowane stoplisty dla niektórych języków, np.: duńskiego, holenderskiego, angielskiego, fińskiego, francuskiego, niemieckiego, węgierskiego, włoskiego, norweskiego, portugalskiego, rosyjskiego, hiszpańskiego, szwedzkiego, katalońskiego, rumuńskiego. Dla języka polskiego niestety brak takiej listy, dlatego autorzy zastosowali stoplistę dostępną na Wikipedii[7], natomiast dla anglojęzycznych tekstów wykorzystano wbudowaną stoplistę;
- *stripWhitespace* – usuwanie pustych znaków (tzw. *spacji*).

6 Pakiet *tm*, https://CRAN.R-project.org/package=tm (dostęp: 20.09.2021).
7 Stoplista, https://pl.wikipedia.org/wiki/Wikipedia:Stopwords (dostęp: 20.09.2021).

Kod ilustrujący wykorzystanie funkcji *tm_map* na korpusie dokumentów zamieszczono poniżej:

```
#obiekt klasy VCorpus tworzony na podstawie wektora znakowego "dokumenty"
dokorpus <- VCorpus(VectorSource(dokumenty))

#przetwarzanie wstępne dokumentów przy pomocy funkcji tm_map
korpus <- tm_map(korpus, content_transformer(tolower))
korpus <- tm_map(korpus, content_transformer(removeNumbers))
korpus <- tm_map(korpus, content_transformer(removePunctuation))
korpus <- tm_map(korpus, removeWords, words=stopwordsPL)
korpus <- tm_map(korpus, content_transformer(stripWhitespace))

```

Funkcję *tm_map* zastosowano również przy lemmatyzacji wyrazów, bowiem dzięki niej można wywołać własną funkcję (podając jej definicję lub nazwę jako argument funkcji *tm_map*), modyfikującą zawartość dokumentów, uprzednio zdefiniowaną w języku *R*, służącą do zamiany wyrazów w całym korpusie na ich podstawowe wersje słownikowe. Autorzy zdecydowali się na lemmatyzację z wykorzystaniem rozbudowanego słownika dla języka polskiego. Logika działania własnej funkcji jest następująca. Przy pomocy pakietu *text2vec*[8] rozdzielono ciągi tekstów w poszczególnych dokumentach na pojedyncze wyrazy. W następnym kroku poszczególne wyrazy wyszukiwane są w słowniku – jeżeli zostaną w nim odszukane, następuje zamiana słowa występującego w dokumencie na formę podstawową pochodzącą ze słownika. W przypadku gdy dane słowo nie zostanie znalezione w słowniku, pozostaje ono w niezmienionej formie w dokumencie tekstowym. W rezultacie otrzymujemy kolekcję dokumentów tekstowych z możliwie wszystkimi słowami sprowadzonymi do ich form podstawowych. Powyższe kroki mogą być zrealizowane przy pomocy następującego programu:

```

#laduje słownik-hashmape z pliku
slownik <- load_hashmap(file="C:/Users/user/Desktop/text mining/kody/slownik")

#funkcja zamieniająca wyrazy na ich formy podstawowe ze słownika
lematyzator = function(x, Slownik, tokenizer = text2vec::word_tokenizer) {
  wyrazy = tokenizer(x)
  for(i in seq_along(wyrazy)) {
    wyraz = wyrazy[[i]]
    lemma_slownik = Slownik[[wyraz]]
    ind = !is.na(lemma_slownik)
    wyrazy[[i]][ind] = lemma_slownik[ind]
  }
  sapply(wyrazy, (function(i){paste(i, collapse = " ")}))
}

#wykorzystanie funkcji lematyzator
korpus <- tm_map(korpus, content_transformer(function(x) {lematyzator(x, Slownik=slownik)}))

```

Ostatnie dwa kroki wstępnej obróbki tekstu zrealizowano przy pomocy pakietów *RWeka*[9] oraz *tm*. Funkcja *NGramTokenizer*, pochodząca z pakietu *RWeka*, umożliwia utworzenie listy *n*-gramów na bazie przetworzonego we wcześniejszych

8 Pakiet *text2vec*, https://CRAN.R-project.org/package=text2vec (dostęp: 20.09.2021).
9 Pakiet *RWeka*, https://CRAN.R-project.org/package=RWeka (dostęp: 21.09.2021).

krokach korpusu (użytkownik ma możliwość wyboru unigramów, bigramów, trigramów etc.). Na bazie *n*-gramów funkcja *DocumentTermMatrix* (z pakietu *tm*) tworzy obiekt, który przechowuje informacje o macierzy DTM. Na etapie wywoływania powyższej funkcji można wyspecyfikować pożądany schemat ważenia, podając parametr *control = list(…, weighting = wybrany_schemat_ważenia)*, gdzie w miejscu *wybrany_schemat_ważenia* można podać m.in. *weightBin* (postać binarna), *weightTf*, (postać częstościowa), *weightTfIdf* (postać TFIDF). Omawiane kroki przedstawione w postaci kodu zawarto poniżej:

```
30
31 #funkcja tworząca unigramy
32 Tokenizer <- function(x) NGramTokenizer(x, Weka_control(min = 1, max = 1))
33 #Tokenizer <- function(x) NGramTokenizer(x, Weka_control(min = 1, max = 2))#bigramy
34
35 #tworzy macierz DTM
36 DTM <- DocumentTermMatrix(korpus, control = list(tokenize = Tokenizer)) #z częstościami wystąpień terminów
37 #DTM <- DocumentTermMatrix(korpus, control = list(tokenize = Tokenizer, weighting = weightTfIdf)) #schemat ważenia TFIDF
38
```

1.7.2. Klasyfikacja

Naiwny klasyfikator Bayesa rozważany w badaniach w niniejszym opracowaniu zaimplementowany jest w pakiecie *naivebayes*[10]. Omawiany klasyfikator wytrenować można przy pomocy funkcji *multinomial_naive_bayes*; argumenty, jakie trzeba podać, to *y* oraz *x*, oznaczające odpowiednio: wektor z etykietami klas oraz macierz zawierającą predyktory. Opcjonalnie można skorzystać m.in. z wygładzania Laplace'a, w tym celu w przeprowadzonych badaniach ustawiono parametr *laplace = 1*.

Drugi z omawianych klasyfikatorów, SVM, wytrenować można, wywołując funkcję *svm* z pakietu *e1071*[11], argumenty, jakie trzeba podać, to *formula* (w postaci *Y*~*X*1 *X*2 *X*3 …, gdzie *Y* to wektor z etykietami klas, zaś *X*1, *X*2, *X*3… to predyktory) oraz *data* (zbiór danych, np. macierz DTM). Pozostałe parametry są opcjonalne, pominięcie ich będzie skutkować wywołaniem procedury z domyślnymi wartościami tych parametrów – na takie rozwiązanie zdecydowano się, wykonując obliczenia do przedstawionych badań.

Parametry trzeciego i ostatniego klasyfikatora, regresji logistycznej, oszacowano przy pomocy funkcji *glm*, pochodzącej z pakietu *stats*. Oprócz parametrów *formula* oraz *data* specyfikowanych jak w przypadku wcześniej omawianej funkcji *svm*, należy podać, do jakiej rodziny rozkładów należy modelowane zjawisko, ustawiając parametr *family = "binominal"*.

Predykcję klasy na zbiorze treningowym przez wytrenowany klasyfikator ułatwia funkcja *predict*, do której należy wskazać nazwę modelu (nazwa obiektu, któ-

10 Pakiet *naivebayes*, https://CRAN.R-project.org/package=naivebayes (dostęp: 21.09.2021).
11 Pakiet *e1071*, https://CRAN.R-project.org/package=e1071 (dostęp: 21.09.2021).

ry został stworzony przy pomocy funkcji *naiveBayes*, *svm*, *glm*) oraz dane testowe. Należy pamiętać, aby nazwy predyktorów oraz zmiennej objaśnianej były takie same, jak w zbiorze treningowym.

Uczenie wybranych klasyfikatorów na zbiorze treningowym oraz przeprowadzenie klasyfikacji na zbiorze testowym można przeprowadzić przy pomocy poniższego fragmentu programu:

```
39
40 #na potrzeby dalszego przetwarzania utworzony zostaje obiekt typu data frame na podstawie obiektu DTM
41 zbior_treningowy <- data.matrix(DTM)
42 zbior_treningowy <- as.data.frame(zbior_treningowy)
43 #do obiektu zbior_df dodany jest wektor przechowujący informację o wydźwięku dokumentu
44 zbior_treningowy$Ocena_label <- Ocena
45
46 #nainwy klasyfikator Bayesa
47 classifier_mnb = multinomial_naive_bayes(y=zbior_treningowy$Ocena_label, x=zbior_treningowy[,-ncol(zbior_treningowy)], laplace=1)
48 y_pred_mnb = predict(classifier_mnb, newdata = as.matrix(zbior_testowy))
49
50 #regresja logistyczna
51 classifier_glm = glm(formula = Ocena_label~., data = zbior_treningowy, family = "binomial")
52 y_pred_glm = factor(ifelse(predict(classifier_glm, newdata = zbior_testowy, type="response")>0.5,"Pozytywna","Negatywna"))
53
54 #metoda wektorów nośnych
55 classifier_svm = svm(formula = Ocena_label ~ ., data = zbior_treningowy)
56 y_pred_svm = predict(classifier_svm, newdata = zbior_testowy)
57
```

Rozdział 2

Metody doboru zmiennych na potrzeby klasyfikacji tekstów

2.1. Podejścia modelowe

W tym podrozdziale przedstawimy przykłady kilku metod badania sentymentu opartych na modelach probabilistycznych. Eyheramendy i Madigan (2007) zaproponowali podejście bayesowskie o nazwie PIP (*posterior inclusion probability*), w którym związek terminu z klasą wyrazili za pomocą prawdopodobieństwa a posteriori dla naiwnego modelu Bayesa. Miara PIP dana jest wzorem, w którym termin musi być numerowany ze względu na konieczność opisania wszystkich terminów za pomocą jednego wektora o M współrzędnych:

$$PIP(f_j, C_k) = \sum_{l:l_j=1} P(M_l|data), \tag{2.1}$$

gdzie $\boldsymbol{l}$ jest zero-jedynkowym wektorem o długości M (liczba wszystkich terminów), w którym na j-tym miejscu jest 1, jeśli termin f_j występuje w modelu, oraz 0 w przeciwnym razie. Wszystkich modeli jest 2^M, bo jest M terminów. Innymi słowy, miara PIP to prawdopodobieństwo a posteriori tego, że każdy termin pojawi się w modelu, dla wszystkich terminów występujących w dokumentach klasy C_k. Autorom udało się wyprowadzić wzór na prawdopodobieństwo a posteriori, gdy prawdopodobieństwa warunkowe z klasyfikatora naiwnego Bayesa mają rozkład Bernoulli'ego (lub Poissona) z parametrem o rozkładzie a priori beta (lub gamma dla rozkładu Poissona). Na przykład dla rozkładu Bernoulli'ego miara PIP ma wzór:

$$PIP(f_j, C_k) = \frac{l_{0,f,k}}{l_{0,f,k} + l_{f,k}}, \tag{2.2}$$

gdzie jednak wyrażenia występujące po prawej stronie mają dość dowolnie dobierane wartości przez, jak piszą autorzy, „praktyka". I tak:

$$l_{0,f,k} = \frac{B(n_{k,f} + a_{k,f}, n_{k,\bar{f}} + b_{k,f})}{B(a_{k,f}, b_{k,f})} \frac{B(n_{\bar{k},f} + a_{\bar{k},f}, n_{\bar{k},\bar{f}} + b_{\bar{k},f})}{B(a_{\bar{k},f}, b_{\bar{k},f})}, \tag{2.3}$$

$$l_{f,k} = \frac{B(n_f + a_f, n_{\bar{f}} + b_f)}{B(a_f, b_f)}, \tag{2.4}$$

gdzie na przykład $a_f = 0{,}2$ $b_f = 2/25$ dla dowolnego terminu f (B oznacza funkcje beta). Dla tak dobranych przez siebie wartości parametrów porównano miarę daną wzorem (2.2) z sześcioma innymi metodami selekcji terminów dla czterech popularnych klasyfikatorów, na zbiorach *Reuters* oraz *Newsgroups*. Dla małej liczby terminów miara PIP wypadła gorzej od kilku innych. Trochę lepiej spisała się przy nieco większej liczbie terminów (około 100). Przy kilkuset terminach efektywność wszystkich miar stabilizowała się i była bardzo podobna. Należy nadmienić, że miara PIP ma tendencję do faworyzowania terminów występujących często, wobec czego autorzy sztucznie wyeliminowali np. 15 najczęstszych terminów ze zbioru *Reuters*. Nie uzyskano wnioskowania statystycznego o klasyfikacji dokumentów.

Davies i Ghahramani (2011) zaproponowali modelowanie probabilistyczne z wykorzystaniem rozkładów prawdopodobieństwa danego słowa dla danego sentymentu opartych na słowach kluczowych, w roli których wystąpiły emotikony. Autorzy modelują prawdopodobieństwo dowolnego tweeta $t_i = (w_{i,1}, ..., w_{i,n_i})$, gdzie n_i jest liczbą słów w tweecie t_i, przy danych rozkładach prawdopodobieństwa słów w_i za pomocą rozkładu:

$$P(t|\vec{\theta}) = \sum_{s \in S} P(t|\vec{\theta}_s) P(s) = \sum_{s \in S} \prod_{w \in t} P(w|\vec{\theta}_s) P(s) = \sum_{s \in S} P(s) \prod_{w \in t} P(w|\vec{\theta}_s), \tag{2.5}$$

gdzie $\vec{\theta}_s$ jest wielowymiarowym rozkładem prawdopodobieństwa słów dla ustalonego sentymentu s (możliwe tylko dwie wartości sentymentu: *happy* albo *sad*). Rozkład a priori $\vec{\theta}_s$ jest rozkładem asymetrycznym Dirichleta, który oznaczamy jako $Dir(\vec{\alpha}_s)$, gdzie $\vec{\alpha}_s$ oznacza zbiór słów kluczowych dla sentymentu s (są to emotikony). Jeśli prawdopodobieństwo sentymentu $P(s)$ oszacujemy empirycznie, to prognozę sentymentu tweeta t, na podstawie twierdzenia Bayesa, można wyrazić wzorem:

$$P(s|t) = \frac{exp(log(P(t|\overrightarrow{\alpha'}_s)))}{\sum_{s' \in S} exp(log(P(t|\overrightarrow{\alpha'}_{s'})))} P(s), \tag{2.6}$$

gdzie $\overrightarrow{\alpha'}_s$ jest rozkładem a posteriori $Dir(\overrightarrow{\alpha'}_s)$ sentymentu. Przybliżony wzór na $log(P(t|\overrightarrow{\alpha'}_s))$ został wyprowadzony. Autorzy twierdzą, że na zbadanym zbiorze tweetów ich model ma precyzję przewyższającą klasyfikator naiwny Bayesa o około 10%. Cechą oryginalną tego modelu jest to, że autorzy podają ujemne zlogarytmowane prawdopodobieństwo, które jest miarą tego, jak dobrze model

jest dopasowany do danych. Co ciekawe, przy lepszych wskazaniach tego zlogarytmowanego prawdopodobieństwa były słabsze wskazania trafności klasyfikacji.

Genkin i inni (2007) zaproponowali model bayesowski w postaci:

$$P(y=+1|\boldsymbol{\beta},\boldsymbol{x}_i)=\psi(\boldsymbol{\beta}^T\boldsymbol{x}_i)=\psi\left(\sum_j \beta_j x_{ij}\right), \tag{2.7}$$

gdzie w roli ψ użyto funkcji logistycznej,

$$\psi(t)=\frac{1+\exp(t)}{\exp(t)}, \tag{2.8}$$

która pod względem obliczeniowym nie nastręcza takich problemów, jak estymacja metodą największej wiarygodności. Prawdopodobieństwo tego, że dokument zostanie zaliczony do klasy dokumentów pozytywnych w takim modelu, to: $P(y = +1|\boldsymbol{x}_i)$. Wektor $\boldsymbol{x}_i = [x_{i,1}, ..., x_{i,N}]^T$ składa się z częstości występowania terminu f_i w poszczególnych dokumentach. Jako rozkład a priori parametru β_j zastosowano rozkład normalny $N(0, \tau_i)$, który jest odporny na nadmierne dopasowanie modelu do danych i stosunkowo prosty obliczeniowo. Wartość oczekiwana równa 0 może być interpretowana jako założenie o tym, że każdy pojedynczy termin ma niewielki wpływ na efekt klasyfikacji. Z kolei τ_i należy zadać, w najprostszym przypadku są one jednakowe dla wszystkich terminów. Oszacowanie a posteriori parametru $\boldsymbol{\beta}$ ustalono (między innymi) za pomocą algorytmu CLG (*cyclic coordinate descent*, Zhang i Oles 2001) zmodyfikowanego dla metody lasso regresji logistycznej. Efektywność modelu zbadano na zbiorach *Reuters* i *OHSUMED*, przy czym dokumenty reprezentowane były w postaci TFxIDF z normalizacją cosinusową. Model regresji logistycznej proponowany przez autorów na ogół był lepszy (choć niekiedy jedynie o około 1%) od klasyfikacji metodą SVM z selekcją zmiennych metodą χ^2 lub BNS (por. podrozdział 3.1).

Podsumowując przedstawione przykłady podejść modelowych, należy zauważyć, że jest w nich wykorzystywany skomplikowany aparat matematyczny, dość dowolny dobór rozkładów a priori, metod estymacji parametrów modeli, natomiast uzyskane efekty są umiarkowanej jakości.

2.2. Podejścia heurystyczne

Ta część opracowania będzie poświęcona metodom selekcji cech-terminów, co stanowi pierwszy z dwóch podstawowych etapów badania sentymentu dokumentów tekstowych. Drugim etapem jest klasyfikacja dokumentów przy użyciu

wybranego klasyfikatora na podstawie terminów wybranych w pierwszym etapie. Siłą rzeczy niemożliwe jest ocenianie metod selekcji zmiennych bez odwołania się do użycia klasyfikatora. Na ogół autorzy oceniają swoje metody selekcji zmiennych, posługując się popularnymi klasyfikatorami takimi jak: klasyfikator Bayesa, *k*-najbliższego sąsiada, klasyfikacja za pomocą metody wektorów nośnych.

Popularną grupą metod selekcji cech są metody wykorzystujące wzajemną informację (*mutual information*). Dla zbioru cech $F = \{f_1, f_2 \ ..., f_M\}$ i cechy C etykiet klas możemy mierzyć związek pomiędzy cechą f_m a cechą C za pomocą następującej relacji pomiędzy rozkładem łącznym $P(f_m, C)$ a rozkładami brzegowymi $P(f_m)$ oraz $P(C)$:

$$A = MI(f_m, C) = \sum_i P(f_{m,i}, C) log \frac{P(f_{m,i}, C)}{P(f_{m,i})P(C)}, \tag{2.9}$$

gdzie sumowanie przebiega po wszystkich możliwych wartościach $f_{m,i}$ cechy f_m. Podobny wzór można zastosować do mierzenia korelacji pomiędzy dwiema różnymi cechami. Wtedy otrzymamy wielkość, która może zostać użyta jako miara zbędności (*redundancy*) cech (im silniejsza korelacja pomiędzy cechami, tym jedna z nich mniej przydatna w klasyfikacji). Dla przypadku cech dyskretnych otrzymamy wzór:

$$B = MI(f, g) = \sum_{i,j} P(f_i, g_j) log \frac{P(f_i, g_j)}{P(f_i)P(g_j)}, \tag{2.10}$$

gdzie sumowanie przebiega po wszystkich możliwych wartościach obu cech. W celu wyselekcjonowania cech, które są jak najsilniej związane z cechą etykiet i jak najsłabiej powiązane między sobą, można stosować różne kryteria, na przykład różnicowe, czyli maksymalizację wielkości *A*–*B* lub ilorazowe, czyli maksymalizację wielkości *A*/*B*. Wykorzystując powyższe wzory, można proponować metody selekcji zmiennych oparte na tej idei – oznaczmy te metody symbolem mRMR (*minimum redundancy maximum relevance*).

Zysk informacji (*information gain*) dla terminu *f* definiujemy jako:

$$\begin{aligned} IG(f) = -\sum_k P(C_k) log(P(C_k)) + P(f) \sum_k P(C_k|f) log(P(C_k|f)) + \\ + P(\overline{f}) \sum_k P(C_k|\overline{f}) log(P(C_k|\overline{f})), \end{aligned} \tag{2.11}$$

gdzie $P(C_k|f)$ (odpowiednio, $P(C_k|\overline{f})$ to odsetek dokumentów z klasy C_k zawierających termin *f* (odpowiednio, niezawierających terminu *f*). Pierwszy składnik powyższego wzoru to entropia dla całego zbioru uczącego dokumentów (por. wzór 2.49). Od entropii dla całego zbioru odejmowana jest entropia dla terminu. Wzór ten możemy więc interpretować jako spodziewany spadek entropii

po podzieleniu zbioru dokumentów w zależności od terminu. Zysk informacji jest nazywany niekiedy wzajemną informacją, ponieważ może być zapisany jako suma wzajemnych informacji pomiędzy cechami a klasami danymi wzorem 2.9 ważonymi prawdopodobieństwami tego, że klasa zawiera cechę (czyli wagami są częstości dokumentów z danej klasy i zawierających daną cechę).

Jeżeli będziemy chcieli zastosować powyższy wzór do selekcji cech, to można użyć go do uporządkowania cech w kolejności od najistotniejszej dla klasyfikacji (tej z najwyższym zyskiem informacji) do najmniej istotnej. Na ogół cechy f, które mają $IG(f) > 0$, uważa się za istotne dla klasyfikacji. Metody selekcji cech oparte na zysku informacji będziemy oznaczać symbolem IG. Metody takie mogą mieć różne postacie w zależności od postaci wyjściowego zbioru cech zdeterminowanej przez sposób opisu tekstu. Agarwal i Mittal (2016) przeprowadzili dość obszerne porównawcze badanie efektywności metod mRMR oraz IG selekcji cech dla kilku powszechnie stosowanych klasyfikatorów, a także dla kilku różnorodnych zbiorów tekstów. Wśród cech oryginalnych najlepiej spisał się zbiór cech opartych na unigramach, co potwierdza wyniki większości innych badań. Wszystkie metody selekcji cech poprawiały efektywność klasyfikacji metod opartych na wyjściowych zbiorach cech dla wszystkich zbiorów dokumentów. Metody typu mRMR, na ogół, były lepsze od metod IG, co świadczy o tym, że samo podkreślenie znaczenia dla poprawnej klasyfikacji nie daje tak dobrych efektów, jak podkreślenie znaczenia dla poprawnej klasyfikacji wraz z jednoczesnym obniżeniem znaczenia cech zbędnych.

Znalezienie wzajemnej informacji dla wszystkich podzbiorów zbioru wszystkich terminów jest obliczeniowo niewykonalne dla dużych lub nawet średnich rozmiarów zbiorów dokumentów, dlatego Battiti (1994) zaproponował chciwy (*greedy*) algorytm, trochę upraszczający obciążenia obliczeniowe. Algorytm został nazwany MIFS (*mutual information for feature selection*) i działa następująco.

Krok 1. Znajdujemy $MI(f, C)$ dla każdego terminu f i wybieramy termin f_0 z największą wartością wzajemnej informacji. Wybrany termin dołączamy do zbioru S terminów wyselekcjonowanych i usuwamy ze zbioru F.

Krok 2. Dla każdego terminu $g \in S$ znajdujemy $MI(f, g)$ dla wszystkich par (f, g), dla których $f \in F$.

Krok 3. Dołączamy do zbioru S (i usuwamy ze zbioru F) ten termin f, który maksymalizuje:

$$MI(f, C) - \beta \sum_{g \in S} MI(f, g). \tag{2.12}$$

Krok 4. Powtarzamy krok 2 oraz krok 3 do momentu wyselekcjonowania k terminów.

W powyższym algorytmie MIFS parametr β „reguluje ważność" kandydata do selekcji w stosunku do wzajemniej informacji pomiędzy tym terminem a całą klasą. Na ogół (por. Battiti 1994) wartość parametru β jest z przedziału [0,5; 1].

Kwak i Choi (2002) zaproponowali modyfikację algorytmu MIFS, która zmniejsza obciążenia obliczeniowe związane ze znajdywaniem wszystkich wzajemnych informacji z kroku drugiego. Inne metody wykorzystujące wzajemną informację można znaleźć m.in. w takich pracach, jak: Bagheri i inni (2013), Chen i inni (2013), Jiang i Shui (2013), Ding i Tang (2013), Gündüz i Çataltepe (2015), Lifang i inni (2017).

Prawdopodobieństwa warunkowe $P(C_k|f)$ są często „zbyt silne" dla wielu terminów (bo równe 0 lub 1), wobec tego Ong i inni (2015) zaproponowali dla przypadku dwóch klas poprawkę, która ma za zadanie złagodzić wpływ na zysk informacji takich skrajnych prawdopodobieństw (*sparsity adjusted information gain*). Prawdopodobieństwa warunkowe $P(C_k|f)$ obliczane były według formuł:

$$P(C_1|f) = \frac{A}{A+B}, \qquad P(C_2|f) = \frac{B}{A+B}, \tag{2.13}$$

gdzie wyrażenia A, B obliczane są ze wzorów:

$$A = \frac{n_{1,f}}{n_f} \cdot \frac{m_{1,f}}{m_f}, \qquad B = \frac{n_{2,f}}{n_f} \cdot \frac{m_{2,f}}{m_f} \tag{2.14}$$

lub $A = 0{,}5/n_{2,f}$ i $B = m_{2,f}$, gdy $n_{1,f} = 0$ oraz $B = 0{,}5/n_{1,f}$ i $A = m_{1,f}$, gdy $n_{2,f} = 0$. W dalszym ciągu metoda selekcji terminów działa tradycyjnie, tzn. porządkujemy terminy według malejących wartości IG (por. wzór 2.11) i wybieramy ustaloną liczbę początkowych terminów. Autorzy zbadali swoją propozycję na części zbioru Amazon, przyjmując za próg rzadkości występowania terminów 0,01 dla trzech popularnych klasyfikatorów. Najlepszą poprawę precyzji klasyfikacji (około 5–6%) uzyskali dla klasyfikatora SVM dla kilkudziesięciu terminów początkowych. Dla pozostałych klasyfikatorów poprawa była dużo mniejsza lub nie było jej wcale.

Dalsze modyfikacje metod opartych na wzajemnej informacji oraz zysku informacji można znaleźć w pracach: Malik i Novovicova (2005), Bakus i Kamel (2006), Chen i inni (2013), Patil i Atique (2013), Zhang i inni (2013), Gao i inni (2014), Jiang i Yu (2015), Wu i Xu (2015), Zhu i inni (2017), Rastogi (2018). Wyniki, które przytaczają autorzy, świadczą o dość istotnym przyspieszeniu metod podstawowych, natomiast zysk efektywności jest niewielki i wynosi około 2%.

Dość często pojawiającym się pojęciem jest orientacja semantyczna (*semantic orientation*) cechy (terminu). To pojęcie jest związane z wzajemną informacją pomiędzy terminem a cechą etykiet (por. wzór 2.9). Orientację semantyczną rozumiemy w tym sensie, że terminy pojawiające się często w dokumentach pozytywnych kojarzone są z orientacją pozytywną, zaś te pojawiające się często w dokumentach negatywnych mają orientację negatywną. Wzajemna informacja punktowa (*pointwise mutual information*) pomiędzy terminem a pojedynczą kla-

są to część składników wzoru 2.9, którą możemy zapisać w zależności od rodzaju orientacji (czyli klasy) następująco:

$$PMI(f, pos) = log_2 \frac{P(f, pos)}{P(f)P(pos)}, \tag{2.15}$$

$$PMI(f, neg) = log_2 \frac{P(f, neg)}{P(f)P(neg)}, \tag{2.16}$$

gdzie:

$P(f, pos)$ – częstość występowania terminu f w dokumentach pozytywnych, tj. liczba wszystkich dokumentów pozytywnych zawierających termin f podzielona przez liczbę wszystkich dokumentów pozytywnych,

$P(f, neg)$ – częstość występowania terminu f w dokumentach negatywnych, tj. liczba wszystkich dokumentów negatywnych zawierających termin f podzielona przez liczbę wszystkich dokumentów negatywnych.

Wzajemna informacja punktowa jest pomocna w zdefiniowaniu orientacji semantycznej terminu:

$$SO(f) = PMI(f, pos) - PMI(f, neg). \tag{2.17}$$

Abbasi i inni (2008) zaproponowali metodę będącą połączeniem zysku informacji IG (wzór 2.11) z algorytmem genetycznym (zob. Holland 1975). Zysk informacji dla każdego terminu jest stosowany w postaci wagi tego terminu. Następnie jest stosowany algorytm genetyczny w swej standardowej postaci, w którym w kroku krzyżówki mieszane są dwa zbiory terminów. Zgodnie z naczelną ideą algorytmu genetycznego polegającą na tym, by krzyżować osobniki różnorodne, zaproponowany algorytm wybiera do krzyżowania terminy z wysoką wartością IG oraz te z niską wartością. Taka selekcja nie ma charakteru „najszybszego spadku", ale zapewnia łagodne dochodzenie do rozwiązania optymalnego, dzięki czemu nie grzęźnie szybko w maksimach lokalnych. Swój algorytm autorzy nazwali EWGA (*Entropy Weighted Genetic Algorithm*), jako że zysk informacji IG (wzór 2.11) można przedstawić jako różnicę dwóch entropii (por. wzór 2.49). Swoją metodę selekcji cech zbadali w połączeniu z klasyfikatorem SVM na zbiorze *Polarity*. Uzyskali nieznacznie lepszą precyzję klasyfikacji równą 91,70% wobec 87,95% dla podstawowego zbioru terminów, 89,85% dla zbioru zmodyfikowanego za pomocą zysku informacji IG, 90,05% dla zbioru zmodyfikowanego za pomocą algorytmu genetycznego i 90,20% dla zbioru zmodyfikowanego za pomocą wag z metody SVM.

Najprostszą miarą charakteryzującą terminy występujące w dokumentach jest częstość słowa (*word frequency*). Miarę tę definiuje się jako średnią liczebności $n_{f,k}$ dokumentów zawierających termin f ważoną liczebnościami względnymi klas:

$$WF(f) = \sum_k \frac{n_k}{N} n_{f,k} \tag{2.18}$$

W początkowym okresie rozwoju klasyfikacji tekstów ze względu na opinię miara ta była wykorzystywana przez kilku badaczy, ale metody oparte tylko na niej nie uzyskały dobrych wyników.

Inną prostą miarą ważności terminów jest iloraz szans (*odds ratio*) porównujący szanse na to, że termin pojawi się w danej klasie z szansami na to, że pojawi się w innej klasie. Jeżeli termin pojawia się częściej w danej klasie, to iloraz szans przyjmuje wartości dodatnie. W przeciwnym przypadku przyjmuje wartości ujemne. Miarę tę definiuje się następująco:

$$OR(f, C_k) = log \frac{P(f|C_k)(1-P(f|\overline{C_k}))}{P(f|\overline{C_k})(1-P(f|C_k))}. \tag{2.19}$$

W przypadku binarnym można ten wzór (zakładając, że klasą odnośną C_k są dokumenty pozytywne) zapisać w postaci:

$$OR(f) = log \frac{P(f|pos)(1-P(f|neg))}{P(f|neg)(1-P(f|pos))}. \tag{2.20}$$

Ghareb i inni (2018) zmodyfikowali miarę OR za pomocą wzoru:

$$OR_{mod}(f) = log \frac{m_{f,pos} + (1-m_{f,neg})}{m_{f,neg} + (1-m_{f,pos})}, \tag{2.21}$$

opierając ją na częstościach terminów, a nie dokumentów. Ponadto mnożenie zamieniono na dodawanie, co pozwala uniknąć zera w mianowniku. Autorzy zbadali tę propozycję tylko na zbiorze tekstów arabskich, uzyskując wyniki lepsze o kilka punktów procentowych od ilorazu szans.

Na bazie ilorazu szans można definiować globalne (niezależne od klasy) miary ważności cechy. Pierwszym przykładem może być poszerzony iloraz szans EOR (*extended odds ratio*):

$$EOR(f) = \sum_k OR(f, C_k). \tag{2.22}$$

Innym przykładem takiej miary może być ważony iloraz szans WOR (*weighted odds ratio*):

$$WOR(f) = \sum_k \frac{n_k}{N} OR(f, C_k). \tag{2.23}$$

Miary EOR oraz WOR preferują terminy o wydźwięku pozytywnym, a zatem w przypadku kategoryzacji wieloklasowej będą spisywały się słabo, gdyż w takiej

sytuacji istotne dla klasyfikacji są również terminy negatywne. Wobec tego Chen i inni (2009) zaproponowali miarę MOR, która miała złagodzić tę wadę:

$$MOR(f) = \sum_k \left| OR(f, C_k) \right| = \sum_k \left| log \frac{P(f|C_k)(1 - P(f|\overline{C_k}))}{P(f|\overline{C_k})(1 - P(f|C_k))} \right|. \tag{2.24}$$

Jeśli jednak MOR zapiszemy w postaci:

$$MOR(f) = \sum_k \left| log \frac{P(f|C_k)}{P(f|\overline{C_k})} + log \frac{(1 - P(f|\overline{C_k}))}{(1 - P(f|C_k))} \right|, \tag{2.25}$$

to łatwo zobaczyć, że drugi składnik ma za zadanie podkreślić różnicę pomiędzy $P(f|C_k)$ a $P(f|\overline{C_k})$, czyli takie samo zadanie, jak składnik pierwszy. Wobec tego możemy pominąć drugi składnik i otrzymamy miarę dyskryminacji klas CDM (*class discriminant measure*) (por. Chen i inni 2009) daną wzorem:

$$CDM(f) = \sum_k \left| log \frac{P(f|C_k)}{P(f|\overline{C_k})} \right|. \tag{2.26}$$

W publikacji źródłowej miara CDM była zaprojektowana pod kątem problemu kategoryzacji wieloklasowej przy zastosowaniu naiwnego klasyfikatora bayesowskiego i, jak twierdzą autorzy, spisała się lepiej od zysku informacji IG, jak również MOR czy WOR. Badania przeprowadzono na zbiorze dokumentów *Reuters*. Jednak precyzja dla metody CDM była wyższa od precyzji dla IG tylko o około 1–2% w zależności od liczby początkowych terminów na zbiorze *Reuters*. Z kolei na zbiorze Ronglu Li zwyciężyła metoda MOR.

Ghareb i inni (2018) zmodyfikowali miarę CDM za pomocą wzoru:

$$MCDM(f) = \sum_k \frac{P(f|C_k) \cdot m_{f,k} + 1}{P(f|\overline{C_k}) + 1}, \tag{2.27}$$

gdzie mnożenie przez $m_{f,k}$ ma za zadanie wykorzystanie informacji o liczbie wystąpień terminów w klasach, zaś dodanie 1 (metoda wygładzania Laplace'a) pozwala uniknąć zera w mianowniku (co dawałoby bardzo duże wartości miar CDM terminom niewystępującym w jednej z klas). Autorzy zbadali tę propozycję tylko na zbiorze tekstów arabskich, uzyskując wyniki lepsze o kilka punktów procentowych od miary CDM.

W analizie sentymentu statystyka χ^2 może być stosowana jako metoda selekcji cech. Statystyka ta mierzy siłę zależności pomiędzy dowolną cechą a cechą etykiet klas. Ważność cechy dla klasyfikacji rośnie wraz ze wzrostem wartości statystyki χ^2. Statystykę tę definiujemy następująco:

$$\tilde{\chi}^2(f) = \sum_k \frac{n_k}{N} \cdot \chi^2(f, C_k) = \sum_k P(C_k) \cdot \chi^2(f, C_k), \tag{2.28}$$

gdzie $\chi^2(f, C_k)$ jest tak obliczane, jak dla tablicy kontyngencyjnej $2 \times k$. W przypadku dwóch klas dokumentów to wyrażenie jest dane wzorem:

$$\chi^2(f, C_{pos}) = \frac{N(ad - cb)^2}{(a+b)(a+c)(c+d)(b+d)}, \tag{2.29}$$

gdzie liczebności a, b, c, d są tak rozumiane, jak w tabeli:

	C_k	$\overline{C}_k$
f	a	b
$\overline{f}$	c	d

Miara NGL (skrót od nazwisk autorów) (Ng i inni 1997) jest lokalnym wariantem miary $\tilde{\chi}^2(f)$ danym wzorem:

$$NGL(f, C_k) = \sum_k \frac{\sqrt{N}(n_{f,k} n_{\overline{f},\overline{k}} - n_{f,\overline{k}} n_{\overline{f},k})}{\sqrt{n_f n_{\overline{f}} n_k n_{\overline{k}}}}. \tag{2.30}$$

Miarę lokalną (wzór 2.30) można uogólnić do miary globalnej za pomocą na przykład średniej ważonej względem klas. W publikacji źródłowej miara ta została zbadana na zbiorze dokumentów *Reuters-21578* i spisała się lepiej od miary CHI. W pracy Ruiz i Srinivasan (1999) miara (wzór 2.30) na niektórych liczbach terminów okazała się lepsza od ilorazu szans oraz wzajemnej informacji, na innych liczbach okazała się jednak gorsza. Godne uwagi jest to, że w tym badaniu zbiorem dokumentów był zbiór *OHSUMED*, który jest bardzo dużym zbiorem (348 543 dokumentów), z którym inne metody sobie nie radzą. Inne metody związane z wykorzystaniem statystyki χ^2 można znaleźć w pracach: Fukumoto i Suzuki (2015), Agnihotri i inni (2016), Bahassine i inni (2016), Sun i inni (2017), Bahassine i inni (2018).

Miara lokalna GSS (skrót od nazwisk autorów) (Galavotti i inni 2000) jest dana wzorem:

$$GSS(f, C_k) = n_{f,k} n_{\overline{f},\overline{k}} - n_{f,\overline{k}} n_{\overline{f},k}. \tag{2.31}$$

Najlepsze rezultaty autorzy osiągnęli dla wersji:

$$GSS(f) = \max_k GSS(f, C_k). \tag{2.32}$$

W badaniu przeprowadzonym na zbiorze *Reuters-21578* dla klasyfikatora *k*-NN autorzy osiągnęli nieznacznie lepszą precyzję od selekcji zrobionej przy użyciu miary $\tilde{\chi}^2(f)$.

Forman (2003) zaproponował miarę BNS (*Bi-Normal Separation*) opartą na założeniu, że pojawianie się terminu można modelować rozkładem normalnym, który przekracza pewien próg. Wtedy związek terminu charakterystycznego dla klasy z tą klasą powinien skutkować większą odległością tego progu od ogona rozkładu niż odległości innych progów. Wobec tego miarę tę zdefiniowano następująco:

$$BNS(f, C_k) = \left| F^{-1}\left(\frac{n_{f,k}}{n_k}\right) - F^{-1}\left(\frac{n_{f,\bar{k}}}{n_{\bar{k}}}\right)\right|, \tag{2.33}$$

gdzie F jest dystrybuantą rozkładu normalnego standaryzowanego. W celu uniknięcia nieciągłości na krańcach przedziału [0; 1] Foreman proponuje zastąpić 0 przez 5/10 000 oraz 1 przez 1-5/10 000. Najefektywniejszą miarą globalną dla cech okazał się wariant sumy ważonej:

$$BNS(f) = \sum_k \frac{n_k}{N} BNS(f, C_k). \tag{2.34}$$

W obszernym badaniu na wielu zbiorach danych (zbiory m.in. *Reuters-21578* i *OHSUMED*) miara BNS okazała się minimalnie lepsza od miary CHI, IG oraz ilorazu szans dla dużej liczby terminów. Dla małej liczby terminów (poniżej 50), na ogół, miara ta przegrywała z miarą IG.

Simeon i Hilderman (2008) zastosowali prostą różnicę proporcji kategorii (*categorical proportion difference*):

$$CPD(f, C_k) = \frac{n_{f,k} - n_{f,\bar{k}}}{n_k} \tag{2.35}$$

i jej globalną wersję:

$$CPD(f) = \max_k CPD(f, C_k). \tag{2.36}$$

Miara $CPD(f)$ przyjmuje wartości z przedziału [0; 1]. Jeśli wartość jest bliska 0, to oznacza to tyle, że termin powinien być nieprzydatny dla klasyfikacji dokumentów (bo jednakowo często pojawia się zarówno w klasie pozytywnej, jak i negatywnej). Gdy wielkość $CPD(f)$ jest bliska 1, to ma to miejsce tylko wtedy, gdy termin f występuje głównie tylko w jednej klasie i taki termin powinien być ważny dla klasyfikacji. Metoda selekcji wykorzystująca $CPD(f)$ polega na tym, że porządkujemy wszystkie terminy f według malejących wartości $CPD(f)$. W artykule źródłowym zbadano efektywność miary CPD w bardzo obszernym badaniu (zbiory *OHSUMED*, *20 Newsgroups*, *Reuters-21578*). Miara CPD uzyskała dobre wyniki, lepsze od miary IG, ilorazu szans i kilku innych, jednak potrzebowała od 60% do 70% wszystkich terminów.

Garnes (2009) zbadał miary CHI, GSS, EOR, IG, NGL, MI i BNS (oraz inne) w bardzo obszernym badaniu na zbiorach *Reuters*, *20 Newsgroups* oraz *OHSUMED*

dla kilku wybranych popularnych klasyfikatorów. Wśród klasyfikatorów, na ogół, metoda wektorów nośnych SVM była lepsza od klasyfikatora bayesowskiego, przy czym oba klasyfikatory zachowywały się bardzo podobnie względem metod selekcji zmiennych. Miara BNS okazała się dobra tylko dla bardzo małych liczebności wyselekcjonowanych cech, tzn. przy bardzo wysokim stopniu agresywności selekcji. Ogólnie we wszystkich eksperymentach najlepsze okazały się metody CHI, EOR oraz GSS. Metoda NGL nie potwierdziła dobrych wyników z innych badań, gdyż wypadła bardzo słabo.

Agarwal i Mittal (2012) zastosowali do ustalania istotności terminów różnicę proporcji prawdopodobieństw (*probability proportion difference*):

$$PPD(f, C_k) = \frac{n_{f,k}}{N + n_k} - \frac{n_{f,\bar{k}}}{N + n_{\bar{k}}}. \tag{2.37}$$

Idea tej miary polega na tym, że jeśli jakiś termin *w* ma wysokie prawdopodobieństwo występowania w jakiejś klasie sentymentu, to powinien być on istotny dla klasyfikacji dokumentów. Gdy zaś termin *w* ma jednakowo wysokie prawdopodobieństwo występowania w obu klasach sentymentu, to nie powinien on być istotny dla klasyfikacji. Metoda selekcji wykorzystująca PPD polega na tym, że porządkujemy wszystkie terminy według malejących wartości bezwzględnych $PPD(f)$.

Agarwal i Mittal (2012) połączyli obie miary $CPD(f)$ oraz $PPD(f)$ w ten sposób, że do uporządkowanego ciągu terminów włączyli tylko te, dla których obie miary przyjmują wartości powyżej wcześniej ustalonych progów. Autorzy nie podali jednak, jak te progi ustalać. Porównano efektywność wszystkich trzech metod (tj. $CPD(f)$, $PPD(f)$ oraz $CPD(f) + PPD(f)$) oraz metody *IG* opartej na zysku informacji z czystą reprezentacją unigramową terminów na zbiorze *Cornell Movie Review Dataset* oraz na zbiorze *Amazon reviews* dla dwóch popularnych klasyfikatorów. Najlepiej spisała się metoda $CPD(f) + PPD(f)$, jednak dla klasyfikatora SVM uzyskała tylko o około 1–2% wyższą wartość miary *F* od metody IG.

Cai i Song (2008) zaproponowali do ustalania istotności terminu miarę różnicy zliczeń (*count difference*), którą zdefiniowali następująco:

$$CD(f, C_k) = \left(\frac{n_{f,k}}{\overline{n}_k} - \frac{n_{f,\bar{k}}}{\overline{n}_{\bar{k}}} \right)^2, \tag{2.38}$$

gdzie $\overline{n}_k$ i $\overline{n}_{\bar{k}}$ są średnimi arytmetycznymi liczby wystąpień terminów *w*, odpowiednio, w klasie *k* i poza nią, tzn.:

$$\overline{n}_k = \frac{1}{M} \sum_j n_{j,k}, \qquad \overline{n}_{\bar{k}} = \frac{1}{M} \sum_j n_{j,\bar{k}}. \tag{2.39}$$

Autorzy zbadali wersję binarną klasyfikacji. Wtedy powyższy wzór (jednakowy dla obu klas, więc indeks klasy można pominąć) upraszcza się do postaci:

$$CD(f) = \left(\frac{n_{f,\,pos}}{\overline{n}_{pos}} - \frac{n_{f,\,neg}}{\overline{n}_{neg}} \right)^2. \tag{2.40}$$

Miara CD została zbadana przez autorów na zbiorze *Reuters-21578* i spisała się lepiej (w sensie miary F1) od takich miar, jak IG czy CHI, ale tylko dla małej liczby terminów początkowych, tj. 100. Wraz ze wzrostem liczby terminów efektywność wszystkich miar (poza jedną) stabilizowała się i miara CD nie była już najlepsza.

Lam i Ho (1998) zaproponowali algorytm GIS (*Generalized Instances Set*), którego ideą klasyfikacyjną jest skonstruowanie zbioru wzorcowych terminów z klasy pozytywnej, który w dalszej kolejności znacznie ułatwia klasyfikowanie terminów. Kluczową rolę w algorytmie odgrywa funkcja *Rep*(*G*) reprezentująca moc przypadku *G* (ze zbioru przypadków uogólnionych) dana wzorem:

$$Rep(G) = \sum_{I^+ \in K} (k - rank(I^+)), \tag{2.41}$$

gdzie:
K – zbiór *k*-najbliższych sąsiadów przypadku *G*,
rank(I^+) – ranga dokumentu pozytywnego I^+ w zbiorze *K*.

Rangowanie jest przeprowadzane przy zastosowaniu odległości cosinusowej (por. np. Hand i inni 2005). Duże wartości funkcji *Rep*(*G*) świadczą o tym, że w zbiorze *K* jest dużo dokumentów pozytywnych. Algorytm rozpoczyna od losowo wybranego pozytywnego dokumentu spośród zbioru uczącego. Ten zbiór początkowy *GIS* jest sekwencyjnie uogólniany do momentu, gdy w zbiorze uczącym nie będzie już dokumentów pozytywnych. Uogólnianie polega na dołączaniu do zbioru już istniejącego *GIS* dokumentów, które zwiększają wartość funkcji *Rep*(*G*). Dołączone dokumenty są w każdym kroku usuwane ze zbioru uczącego. Uogólnianie zbioru *GIS* może być wykonywane różnymi metodami. Na przykład metoda Rocchio polega na tym, że dokument G' kandydat do dołączenia ma postać:

$$G' = \frac{1}{|Pos_k|} \sum_{\boldsymbol{x} \in Pos_k} \boldsymbol{x} - \alpha \frac{1}{|Neg_k|} \sum_{\boldsymbol{x} \in Neg_k} \boldsymbol{x}, \tag{2.42}$$

gdzie:
$|Pos_k|$ – moc zbioru Pos_k dokumentów pozytywnych w zbiorze *K*,
α – parametr wyważający rolę dokumentów pozytywnych i negatywnych.

Cały algorytm GIS zależy, jak widać, od kilku wzorcowych, osłabia wpływ szumu. Wpływ niektórych cech/terminów jest uwypuklony, niektórych zaś osłabiony, zatem można uznać, że algorytm ten jest połączeniem klasyfikacji z selekcją

terminów. Autorzy zbadali algorytm na zbiorach *Reuters-21578* oraz *OHSUMED*. Dla niektórych wartości parametrów algorytm CIS okazał się lepszy od klasyfikatora *k*-NN oraz klasyfikatorów liniowych.

Fragoudis i inni (2005) zaproponowali selekcję terminów polegającą na wybieraniu najlepszych cech pozytywnych w danej klasie i najlepszych cech negatywnych poza daną klasą. Kluczową rolę odgrywa w tym podejściu definicja terminu pozytywnego *f* dla klasy *C*:

$$P(C|f) > \frac{1}{2} \cdot p + \frac{1}{2} \cdot P(C), \tag{2.43}$$

gdzie *p* jest parametrem, który należy zadać. Celem użycia parametru *p* jest wyeliminowanie prostych akceptacji bądź odrzuceń wtedy, gdy $P(C|f) > P(C)/2$ z powodu bardzo małych/dużych wartości $P(C)$. Podobnie definiuje się termin *f* negatywny dla klasy *C*:

$$P(C|f) < \frac{1}{2} \cdot (1-p) + \frac{1}{2} \cdot P(C). \tag{2.44}$$

Dokumenty niezawierające żadnego terminu pozytywnego są pomijane w dalszej analizie. Algorytm rozpoczyna od wybrania wszystkich cech pozytywnych dla danej klasy (w przypadku binarnym jest to tylko jedna klasa dokumentów pozytywnych) oraz wszystkich cech negatywnych spoza tej klasy. Następnie, dla każdego terminu, można obliczać wartości funkcji scoringowej, którą może być dowolna z popularnych miar istotności terminu, takich jak MI, IG, iloraz szans etc. Autorzy zbadali efektywność swojej metody na zbiorze *Reuters-21578*. Najpierw oszacowali na zbiorach treningowych wartość parametru *p* odpowiedzialnej za liczbę terminów pozostawionych do analizy. Okazało się, że wartość ta bardzo różni się w zależności od miary istotności terminu (od 0,25 do 0,75). Ogólnie ujmując wyniki, największa poprawa (około 10%) efektywności miar przy zastosowaniu metody Fragoudisa miała miejsce dla naiwnego algorytmu Bayesa (podobnie jak w innych badaniach), znacznie mniejsza poprawa (około 2%) była uzyskana dla klasyfikatora SVM. W kilku przypadkach (na kilkanaście) metoda Fragoudisa pogorszyła efektywność miary oryginalnej.

Combarro i inni (2005) zaproponowali do selekcji terminów cech rodzinę miar liniowych, zależnych tylko od liczebności terminów w klasach, postaci:

$$LM_a(f, k) = a \cdot n_{f,k} - n_{f,\bar{k}}, \tag{2.45}$$

gdzie parametr *a* może przyjmować dowolną wartość rzeczywistą. W swojej pracy autorzy wykazali szereg własności optymalizacyjnych, które ułatwiają posługiwanie się miarami postaci (wzór 2.45). Kluczową rolę odgrywa własność stwierdzająca, że jeśli *f*, *g* są dwoma różnymi terminami z ustalonej klasy *k*, to:

$$LM_a(f, k) < LM_a(g, k), \tag{2.46}$$

wtedy i tylko wtedy, gdy:

$$k < \tan\alpha, \tag{2.47}$$

gdzie α jest kątem pomiędzy punktami $(n_{f,\bar{k}}, n_{f,k})$, $(n_{g,\bar{k}}, n_{g,k})$. Ta i inne własności pozwalają znacznie uprościć poszukiwanie dobrej wartości parametru a do tego stopnia, że pozostaje skończona liczba wartości tego parametru dająca różne uporządkowania ważności terminów. Autorzy rozważali dwa różne sposoby poszukiwania wartości parametru a: bezwzględny, tj. taką samą wartość parametru dla wszystkich klas, oraz względny w stosunku do klasy. Badania przeprowadzono na zbiorach *Reuters-21578* oraz *OHSUMED*. Rolę benchmarku odgrywała miara IG. Uzyskane wyniki pozwalają twierdzić autorom, że dla obu zbiorów dokumentów można zoptymalizować na zbiorze uczącym wybór parametru a tak, by efektywność klasyfikacji była taka, jak przy użyciu miary IG lub nawet wyższa (w sensie miary F1). Metoda wartości bezwzględnych spisała się lepiej na zbiorze *Reuters-21578*, zaś metoda wartości względnych wypadła lepiej na zbiorze *OHSUMED*.

Shang i inni (2007) zaproponowali do selekcji terminów miarę GT (*Gini Text*) bardzo blisko związaną z indeksem Giniego. Miara ta dla terminu f dana jest wzorem:

$$GT(f) = \sum_k (P(C_k|f))^2 (P(f|C_k))^2. \tag{2.48}$$

Wzór różni się od wyjściowej postaci współczynnika Giniego dla zbioru po podziale na kilka rozłącznych klas zastąpieniem prawdopodobieństw występowania terminu f prawdopodobieństwami warunkowymi $P(f|C_k)$ podniesionymi do kwadratu. Autorzy uzasadniają ten zabieg tym, że przy niezrównoważonych liczebnościach klas otrzymywalibyśmy bardzo różne wartości miar GT dla, na przykład, dwóch jednakowo ważnych terminów pojawiających się jednakowo często w swoich klasach. Miara GT została zbadana na zbiorze *Reuters-21578* oraz na zbiorze dokumentów chińskich dla trzech popularnych klasyfikatorów (k-NN, rozmytej wersji k-NN, SVM). Miara na ogół plasowała się na pierwszym lub drugim miejscu, ustępując niekiedy mierze IG, ale jej zaletą, jak twierdzą autorzy, jest prostota obliczeniowa i, co za tym idzie, szybkość działania. Inne prace dotyczące indeksu Giniego można znaleźć w pracach: Chen i inni (2014), Wu i inni (2017), Ortega-Mendoza i inni (2018).

Mladenic i Grobelnik (1999) jako pierwsi zaproponowali użycie entropii do pomiaru istotności terminu dla klasyfikacji. Swoją metodę nazwali oczekiwaną entropią przekrojową (*expected cross entropy*), która jest dana wzorem:

$$ECE(f) = P(f)\sum_k P(C_k|f)\frac{log(P(C_k|f))}{P(C_k)}. \tag{2.49}$$

Ta metoda była wielokrotnie badana (np. Wang i inni 2008; Shan i inni 2011) oraz w wielu publikacjach przekrojowych, w których pełniła funkcję benchmarku, była

kilkakrotnie modyfikowana. Wu i inni (2015) zaproponowali modyfikację mającą na celu uwzględnienie możliwej zerowej wartości prawdopodobieństwa $P(C_k|f)$ (dodając do tego prawdopodobieństwa 0,001) oraz możliwy różny rozkład liczebności terminów w poszczególnych klasach, wprowadzając pomocniczą entropię $I(f) = -\sum_k \frac{m_{f,k}}{m_f} P\left(\frac{m_{f,k}}{m_f}\right)$. Otrzymali wzór wstępnej modyfikacji:

$$ECE^*(f) = P(f)\frac{1}{I(f)+0{,}001}\sum_k (P(C_k|f)+0{,}001)\frac{log(P(C_k|f)+0{,}001)}{P(C_k)}. \quad \textbf{(2.50)}$$

We wzorze końcowym:

$$ECE_{mod}(f) = \gamma \cdot max_k\{\alpha_k \cdot \beta_k\} \cdot ECE^*(f) \quad \textbf{(2.51)}$$

γ jest względnym odchyleniem standardowym liczebności terminu *f* względem klas, zaś β_k jest znormalizowanym odchyleniem liczebności terminu *f* w ustalonej klasie C_k i α_k jest wagą liczebności terminu *f* w ustalonej klasie C_k. Autorzy porównali swoją metodę z tradycyjną oczekiwaną entropią przekrojową na zbiorze tekstów chińskich z wyróżnionymi siedmioma klasami. W sensie miary F1 modyfikacja okazała się lepsza we wszystkich klasach o około 5–7%.

Largeron i inni (2011) zaproponowali do selekcji terminów algorytm oparty na entropii Shannona:

$$E(f) = -\sum_k \frac{m_{f,k}}{m_f} \cdot log_2\left(\frac{m_{f,k}}{m_f}\right). \quad \textbf{(2.52)}$$

Entropia osiąga wartość minimalną zero wtedy, gdy termin *f* występuje tylko w jednej klasie i taka sytuacja świadczy oczywiście o ważności terminu *f* dla klasyfikacji. Inaczej jest w przeciwnym razie, tj. gdy entropia osiąga wartość maksymalną E_{max}, co ma miejsce, gdy termin *f* występuje we wszystkich klasach z jednakową częstością. Autorzy proponują do ustalenia ważności terminu *f* dla klasyfikacji miarę ECCD (*Entropy based Category Coverage Difference*) daną wzorem:

$$ECCD(f, C_k) = (P(f|C_k) - P(f|\overline{C_k}))\frac{E_{max} - E(f)}{E_{max}}. \quad \textbf{(2.53)}$$

W przypadku podziału na dwie klasy otrzymujemy wzór (por. wzór 2.19):

$$ECCD(f, C_{pos}) = \frac{ad-bc}{(a+c)(b+d)} \cdot \frac{E_{max} - E(f)}{E_{max}}. \quad \textbf{(2.54)}$$

Obie miary autorzy zbadali na zbiorach *Reuters-21578* oraz pięciokrotnie większym zbiorze *INEX XML Mining collection* podzielonym na 15 klas. Selekcja ter-

minów polegała na tym, że oddzielnie dla każdej klasy zostały one uporządkowane według malejących wartości miar. Następnie wybrano ustaloną liczbę słów początkowych. Miara ECCD okazała się lepsza od miar IG, GSS, CC, DGL, $\widetilde{\chi}^2$, osiągając średnią precyzję klasyfikacji około 75%. Z tego badania autorzy wywnioskowali, że miara ta nadaje się do klasyfikacji zbiorów zarówno małych, jak i dużych rozmiarów. Nie potwierdzono wniosków z badania Sebastiani (2002) o tym, że miara CHI spisuje się podobnie do miary IG (okazała się słabsza). Inną modyfikację dotyczącą entropii można znaleźć w pracy Mladenović i inni (2016).

Zhen i inni (2011) zaproponowali metodę selekcji terminów opartą na odległości Kullbacka-Leiblera dwóch rozkładów. Odległość, a właściwie rozbieżność Kullbacka-Leiblera (miara nie spełnia warunku trójkąta ani symetrii odległości), dwóch rozkładów prawdopodobieństwa dana jest wzorem:

$$KL(p, q) = \sum_x p(x) log \frac{p(x)}{q(x)}. \tag{2.55}$$

Zhen i inni zastosowali tę miarę do wyrażenia rozproszenia terminu *f* względem wszystkich klas:

$$S(f) = KL(s(f, C_k), s(f, C_l)) = \sum_{s(f,\, C_k) > s(f,\, C_l)} s(f, C_k) log \frac{s(f, C_k)}{s(f, C_l)}, \tag{2.56}$$

gdzie rolę $s(f, C_k)$ może odgrywać dowolna miara związku pomiędzy terminem *f* a klasą C_k. Autorzy zastosowali w tym miejscu miarę $\widetilde{\chi}^2(f, C_k)$ (por. wzór 2.28). Miara *S* została zbadana na zbiorze *Reuters-21578* dla klasyfikatora *k*-NN. Do liczby około 1000 początkowych terminów miara była lepsza (o około 2–3%) od miary $\widetilde{\chi}^2(f, C_k)$.

Sathic Ali i Venkateswaran (2014) zastosowali matematyczną teorię dowodu Dempstera-Shafera (por. Wang i Bell 2004) do charakterystyki podzbiorów terminów. Na zbiorze 2^F wszystkich podzbiorów zbioru *F* wszystkich terminów określamy funkcję *m* przypisania podstawowego prawdopodobieństwa (*basic probability assignment*). Dwie podstawowe własności tej funkcji przypominają własności prawdopodobieństwa:

$$m(\varnothing) = 0 \text{ oraz } \sum_{A \in 2^F} m(A) = 1. \tag{2.57}$$

Liczba $m(A)$ jest miarą „poparcia" zbioru *A*. Następnie, na zbiorze 2^F definiujemy funkcję wiary (*belief function*):

$$Bel(A) = \sum_{B \subset A} m(B). \tag{2.58}$$

Liczba $Bel(A)$ wyraża całkowite „poparcie" dla zdarzeń ze zbioru *A*, ale nie dla żadnego z tych zdarzeń oddzielnie. Miarę $Bel(A)$ stosuje się w dalszej kolejności

do oceny „poparcia" dla grupy terminów *f*, z których każdy ma swoje przypisane podstawowe prawdopodobieństwo $m(f)$. Przypisane podstawowe prawdopodobieństwa mogą być generowane przez dowolne metody oceniające ważność terminu dla klasyfikacji (autorzy zastosowali metody IG, OR i $\widetilde{\chi}^2$). Wyrażenie:

$$m_i(d_j) = x_{i,j} \cdot log \frac{N}{n_{i,j}} \tag{2.59}$$

jest używane do przypisania podstawowego prawdopodobieństwa dla terminu f_i i dla dokumentu d_j. W efekcie końcowym wartość $Bel(A)$ jest wspólną dla wszystkich pojedynczych metod oceną ważności terminu dla klasyfikacji. Wszystkie terminy należy uporządkować względem tych wartości i wybrać ustaloną liczbę początkowych terminów. Autorzy podają wyniki badania na trzech zbiorach tekstów: *Reuters-21578*, *WebKB* i *20 NewsGroups* dla klasyfikatora *k*-NN oraz SVM. W większości przypadków ich miara wygrywa (na ogół z niewielką przewagą 1–2%) z pojedynczo stosowanymi miarami IG, OR i $\widetilde{\chi}^2$.

2.3. Metody inspirowane naturą

W ostatnich latach dość dużą popularność zdobywają metody naśladujące procesy, które można zaobserwować w przyrodzie, na ogół wśród zwierząt. Algorytm robaczków świętojańskich (*firefly algorithm*) został wprowadzony przez Yang (2009) i jego ideą jest naśladowanie zachowania robaczków w celu badania przestrzeni rozwiązań. Podstawą analizy wykorzystującej tego typu algorytmy są następujące dwa założenia. Po pierwsze, robaczki świętojańskie są tej samej płci i nie ma ona wpływu na ich wzajemne przyciąganie się. Przyciąganie się robaczków jest uzależnione od jasności światła emitowanego przez nie. Robaczek emitujący światło mniej jasne porusza się w kierunku robaczka z jaśniejszym światłem, gdy zaś nie ma jaśniejszego robaczka, to porusza się losowo. Po drugie, w problemie optymalizacyjnym jasność oblicza się, używając funkcji celu. Przyciąganie się dwóch robaczków oblicza się przy użyciu następującego równania:

$$\beta_r = \beta_0 exp(-\gamma r_{jk}^2), \tag{2.60}$$

gdzie:
β_0 – przyciąganie się przy odległości $r = 0$,
γ – współczynnik absorpcji,
r_{jk}^2 – odległość pomiędzy robaczkiem *j* a robaczkiem *k*.

Odległość pomiędzy robaczkiem *j* a robaczkiem *k* jest rozumiana w sensie euklidesowym:

$$r_{jk} = \sqrt{\sum_{i=1}^{n}(x_{j,i} - x_{k,i})^2}, \tag{2.61}$$

gdzie $x_{j,i}$ oraz $x_{k,i}$ to współrzędne robaczków w przestrzeni *n*-wymiarowej. Ruch robaczka od jednej pozycji do drugiej opisany jest równaniem:

$$x_j = x_j + \beta_0 exp(-\gamma r_{jk}^2)(x_j - x_k) + \alpha\left(random - \frac{1}{2}\right), \tag{2.62}$$

w którym pierwszy składnik po prawej stronie to bieżąca pozycja robaczka, drugi składnik opisuje ruch w kierunku robaczka jaśniejszego, zaś trzeci składnik opisuje ruch losowy względem α. Zastosowanie opisanej powyżej techniki obliczeniowej do badania tekstu polega na tym, że każdy dokument jest rozumiany jako robaczek. Robaczek to ciąg zer i jedynek związanych z przytoczonymi oznaczeniami zależnością:

$$x_{j,i} = \begin{cases} 1 & \text{gdy termin } j \text{ występuje w dokumencie } i \\ 0 & \text{w przeciwnym przypadku} \end{cases} \tag{2.63}$$

Jasność światła robaczka to częstość występowania danego terminu. Dokumenty są przesuwane względem siebie iteracyjnie, po obliczeniu siły przyciągania, odległości i po wykonaniu ustalonej liczby iteracji każdy dokument jest oceniany za pomocą intensywności światła, jaką emituje w danej pozycji dokument-robaczek. Po takiej transformacji dokumentów można ustalić dokument najlepszy, tj. ten zajmujący pierwszą pozycję w rankingu, i wraz z nim określony zbiór terminów.

Wadą podstawowej wersji algorytmu *firefly* jest łatwość grzęźnięcia w maksimach lokalnych, wobec czego opracowano kilka modyfikacji mających na celu wyeliminowanie tej wady. Xu i inni (2018) zaproponowali modyfikację polegającą na połączeniu algorytmu z ideą uczenia się na przeciwieństwie (*opposition based learning*). Jeśli $x \in [m; n]$ jest liczbą rzeczywistą, to liczbę opozycyjną do *x* nazywamy $\tilde{x} = m + n - x$. W przypadku wielowymiarowym tak postępujemy na każdej współrzędnej. Modyfikacja Xu polega na tym, by liczby opozycyjne zastosować w odniesieniu do jasności dokumentu-robaczka. Powiększamy w ten sposób liczbę analizowanych wariantów, ale okazuje się, że modyfikacja jest dużo odporniejsza na możliwość ugrzęźnięcia w maksimach lokalnych, gdyż takie są efekty uczenia się na przeciwieństwie, sprawdzone wcześniej w innych działach statystyki. Modyfikacja ta została zbadana jednak tylko na kilku bardzo małych zbiorach, nie zbadano jej na żadnym benchmarkowym zbiorze tekstowym.

Elakkiya i inni (2020) zwrócili uwagę na to, że w przedstawionym algorytmie podstawowym, w każdej aktualizacji pozycji robaczka-dokumentu konieczne jest obliczanie odległości do wszystkich jaśniejszych robaczków, co powoduje wydłużenie czasu działania algorytmu. Zaproponowali oni, by obliczać tylko odległości do ustalonej liczby robaczków ograniczonej przynależnością do jakiejś grupy społecznej (*community inspired firefly algorithm*). Takie podejście może mieć logiczne uzasadnienie, jeżeli będziemy rozważać, na przykład, klasyfikację wpisów z forów internetowych. Wadą takiego wariantu może być, podobnie jak w podstawowej wersji algorytmu, grzęźnięcie w optimach lokalnych, wobec czego autorzy zaproponowali formułę optymalnego rozmiaru c społeczności:

$$c = \delta \times \left| \frac{t - m}{t_i + t_{best}} \right|, \tag{2.64}$$

gdzie:
t – iteracja bieżąca,
m – maksymalna liczba iteracji,
t, t_{best} – jasności dokumentu w, odpowiednio, bieżącej lokalizacji oraz najlepszego znalezionego dotychczas;
δ – współczynnik społeczności.

Modyfikacja została przetestowana na standardowym zbiorze tweetów pod kątem wykrywania tweedów zawierających spam. W sensie dokładności klasyfikacji okazała się nieznacznie lepsza (1–2%) od podstawowej wersji algorytmu *firefly*.

2.4. Metody z grupy *ensamble*

Jak zauważają Pintas i inni (2021), metody typu *ensamble* można podzielić na trzy główne nurty:

1. Metody scalające zbiory cech otrzymane różnymi metodami pojedynczymi. Wykorzystanie takich metod polega na tym, że stosuje się kilka metod pojedynczych i ostateczny zbiór wybranych cech jest efektem scalenia zbiorów otrzymanych z metod pojedynczych.
2. Metody typu łańcuchowego. Działanie takich metod polega na tym, że używa się kilku metod pojedynczych w ten sposób, że zbiór cech wyselekcjonowany przez jedną metodę jest zbiorem wejściowym do następnej.
3. Metody łączące rankingi. Stosując kilka metod pojedynczych tworzonych jest kilka rankingów cech opartych na różnych metrykach, które następnie są scalane w jeden ranking i wybierane są cechy powyżej pewnego progu.

Jedną z najwcześniejszych prób włączenia metodologii *ensamble* do analizy sentymentu był z pewnością artykuł Dai (Dai i inni 2011). W tej pracy zastosowano dwie nowatorskie techniki klasyfikacji dokumentów. Pierwsza to uwypuklanie znaczenia wybranych terminów (*feature highlighting*), druga to wrzucanie terminów do jednego worka (*feature bagging*). Uwypuklanie znaczenia niektórych terminów to prosta technika polegająca na nadawaniu wyższych wag terminom z ładunkiem sentymentalnym. Z kolei *feature bagging* polega w przypadku klasyfikacji dokumentów ze względu na ich sentyment na tym, że klasyfikujemy w oparciu o różne zbiory cech wylosowanych bez zwracania ze zbioru wszystkich cech, po czym ostateczna klasyfikacja odbywa się na zasadzie najwyższej zdobytej liczby głosów. Tę metodę należy zaliczyć do pierwszej grupy metod. W swoim opracowaniu autorzy wyjaśniają od strony teoretycznej, dlaczego technika *feature bagging* powinna dawać dobre rezultaty. Testowanie przeprowadzili na zbiorze *Polarity*, uzyskując dokładność klasyfikacji lepszą o około 2% od standardowych metod pojedynczych. Pewną słabością opracowania jest to, że wykorzystywane były leksykony do ustalenia ładunku sentymentalnego terminów.

Przykładem metody z ostatniej grupy jest metoda *soft vote* zaproponowana przez Agnihotri (Agnihotri i inni 2019), która ma dwie charakterystyczne cechy. Pierwsza polega na tym, że stosujemy kilka standardowych metod selekcji zmiennych (autorzy użyli MI (por. wzór 2.9), Odds Ratio (por. wzór 2.22), IG (por. wzór 2.11), GSS (por. wzór 2.32), GT (por. wzór 2.48)) i ostatecznym rankingiem jest ranking dany średnią ważoną punktów zdobytych przez terminy w każdej z kilku metod. Druga cecha, różniąca tę procedurę od innych tego typu, to wprowadzenie wag uzależnionych od relacji liczby terminów nacechowanych pozytywnie do liczby terminów negatywnych w klasach. To było słabością wspomnianych metod standardowych, tzn. działały one lepiej, gdy liczby terminów pozytywnych i negatywnych były we wszystkich klasach zrównoważone. Metoda została zbadana na pięciu zbiorach danych (m.in. *Reuters* i *OHSUMED*) dla czterech różnych klasyfikatorów i zwyciężyła pojedyncze standardowe metody dla każdego z klasyfikatorów (zwycięstwo nad najlepszą z metod pojedynczych było zawsze minimalne, tj. około 1%).

Do trzeciej grupy należy też zaliczyć metodę zaproponowaną przez Shena i innych (2013), która polega na znalezieniu wagi dla każdej pojedynczej metody S selekcji terminów za pomocą miary:

$$W_S = 0.5 \cdot log\left(\frac{P_S}{1-P_S}\right), \tag{2.65}$$

gdzie P_S jest oceną jakości pojedynczej metody, a konkretnie precyzją (por. wzór 1.36) tej metody uzyskaną na zbiorze uczącym. Następnie znajdujemy wagi w_f dla każdego terminu f według formuły:

$$w_f = \sum_S W_S \cdot g(f,S), \tag{2.66}$$

gdzie $g(f, S)$ jest dane wzorem:

$$g(f,S) = \begin{cases} 0 & \text{gdy termin } f \text{ nie jest wybrany przez metodę } S \\ 1 & \text{gdy termin } f \text{ nie jest wybrany przez metodę } S \end{cases} \quad \textbf{(2.67)}$$

Po znalezieniu wag dla wszystkich terminów wybieramy spośród nich ustaloną liczbę najlepszych. W roli metod pojedynczych użyto cztery standardowe metody: częstość terminów (por. wzór 2.18), metodę MI (por. wzór 2.10), metodę IG (por. wzór 2.11) oraz CHI (por. wzór 2.28). Zaproponowana metoda została sprawdzona na zbiorze chińskich recenzji hotelowych, dla ustalonych liczb terminów początkowych i okazała się nieco lepsza od metod pojedynczych, ale tylko dla dużych liczb terminów (dla małych liczb przegrywała rywalizację z najlepszą spośród pojedynczych metod).

Do trzeciej grupy zaliczyć należy również metodę zaproponowaną przez Hai i innych (2015), która polega na znalezieniu wagi dla każdej pojedynczej metody S selekcji terminów za pomocą miary:

$$SIGCHI(f, C_k) = (\alpha \cdot Norm(IG(f, C_k)) + \beta \cdot Norm(CHI(f, C_k)))^2, \quad \textbf{(2.68)}$$

gdzie $Norm(IG(f, c_k))$ oznacza wartość $IG(f, c_k)$ (por. wzór 2.11) przeniesioną na przedział [0; 1] za pomocą klasycznej unitaryzacji, analogicznie dla $CHI(f, c_k)$. Metoda została przez autorów nazwana SIGCHI (*Square of Information Gain and Chi-square*). Współczynniki spełniają warunek $\alpha + \beta = 1$, przy czym rozważano tylko wartości skokowe $\alpha \in \{0{,}1; 0{,}2; ...; 0{,}9\}$. Miara lokalna (wzór 2.68) została następnie przekształcona do globalnej za pomocą średniej ważonej częstościami klas. Wszystkie terminy porządkujemy malejąco względem zaproponowanej miary i wybieramy ustaloną liczbę początkowych. Propozycja została zbadana na kilku zbiorach dokumentów w języku wietnamskim i, na ogół, była lepsza od pojedynczych metod IG oraz CHI o 1–2%. W kilku przypadkach przegrała z najlepszą z metod konkurencyjnych, jednak należy zaznaczyć, że metoda jest bardzo odporna na dobór wartości α. Niezależnie od α efektywność metody mierzona dokładnością klasyfikacji oraz miarą F1 była jednakowa z dokładnością do 0,05.

2.5. Wybrane metody wykorzystujące źródła zewnętrzne

Przykładem metody hybrydowej oraz, poniekąd, *ensamble* jest metoda Govindarajana (2013). Liczba cech jest redukowana za pomocą algorytmu *best first search*. Autorzy nie podają jednak kryteriów optymalizacji tego algorytmu. Następnie

jest stosowana metoda typu *ensamble* (ściślej *arcing classifier*) z użyciem dwóch klasyfikatorów: naiwnego Bayesa oraz klasyfikatora opartego na algorytmie genetycznym. Efektywność metody została zbadana na zbiorze recenzji filmowych. Zysk precyzji klasyfikacji w porównaniu do pojedynczych klasyfikatorów był jednak niewielki (93,8% dla metody *ensamble* wobec 91,15% dla naiwnego Bayesa i 91,25% dla algorytmu genetycznego).

Przykładem metody hybrydowej podobnej do pomysłu Govindarajana może być metoda Iqbala i innych (2019), którą można scharakteryzować następująco. Metoda jest hybrydą łączącą maksymalizację funkcji wiarygodności z podejściem leksykonowym. Podejście leksykonowe używające sieci SentiWordNet zostało wykorzystane do redukcji liczby cech za pomocą algorytmu genetycznego. Idea stosująca algorytm genetyczny polega na tym, by wyznaczyć zbiór terminów $S \subset F$ taki, by $\sum_i S_i = Sent_j(lub\ najbliżej)$, gdzie S_i jest punktacją sentymentu i-tego terminu ($i = 1, 2, ...d \leq I$) ze zbioru S odczytaną z SentiWordNet. Maksymalizację warunku $\sum_i S_i = Sent_j(lub\ najbliżej)$ przeprowadza się za pomocą wektora wag (zero-jedynkowych) terminów, na których algorytm genetyczny może pracować. Następnie na zredukowanym zbiorze cech stosowane są typowe klasyfikatory oparte na maksymalizacji funkcji wiarygodności.

Dunning (1993) podał wzór funkcji ilorazu wiarygodności dla porównania dwóch rozkładów dwumianowych lub wielomianowych. Wzory te zostały wykorzystane przez Gamona (2004) do wykrywania najistotniejszych dla klasyfikacji dokumentów cech unigramowych, dwugramowych oraz trójgramowych. Ciekawym wynikiem, jaki uzyskał, było to, że oprócz typowych słów posiadających ładunek emocjonalny za bardzo istotne zostały uznane cechy, których człowiek raczej by o to nie podejrzewał. Na przykład takimi cechami jednogramowymi lub dwugramowymi okazały się:

try the, of, off, @@przymiotnik, your

Gamon spośród około 30 000 cech wybrał pierwszych 2000 z początku listy cech istotnych (ustalonej za pomocą ilorazu wiarygodności Dunninga) i dla tradycyjnego klasyfikatora metodą wektorów nośnych uzyskał trafność klasyfikacji 85,47%, co można uznać za dobry wynik dla tekstów z (jak autor ocenia) niewielkim szumem.

Agarwal i inni (2011) zaproponowali dwie metody oceny sentymentu wpisów na Twitterze, które można zaliczyć do metod leksykonowych. W pierwszej metodzie wykorzystano kilka różnych baz leksykonowych do oceny sentymentu słów. Następnie każdy tweet przedstawiono za pomocą drzewa i mierzono podobieństwo dwóch drzew (potrzebne na etapie klasyfikacji) za pomocą częściowego jądra drzewa (*partial tree kernel*), które działa na zasadzie analizowania wszystkich możliwych poddrzew danego drzewa. Elementami drzewa mogą być słowa lub etykiety (*tags*) innych rodzajów czy elementów wpisu twitterowego. Idea takiego

podejścia polega na tym, że analizowanie wszystkich możliwych poddrzew powinno wychwycić różne zależności pomiędzy słowami lub etykietami, które niekiedy trudno opisać inaczej. W drugiej metodzie rozwinięto pomysł Gamona (2004), oceniając sentyment wpisu za pomocą 100 dość sztucznie zdefiniowanych cech. Cechy te można podzielić na kilka grup. Na przykład jedną grupę cech przyjmujących wartości ze zbioru liczb naturalnych stanowią cechy, które są zliczeniami wystąpień np. pozytywnych (ocena z tych samych, co w pierwszej metodzie kilku źródeł leksykonowych) przymiotników, negatywnych przysłówków itp. Inną grupą są cechy o wartościach rzeczywistych, które są zsumowanymi wartościami sentymentu rzeczowników, przymiotników, przysłówków itp. lub wszystkich słów. Rezultaty obu metod (oraz różne kombinacje ich i metody unigramowej) modelowania wpisów poddano klasyfikacji za pomocą metody wektorów nośnych. Obie metody dały marginalnie lepszą poprawność klasyfikacji, odpowiednio 73,93% i 71,27%, wobec 71,35% dla metody unigramowej. Nieco lepiej spisały się hybrydy: metoda pierwsza plus druga – 74,61% oraz unigram plus metoda druga – 75,39%.

Rozdział 3

Autorska propozycja metody klasyfikacji tekstów

3.1. Wnioski z przeglądu literatury – zadania badawcze

Z dokonanego w poprzednich rozdziałach przeglądu literatury można wyprowadzić kilka, naszym zdaniem, zdecydowanych wniosków. Po pierwsze, niemożliwe jest rzetelne porównanie wszystkich metod badania sentymentu dokumentów z powodu olbrzymiej liczby zaproponowanych metod. Co więcej, uważamy, że niemożliwe jest nawet porównanie pośrednie wszystkich metod z powodu różnych założeń dotyczących organizacji eksperymentu czynionych przez badaczy. Należy tu wymienić różnorodność wielu podejść do problemu analizy sentymentu dokumentów. Nawet jeśli ograniczymy się do oceny jednego, najpopularniejszego podejścia heurystycznego, tzn. podejścia dwuetapowego składającego się z filtrowania zbioru wszystkich terminów w pierwszym etapie oraz klasyfikacji dokumentów w drugim etapie, to i tak pozostanie wiele czynników nastręczających wielu trudności w ocenie metod. Należy tu wymienić: wzgląd na metody i oprogramowanie używane do wstępnej obróbki tekstu, dobór tekstów testowych, dobór liczebności zbioru uczącego czy też dobór klasyfikatorów. W takich okolicznościach trudno traktować każdy wynik badawczy, który postuluje znalezienie metody o efektywności o 1% wyższej od (niektórej) konkurencji jako oznakę dużego sukcesu badawczego. Uważamy, że o wiele ważniejsze jest, na przykład, poszukiwanie metod jak najmniej uzależnionych od zbioru uczącego lub szybkich.

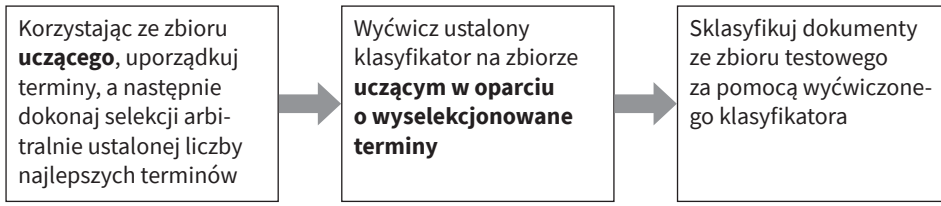
Przegląd osiągnięć zawarty w poprzednich rozdziałach pozostawia pewien niedosyt w ocenie efektywności metod badania sentymentu dokumentów. Najefektywniejsze są metody z grupy *ensamble*, ale one opierają się na wykorzystaniu dobrych własności metod indywidualnych. Oceniając metody indywidualne, można zauważyć, że te charakteryzujące się najwyższą efektywnością są często uzależnione od wartości parametrów trudnych do precyzyjnego dostrojenia (np. metody wykorzystujące sieci neuronowe, metody naśladujące naturę). Z kolei metody prostsze, mające naturalną interpretację swoich rozwiązań technicznych i parametrów,

charakteryzują się niewiele niższą efektywnością (np. metody oparte na zysku informacji IG). Proponujemy zbadanie efektywności dwóch wybranych metod filtrujących zbiory terminów, które mają dobrą opinię w literaturze przedmiotu dotyczącej tekstów w języku angielskim. Takie badanie powinno być wystarczające do odniesienia się do efektywności innych metod, wielu modyfikacji metod standardowych, gdyż te zawsze były porównywane z metodami standardowymi. Na tym etapie istotne jest ustalenie stopnia skrócenia zbioru wszystkich terminów (*level of aggressiveness*). Większość omawianych w poprzednich rozdziałach metod nie podaje żadnych kryteriów, które mogłyby być pomocne w wyborze liczby terminów. Wobec tego jednym z zadań badawczych będzie próba ustalenia wniosków dotyczących tego problemu. W naszej opinii ważne jest także to, żeby badanie efektywności uwzględniało, w odróżnieniu od przyjętych schematów, możliwości ograniczenia rozmiaru zbioru uczącego. Ponadto chcielibyśmy zaproponować swoją metodę prostą intuicyjnie, bo opartą na skorelowaniu częstości terminów oraz zdolności dyskryminacyjnej terminów, efektywną, jak również posiadającą cechy nauczania bez nadzoru.

3.2. Sformułowanie nowej metody

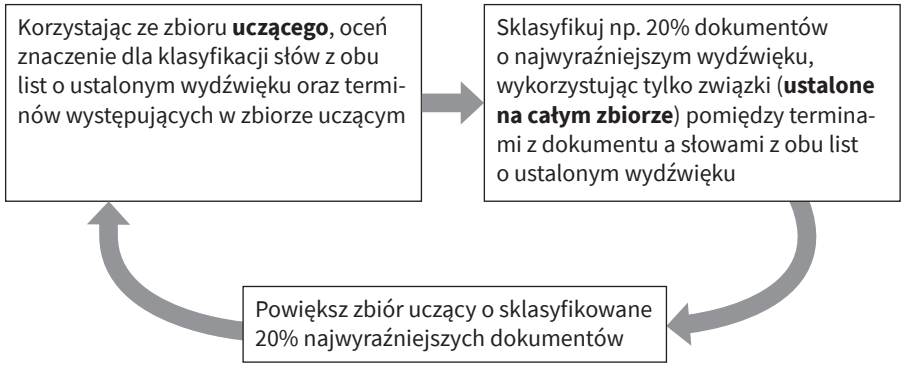
Chcielibyśmy zaproponować metodę, która przenosiłaby ciężar klasyfikacji dokumentów z wykorzystania obszernych zbiorów uczących na dokładniejsze analizowanie związków pomiędzy terminami występującymi w dokumentach. Należy mieć świadomość, że wkraczamy wówczas w obszar klasyfikacji bez nadzoru, czyli w analizę skupień. Ta dziedzina statystyki jest bogata w różnego rodzaju metody i algorytmy, ale tworzy swego rodzaju błędne koło, wewnątrz którego wybór zmiennych może być uzależniony od metody grupowania i od liczby klas, podobnie optymalna liczba skupień jest uzależniona od metody grupowania, z kolei grupowanie jest oczywiście zależne od doboru zmiennych. Ta sytuacja powoduje na ogół możliwość powstawania istotnie różniących się wyników klasyfikacji. Zauważmy jednak, że w naszym problemie istnieją dwa zasadnicze ułatwienia, które pozwalają na wybrnięcie z tych trudności. Po pierwsze, znamy liczbę klas – zawsze będą to dwie klasy dokumentów. To pozwala na lepszy dobór zmiennych. Po drugie, zbiór uczący wcale nie jest potrzebny do identyfikacji tego, która klasa będzie pozytywna, a która negatywna. Do ustalenia wydźwięku klas wystarczy impuls uzyskany, na przykład, na podstawie nagromadzenia słów o wydźwięku pozytywnym lub negatywnym. Nie musimy w tym celu korzystać ze źródeł zewnętrznych – wystarczy zadać listy kilkudziesięciu powszechnie używanych słów o wydźwiękach w oczywisty sposób pozytywnych lub negatywnych niezależnie od kontek-

stu leksykalnego badania. Mając wstępną klasyfikację pewnej części (na przykład jednej czwartej) zbioru testowego, tej którą można sklasyfikować poprzez wysokie nagromadzenie słów o ustalonym wydźwięku, można, korzystając z niewielkiego zbioru uczącego, ustalić miary istotności dla klasyfikacji zdefiniowanych list słów. Taką klasyfikację można powtarzać etapami aż do sklasyfikowania całego zbioru testowego. Zasadę działania standardowych metod klasyfikacji tekstów oraz ideę naszej propozycji przedstawiamy na rysunkach 2 i 3.



Rysunek 2. Zasada działania standardowej metody klasyfikacji tekstów.

Źródło: opracowanie własne.



Rysunek 3. Schemat przedstawiający ideę nowej metody.

Źródło: opracowanie własne.

Zaproponujemy zatem algorytm, który będzie bazował na następujących założeniach:

- wykorzystując **cały zbiór dokumentów**, ustalamy związki pomiędzy wszystkimi terminami występującymi w całym zbiorze dokumentów oraz wszystkimi słowami o określonym sentymencie występującymi w dwóch zdefiniowanych grupach słów – **krótkich słownikach sentymentu**;

- wykorzystując zbiór uczący, ustalamy wagi terminów z każdej z obu grup słów oraz istotność wszystkich terminów występujących w zbiorze uczącym dla separowalności klas;
- w wersji sekwencyjnej algorytmu w kolejnych krokach klasyfikujemy grupy, np. 20% dokumentów, i wykorzystujemy te dokumenty do aktualizacji wag terminów ze słowników sentymentu oraz istotności terminów dla separowalności klas.

Pomiar siły związku pomiędzy terminem a terminem jest utrudniony ze względu na nieliczbowy charakter cechy-terminu, ale można zastosować pomiar w postaci współczynnika korelacji liniowej pomiędzy częstościami występowania terminów w dokumentach. Sentyment pozytywny bądź negatywny można ustalić na podstawie skorelowania terminów z dokumentów ze słowami występującymi w słownikach sentymentu. Jeżeli większe będzie średnie (ze wszystkich terminów dokumentu) przywiązanie do grupy słów pozytywnych, to uznamy, że termin przemawia za zaliczeniem dokumentu do pozytywnych. Takie podejście trochę przypomina to zaproponowane przez Dai (por. Dai i inni 2011), przy czym chcemy uniknąć konieczności korzystania z obszernych słowników sentymentu, co byłoby związane z koniecznością korzystania ze źródeł zewnętrznych. W miejsce obszernych źródeł zewnętrznych proponujemy dwa krótkie słowniki sentymentu, które będą punktem startowym do ustalania sentymentu terminów oraz do ewentualnego uczenia się algorytmu na tych grupach i, w pewnym sensie, powiększania ich. Z kolei znaczenie terminu dla siły separowalności dwóch klas można mierzyć, tak jak w wielu innych znanych metodach, za pomocą większego z dwóch odsetków przynależności terminu dla klasy, czyli większej z dwóch liczb: $\frac{n_{f,1}}{n_f}, \frac{n_{f,0}}{n_f}$.

Współczynnik korelacji liniowej pomiędzy częstościami (liczbami wystąpień) terminów *f* oraz *g* będziemy obliczać ze wzoru:

$$WK(f,g,l) = \frac{1}{s_f s_g}\left(\frac{1}{l}\sum_{i=1}^{l} n_{f,i} n_{g,i} - \overline{n}_f \overline{n}_g\right), \tag{3.1}$$

gdzie:
l – liczba rozważanych dokumentów,
$n_{f,i}$, $n_{g,i}$ – częstość terminów *f* oraz *g* w dokumencie *i*-tym,
$\overline{n}_f$, $\overline{n}_g$ – średnie arytmetyczne częstości,
s_f, s_g – odchylenia standardowe częstości.

Zauważmy, że wykorzystanie współczynnika korelacji pomiędzy częstościami terminów rozwiązuje problem niewykorzystywania częstości występowania terminów w wielu popularnych i efektywnych metodach selekcji terminów, takich jak metoda IG czy CHI-kwadrat.

W celu ustalenia sentymentu terminu będziemy posługiwać się skorelowaniem częstościowym pomiędzy terminem a słowami z następującego słownika sentymentu.

Grupa *SP* słów pozytywnych po stemmatyzacji (por. podrozdział 1.3):

bezproblemowo; *ciekawa*; *ciekawy*; *dobra*; *dobry*; *fajna*; *fajny*; *godna*; *godny*; *korzysc*; *korzystna*; *korzystny*; *ladna*; *ladnie*; *ladny*; *lepsza*; *lepszy*; *mila*; *milo*; *mily*; *najlepsza*; *najlepszy*; *piekna*; *piekny*; *pozytywna*; *pozytywnie*; *pozytywny*; *profesjonalna*; *profesjonalny*; *rzetelna*; *rzetelny*; *sprawna*; *sprawnie*; *sprawny*; *super*; *swietna*; *swietny*; *szybka*; *szybki*; *zadowolona*; *zadowolony*.

Grupa *SN* słów negatywnych po stemmatyzacji:

bezplciowa; *bezplciowy*; *brak*; *brakowac*; *brud*; *brudna*; *brudny*; *brzydka*; *brzydki*; *dopiero*; *dramat*; *dramatyczna*; *dramatyczny*; *glupi*; *glupia*; *gorsza*; *gorszy*; *gowno*; *licha*; *lichy*; *najgorsza*; *najgorszy*; *najslabsza*; *najslabszy*; *nedzna*; *nedzny*; *nieciekawa*; *nieciekawy*; *niedobra*; *niedobry*; *nieladna*; *nieladnie*; *nieladny*; *niemila*; *niemily*; *nieprofesjonalna*; *nieprofesjonalny*; *nierzetelna*; *nierzetelny*; *niesprawna*; *niesprawny*; *niestety*; *nijaka*; *nijaki*; *nudna*; *nudny*; *oszust*; *oszustka*; *porazka*; *razic*; *reklamacja*; *slaba*; *slabsza*; *slabszy*; *slaby*; *syf*; *syfiasty*; *szkoda*; *tepa*; *tepy*; *uszkodzic*; *uszkodzona*; *uszkodzony*; *zepsuc*; *zepsuta*; *zepsuty*; *zla*; *zly*; *zniszczona*; *zniszczony*; *zniszczyc*.

Zauważmy, że obie powyższe grupy słów zawierają tylko niektóre spośród powszechnie znanych i używanych słów niezależnie od kontekstu tematycznego czy okresu czasowego. Uważny Czytelnik z łatwością uzupełniłby te listy kilkudziesięcioma słowami o zdecydowanie określonym wydźwięku bez posługiwania się słownikami. Mocne ograniczenie liczby słów w obu grupach jest powodowane dbałością o to, by proponowana metoda mogła być zaliczona do grupy metod maszynowego uczenia się, nie zaś do grupy opartej na wykorzystaniu źródeł zewnętrznych.

Proponowaną podstawową wersję algorytmu można dokładniej zapisać w postaci następującego algorytmu.

Krok 1. Znajdź korelację częstościową daną wzorem 3.1 dla wszystkich par (f, g) terminów, gdzie f to dowolny termin z występujących w całym zbiorze dokumentów, zaś g dowolny termin z obu list.

Krok 2. Dla każdego terminu f, na podstawie zbioru uczącego, znajdź miarę jego istotności dla separowalności klas w postaci większego z odsetków wystąpień tego terminu w dokumentach z obu klas, tzn. w postaci:

$$max_k P(C_k|f) = max_k \frac{n_{f,k}}{n_f}. \tag{3.2}$$

Dla każdego terminu *g* należącego do grupy słów pozytywnych (negatywnych) znajdź wagę tego terminu dla właściwej grupy w postaci odsetka wystąpień tego terminu w dokumentach pozytywnych $w_{g,1}$ oraz negatywnych $w_{g,0}$ w zbiorze uczącym.

Krok 3. Dla każdego terminu *f* wyznacz siłę związku $s(P, f, w_{g,1})$ pomiędzy terminem a grupą słów pozytywnych oraz $s(N, f, w_{g,0})$ pomiędzy terminem a grupą słów negatywnych za pomocą średniej arytmetycznej ze wszystkich współczynników korelacji częstościowych (wzór 3.1) pomiędzy terminem *f* a poszczególnymi słowami z właściwej listy pomnożonych przez odpowiednią wagę $w_{g,i}$.

Krok 4. Uporządkuj wszystkie dokumenty malejąco według średniej arytmetycznej miary $M(f)$ spośród wszystkich terminów *f* występujących w dokumencie, przy czym:

$$M(f) = (s(P, f, w_{g,1}) - s(N, f, w_{g,0})) \cdot max_k \frac{n_{f,k}}{n_f}. \tag{3.3}$$

Krok 5. Zaklasyfikuj dokumenty do jednej z dwóch klas w zależności od miejsca zajmowanego w ciągu otrzymanym w kroku 4, proporcjonalnie do odsetków dokumentów pozytywnych i negatywnych w zbiorze uczącym.

Przedstawiona propozycja algorytmu to wariant podstawowy. Można łatwo modyfikować ten algorytm w zależności od tego, jakim czasem na wykonanie klasyfikacji dysponujemy. Na przykład można zaproponować wersję z sekwencyjną aktualizacją odsetka $max_k(n_{f,k} / n_f)$ – tak jak zostało przedstawione na rysunku 3. Ideą takiej aktualizacji jest to, że niektóre dokumenty mają tak silne wskazania do klasyfikacji, że prawdopodobieństwo błędnej klasyfikacji jest nikłe, wobec czego możemy niewielkim kosztem powiększać zbiór uczący i w oparciu o coraz większy zbiór aktualizować odsetek $max_k(n_{f,k} / n_f)$ oraz wagi $w_{g,1}$, $w_{g,0}$ terminów. Taką modyfikację można oprzeć na klasyfikowaniu najłatwiejszych dokumentów, czyli tych o najbardziej zdecydowanych wskazaniach (znajdujących się na samym początku lub samym końcu ciągu z kroku 4). Po sklasyfikowaniu np. 20% dokumentów aktualizujemy zbiór uczący, powiększając go o nowo sklasyfikowane dokumenty, co pozwala na aktualizację miar (wzór 3.2) istotności terminów dla separowalności oraz wag słów z obu grup. Po zaktualizowaniu zbioru uczącego w kolejnym kroku klasyfikujemy następne 20% najłatwiejszych do sklasyfikowania dokumentów, kontynuując ten proces do momentu sklasyfikowania wszystkich dokumentów. Wersję sekwencyjną algorytmu można zapisać następująco.

Krok 4a. Uporządkuj wszystkie dokumenty malejąco według średniej arytmetycznej miary $M(f)$ spośród wszystkich terminów *f* występujących w dokumencie.

Krok 5a. Sklasyfikuj $p = 20\%$ wszystkich dokumentów do jednej z dwóch klas w zależności od miejsca zajmowanego w ciągu z kroku 4a, proporcjonalnie do

odsetków dokumentów pozytywnych i negatywnych w zbiorze uczącym. Dołącz zaklasyfikowane dokumenty do zbioru uczącego.

Krok 6a. Dla każdego terminu *f*, na podstawie bieżącego zbioru uczącego zaktualizuj miary (wzór 3.2) istotności dla separowalności oraz wagi $w_{g,1}$, $w_{g,0}$ dla każdego słowa z obu list.

Krok 7a. Powtarzaj kroki 4a–6a do momentu sklasyfikowania wszystkich dokumentów, które mają miarę (wzór 3.3) różną od zera.

Krok 8a. Pozostałe dokumenty, tzn. te, dla których miara (wzór 3.3) równa jest zeru, zaklasyfikuj losowo do jednej z klas proporcjonalnie do frakcji klas w zbiorze uczącym.

Poniżej przestawiamy najistotniejszą część pseudokodu wersji sekwencyjnej algorytmu:

D – set of all *N* documents; $D = DTRAIN \cup DTEST$
$D_1 \leftarrow \#DTRAIN$; $D_2 \leftarrow \#DTEST$; $N = D_1 + D_2$;
$P_1 \leftarrow$ *#Positive Documents in DTRAIN*;
CurrentDTEST ← *DTEST*;
CurrentDTRAIN ← *DTRAIN*;
for f ∈ *D do begin*
for g ∈ *SP* ∪ *SN do begin Find WK*(*f*, *g*) *end*; *end*;
Repeat until #*CurrentDTEST* = 0 {

for f ∈ *TermsOfCurrentDTRAIN do begin* $Find\frac{n_{f,1,train}}{n_{f,train}}$; $Find\frac{n_{f,0,train}}{n_{f,train}}$; *end*;

for (*g* ∈ *TermsOfCurrentDTRAIN*) *and* (*g* ∈ *SP* ∪ *SN*) *do begin*

$$w_{g,1} \leftarrow \frac{n_{g,1,train}}{n_{g,train}}; \qquad w_{g,0} \leftarrow \frac{n_{g,0,train}}{n_{g,train}}; \quad end;$$

for d ∈ *CurrentDTEST do begin*
for f ∈ *TermsOf d do begin*

$$for\ g \in SP\ do\ begin\ \ s(P,f,w_{g,1}) \leftarrow \frac{1}{m_{SP,f,d}} \sum WK(f,g) \quad end;$$

$$for\ g \in SN\ do\ begin\ \ s(N,f,w_{g,0}) \leftarrow \frac{1}{m_{SN,f,d}} \sum WK(f,g) \quad end;$$

$$M(f) \leftarrow (s(P,f,w_{g,1}) - s(N,f,w_{g,0})) \cdot max_k \frac{n_{f,k}}{n_f}$$

$$V(d) \leftarrow V(d) + M(f)$$

end; // *for f* ∈ *TermsOf d*

$$V(d) \leftarrow \frac{V(d)}{m_d}$$

end; // *for* $d \in$ *CurrentDTEST*

Rank CurrentDTEST with respect to decreasing $V(d)$;

$$SetA \leftarrow entier\left[\frac{P_1}{D_1} \cdot 0.2 \cdot \#CurrentDTEST\right]$$

of initial documents of CurrentDTEST;

$$SetB \leftarrow entier\left[\left(1 - \frac{P_1}{D_1}\right) \cdot 0.2 \cdot \#CurrentDTEST\right]$$

of end_of_listdocuments of CurrentDTEST;

$CurrentDTEST \leftarrow CurrentDTEST \setminus (Set\ A \cup Set\ B)$;

$CurrentDTRAIN \leftarrow CurrentDTRAIN \cup SetA \cup SetB$;

}

Zauważmy, że przedstawiona wersja sekwencyjna ma bardzo istotną zaletę uwolnienia procedury od konieczności określania wartości tajemniczych parametrów, na ogół bardzo istotnych dla efektów działania procedury. Jedynym parametrem jest wartość $p = 20\%$ decydująca o szybkości uczenia się algorytmu na zbiorze uczącym czy też o szybkości powiększania zbioru uczącego. W eksperymencie badawczym zostanie zbadany wpływ zamiany 20% na przykład na 10%. Jeśli okaże się, że wpływ tej zmiany na efektywność klasyfikacji dokumentów będzie znikomy, to algorytm będzie można nazwać odpornym na ten parametr. Inną zaletą algorytmu jest to, że ma on cechy uczenia się, paradoksalnie, bez nadzoru. Po sklasyfikowaniu najłatwiejszych dokumentów powiększamy zbiór uczący, ale możemy też powiększać liczby słów w obu grupach o zdecydowanym zabarwieniu sentymentalnym poprzez dołączanie do jednej z tych grup terminu f, który jest odpowiednio silnie skorelowany z jedną z grup.

Należy doprecyzować jeszcze pewne szczegóły techniczne zmodyfikowanego algorytmu, zanim przystąpimy do badania. Liczbę l rozważanych dokumentów należy dobrać tak, by wyznaczane wartości miary skorelowania były stabilne. Gdy l jest małe (np. $l = 10$), to dokumenty trzeba wielokrotnie losować. Wobec tego w celu obniżenia kosztów czasowych zdecydowaliśmy się na obliczanie jednokrotne na całym zbiorze dokumentów. Innym szczegółem obliczeniowym niezależącym od użytkownika jest odsetek zbiorów, które nie dadzą się zaliczyć do żadnej klasy, ponieważ wartości miary $M(f)$ są zerowe dla wszystkich terminów występujących w dokumencie. Ten odsetek podajemy w wynikach badania w postaci średniej arytmetycznej odsetka na wszystkich etapach algorytmu. Wszystkie dokumenty, których nie można sklasyfikować na podstawie miary (wzór 3.3), są klasyfikowane losowo do jednej z klas, z prawdopodobieństwem proporcjonalnym do frakcji klas w zbiorze uczącym.

3.3. Organizacja badania

Badanie będzie polegało na zastosowaniu zaproponowanej metody do 13 zbiorów polskojęzycznych dokumentów przedstawionych w podrozdziale 1.6. Jako porównawcze metody benchmarkowe zastosowaliśmy dwie metody dwuetapowe – podejście najpopularniejsze wśród badaczy sentymentu tekstów. Pierwszy etap to filtrowanie zbioru wszystkich terminów do zbiorów mniejszych. Drugi etap to klasyfikacja wszystkich dokumentów ze zbioru testowego do jednej z dwóch klas. W pierwszym etapie metodami benchmarkowymi będą metoda IG (por. wzór 2.11) oraz CHI-kwadrat (por. wzory 2.28, 2.29). Obie te metody można uznać za stabilne w tym sensie, że w każdym badaniu spisywały się dobrze lub bardzo dobrze. Ponadto obie metody, lub przynajmniej jedna z nich, odgrywały w większości badań naukowych rolę metod benchmarkowych, co pozwala na dokonywanie oceny porównawczej z innymi metodami. W roli metod klasyfikacyjnych zastosujemy trzy klasyfikatory (por. podrozdział 1.4): naiwny klasyfikator bayesowski, metodę wektorów nośnych SVM oraz regresję logistyczną. Miarami efektywności klasyfikacji będą dwie popularne miary (dokładność oraz miara F1). Te dwie miary pojawiają się w prawie wszystkich badaniach, z kolei im większa liczba analizowanych miar, tym trudniejsze ocenianie końcowe metod. Badanie będzie polegało na powtórzeniu losowania zbioru uczącego o zadanym rozmiarze (3%, 6%, 10%, 15%, 25% całego zbioru) 100 razy. Po wylosowaniu zbioru uczącego będzie przeprowadzona klasyfikacja dokumentów zbioru testowego (dokumenty pozostałe po losowaniu) za pomocą wszystkich badanych metod. Ostateczne wyniki to średnie arytmetyczne dwóch miar efektywności klasyfikacji ze 100 powtórzeń.

3.4. Wyniki badania i wnioski

W tabeli 1 są zebrane wyniki nowej metody dla wszystkich badanych zbiorów dokumentów oraz wszystkich rozważanych rozmiarów zbioru uczącego. Nową metodę oznaczono symbolem KOR ze względu na jej głównie korelacyjny charakter. Wykresy dla poszczególnych zbiorów, jako najbardziej sugestywne, są przedstawione w dalszej kolejności dla poszczególnych zbiorów, natomiast dokładne wartości liczbowe są przedstawione w tabelach w *Załączniku*. Wartości miar dla metody KOR zostały zaznaczone na długości całej osi poziomej w celu ułatwienia porównania z innymi metodami, ale pamiętać należy o tym, że nie odnoszą się one do konkretnego odsetka liczby wykorzystanych terminów.

Tabela 1. Jakość (dokładność / miara F1) klasyfikacji nowej metody. ŚODKL – średnia (ze wszystkich rozmiarów zbioru uczącego i ze wszystkich iteracji klasyfikacji) arytmetyczna odsetka dokumentów klasyfikowanych losowo w ostatnim etapie algorytmu.

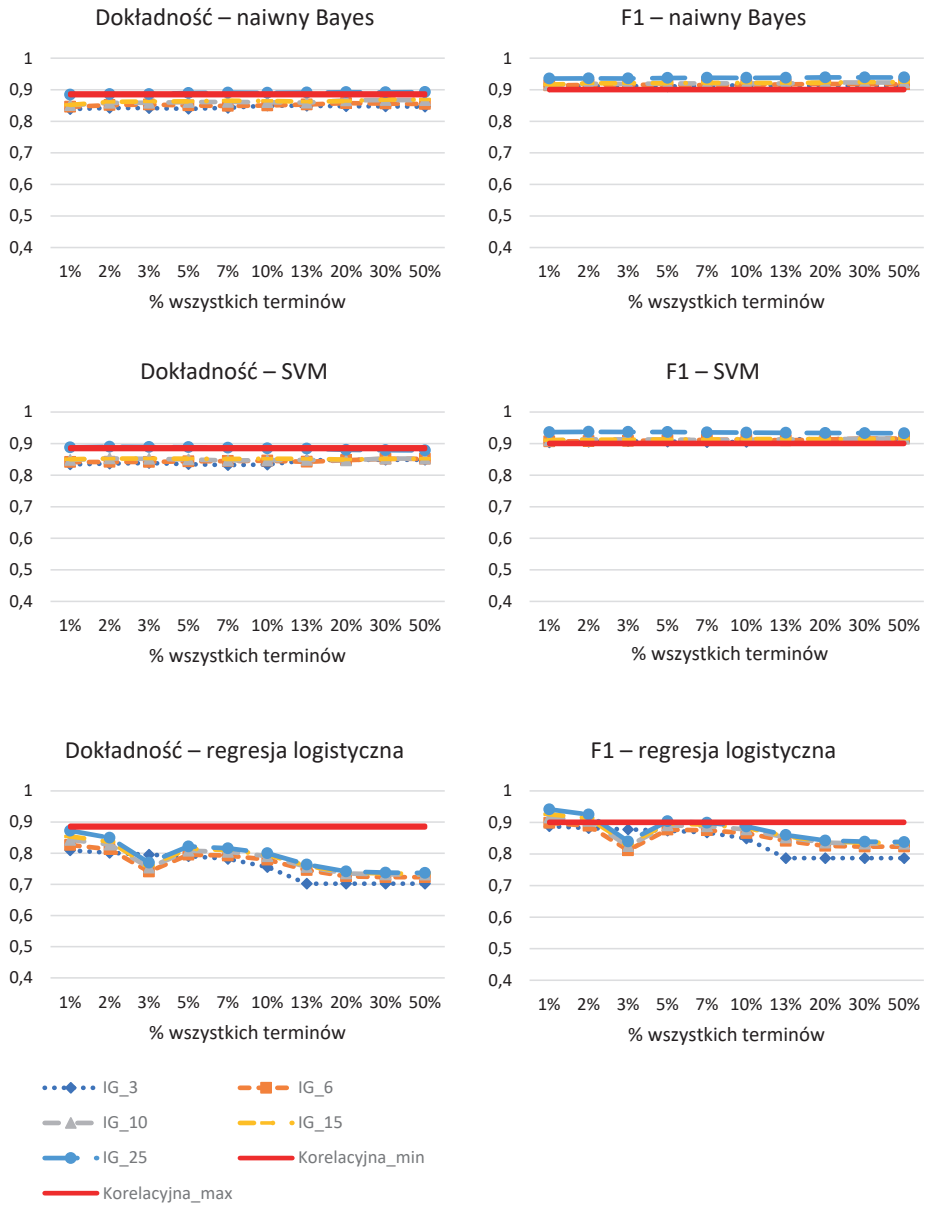
Nazwa zbioru	ŚODKL	3%	6%	10%	15%	25%
ksiazki.txt	0,062	0,831 / 0,860	0,829 / 0,859	0,835 / 0,860	0,834 / 0,854	0,835 / 0,855
ksiazkizbil.txt	0,047	0,820 / 0,849	0,833 / 0,853	0,831 / 0,854	0,835 / 0,860	0,835 / 0,852
hotele.txt	0,055	0,848 / 0,930	0,852 / 0,931	0,851 / 0,938	0,850 / 0,940	0,851 / 0,940
hotelezbil.txt	0,052	0,845 / 0,939	0,846 / 0,938	0,847 / 0,940	0,849 / 0,937	0,848 / 0,938
esklepy.txt	0,043	0,890 / 0,900	0,895 / 0,891	0,892 / 0,891	0,895 / 0,890	0,894 / 0,891
esklepyzbil.txt	0,044	0,880 / 0,896	0,890 / 0,891	0,892 / 0,893	0,890 / 0,890	0,892 / 0,893
apteki.txt	0,085	0,883 / 0,899	0,886 / 0,902	0,886 / 0,901	0,887 / 0,900	0,888 / 0,902
aptekizbil.txt	0,052	0,881 / 0,901	0,885 / 0,904	0,884 / 0,904	0,885 / 0,905	0,885 / 0,904
bank.txt	0,045	0,810 / 0,880	0,820 / 0,890	0,821 / 0,891	0,822 / 0,893	0,823 / 0,893
kurier.txt	0,051	0,897 / 0,930	0,900 / 0,932	0,901 / 0,931	0,901 / 0,932	0,902 / 0,933
kurierzbil.txt	0,022	0,880 / 0,917	0,887 / 0,920	0,886 / 0,919	0,888 / 0,919	0,889 / 0,920
perfumy.txt	0,052	0,899 / 0,940	0,912 / 0,947	0,915 / 0,948	0,916 / 0,948	0,915 / 0,947
perfumyzbil.txt	0,058	0,820 / 0,840	0,878 / 0,899	0,883 / 0,903	0,885 / 0,904	0,890 / 0,905

Źródło: obliczenia własne.

Pierwsze cztery wykresy dotyczą zbiorów *apteki* oraz *aptekizbil*. W przypadku zbioru *apteki* jakość metody KOR jest bardzo zbliżona do konkurencji niezależnie od tego, czy selekcja terminów wykonywana była metodą IG czy CHI oraz dla obu klasyfikatorów NB oraz SVM. Regresja logistyczna spisała się bardzo słabo. Zupełnie inaczej wyglądają wyniki w przypadku skrócenia zbioru do mniej więcej takich samych liczebności obu klas, tj. dla zbioru *aptekizbil*. W tym przypadku we wszystkich czterech wariantach (dwie metody IG i CHI oraz dwa klasyfikatory NB i SVM) widoczna jest wyraźna przewaga metody KOR.

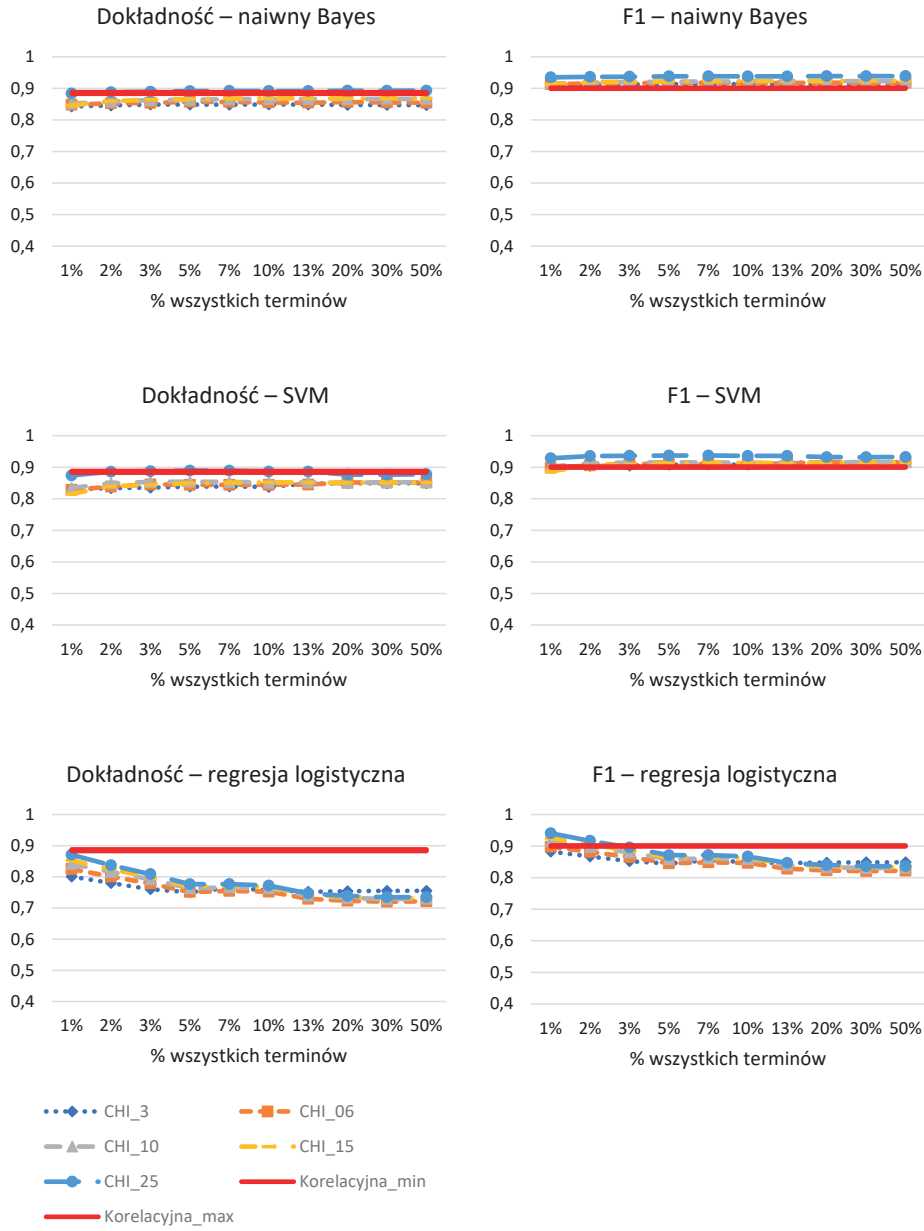
Ponadto oba estymatory były dość niestabilne względem odsetka początkowych terminów z uporządkowanych list uwzględnianego w klasyfikacji. Na ogół nieco niższa była jakość dla najmniejszych odsetków, ale zdarzył się też przypadek trudnego do wyjaśnienia spadku jakości przy wyższych odsetkach początkowych terminów.

Wykresy 5 i 6 dotyczą zbioru *bank*. Ten zbiór miał bardzo podobne liczebności klas. Wyniki metod konkurencyjnych są podobne do wyników metody KOR, przy czym uwagę zwraca duża stabilność klasyfikatora NB (jakość lekko wzrasta wraz z odsetkiem początkowych terminów) i zaskakująco wysoka niestabilność klasyfikatora SVM. Co więcej, wygląda na to, że klasyfikator SVM nie poradził sobie przy dużej liczbie początkowych terminów i przy niewielkim rozmiarze zbioru uczącego (3%, 6% i 10%). Wykresy 7 i 8 dotyczą największego z badanych zbiorów,



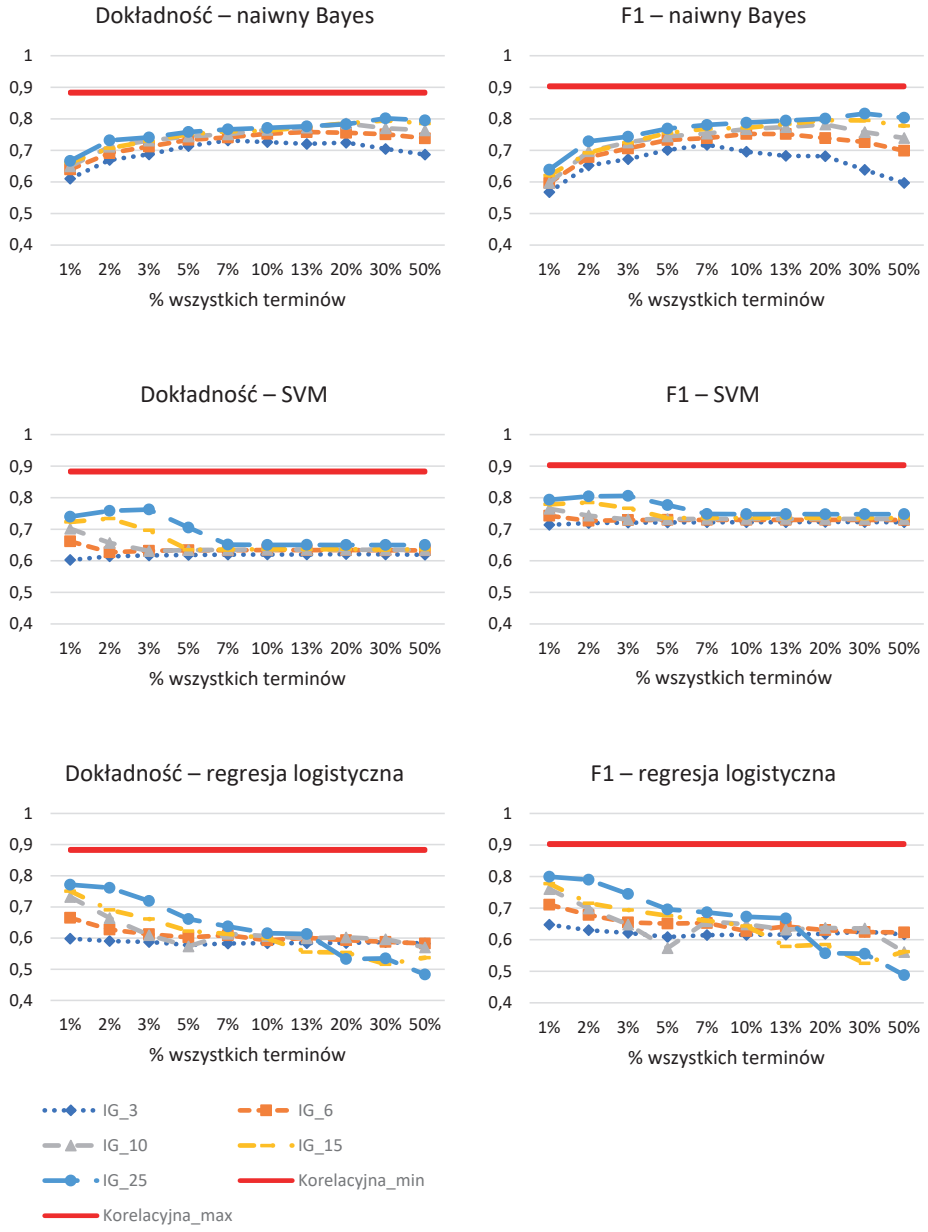
Wykres 1. Jakość klasyfikacji dla zbioru *apteki* (metody KOR oraz IG).

Źródło: obliczenia własne.



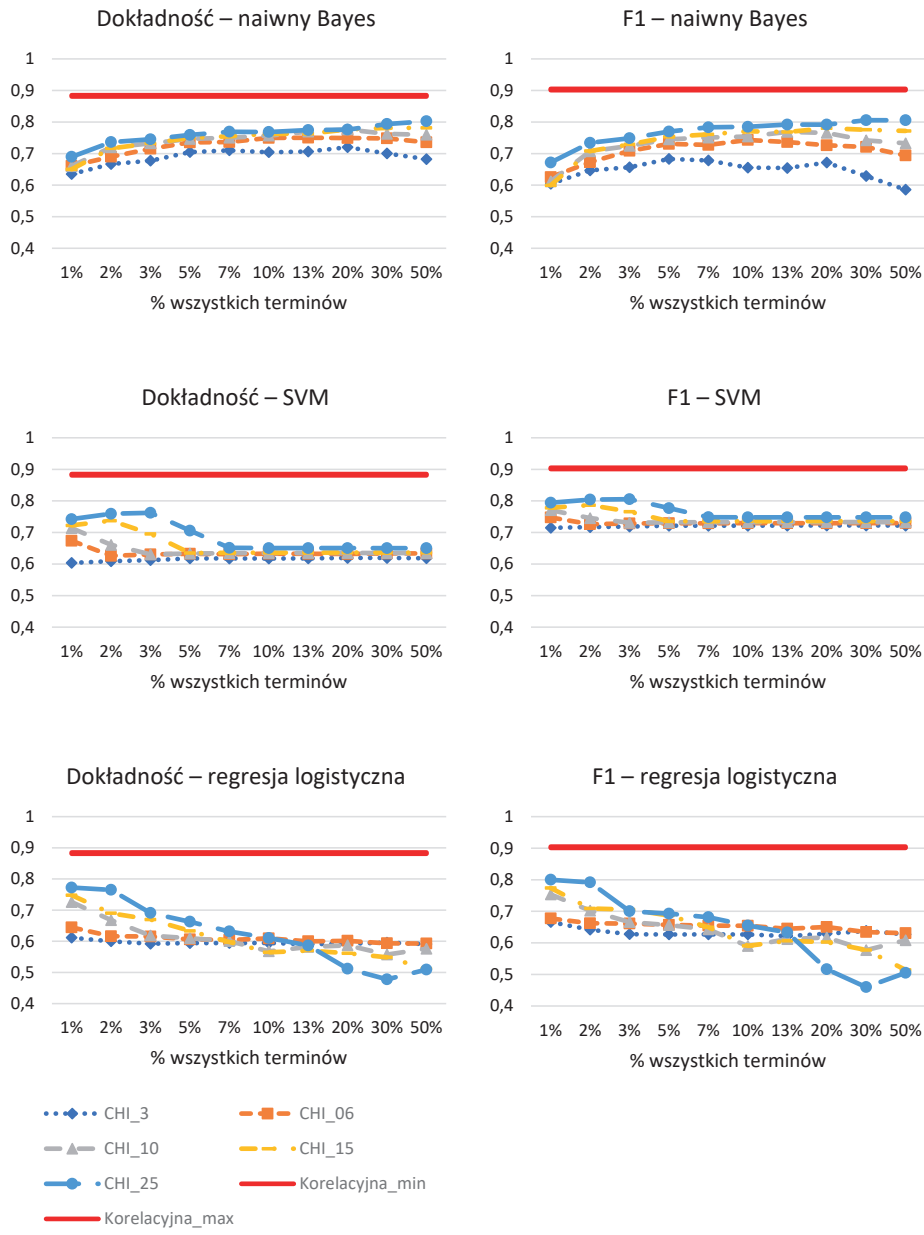
Wykres 2. Jakość klasyfikacji dla zbioru *apteki* (metody KOR oraz CHI).

Źródło: obliczenia własne.



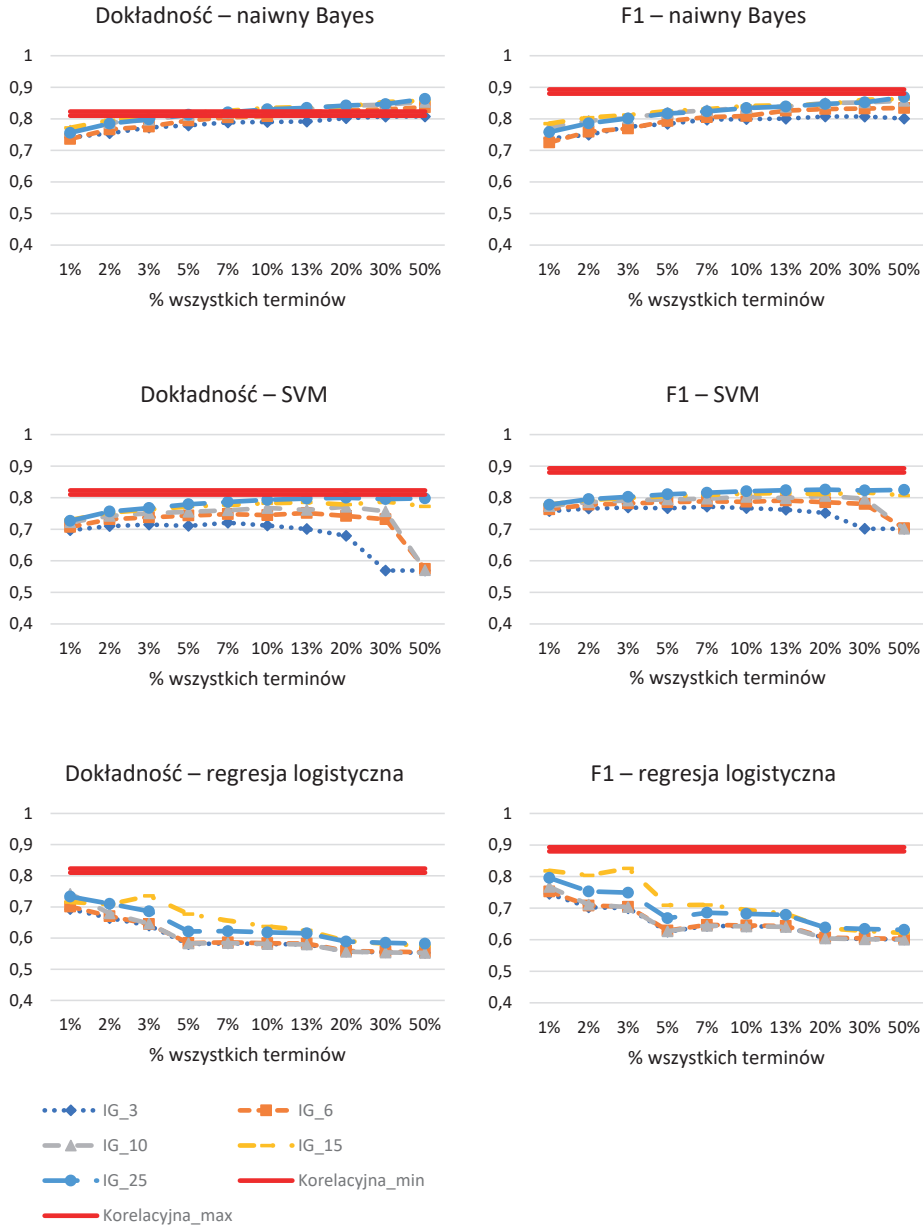
Wykres 3. Jakość klasyfikacji dla zbioru *aptekizbil* (metody KOR oraz IG).

Źródło: obliczenia własne.



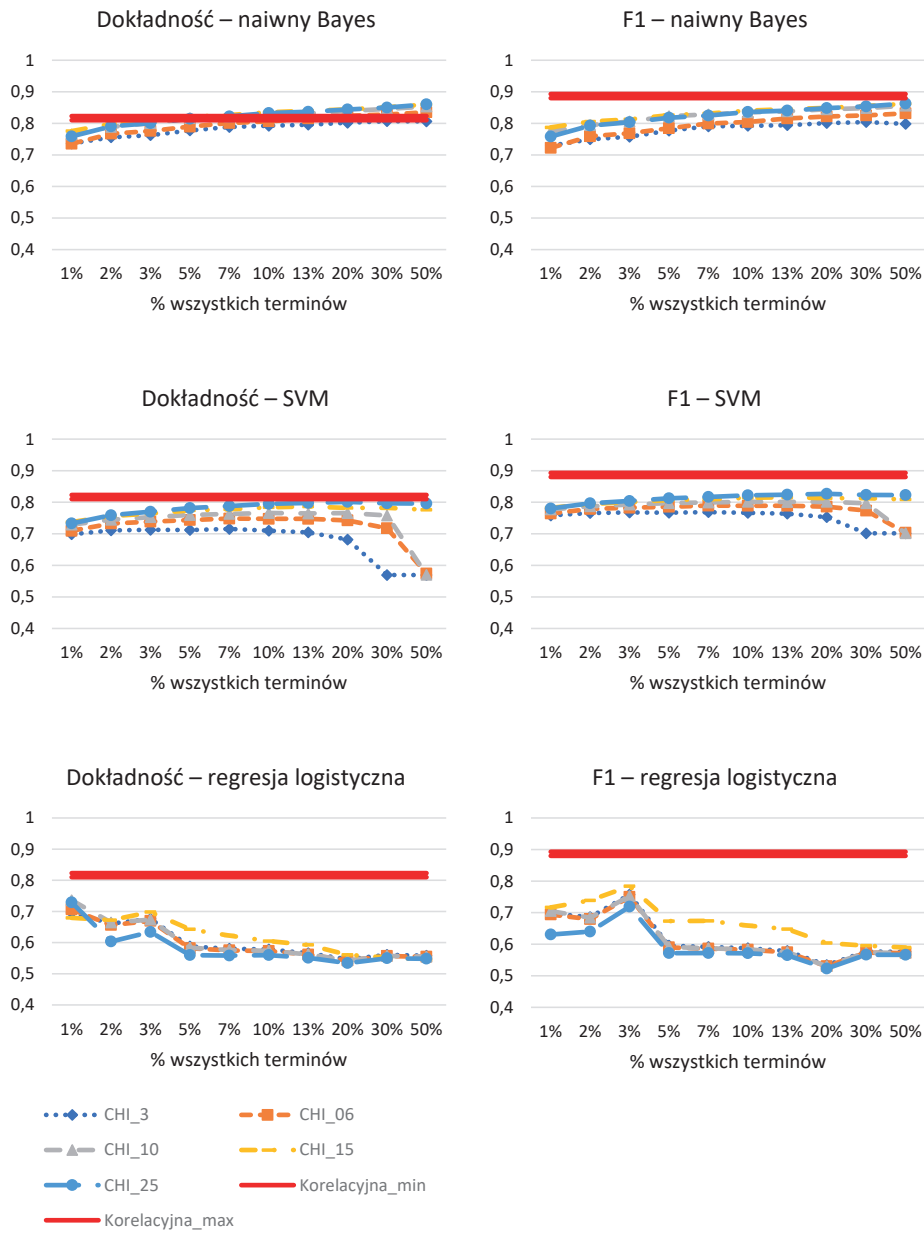
Wykres 4. Jakość klasyfikacji dla zbioru *aptekizbil* (metody KOR oraz CHI).

Źródło: obliczenia własne.



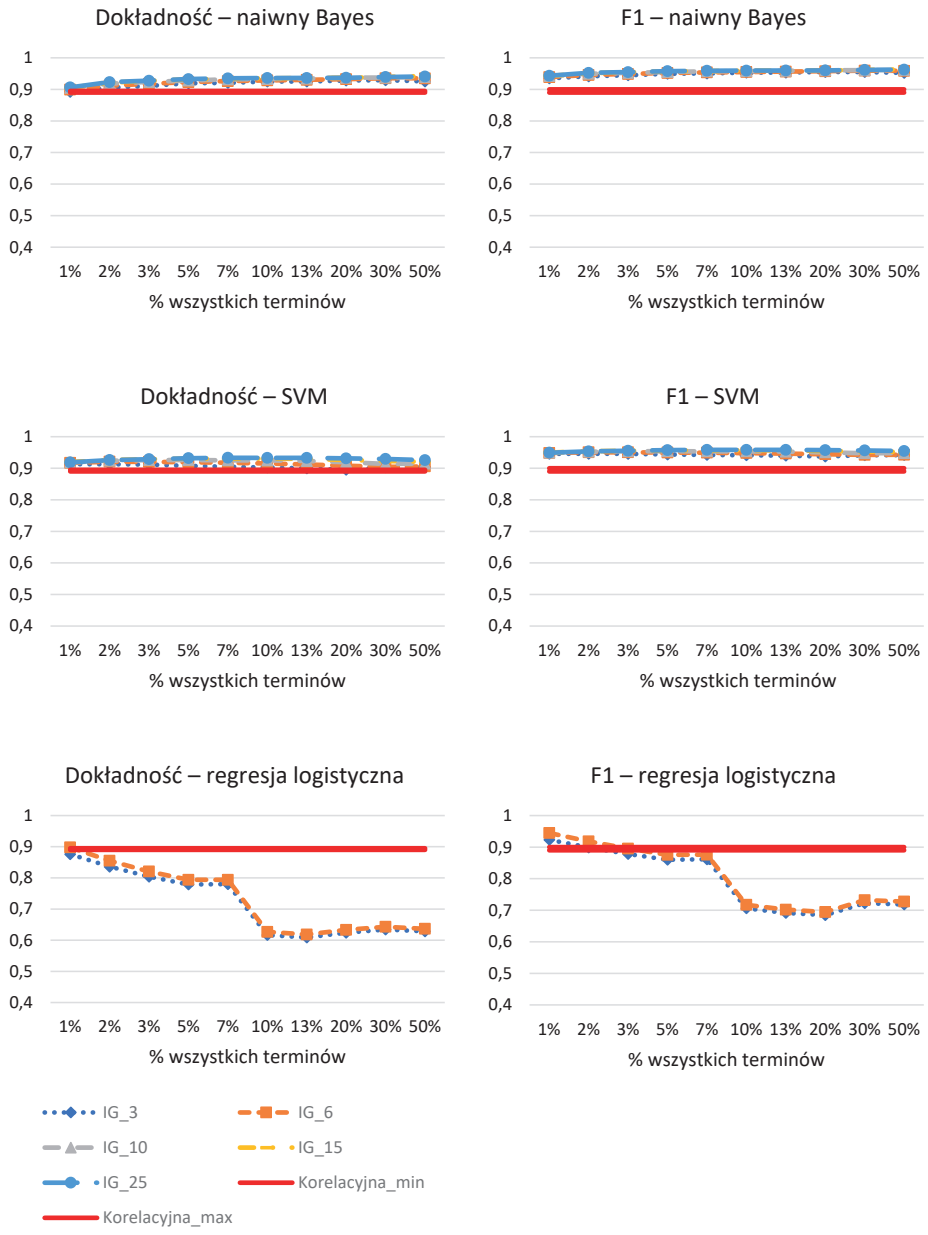
Wykres 5. Jakość klasyfikacji dla zbioru *bank* (metody KOR oraz IG).

Źródło: obliczenia własne.



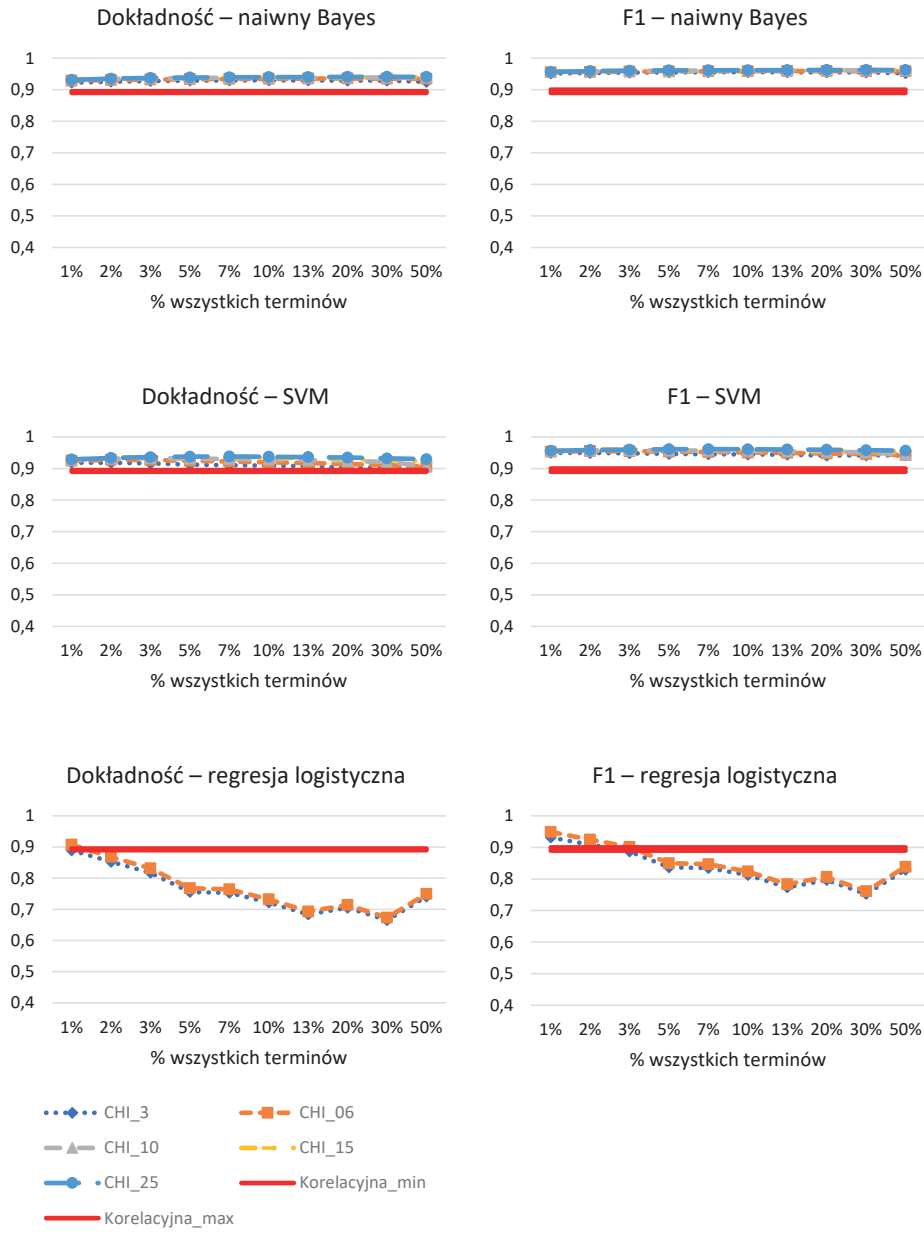
Wykres 6. Jakość klasyfikacji dla zbioru *bank* (metody KOR oraz CHI).

Źródło: obliczenia własne.



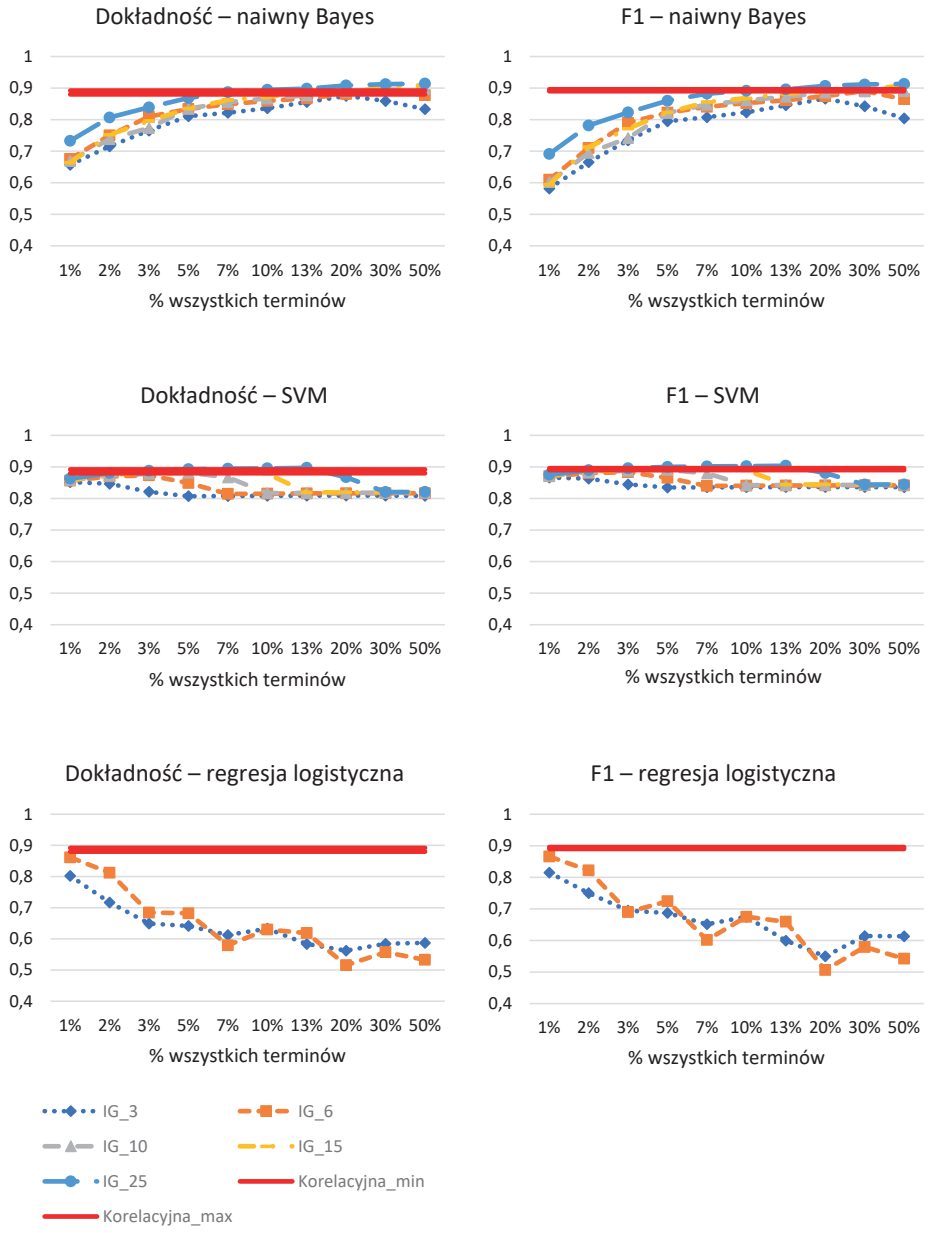
Wykres 7. Jakość klasyfikacji dla zbioru *esklepy* (metody KOR oraz IG).

Źródło: obliczenia własne.



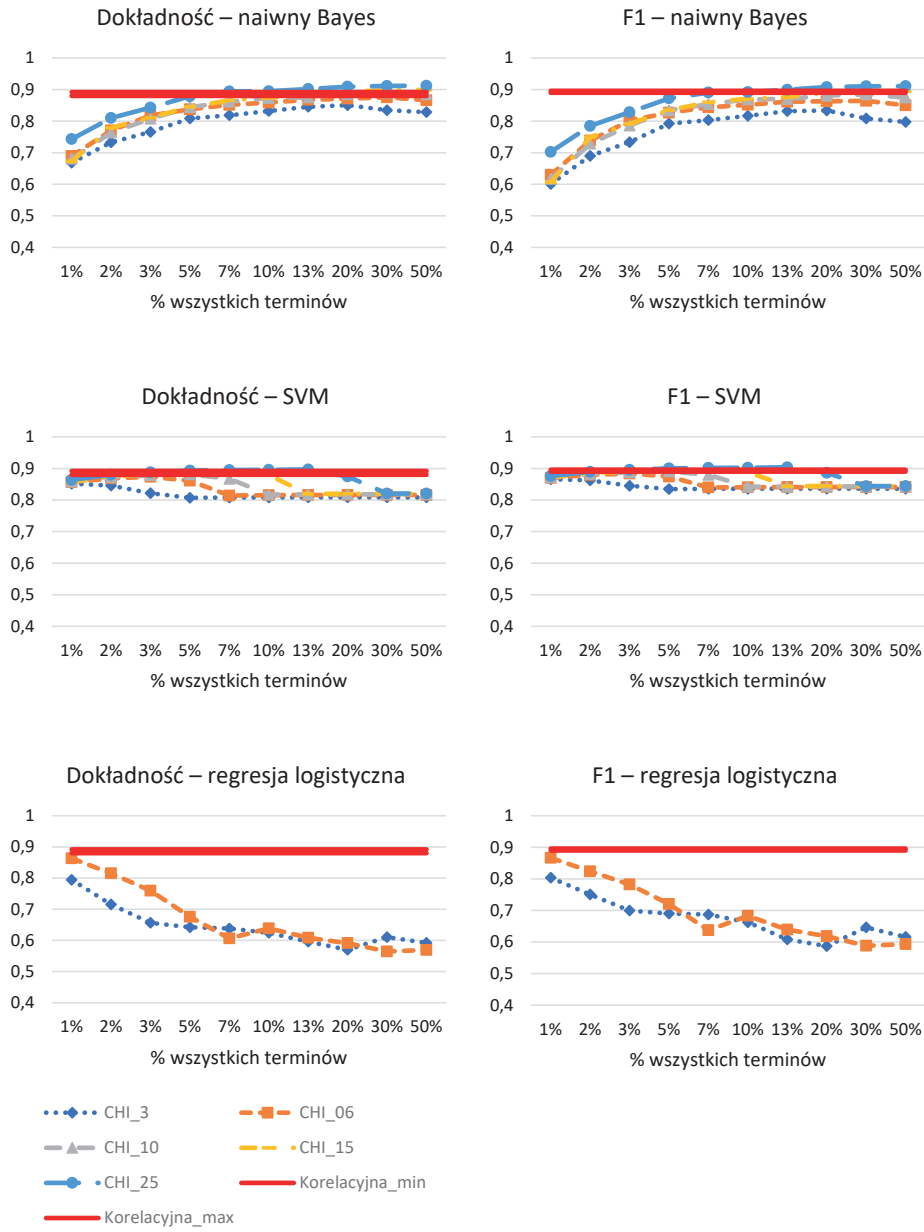
Wykres 8. Jakość klasyfikacji dla zbioru *esklepy* (metody KOR oraz CHI).

Źródło: obliczenia własne.



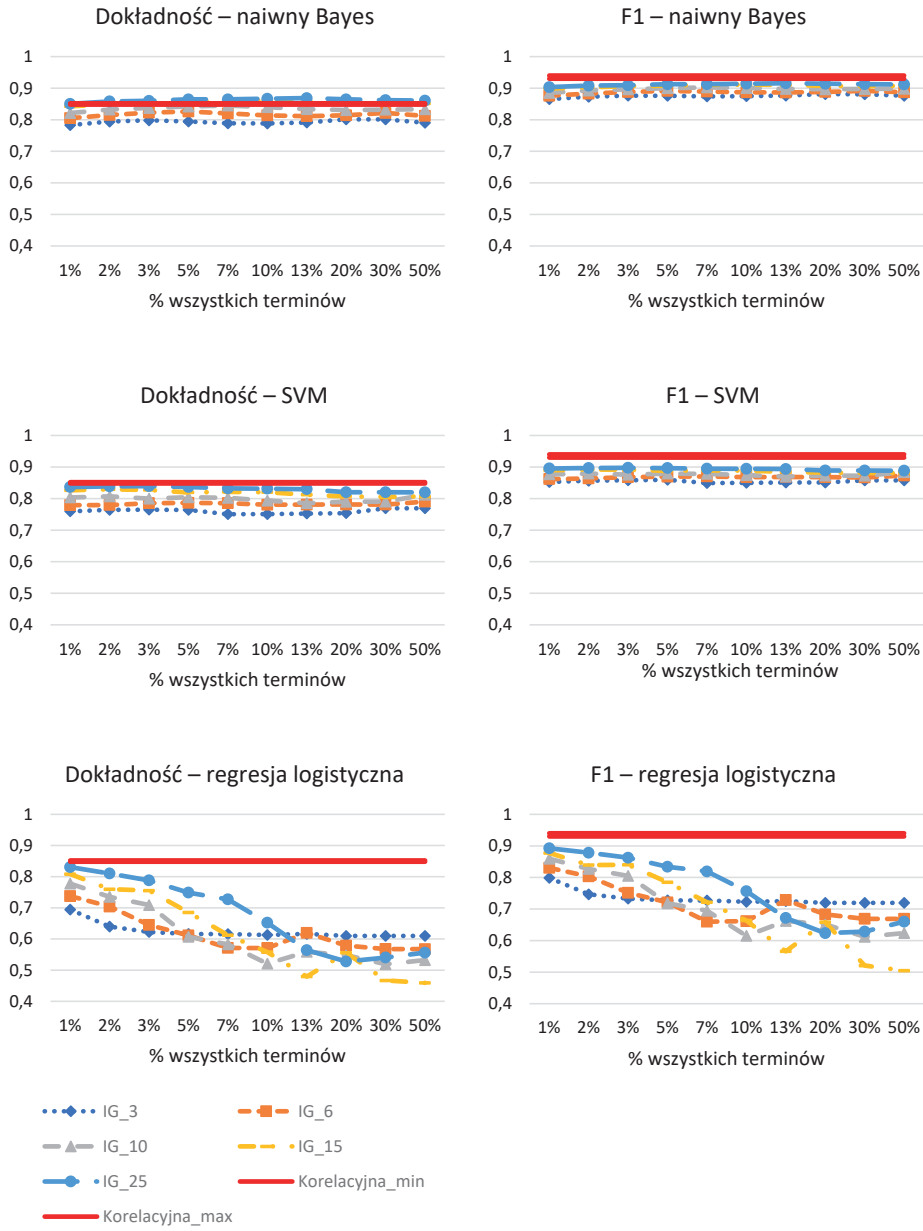
Wykres 9. Jakość klasyfikacji dla zbioru *esklepyzbil* (metody KOR oraz IG).

Źródło: obliczenia własne.



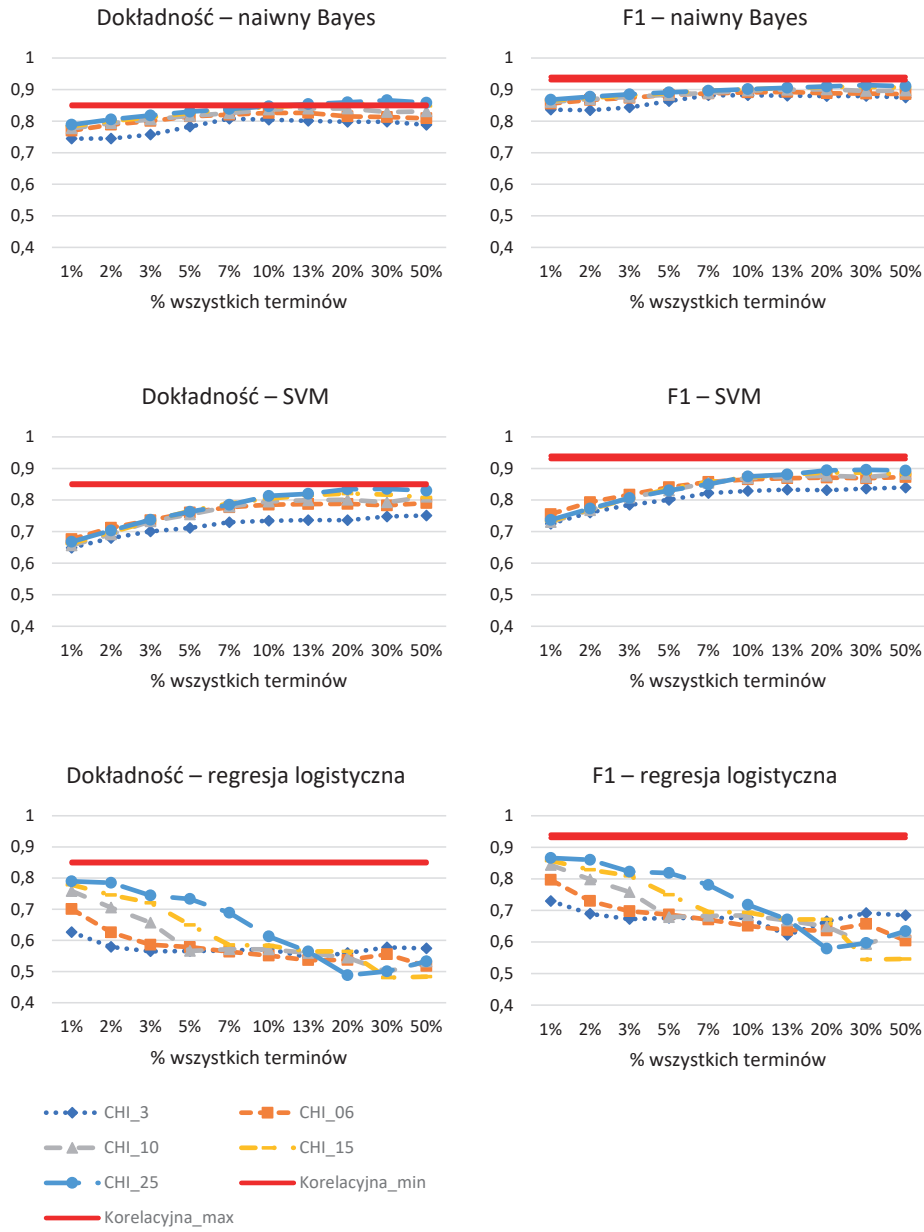
Wykres 10. Jakość klasyfikacji dla zbioru *esklepyzbil* (metody KOR oraz CHI).

Źródło: obliczenia własne.



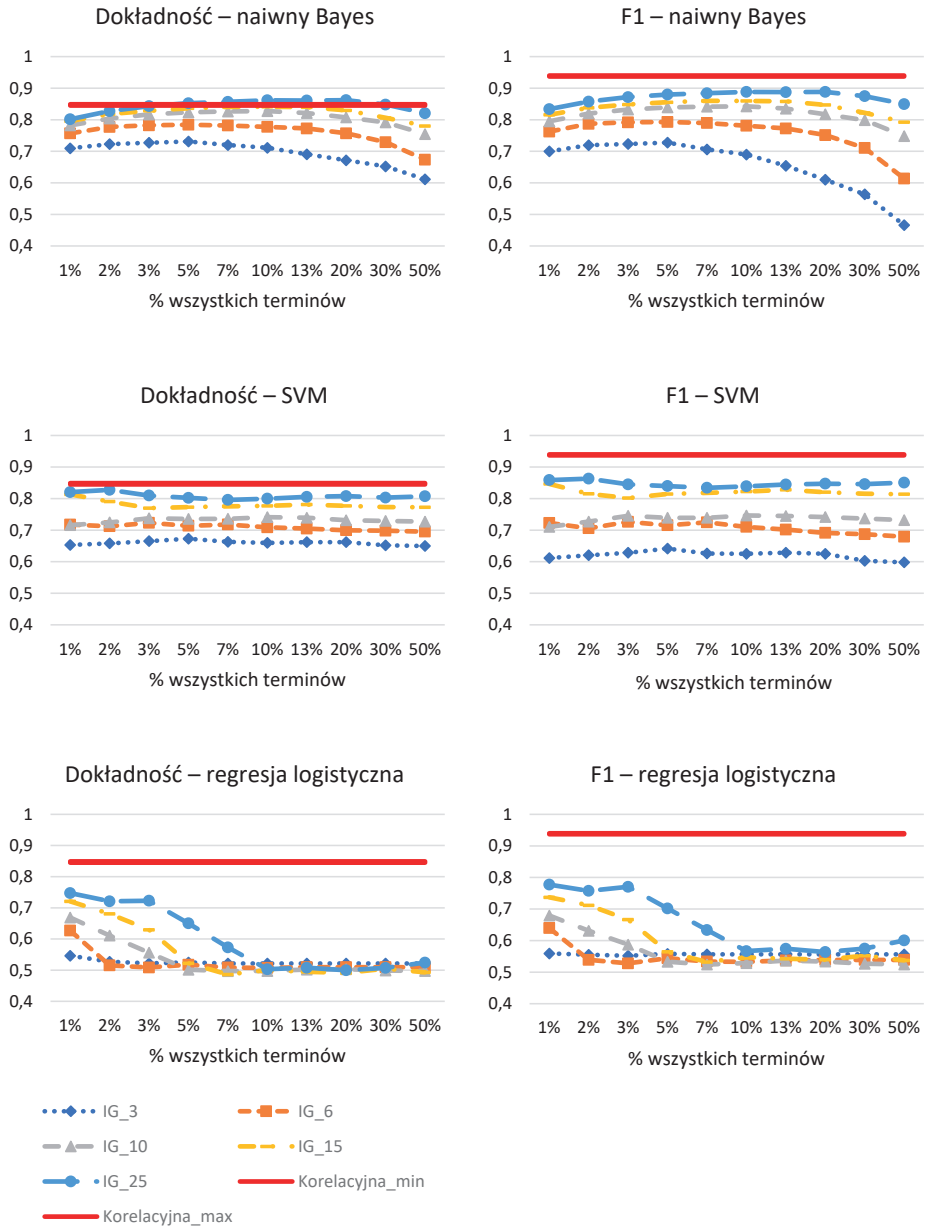
Wykres 11. Jakość klasyfikacji dla zbioru *hotele* (metody KOR oraz IG).

Źródło: obliczenia własne.



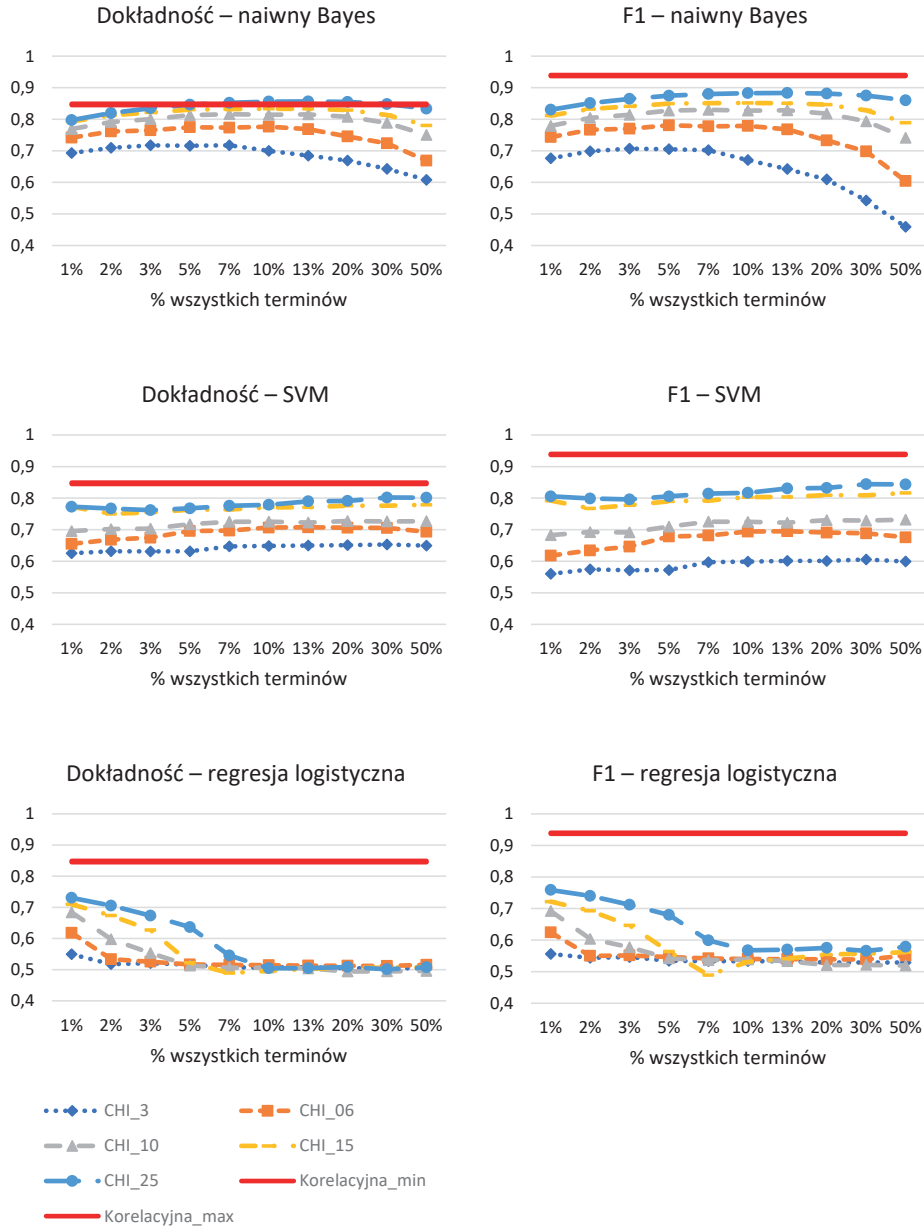
Wykres 12. Jakość klasyfikacji dla zbioru *hotele* (metody KOR oraz CHI).

Źródło: obliczenia własne.



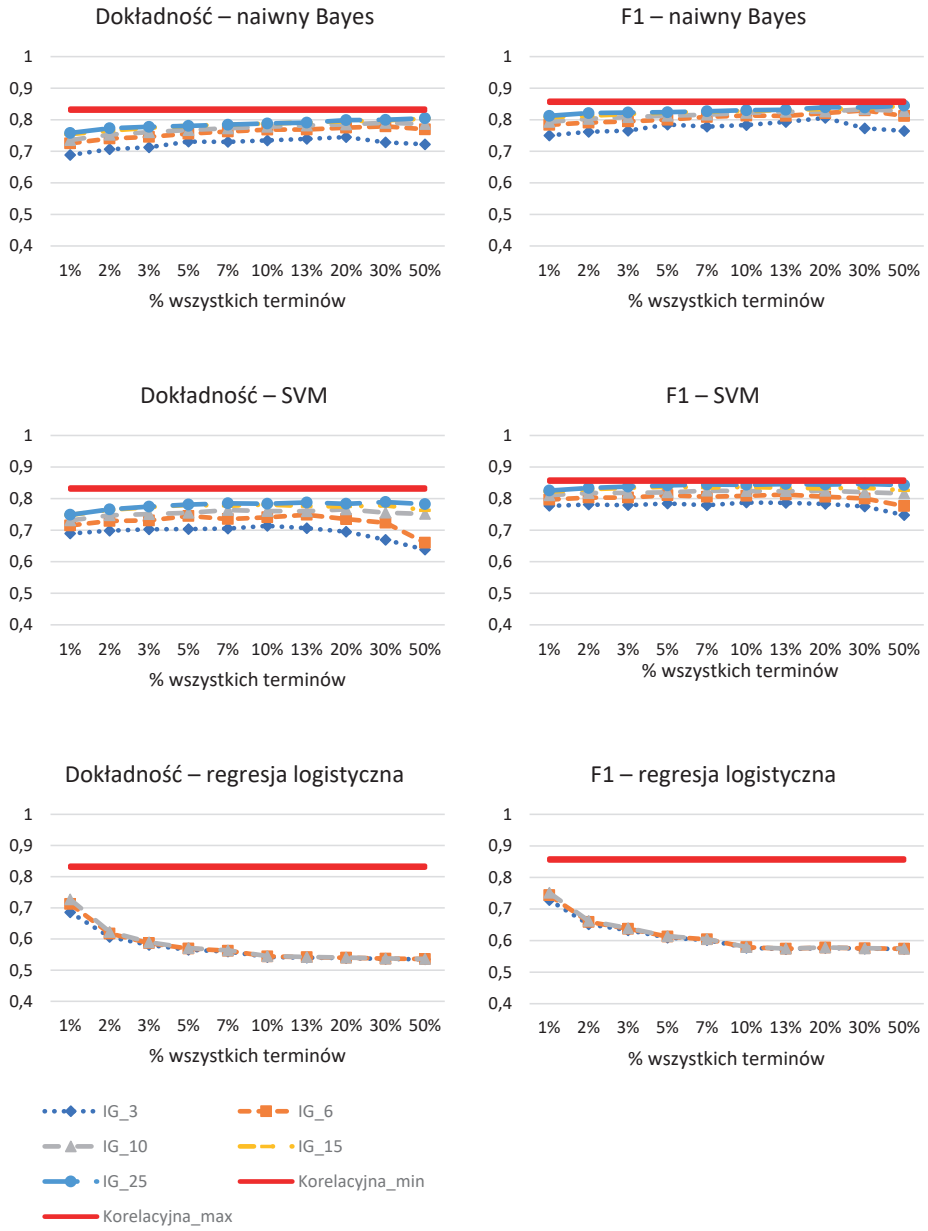
Wykres 13. Jakość klasyfikacji dla zbioru *hotelezbil* (metody KOR oraz IG).

Źródło: obliczenia własne.



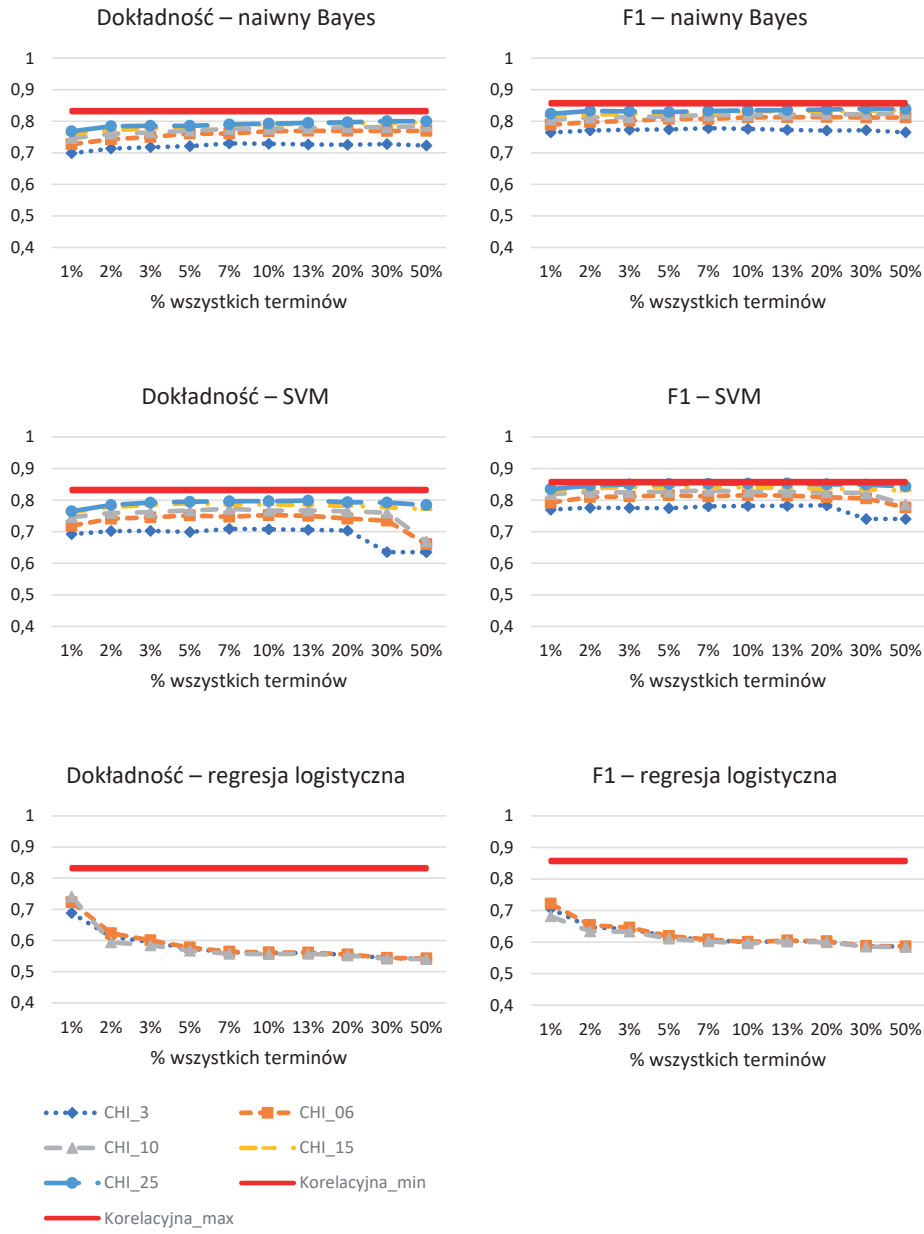
Wykres 14. Jakość klasyfikacji dla zbioru *hotelezbil* (metody KOR oraz CHI).

Źródło: obliczenia własne.



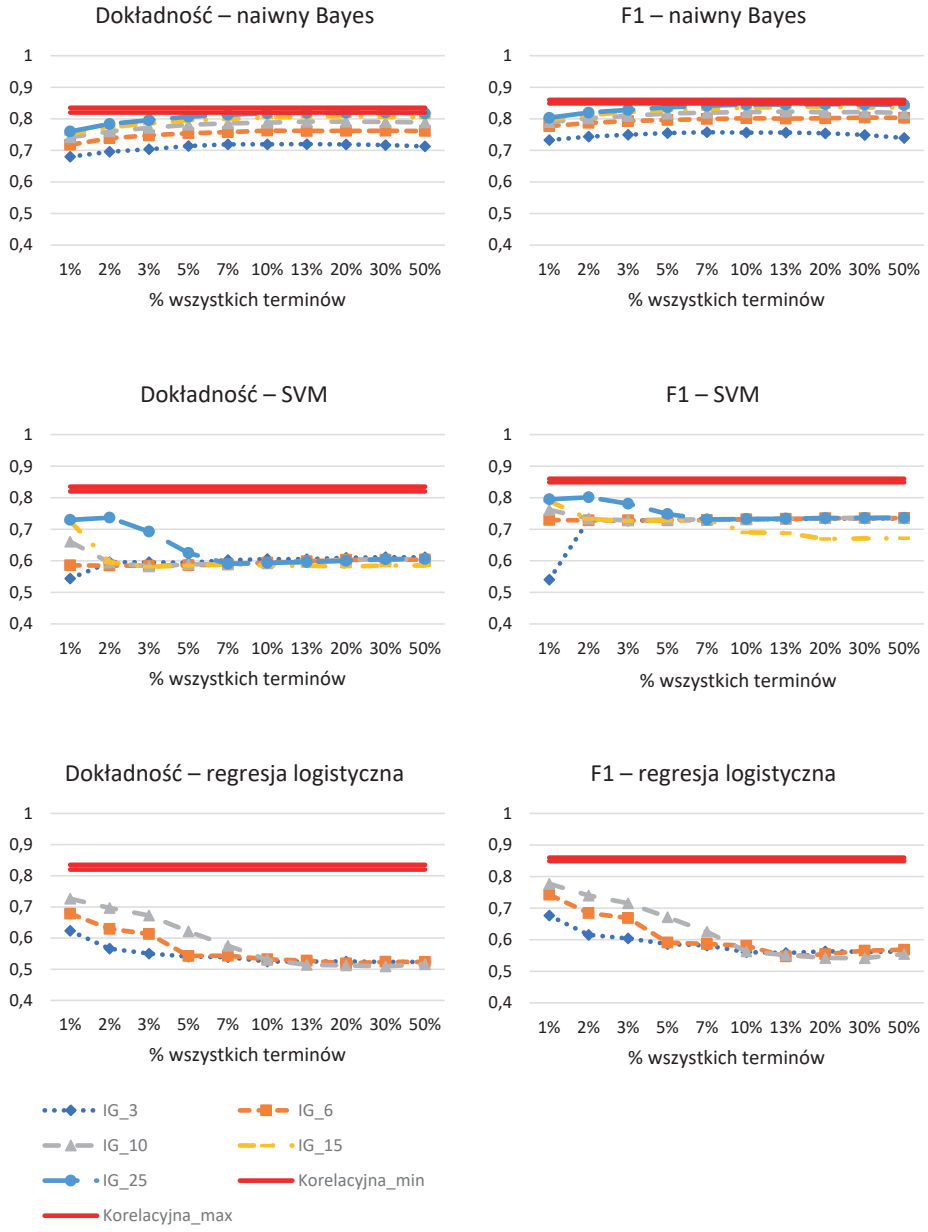
Wykres 15. Jakość klasyfikacji dla zbioru *ksiazki* (metody KOR oraz IG).

Źródło: obliczenia własne.



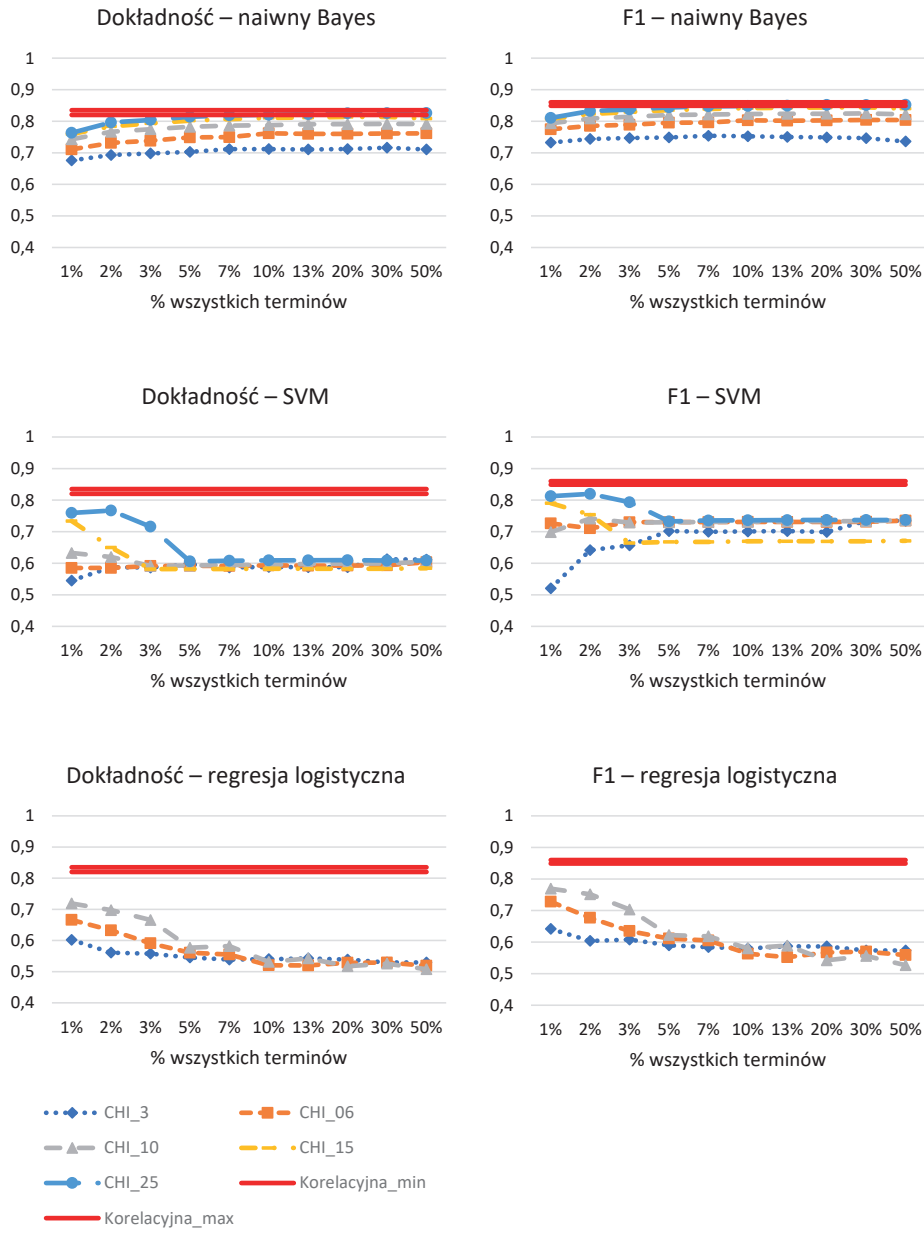
Wykres 16. Jakość klasyfikacji dla zbioru *ksiazki* (metody KOR oraz CHI).

Źródło: obliczenia własne.



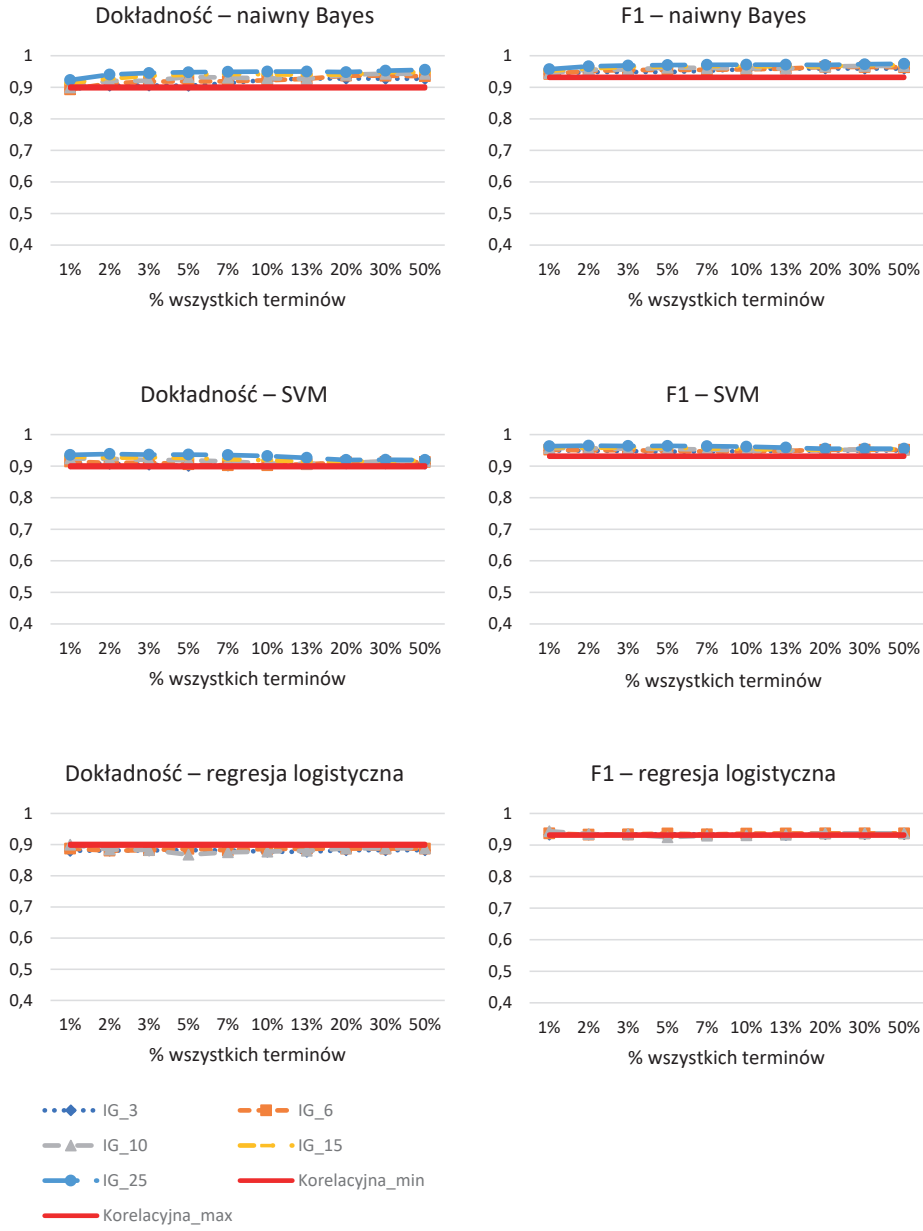
Wykres 17. Jakość klasyfikacji dla zbioru *ksiazkizbil* (metody KOR oraz IG).

Źródło: obliczenia własne.



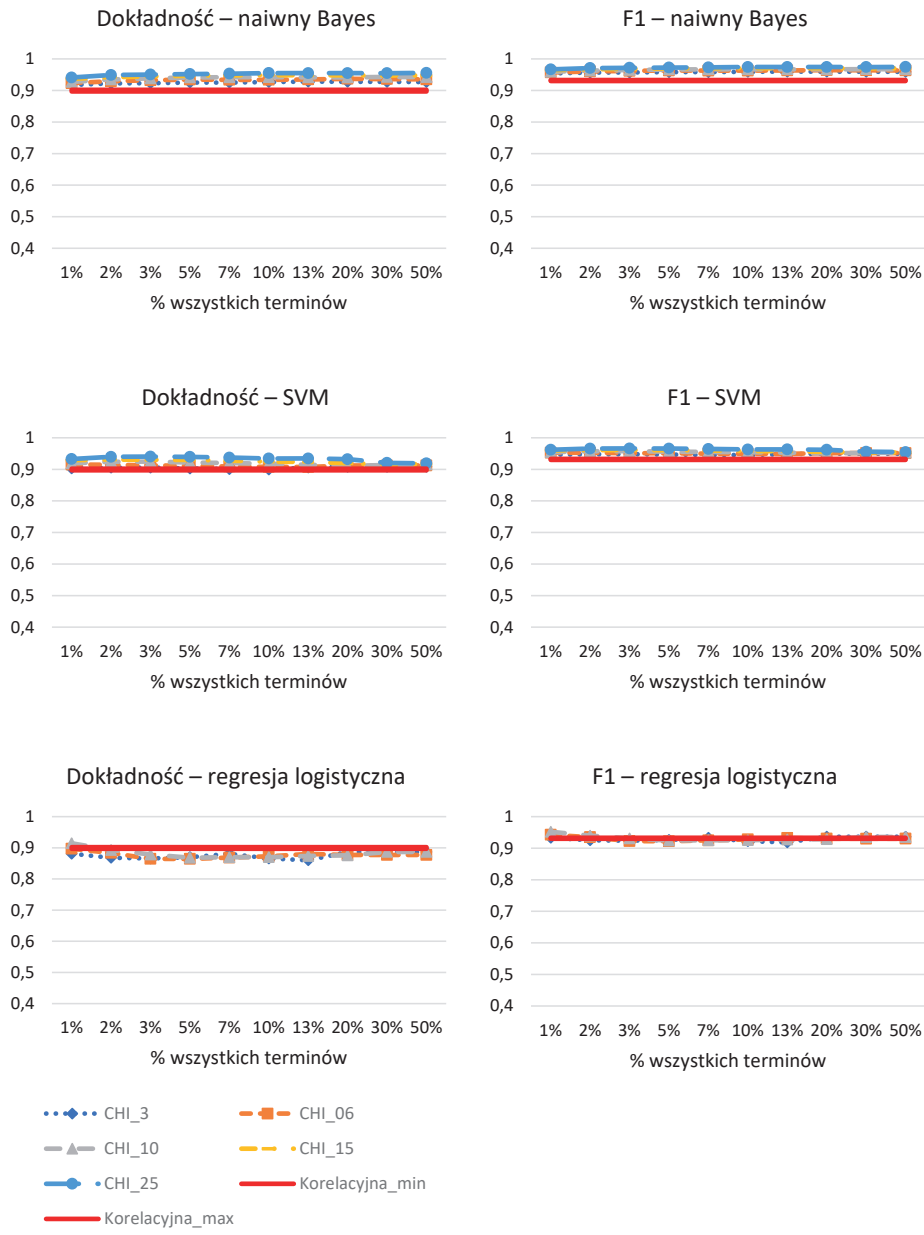
Wykres 18. Jakość klasyfikacji dla zbioru *ksiazkizbil* (metody KOR oraz CHI).

Źródło: obliczenia własne.



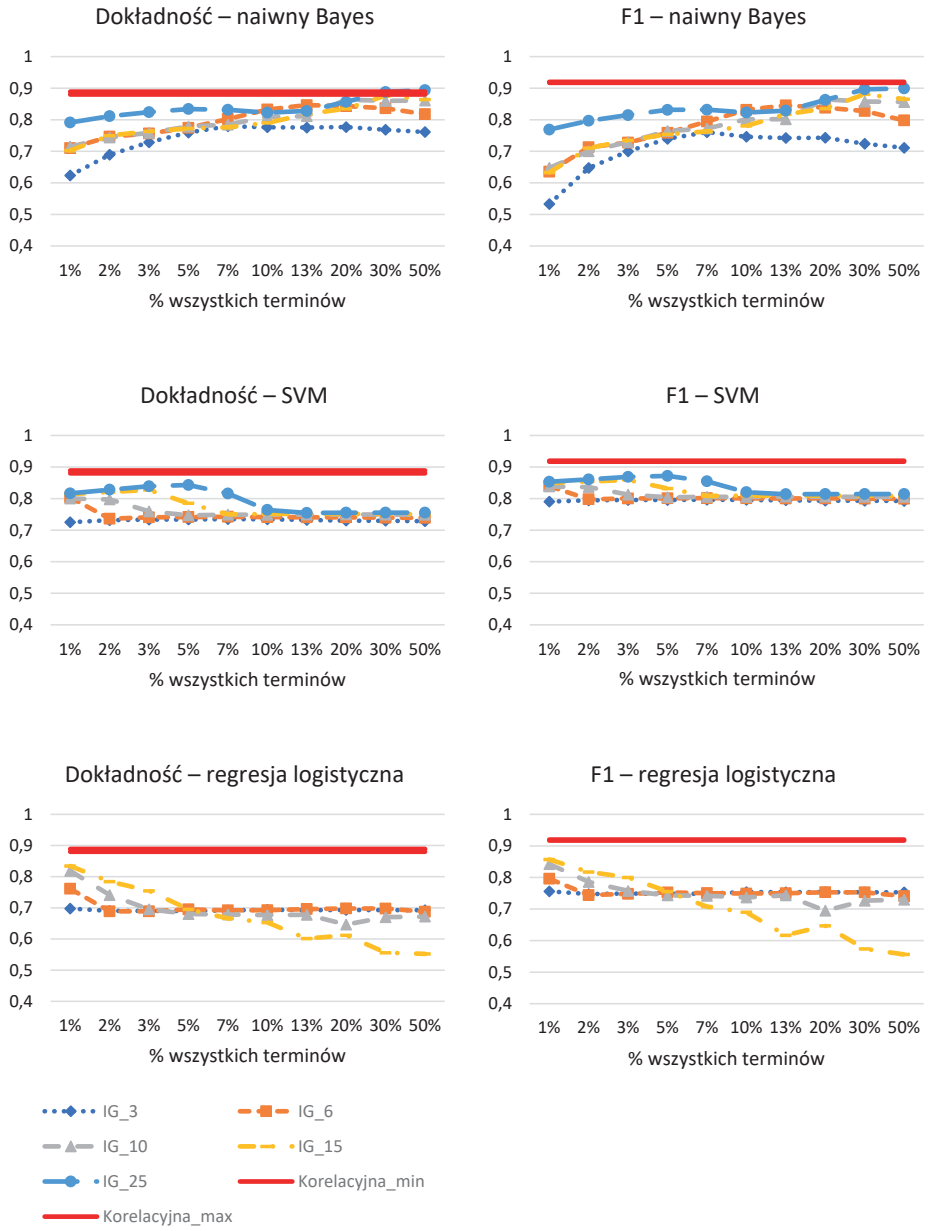
Wykres 19. Jakość klasyfikacji dla zbioru *kurier* (metody KOR oraz IG).

Źródło: obliczenia własne.



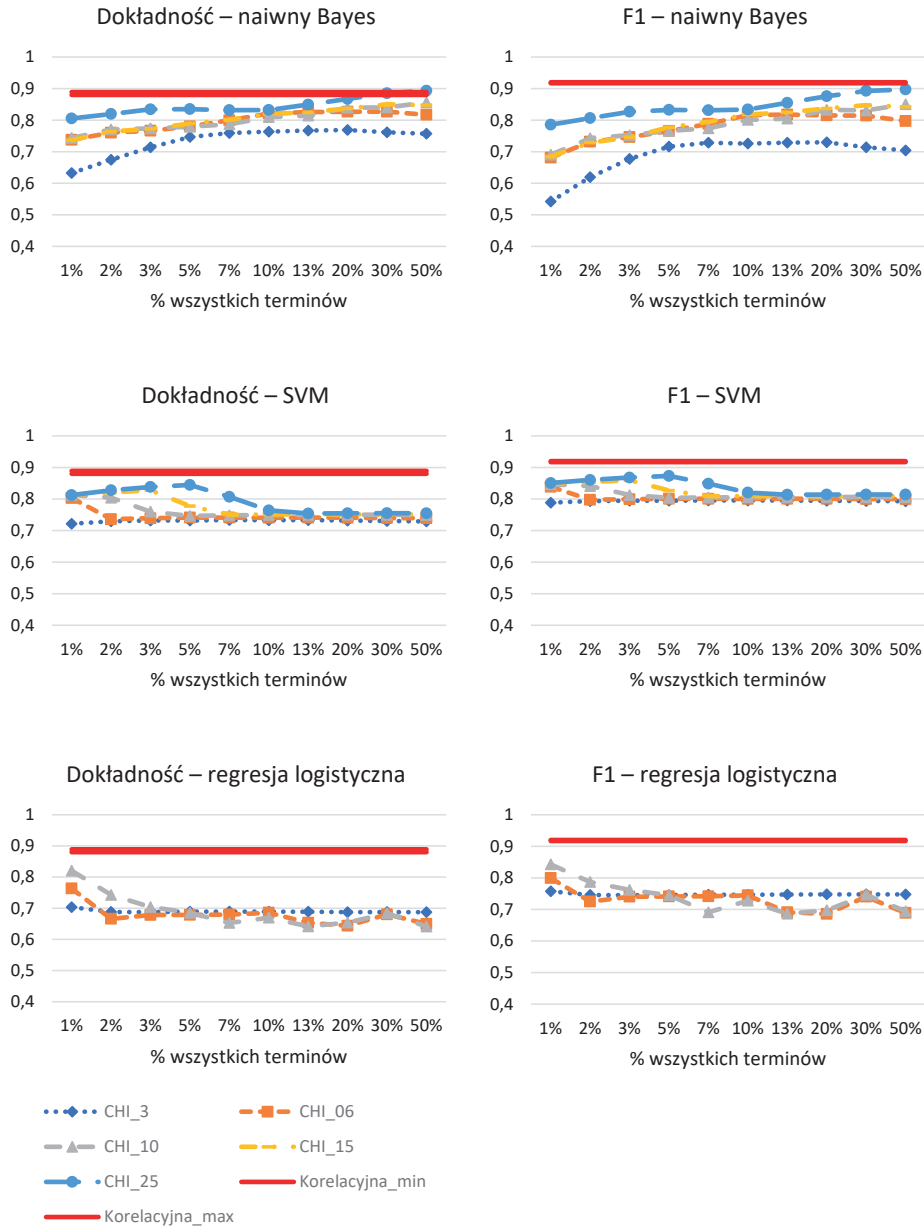
Wykres 20. Jakość klasyfikacji dla zbioru *kurier* (metody KOR oraz CHI).

Źródło: obliczenia własne.



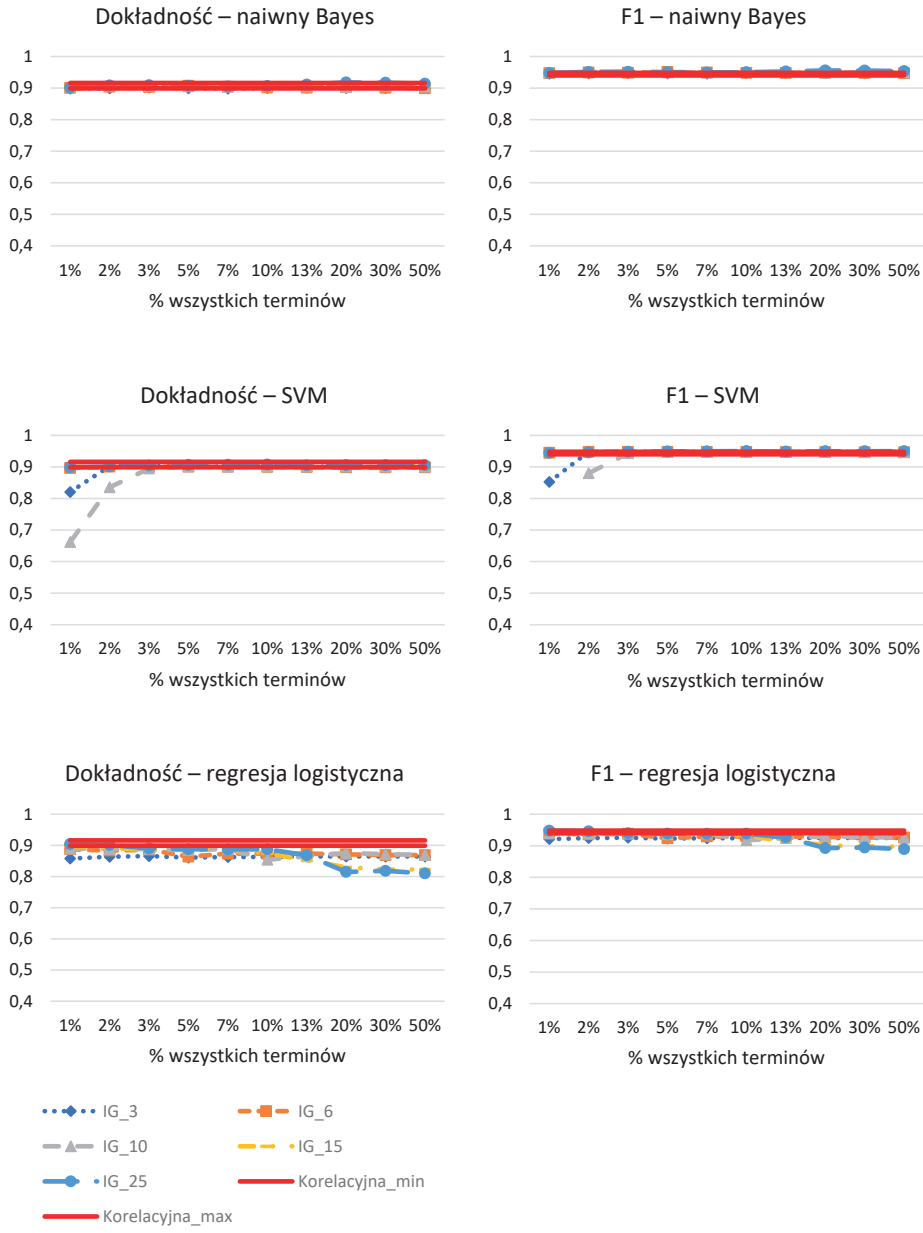
Wykres 21. Jakość klasyfikacji dla zbioru *kurierzbil* (metody KOR oraz IG).

Źródło: obliczenia własne.



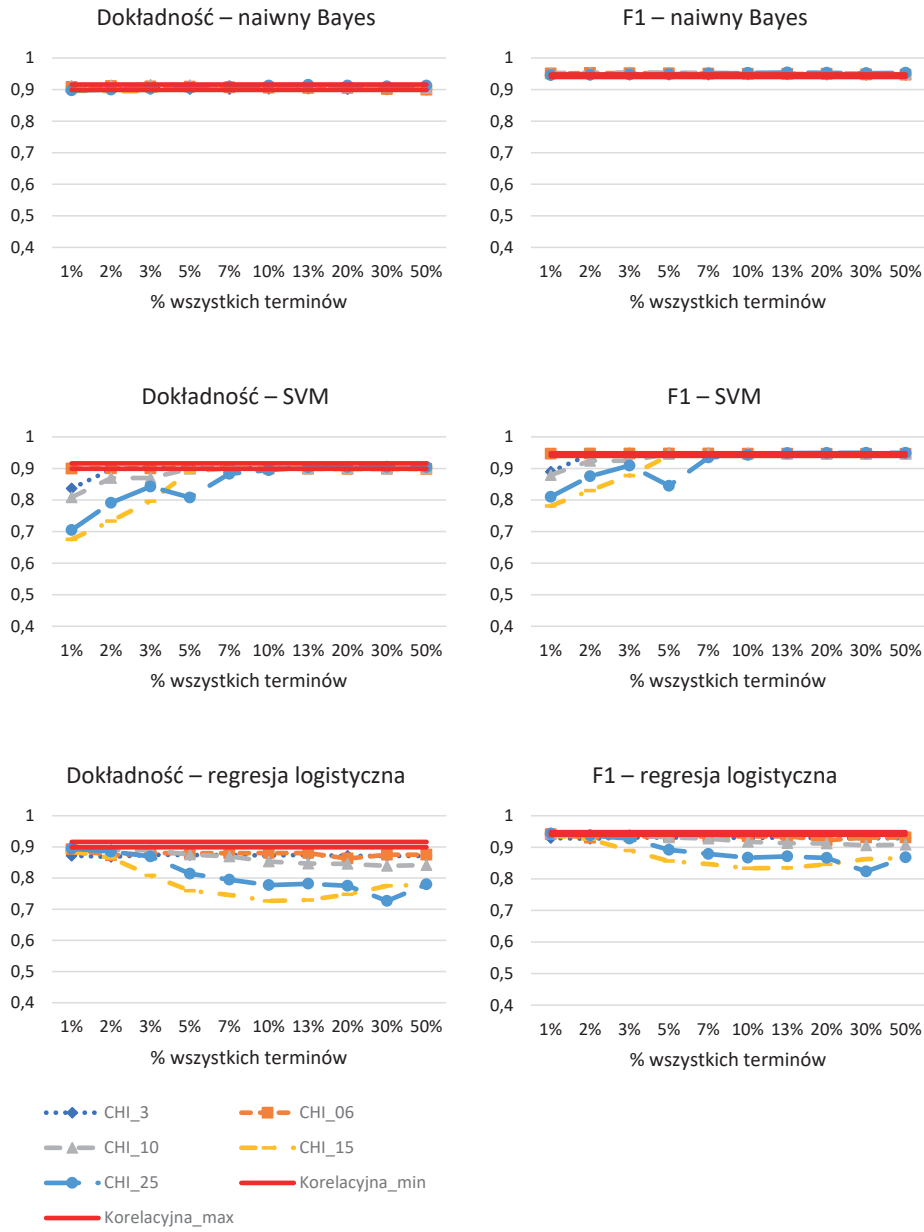
Wykres 22. Jakość klasyfikacji dla zbioru *kurierzbil* (metody KOR oraz CHI).

Źródło: obliczenia własne.



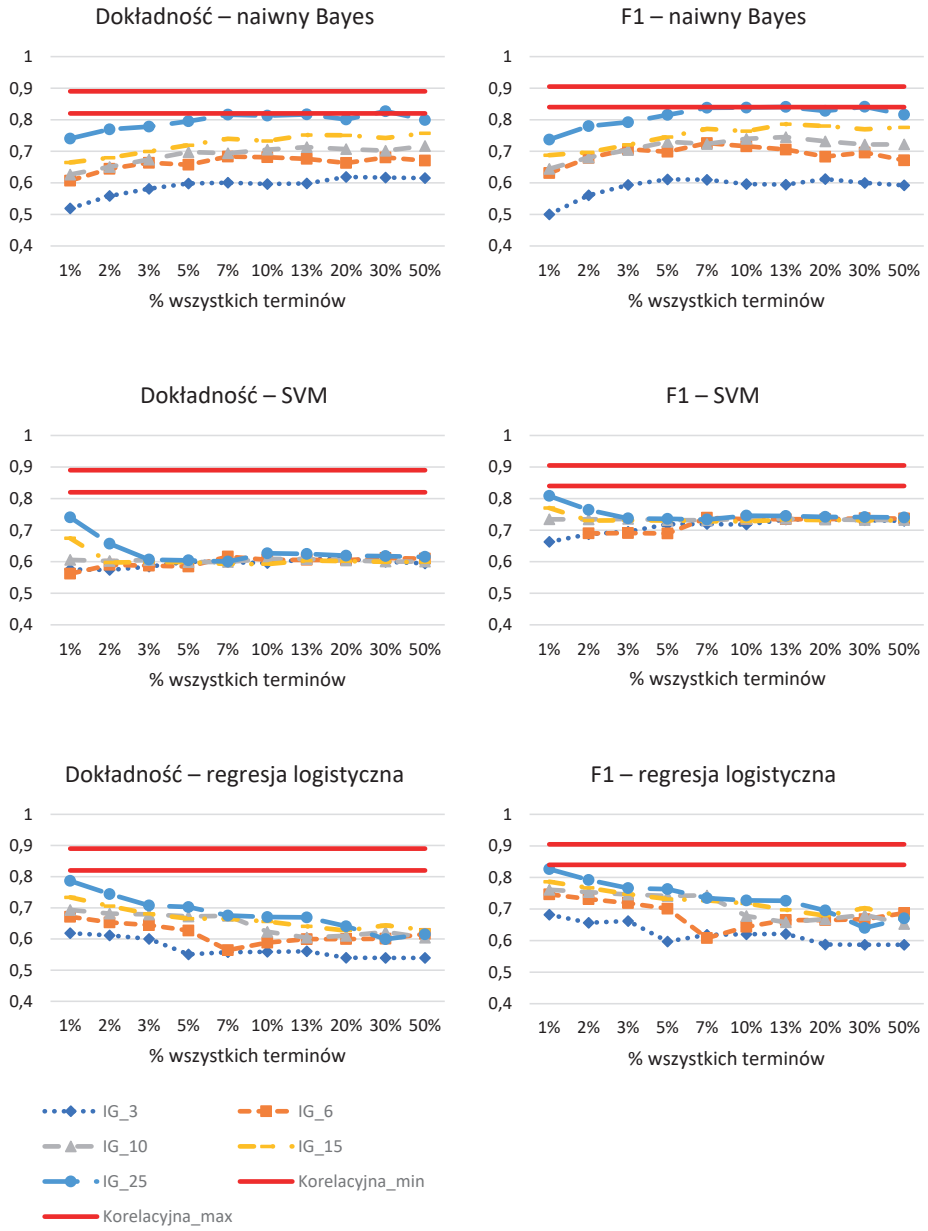
Wykres 23. Jakość klasyfikacji dla zbioru *perfumy* (metody KOR oraz IG).

Źródło: obliczenia własne.



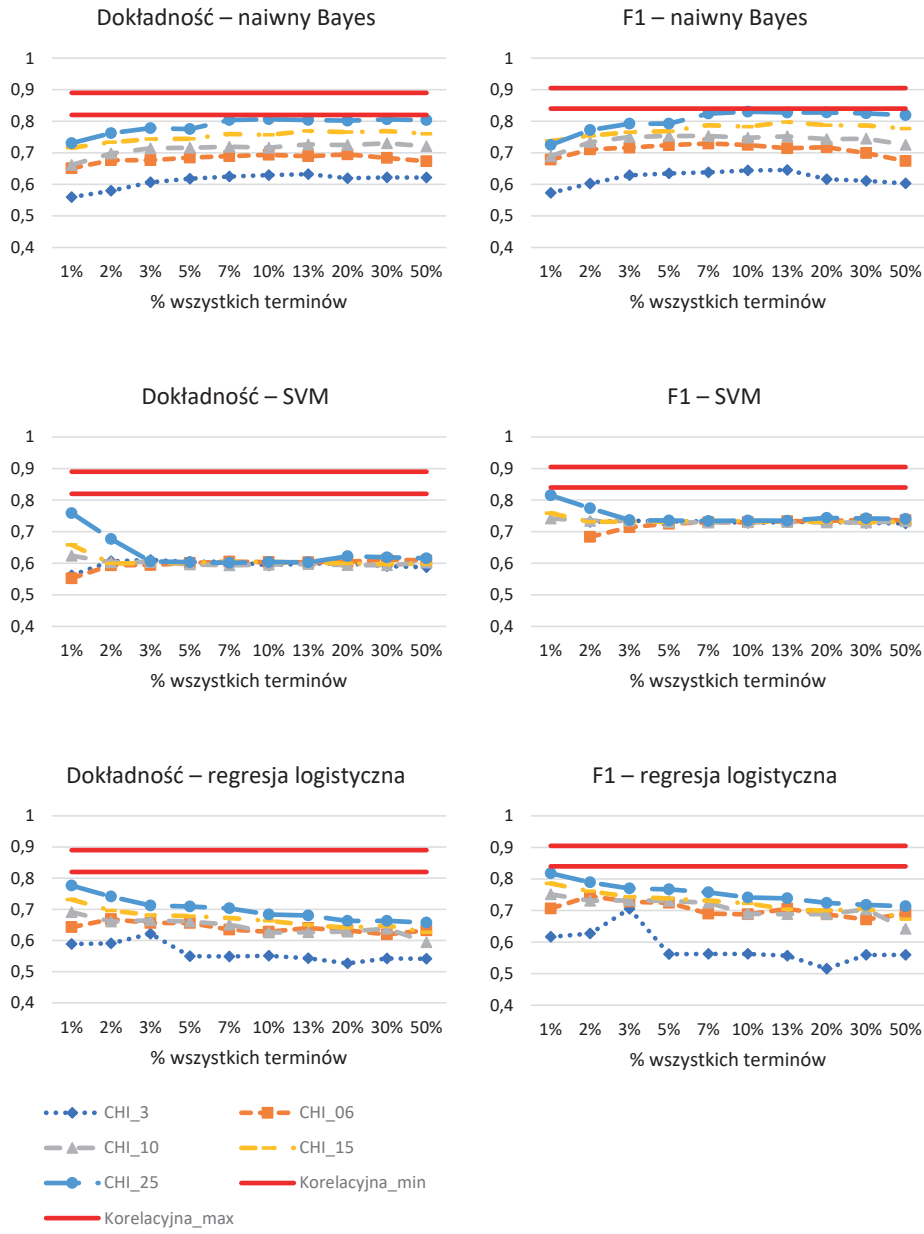
Wykres 24. Jakość klasyfikacji dla zbioru *perfumy* (metody KOR oraz CHI).

Źródło: obliczenia własne.



Wykres 25. Jakość klasyfikacji dla zbioru *perfumyzbil* (metody KOR oraz IG).

Źródło: obliczenia własne.



Wykres 26. Jakość klasyfikacji dla zbioru *perfumyzbil* (metody KOR oraz CHI).

Źródło: obliczenia własne.

tj. zbioru *esklepy*, zaś wykresy 9 i 10 jego wersji *esklepyzbil* z wyrównaną liczebnością obu klas. W przypadku dużego zbioru *esklepy* metoda KOR wypadła odrobinę słabiej od konkurencji. Ten wynik nie dziwi, w przypadku dużej liczby terminów spada efektywność metody opartej na korelacjach pomiędzy pojedynczymi terminami. Z kolei w przypadku zbioru *esklepyzbil* metody tradycyjne znowu miały trudności z osiągnięciem dobrych wyników. Klasyfikator NB okazał się dość stabilny i jego wyniki rosły wraz ze wzrostem odsetka wykorzystywanych początkowych terminów, ale dla małych odsetków miała słabsze wyniki od metody KOR. Klasyfikator SVM był dość niestabilny, uzależniony od rozmiaru zbioru uczącego i tylko dla dużego rozmiaru tego zbioru (25%) osiągał wyniki lepsze od metody KOR. Wykresy 11 i 12 dotyczą zbioru *hotele*, zaś wykresy 13 i 14 jego wersji *hotelezbil* z wyrównaną liczebnością obu klas. Dla zbioru *hotele* metoda KOR spisała się nieco lepiej od konkurencji, która dorównała jej tylko dla dużego rozmiaru zbioru uczącego (25%). Ciekawe jest to, że jest to jedyny zbiór, dla którego metoda CHI daje inne wyniki od metody SVM. Trochę lepiej wypadła metoda IG. Wśród klasyfikatorów ponownie lepszy i stabilniejszy był klasyfikator NB. Konkurencyjne metody dorównują metodzie KOR tylko dla dużych rozmiarów zbioru uczącego (25%) i dla większego niż kilkuprocentowy wykorzystania listy wszystkich terminów. Taka sytuacja miała miejsce dla zbioru *hotele*. Dla zbioru *hotelezbil* metoda KOR była lepsza. Wykresy 15 i 16 odnoszą się do zbioru *ksiazki*, zaś wykresy 17 i 18 do jego wersji *ksiazkizbil* z wyrównaną liczebnością obu klas. Dla zbioru *ksiazki* metoda KOR spisała się nieznacznie lepiej od konkurencji, która nie dorównała jej. Jeszcze większą przewagę nad konkurencją metoda KOR uzyskała na zbiorze *ksiazkizbil*, choć różnice pomiędzy oboma zbiorami nie są tak duże, jak w przypadku innych par zbiorów. Ponownie klasyfikator SVM był znacznie mniej stabilny od klasyfikatora NB. Klasyfikator NB osiągnął stabilne wyniki na obu zbiorach, dorównując metodzie KOR przy największych rozmiarach zbioru uczącego (25%). Natomiast klasyfikator SVM wykazał się bardzo dziwnymi wynikami, osiągając dla początkowych, kilkuprocentowych wartości odsetka wykorzystywanych terminów wyniki lepsze lub znacznie lepsze od wyników dla kilkunastoprocentowych odsetków wykorzystanych terminów. Taka sytuacja miała miejsce we współpracy z obiema metodami porządkującymi terminy – zarówno IG, jak i CHI. Wykresy 19 i 20 odnoszą się do zbioru *kurier*, zaś wykresy 21 i 22 do jego wersji *kurierzbil* z wyrównaną liczebnością obu klas. Dla zbioru *kurier* metoda KOR spisała się nieznacznie słabiej od konkurencji, zaś dla zbioru *kurierzbil* nieznacznie lepiej. Ten zbiór to jedyny w całym badaniu, dla którego klasyfikator NB wykazał pewną niestabilność względem odsetka wykorzystanych terminów. Taka sytuacja miała miejsce we współpracy z obiema metodami porządkowania terminów – zarówno IG, jak i CHI. Klasyfikator SVM znów zwracał bardzo niestabilne wyniki względem odsetka wykorzystywanych terminów, ale tylko na zbiorze *kurierzbil*, na zbiorze *kurier* był bardzo stabilny. Wykresy 23 i 24 odnoszą się do zbioru *perfumy*, zaś wykresy 25 i 26 do jego wersji

perfumyzbil z wyrównaną liczebnością obu klas. Wynikiem godnym odnotowania jest to, że w całym badaniu są to jedyne zbiory, na których w miarę poprawnie w roli klasyfikatora spisała się regresja logistyczna. Przyczyną jest zapewne to, że relacja liczebności terminów do dokumentów jest tu najniższa. Regresja logistyczna osiągnęła prawie taką jakość, jak klasyfikator SVM i nieco gorszą od klasyfikatora NB. Po raz kolejny klasyfikator SVM okazał się bardzo niestabilny względem wykorzystywanego odsetka początkowych terminów. Ta wada miała miejsce we współpracy z metodą CHI porządkowania terminów i uwidoczniła się szczególnie dla większych rozmiarów zbioru uczącego (15% i 25%). Metoda KOR na zbiorze *perfumy* spisała się na równi z klasyfikatorem NB, natomiast na zbiorze *perfumyzbil* wypadła lepiej od pozostałych klasyfikatorów.

Podsumowując wyniki uzyskane przez porównywane metody na wszystkich zbiorach i odnosząc się do drugiego celu monografii, czyli do oceny jakości klasyfikacji dla zbiorów tekstów polskojęzycznych, należy stwierdzić, że, porównując wyniki intuicyjnie, jakość jest podobna do jakości uzyskiwanej przez te same metody dla tekstów anglojęzycznych. Mamy na myśli to, że dla łatwiejszych zbiorów jakość (obu miar) oscyluje w granicach 0,9, natomiast dla trudniejszych zbiorów jest, *ceteris paribus*, o kilkanaście setnych niższa. Wszystkie metody lepiej radzą sobie w przypadku zbiorów niezbilansowanych, czyli tam, gdzie liczebności klas są wysoko zróżnicowane. Znacznie trudniejsza jest klasyfikacja wtedy, gdy liczebności obu klas są mniej więcej jednakowe.

Odnosząc się do uzależnienia jakości klasyfikacji od zastosowanego klasyfikatora, należy stwierdzić, że to uzależnienie jest dość mocne. Zdecydowanie najsłabszym klasyfikatorem okazała się regresja logistyczna. Przyczyną tak słabych wyników jest zapewne mała liczba klasyfikowanych obiektów (dokumentów) w stosunku do liczby zmiennych (terminów). Jeśli pominiemy w rozważaniach regresję logistyczną, to okaże się, że pozostałe dwa klasyfikatory również zachowują się w niektórych przypadkach inaczej, tzn. dla niektórych zbiorów lepszy był klasyfikator bayesowski, zaś dla innych SVM. Znacznie częściej jednak lepsze wyniki uzyskał klasyfikator NB. W tym kontekście ważne jest też uzależnienie jakości od odsetka wykorzystywanych terminów początkowych z list uporządkowanych terminów. Trudno w literaturze przedmiotu znaleźć jakąś rozsądną, ogólnego zastosowania metodę, która wskazywałaby, jak ustalać ten odsetek. W przeprowadzonym badaniu okazało się, że odsetek wykorzystywanych terminów ma bardzo istotny wpływ na jakość klasyfikacji. Z pewnością klasyfikacja jest dłuższa czasowo, gdy ten odsetek rośnie, ale dość nieoczekiwanie okazało się, że nie musi się to wiązać z lepszymi wynikami. Dość stabilny pod tym względem był klasyfikator NB – w jego przypadku jakość na ogół rosła wraz ze wzrostem odsetka. Natomiast klasyfikator SVM w kilku przypadkach był bardzo niestabilny i, co więcej, wyraźnie tracił jakość dla większych (15% i 25%) odsetków wszystkich terminów wykorzystywanych w klasyfikacji. Uogólniając wyniki z całego badania, należy stwierdzić, że z dwóch klasyfikatorów benchmarkowych lepszy okazał się naiwny klasyfikator bayesowski.

Odnosząc się do uzależnienia efektywności klasyfikacji od metody filtrowania zbioru terminów należy stwierdzić, że to uzależnienie jest słabe. Zdecydowanie najistotniejszy jest wybór klasyfikatora, natomiast metoda filtrowania terminów ma drugorzędne znaczenie. Dla obu użytych metod filtrujących (IG oraz CHI) ten sam klasyfikator uzyskuje podobne wyniki, choć, biorąc pod uwagę stabilność względem odsetka wykorzystywanych terminów, trochę lepiej spisała się metoda IG. Uogólniając wnioski dla całego badania, należy stwierdzić, że niestabilność najbardziej uwidacznia się w przypadku mniej więcej jednakowych liczebności obu klas (zbiory zbilansowane). Jakość rośnie bądź spada wraz ze zmianą odsetka użytych terminów, niekiedy (zwłaszcza dla metody SVM) w niewytłumaczalny sposób gwałtownie załamuje się dla większej liczby terminów, niekiedy jest bardzo stabilna lub wykazuje niewielki wzrost. Ta wada była najbardziej widoczna na zbiorach zbilansowanych, chociaż dla niezbilansowanych również występowała.

Odnosząc się do efektywności obu metod benchmarkowych w zależności od relacji liczebności zbioru uczącego do liczebności całego zbioru, należy stwierdzić, że dla zbiorów niezbilansowanych, na ogół, to uzależnienie istniało, ale nie było silne. Jakość klasyfikacji rośnie bardzo powoli wraz ze wzrostem liczebności zbioru uczącego i dla najmniejszych rozmiarów zbioru uczącego (3%) ma relatywnie wysokie wartości. Niemniej jednak w niektórych przypadkach zdarzyło się, że różnice pomiędzy wynikami dla dużych rozmiarów zbioru uczącego, a tymi dla małych rozmiarów, były rzędu kilku procent. Inaczej wygląda sytuacja dla zbiorów zbilansowanych. Jakość klasyfikacji była wtedy niestabilna, przy czym trudno nawet stwierdzić, dla jakiego odsetka wykorzystanych terminów rosła (wraz ze wzrostem rozmiaru zbioru uczącego) szybciej, a dla jakiego wolniej. Dość silny był w tym kontekście wpływ rodzaju klasyfikatora. Klasyfikator SVM niekiedy nie radził sobie zupełnie ze zbiorami uczącymi większych rozmiarów.

Oceniając nową metodę KOR, należy stwierdzić, że:

1) metoda jest samodzielna w tym sensie, że klasyfikuje dokumenty ze zbioru testowego bez wspomagania jej metodami selekcjonującymi cechy;
2) jakość metody nie zależy od odsetka terminów wybranych do badania, bo pracuje na wszystkich terminach;
3) metoda bardzo dobrze spisała się dla najmniejszych rozmiarów zbiorów uczących – efektywność dla 3% zbiorów uczących jest prawie taka sama, jak dla większych zbiorów uczących;
4) jakość metody jest, na ogół, wyższa od jakości metod porównywanych, zwłaszcza dla najmniejszych 3% zbiorów uczących; wyjątkiem od tej reguły są tylko nieliczne zbiory wysoko niezbilansowane; dla zbiorów zbilansowanych przewaga nad konkurencją jest największa;
5) metoda nie wymaga zadawania wartości startowych, trudnych do ustalenia, dla żadnych parametrów warunkujących jej działanie; wyniki są niemalże identyczne niezależnie od tego, czy dokumenty są klasyfikowane w grupach z krokiem $p = 10\%$ czy $p = 20\%$;

6) metoda posiada potencjał rozwojowy, może być dalej modyfikowana; można:
   a) aktualizować w kolejnych krokach zbiory słów z obu grup o znanym sentymencie, by włączać do tych grup słowa i terminy charakterystyczne dla danej tematyki, ale bardziej istotne dla separowalności klas od słów w tych grupach będących – metoda nabierze wtedy wyraźniejszych cech nauczania bez nadzoru,
   b) wprowadzać modyfikacje polegające na usuwaniu z obu grup słów, których wagi wskazują na to, że powinny być w innej grupie sentymentalnej od tej nominalnej; początkowe zbiory słów o zdecydowanym zabarwieniu sentymentalnym stają się wówczas tylko punktem startowym dla klasyfikacji dokumentów,
   c) opracowywać lepsze niż losowe klasyfikowanie niewielkiego odsetka dokumentów, które nie dadzą się sklasyfikować za pomocą miary (wzór 3.3) – to będzie przedmiotem dalszych badań,
   d) próbować zmieniać frakcję dokumentów pozytywnych do negatywnych na lepszą, tzn. bardziej reprezentatywną, co dla małych liczebności zbioru uczącego (tych, które najbardziej nas interesują) może mieć znaczenie,
   e) przekształcić w prosty sposób metodę klasyfikacyjną KOR w metodę klasyfikacyjną bez nadzoru; zauważmy, że zbiór uczący jest wykorzystywany tylko do aktualizacji miar separowalności i wag terminów, ale najistotniejszy wpływ na przyporządkowanie dokumentu do klasy mają miary skorelowania ze słowami o ustalonym sentymencie, w takiej sytuacji charakter klas jest ustalony i przekształcenie metody klasyfikacyjnej w metodę analizy skupień wymagałoby jedynie pominięcia miar separowalności i wag terminów;
7) zaproponowana metoda powinna działać też dla dużych korpusów tekstów, gdyż zbiory słów z przyjętego słownika sentymentu są stałe i czas pracy algorytmu zależy (liniowo) tylko od liczby terminów w korpusie dokumentów.

Zakończenie

Klasyfikacja polskojęzycznych dokumentów tekstowych ze względu na ich sentyment może przebiegać przy zastosowaniu tych samych metod, które są używane dla tekstów w innych językach, w szczególności w języku angielskim. Te metody osiągają podobną jakość zarówno dla języka polskiego, jak i dla angielskiego, mimo że oba języki są zupełnie różne w swej naturze. Metoda SVM w roli klasyfikatora nie okazała się na badanych zbiorach lepsza od naiwnego klasyfikatora bayesowskiego. Regresja logistyczna, polecana przez niektórych autorów, okazała się zupełnie nieprzydatna do klasyfikacji tekstów. Pomimo obszernego zbioru metod dostępnych w literaturze przedmiotu można w tej dziedzinie zaproponować nowe rozwiązania. Najlepsze znane metody nie są nieparametryczne, tzn. wymagają konieczności dostosowania wartości parametrów do konkretnego korpusu dokumentów. Przykładem nowego rozwiązania może być prosta w interpretacji, zaproponowana w niniejszej monografii metoda korelacyjno-częstościowa, praktycznie nieparametryczna, która osiąga bardzo dobre wyniki w porównaniu z metodami standardowymi. Można stąd wnioskować, że w porównaniu z innymi modyfikacjami metod standardowych te wyniki będą również dobre. Ponadto należy stwierdzić, że zaproponowana metoda jest szybka, gdyż wymagane są miary skorelowania każdego terminu tylko z kilkudziesięcioma innymi słowami. Uważamy, że są podstawy do tego, by twierdzić, że nowa metoda może być dalej rozwijana w kierunku ustalania sentymentu dokumentów bez konieczności korzystania ze zbioru uczącego. To będzie przedmiotem dalszych badań.

Załącznik

Tabela 2. Jakość klasyfikacji dla zbioru *apteki*, dla metody IG w pierwszym etapie.

Miara	% terminów	Zbiór uczący 3%				Zbiór uczący 6%				Zbiór uczący 10%				Zbiór uczący 15%				Zbiór uczący 25%			
		NB	SVM	GLM	KOR	NB	SVM	GLM	KOR	NB	SVM	GLM	KOR	NB	SVM	GLM	KOR	NB	SVM	GLM	KOR
Dokładność	1%	0,838	0,834	0,809	0,883	0,846	0,841	0,826	0,886	0,851	0,848	0,843	0,886	0,852	0,850	0,854	0,887	0,884	0,888	0,872	0,888
	2%	0,842	0,837	0,801	0,883	0,851	0,842	0,813	0,886	0,858	0,853	0,828	0,886	0,861	0,852	0,837	0,887	0,886	0,890	0,850	0,888
	3%	0,842	0,838	0,795	0,883	0,853	0,842	0,741	0,886	0,859	0,852	0,754	0,886	0,863	0,853	0,761	0,887	0,886	0,889	0,770	0,888
	5%	0,840	0,835	0,791	0,883	0,850	0,844	0,795	0,886	0,861	0,850	0,808	0,886	0,863	0,852	0,814	0,887	0,890	0,889	0,822	0,888
	7%	0,842	0,832	0,782	0,883	0,849	0,845	0,793	0,886	0,861	0,846	0,805	0,886	0,865	0,851	0,810	0,887	0,891	0,887	0,816	0,888
	10%	0,850	0,834	0,756	0,883	0,850	0,846	0,779	0,886	0,860	0,847	0,791	0,886	0,864	0,852	0,795	0,887	0,890	0,885	0,800	0,888
	13%	0,849	0,847	0,702	0,883	0,853	0,842	0,746	0,886	0,859	0,848	0,756	0,886	0,862	0,852	0,760	0,887	0,891	0,884	0,764	0,888
	20%	0,848	0,849	0,702	0,883	0,858	0,847	0,726	0,886	0,866	0,847	0,735	0,886	0,864	0,849	0,738	0,887	0,892	0,880	0,742	0,888
	30%	0,847	0,849	0,702	0,883	0,856	0,852	0,723	0,886	0,867	0,852	0,732	0,886	0,869	0,852	0,735	0,887	0,892	0,879	0,738	0,888
	50%	0,846	0,849	0,702	0,883	0,855	0,852	0,723	0,886	0,867	0,852	0,732	0,886	0,869	0,852	0,734	0,887	0,892	0,878	0,737	0,888
F1	1%	0,910	0,904	0,888	0,899	0,914	0,907	0,899	0,902	0,916	0,910	0,916	0,901	0,916	0,911	0,926	0,900	0,935	0,936	0,942	0,902
	2%	0,911	0,906	0,882	0,899	0,916	0,908	0,890	0,902	0,919	0,913	0,905	0,901	0,920	0,912	0,914	0,900	0,936	0,937	0,925	0,902
	3%	0,911	0,908	0,878	0,899	0,916	0,908	0,812	0,902	0,919	0,913	0,825	0,901	0,921	0,913	0,832	0,900	0,936	0,937	0,840	0,902
	5%	0,911	0,906	0,874	0,899	0,915	0,911	0,877	0,902	0,920	0,913	0,891	0,901	0,921	0,913	0,897	0,900	0,937	0,937	0,904	0,902
	7%	0,912	0,905	0,868	0,899	0,915	0,912	0,875	0,902	0,920	0,911	0,888	0,901	0,922	0,913	0,893	0,900	0,938	0,935	0,899	0,902
	10%	0,915	0,905	0,849	0,899	0,916	0,912	0,866	0,902	0,920	0,912	0,878	0,901	0,922	0,914	0,883	0,900	0,938	0,935	0,887	0,902
	13%	0,913	0,913	0,787	0,899	0,917	0,910	0,842	0,902	0,920	0,913	0,853	0,901	0,921	0,915	0,856	0,900	0,938	0,934	0,860	0,902
	20%	0,913	0,914	0,787	0,899	0,918	0,913	0,826	0,902	0,923	0,912	0,836	0,901	0,922	0,914	0,839	0,900	0,939	0,933	0,843	0,902
	30%	0,913	0,914	0,787	0,899	0,918	0,916	0,824	0,902	0,923	0,916	0,833	0,901	0,924	0,915	0,836	0,900	0,939	0,933	0,839	0,902
	50%	0,912	0,914	0,787	0,899	0,917	0,916	0,824	0,902	0,923	0,916	0,833	0,901	0,924	0,915	0,836	0,900	0,939	0,933	0,838	0,902

Źródło: obliczenia własne.

Tabela 3. Jakość klasyfikacji dla zbioru *apteki*, dla metody CHI w pierwszym etapie.

Miara	% terminów	Zbiór uczący 3%				Zbiór uczący 6%				Zbiór uczący 10%				Zbiór uczący 15%				Zbiór uczący 25%			
		NB	SVM	GLM	KOR	NB	SVM	GLM	KOR	NB	SVM	GLM	KOR	NB	SVM	GLM	KOR	NB	SVM	GLM	KOR
Dokładność	1%	0,843	0,830	0,801	0,883	0,848	0,828	0,825	0,886	0,852	0,834	0,841	0,886	0,846	0,815	0,852	0,887	0,884	0,874	0,871	0,888
	2%	0,846	0,834	0,780	0,883	0,853	0,839	0,801	0,886	0,858	0,849	0,816	0,886	0,860	0,840	0,824	0,887	0,888	0,886	0,837	0,888
	3%	0,849	0,835	0,760	0,883	0,854	0,846	0,779	0,886	0,863	0,853	0,792	0,886	0,864	0,844	0,799	0,887	0,889	0,887	0,809	0,888
	5%	0,849	0,839	0,751	0,883	0,857	0,844	0,751	0,886	0,864	0,854	0,763	0,886	0,866	0,849	0,769	0,887	0,891	0,889	0,776	0,888
	7%	0,848	0,839	0,765	0,883	0,857	0,844	0,755	0,886	0,866	0,853	0,766	0,886	0,868	0,851	0,771	0,887	0,892	0,889	0,777	0,888
	10%	0,849	0,838	0,757	0,883	0,855	0,844	0,752	0,886	0,866	0,851	0,763	0,886	0,868	0,852	0,767	0,887	0,891	0,886	0,772	0,888
	13%	0,849	0,848	0,752	0,883	0,855	0,845	0,730	0,886	0,866	0,852	0,739	0,886	0,868	0,852	0,743	0,887	0,891	0,886	0,747	0,888
	20%	0,847	0,849	0,754	0,883	0,857	0,851	0,723	0,886	0,866	0,851	0,732	0,886	0,867	0,851	0,736	0,887	0,893	0,877	0,739	0,888
	30%	0,847	0,849	0,755	0,883	0,855	0,851	0,720	0,886	0,867	0,852	0,729	0,886	0,869	0,852	0,732	0,887	0,892	0,876	0,735	0,888
	50%	0,847	0,849	0,755	0,883	0,854	0,852	0,722	0,886	0,867	0,852	0,730	0,886	0,869	0,851	0,732	0,887	0,893	0,878	0,735	0,888
F1	1%	0,910	0,901	0,882	0,899	0,913	0,898	0,898	0,902	0,915	0,902	0,915	0,901	0,912	0,887	0,925	0,900	0,935	0,928	0,941	0,902
	2%	0,912	0,904	0,867	0,899	0,915	0,906	0,881	0,902	0,918	0,912	0,897	0,901	0,919	0,905	0,905	0,900	0,936	0,936	0,917	0,902
	3%	0,913	0,905	0,852	0,899	0,916	0,911	0,865	0,902	0,920	0,915	0,880	0,901	0,921	0,908	0,887	0,900	0,937	0,936	0,895	0,902
	5%	0,913	0,908	0,845	0,899	0,918	0,910	0,845	0,902	0,921	0,916	0,858	0,901	0,922	0,912	0,864	0,900	0,938	0,937	0,871	0,902
	7%	0,913	0,909	0,856	0,899	0,918	0,911	0,848	0,902	0,922	0,915	0,861	0,901	0,923	0,913	0,866	0,900	0,938	0,937	0,871	0,902
	10%	0,914	0,908	0,849	0,899	0,917	0,911	0,846	0,902	0,922	0,914	0,858	0,901	0,923	0,914	0,862	0,900	0,938	0,936	0,867	0,902
	13%	0,913	0,914	0,846	0,899	0,917	0,912	0,828	0,902	0,922	0,915	0,839	0,901	0,924	0,914	0,843	0,900	0,938	0,936	0,847	0,902
	20%	0,913	0,915	0,848	0,899	0,918	0,916	0,823	0,902	0,923	0,915	0,833	0,901	0,923	0,914	0,836	0,900	0,939	0,932	0,840	0,902
	30%	0,913	0,915	0,849	0,899	0,918	0,916	0,821	0,902	0,923	0,916	0,830	0,901	0,924	0,915	0,833	0,900	0,939	0,932	0,836	0,902
	50%	0,912	0,915	0,849	0,899	0,917	0,916	0,822	0,902	0,923	0,916	0,831	0,901	0,924	0,915	0,834	0,900	0,939	0,933	0,836	0,902

Źródło: obliczenia własne.

Tabela 4. Jakość klasyfikacji dla zbioru *aptekizbil*, dla metody IG w pierwszym etapie.

Miara	% terminów	Zbiór uczący 3%				Zbiór uczący 6%				Zbiór uczący 10%				Zbiór uczący 15%				Zbiór uczący 25%			
		NB	SVM	GLM	KOR	NB	SVM	GLM	KOR	NB	SVM	GLM	KOR	NB	SVM	GLM	KOR	NB	SVM	GLM	KOR
Dokładność	1%	0,611	0,603	0,598	0,881	0,640	0,662	0,665	0,885	0,650	0,702	0,732	0,884	0,662	0,723	0,751	0,885	0,667	0,740	0,771	0,885
	2%	0,670	0,614	0,590	0,881	0,691	0,627	0,627	0,885	0,712	0,657	0,665	0,884	0,707	0,734	0,691	0,885	0,732	0,758	0,761	0,885
	3%	0,687	0,617	0,587	0,881	0,712	0,632	0,613	0,885	0,731	0,631	0,610	0,884	0,729	0,697	0,662	0,885	0,741	0,762	0,719	0,885
	5%	0,714	0,619	0,578	0,881	0,733	0,633	0,603	0,885	0,744	0,634	0,573	0,884	0,751	0,635	0,623	0,885	0,758	0,705	0,661	0,885
	7%	0,732	0,620	0,583	0,881	0,741	0,633	0,609	0,885	0,751	0,635	0,614	0,884	0,760	0,636	0,614	0,885	0,767	0,651	0,637	0,885
	10%	0,726	0,620	0,583	0,881	0,753	0,634	0,593	0,885	0,765	0,635	0,607	0,884	0,762	0,637	0,599	0,885	0,771	0,650	0,616	0,885
	13%	0,721	0,620	0,584	0,881	0,758	0,634	0,601	0,885	0,772	0,635	0,600	0,884	0,774	0,637	0,555	0,885	0,776	0,651	0,613	0,885
	20%	0,724	0,621	0,584	0,881	0,756	0,634	0,593	0,885	0,785	0,635	0,602	0,884	0,786	0,637	0,554	0,885	0,783	0,650	0,533	0,885
	30%	0,705	0,620	0,587	0,881	0,751	0,633	0,588	0,885	0,770	0,635	0,596	0,884	0,795	0,636	0,516	0,885	0,802	0,650	0,535	0,885
	50%	0,687	0,619	0,583	0,881	0,738	0,633	0,583	0,885	0,764	0,634	0,570	0,884	0,785	0,636	0,537	0,885	0,795	0,650	0,483	0,885
F1	1%	0,568	0,714	0,648	0,901	0,597	0,743	0,711	0,904	0,597	0,765	0,760	0,904	0,621	0,779	0,778	0,905	0,639	0,793	0,800	0,904
	2%	0,652	0,719	0,630	0,901	0,679	0,727	0,678	0,904	0,698	0,743	0,699	0,904	0,692	0,785	0,716	0,905	0,729	0,804	0,790	0,904
	3%	0,672	0,721	0,621	0,901	0,706	0,729	0,655	0,904	0,725	0,731	0,648	0,904	0,725	0,767	0,694	0,905	0,743	0,806	0,745	0,904
	5%	0,701	0,722	0,609	0,901	0,733	0,730	0,651	0,904	0,745	0,732	0,573	0,904	0,755	0,735	0,676	0,905	0,769	0,777	0,696	0,904
	7%	0,718	0,722	0,615	0,901	0,739	0,730	0,653	0,904	0,753	0,733	0,658	0,904	0,769	0,736	0,660	0,905	0,781	0,748	0,687	0,904
	10%	0,696	0,722	0,615	0,901	0,752	0,730	0,627	0,904	0,768	0,733	0,649	0,904	0,771	0,736	0,644	0,905	0,788	0,748	0,673	0,904
	13%	0,683	0,722	0,616	0,901	0,752	0,730	0,642	0,904	0,774	0,733	0,632	0,904	0,784	0,736	0,579	0,905	0,794	0,748	0,667	0,904
	20%	0,682	0,723	0,618	0,901	0,739	0,730	0,630	0,904	0,782	0,733	0,637	0,904	0,795	0,736	0,585	0,905	0,801	0,748	0,557	0,904
	30%	0,639	0,723	0,626	0,901	0,727	0,729	0,624	0,904	0,758	0,733	0,636	0,904	0,795	0,736	0,525	0,905	0,817	0,748	0,556	0,904
	50%	0,597	0,722	0,617	0,901	0,699	0,730	0,623	0,904	0,739	0,732	0,561	0,904	0,778	0,735	0,562	0,905	0,804	0,748	0,488	0,904

Źródło: obliczenia własne.

Tabela 5. Jakość klasyfikacji dla zbioru *aptekizbil*, dla metody CHI w pierwszym etapie.

Miara	% terminów	Zbiór uczący 3%				Zbiór uczący 6%				Zbiór uczący 10%				Zbiór uczący 15%				Zbiór uczący 25%			
		NB	SVM	GLM	KOR	NB	SVM	GLM	KOR	NB	SVM	GLM	KOR	NB	SVM	GLM	KOR	NB	SVM	GLM	KOR
Dokładność	1%	0,635	0,604	0,612	0,881	0,661	0,674	0,645	0,885	0,660	0,714	0,726	0,884	0,650	0,722	0,748	0,885	0,690	0,742	0,772	0,885
	2%	0,667	0,609	0,600	0,881	0,690	0,626	0,616	0,885	0,720	0,661	0,668	0,884	0,717	0,737	0,690	0,885	0,736	0,759	0,765	0,885
	3%	0,677	0,612	0,592	0,881	0,715	0,631	0,615	0,885	0,732	0,631	0,618	0,884	0,732	0,696	0,670	0,885	0,745	0,762	0,691	0,885
	5%	0,705	0,618	0,594	0,881	0,735	0,632	0,606	0,885	0,746	0,633	0,610	0,884	0,750	0,635	0,633	0,885	0,759	0,706	0,663	0,885
	7%	0,710	0,618	0,594	0,881	0,737	0,631	0,604	0,885	0,750	0,634	0,601	0,884	0,755	0,635	0,598	0,885	0,769	0,651	0,632	0,885
	10%	0,704	0,618	0,595	0,881	0,750	0,633	0,609	0,885	0,757	0,635	0,568	0,884	0,762	0,636	0,565	0,885	0,768	0,651	0,611	0,885
	13%	0,707	0,618	0,591	0,881	0,750	0,632	0,600	0,885	0,770	0,635	0,582	0,884	0,763	0,636	0,570	0,885	0,775	0,650	0,586	0,885
	20%	0,720	0,619	0,592	0,881	0,749	0,632	0,601	0,885	0,776	0,635	0,587	0,884	0,776	0,636	0,562	0,885	0,776	0,650	0,512	0,885
	30%	0,700	0,619	0,596	0,881	0,748	0,633	0,593	0,885	0,762	0,635	0,557	0,884	0,781	0,636	0,549	0,885	0,794	0,650	0,478	0,885
	50%	0,682	0,618	0,591	0,881	0,736	0,633	0,593	0,885	0,760	0,635	0,576	0,884	0,781	0,636	0,512	0,885	0,802	0,650	0,509	0,885
F1	1%	0,604	0,715	0,666	0,901	0,625	0,748	0,677	0,904	0,612	0,771	0,753	0,904	0,600	0,778	0,774	0,905	0,672	0,794	0,800	0,904
	2%	0,647	0,717	0,642	0,901	0,673	0,726	0,662	0,904	0,707	0,745	0,702	0,904	0,708	0,787	0,709	0,905	0,734	0,804	0,792	0,904
	3%	0,657	0,719	0,628	0,901	0,709	0,729	0,661	0,904	0,724	0,731	0,666	0,904	0,729	0,766	0,706	0,905	0,749	0,805	0,700	0,904
	5%	0,683	0,721	0,627	0,901	0,731	0,729	0,657	0,904	0,745	0,732	0,656	0,904	0,753	0,735	0,687	0,905	0,770	0,777	0,693	0,904
	7%	0,678	0,721	0,627	0,901	0,727	0,729	0,655	0,904	0,751	0,733	0,643	0,904	0,761	0,736	0,649	0,905	0,783	0,748	0,681	0,904
	10%	0,655	0,721	0,627	0,901	0,743	0,730	0,654	0,904	0,754	0,733	0,590	0,904	0,771	0,736	0,590	0,905	0,785	0,748	0,655	0,904
	13%	0,654	0,722	0,622	0,901	0,737	0,729	0,645	0,904	0,769	0,733	0,612	0,904	0,768	0,736	0,607	0,905	0,792	0,748	0,633	0,904
	20%	0,672	0,722	0,629	0,901	0,726	0,729	0,650	0,904	0,766	0,733	0,617	0,904	0,781	0,736	0,603	0,905	0,792	0,748	0,517	0,904
	30%	0,629	0,722	0,638	0,901	0,721	0,730	0,634	0,904	0,742	0,733	0,577	0,904	0,775	0,736	0,577	0,905	0,806	0,748	0,460	0,904
	50%	0,586	0,722	0,627	0,901	0,694	0,730	0,631	0,904	0,733	0,733	0,608	0,904	0,772	0,735	0,514	0,905	0,806	0,748	0,505	0,904

Źródło: obliczenia własne.

Tabela 6. Jakość klasyfikacji dla zbioru *bank*, dla metody IG w pierwszym etapie.

Miara	% terminów	Zbiór uczący 3%				Zbiór uczący 6%				Zbiór uczący 10%				Zbiór uczący 15%				Zbiór uczący 25%			
		NB	SVM	GLM	KOR	NB	SVM	GLM	KOR	NB	SVM	GLM	KOR	NB	SVM	GLM	KOR	NB	SVM	GLM	KOR
Dokładność	1%	0,741	0,697	0,693	0,81	0,736	0,707	0,701	0,82	0,763	0,722	0,742	0,821	0,772	0,730	0,716	0,822	0,756	0,727	0,733	0,823
	2%	0,754	0,710	0,664	0,81	0,766	0,732	0,670	0,82	0,787	0,742	0,680	0,821	0,796	0,752	0,708	0,822	0,785	0,756	0,710	0,823
	3%	0,772	0,715	0,640	0,81	0,776	0,738	0,645	0,82	0,797	0,750	0,647	0,821	0,806	0,762	0,736	0,822	0,798	0,767	0,686	0,823
	5%	0,779	0,711	0,580	0,81	0,795	0,744	0,584	0,82	0,812	0,755	0,583	0,821	0,819	0,772	0,677	0,822	0,813	0,779	0,621	0,823
	7%	0,789	0,721	0,583	0,81	0,805	0,748	0,586	0,82	0,820	0,760	0,584	0,821	0,826	0,777	0,656	0,822	0,820	0,786	0,622	0,823
	10%	0,790	0,712	0,581	0,81	0,809	0,745	0,584	0,82	0,827	0,768	0,581	0,821	0,835	0,782	0,637	0,822	0,830	0,793	0,618	0,823
	13%	0,791	0,701	0,580	0,81	0,821	0,752	0,582	0,82	0,831	0,761	0,580	0,821	0,837	0,785	0,625	0,822	0,835	0,797	0,615	0,823
	20%	0,802	0,679	0,557	0,81	0,827	0,742	0,559	0,82	0,842	0,771	0,556	0,821	0,839	0,778	0,591	0,822	0,842	0,800	0,589	0,823
	30%	0,807	0,569	0,554	0,81	0,830	0,732	0,556	0,82	0,845	0,757	0,554	0,821	0,854	0,788	0,582	0,822	0,847	0,795	0,585	0,823
	50%	0,808	0,569	0,553	0,81	0,837	0,574	0,555	0,82	0,854	0,570	0,552	0,821	0,858	0,773	0,575	0,822	0,863	0,798	0,582	0,823
F1	1%	0,738	0,757	0,745	0,88	0,725	0,763	0,753	0,89	0,773	0,773	0,767	0,891	0,785	0,779	0,818	0,893	0,758	0,778	0,796	0,893
	2%	0,750	0,766	0,703	0,88	0,759	0,778	0,708	0,89	0,794	0,786	0,711	0,891	0,804	0,793	0,804	0,893	0,786	0,795	0,753	0,893
	3%	0,775	0,769	0,700	0,88	0,770	0,782	0,704	0,89	0,802	0,791	0,703	0,891	0,812	0,799	0,826	0,893	0,801	0,803	0,749	0,893
	5%	0,784	0,766	0,625	0,88	0,793	0,786	0,629	0,89	0,820	0,795	0,627	0,891	0,825	0,805	0,709	0,893	0,817	0,811	0,668	0,893
	7%	0,797	0,772	0,644	0,88	0,806	0,789	0,647	0,89	0,827	0,797	0,644	0,891	0,831	0,809	0,710	0,893	0,824	0,816	0,686	0,893
	10%	0,799	0,767	0,642	0,88	0,809	0,787	0,645	0,89	0,835	0,803	0,642	0,891	0,841	0,813	0,695	0,893	0,834	0,821	0,682	0,893
	13%	0,800	0,762	0,641	0,88	0,826	0,791	0,644	0,89	0,839	0,799	0,641	0,891	0,843	0,815	0,683	0,893	0,839	0,824	0,679	0,893
	20%	0,807	0,752	0,605	0,88	0,831	0,786	0,607	0,89	0,850	0,806	0,604	0,891	0,846	0,811	0,636	0,893	0,847	0,826	0,638	0,893
	30%	0,807	0,701	0,602	0,88	0,833	0,780	0,604	0,89	0,852	0,797	0,601	0,891	0,861	0,817	0,628	0,893	0,852	0,823	0,634	0,893
	50%	0,801	0,701	0,601	0,88	0,834	0,703	0,603	0,89	0,857	0,702	0,600	0,891	0,863	0,807	0,621	0,893	0,869	0,825	0,631	0,893

Źródło: obliczenia własne.

Tabela 7. Jakość klasyfikacji dla zbioru *bank*, dla metody CHI w pierwszym etapie.

Miara	% terminów	Zbiór uczący 3%				Zbiór uczący 6%				Zbiór uczący 10%				Zbiór uczący 15%				Zbiór uczący 25%			
		NB	SVM	GLM	KOR	NB	SVM	GLM	KOR	NB	SVM	GLM	KOR	NB	SVM	GLM	KOR	NB	SVM	GLM	KOR
Dokładność	1%	0,736	0,698	0,696	0,81	0,735	0,709	0,707	0,82	0,765	0,729	0,736	0,821	0,775	0,737	0,679	0,822	0,759	0,733	0,729	0,823
	2%	0,756	0,711	0,661	0,81	0,767	0,733	0,656	0,82	0,789	0,745	0,665	0,821	0,798	0,756	0,672	0,822	0,790	0,759	0,604	0,823
	3%	0,763	0,713	0,676	0,81	0,776	0,739	0,670	0,82	0,800	0,753	0,674	0,821	0,807	0,764	0,698	0,822	0,801	0,770	0,634	0,823
	5%	0,777	0,712	0,588	0,81	0,790	0,744	0,582	0,82	0,816	0,760	0,584	0,821	0,821	0,774	0,643	0,822	0,814	0,782	0,560	0,823
	7%	0,788	0,715	0,580	0,81	0,801	0,748	0,575	0,82	0,822	0,763	0,576	0,821	0,826	0,778	0,623	0,822	0,822	0,788	0,558	0,823
	10%	0,792	0,710	0,578	0,81	0,806	0,747	0,573	0,82	0,828	0,766	0,574	0,821	0,835	0,784	0,605	0,822	0,833	0,795	0,560	0,823
	13%	0,796	0,704	0,566	0,81	0,816	0,747	0,562	0,82	0,832	0,764	0,562	0,821	0,840	0,786	0,593	0,822	0,837	0,798	0,551	0,823
	20%	0,802	0,682	0,547	0,81	0,824	0,742	0,543	0,82	0,839	0,766	0,543	0,821	0,844	0,782	0,561	0,822	0,844	0,801	0,534	0,823
	30%	0,807	0,569	0,561	0,81	0,828	0,718	0,557	0,82	0,846	0,758	0,557	0,821	0,851	0,782	0,552	0,822	0,850	0,796	0,550	0,823
	50%	0,806	0,569	0,558	0,81	0,835	0,574	0,555	0,82	0,853	0,570	0,555	0,821	0,860	0,776	0,546	0,822	0,860	0,796	0,548	0,823
F1	1%	0,729	0,758	0,697	0,88	0,722	0,764	0,694	0,89	0,774	0,777	0,705	0,891	0,787	0,782	0,717	0,893	0,758	0,780	0,631	0,893
	2%	0,750	0,766	0,686	0,88	0,759	0,779	0,679	0,89	0,794	0,787	0,684	0,891	0,806	0,795	0,739	0,893	0,793	0,797	0,640	0,893
	3%	0,758	0,768	0,757	0,88	0,769	0,783	0,750	0,89	0,806	0,793	0,753	0,891	0,813	0,800	0,784	0,893	0,804	0,804	0,719	0,893
	5%	0,777	0,767	0,595	0,88	0,785	0,786	0,590	0,89	0,823	0,797	0,591	0,891	0,826	0,807	0,673	0,893	0,818	0,813	0,572	0,893
	7%	0,790	0,769	0,591	0,88	0,799	0,789	0,586	0,89	0,828	0,799	0,587	0,891	0,831	0,809	0,674	0,893	0,825	0,817	0,572	0,893
	10%	0,792	0,766	0,587	0,88	0,804	0,789	0,583	0,89	0,835	0,801	0,583	0,891	0,840	0,814	0,659	0,893	0,836	0,822	0,571	0,893
	13%	0,794	0,764	0,579	0,88	0,815	0,789	0,574	0,89	0,838	0,801	0,575	0,891	0,845	0,816	0,648	0,893	0,840	0,824	0,565	0,893
	20%	0,801	0,753	0,534	0,88	0,822	0,786	0,530	0,89	0,844	0,802	0,531	0,891	0,849	0,813	0,604	0,893	0,848	0,827	0,523	0,893
	30%	0,804	0,702	0,577	0,88	0,825	0,774	0,573	0,89	0,848	0,797	0,574	0,891	0,855	0,813	0,596	0,893	0,854	0,823	0,567	0,893
	50%	0,798	0,701	0,576	0,88	0,832	0,703	0,572	0,89	0,856	0,702	0,572	0,891	0,862	0,809	0,589	0,893	0,863	0,823	0,566	0,893

Źródło: obliczenia własne.

Tabela 8. Jakość klasyfikacji dla zbioru *esklepy*, dla metody IG w pierwszym etapie.

Miara	% terminów	Zbiór uczący 3%				Zbiór uczący 6%				Zbiór uczący 10%				Zbiór uczący 15%				Zbiór uczący 25%			
		NB	SVM	GLM	KOR	NB	SVM	GLM	KOR	NB	SVM	GLM	KOR	NB	SVM	GLM	KOR	NB	SVM	GLM	KOR
Dokładność	1%	0,892	0,913	0,876	0,89	0,899	0,917	0,898	0,895	0,905	0,919	0,000	0,892	0,905	0,920	0,000	0,895	0,906	0,919	0,000	0,894
	2%	0,905	0,913	0,836	0,89	0,912	0,921	0,855	0,895	0,918	0,924	0,000	0,892	0,921	0,926	0,000	0,895	0,923	0,926	0,000	0,894
	3%	0,911	0,912	0,804	0,89	0,918	0,920	0,821	0,895	0,923	0,926	0,000	0,892	0,927	0,929	0,000	0,895	0,927	0,929	0,000	0,894
	5%	0,919	0,908	0,779	0,89	0,923	0,919	0,794	0,895	0,928	0,926	0,000	0,892	0,932	0,931	0,000	0,895	0,933	0,932	0,000	0,894
	7%	0,921	0,905	0,780	0,89	0,926	0,917	0,794	0,895	0,932	0,925	0,000	0,892	0,933	0,930	0,000	0,895	0,935	0,933	0,000	0,894
	10%	0,925	0,903	0,617	0,89	0,928	0,915	0,627	0,895	0,933	0,925	0,000	0,892	0,935	0,931	0,000	0,895	0,936	0,933	0,000	0,894
	13%	0,925	0,900	0,609	0,89	0,930	0,912	0,618	0,895	0,934	0,923	0,000	0,892	0,936	0,930	0,000	0,895	0,936	0,933	0,000	0,894
	20%	0,928	0,895	0,624	0,89	0,932	0,908	0,633	0,895	0,937	0,919	0,000	0,892	0,937	0,926	0,000	0,895	0,937	0,931	0,000	0,894
	30%	0,929	0,903	0,634	0,89	0,936	0,904	0,643	0,895	0,937	0,914	0,000	0,892	0,939	0,923	0,000	0,895	0,939	0,929	0,000	0,894
	50%	0,925	0,904	0,628	0,89	0,935	0,906	0,637	0,895	0,939	0,915	0,000	0,892	0,942	0,920	0,000	0,895	0,941	0,926	0,000	0,894
F1	1%	0,935	0,946	0,923	0,9	0,939	0,948	0,945	0,891	0,942	0,949	0,000	0,891	0,943	0,950	0,000	0,89	0,943	0,949	0,000	0,891
	2%	0,941	0,946	0,899	0,9	0,946	0,951	0,918	0,891	0,950	0,952	0,000	0,891	0,951	0,954	0,000	0,89	0,952	0,953	0,000	0,891
	3%	0,945	0,946	0,878	0,9	0,949	0,950	0,895	0,891	0,952	0,954	0,000	0,891	0,954	0,955	0,000	0,89	0,955	0,955	0,000	0,891
	5%	0,950	0,944	0,860	0,9	0,952	0,950	0,876	0,891	0,955	0,954	0,000	0,891	0,957	0,957	0,000	0,89	0,958	0,957	0,000	0,891
	7%	0,950	0,942	0,861	0,9	0,954	0,949	0,876	0,891	0,957	0,953	0,000	0,891	0,958	0,956	0,000	0,89	0,959	0,958	0,000	0,891
	10%	0,953	0,941	0,706	0,9	0,955	0,948	0,717	0,891	0,958	0,954	0,000	0,891	0,959	0,957	0,000	0,89	0,960	0,958	0,000	0,891
	13%	0,953	0,940	0,692	0,9	0,956	0,946	0,702	0,891	0,959	0,953	0,000	0,891	0,960	0,957	0,000	0,89	0,960	0,958	0,000	0,891
	20%	0,955	0,937	0,685	0,9	0,958	0,944	0,694	0,891	0,960	0,950	0,000	0,891	0,960	0,955	0,000	0,89	0,960	0,957	0,000	0,891
	30%	0,955	0,941	0,722	0,9	0,960	0,942	0,732	0,891	0,961	0,948	0,000	0,891	0,962	0,953	0,000	0,89	0,962	0,956	0,000	0,891
	50%	0,952	0,942	0,718	0,9	0,959	0,943	0,727	0,891	0,961	0,948	0,000	0,891	0,964	0,951	0,000	0,89	0,963	0,954	0,000	0,891

Źródło: obliczenia własne.

Tabela 9. Jakość klasyfikacji dla zbioru *esklepy*, dla metody CHI w pierwszym etapie.

Miara	% terminów	Zbiór uczący 3%				Zbiór uczący 6%				Zbiór uczący 10%				Zbiór uczący 15%				Zbiór uczący 25%			
		NB	SVM	GLM	KOR	NB	SVM	GLM	KOR	NB	SVM	GLM	KOR	NB	SVM	GLM	KOR	NB	SVM	GLM	KOR
Dokładność	1%	0,922	0,918	0,889	0,89	0,928	0,925	0,907	0,895	0,931	0,927	0,000	0,892	0,931	0,928	0,000	0,895	0,931	0,929	0,000	0,894
	2%	0,926	0,918	0,853	0,89	0,931	0,928	0,868	0,895	0,935	0,932	0,000	0,892	0,935	0,935	0,000	0,895	0,935	0,934	0,000	0,894
	3%	0,927	0,916	0,817	0,89	0,932	0,927	0,832	0,895	0,936	0,933	0,000	0,892	0,937	0,936	0,000	0,895	0,937	0,936	0,000	0,894
	5%	0,929	0,912	0,756	0,89	0,933	0,925	0,768	0,895	0,937	0,931	0,000	0,892	0,938	0,936	0,000	0,895	0,939	0,938	0,000	0,894
	7%	0,929	0,911	0,753	0,89	0,934	0,922	0,764	0,895	0,938	0,930	0,000	0,892	0,939	0,936	0,000	0,895	0,939	0,938	0,000	0,894
	10%	0,930	0,909	0,722	0,89	0,935	0,920	0,732	0,895	0,938	0,929	0,000	0,892	0,939	0,934	0,000	0,895	0,940	0,938	0,000	0,894
	13%	0,929	0,907	0,684	0,89	0,935	0,918	0,693	0,895	0,938	0,927	0,000	0,892	0,940	0,933	0,000	0,895	0,940	0,936	0,000	0,894
	20%	0,929	0,903	0,705	0,89	0,935	0,914	0,714	0,895	0,939	0,923	0,000	0,892	0,940	0,930	0,000	0,895	0,940	0,935	0,000	0,894
	30%	0,929	0,904	0,665	0,89	0,935	0,911	0,673	0,895	0,938	0,921	0,000	0,892	0,940	0,927	0,000	0,895	0,941	0,932	0,000	0,894
	50%	0,925	0,904	0,741	0,89	0,935	0,906	0,749	0,895	0,938	0,908	0,000	0,892	0,940	0,923	0,000	0,895	0,940	0,930	0,000	0,894
F1	1%	0,951	0,950	0,931	0,9	0,955	0,954	0,949	0,891	0,957	0,955	0,000	0,891	0,957	0,956	0,000	0,89	0,957	0,956	0,000	0,891
	2%	0,953	0,950	0,909	0,9	0,956	0,955	0,925	0,891	0,959	0,958	0,000	0,891	0,959	0,959	0,000	0,89	0,959	0,959	0,000	0,891
	3%	0,954	0,948	0,887	0,9	0,957	0,955	0,902	0,891	0,959	0,958	0,000	0,891	0,960	0,960	0,000	0,89	0,960	0,960	0,000	0,891
	5%	0,955	0,946	0,837	0,9	0,958	0,954	0,850	0,891	0,960	0,957	0,000	0,891	0,961	0,960	0,000	0,89	0,961	0,961	0,000	0,891
	7%	0,955	0,945	0,835	0,9	0,958	0,952	0,847	0,891	0,961	0,957	0,000	0,891	0,961	0,960	0,000	0,89	0,961	0,962	0,000	0,891
	10%	0,956	0,945	0,813	0,9	0,959	0,951	0,824	0,891	0,961	0,956	0,000	0,891	0,961	0,959	0,000	0,89	0,962	0,961	0,000	0,891
	13%	0,955	0,944	0,774	0,9	0,959	0,950	0,784	0,891	0,961	0,955	0,000	0,891	0,962	0,958	0,000	0,89	0,962	0,960	0,000	0,891
	20%	0,955	0,941	0,796	0,9	0,959	0,948	0,806	0,891	0,961	0,953	0,000	0,891	0,962	0,957	0,000	0,89	0,962	0,960	0,000	0,891
	30%	0,955	0,942	0,753	0,9	0,959	0,946	0,761	0,891	0,961	0,952	0,000	0,891	0,962	0,955	0,000	0,89	0,962	0,958	0,000	0,891
	50%	0,952	0,942	0,829	0,9	0,958	0,943	0,838	0,891	0,961	0,944	0,000	0,891	0,962	0,953	0,000	0,89	0,962	0,957	0,000	0,891

Źródło: obliczenia własne.

Tabela 10. Jakość klasyfikacji dla zbioru *esklepyzbil*, dla metody IG w pierwszym etapie.

Miara	% terminów	Zbiór uczący 3%				Zbiór uczący 6%				Zbiór uczący 10%				Zbiór uczący 15%				Zbiór uczący 25%			
		NB	SVM	GLM	KOR	NB	SVM	GLM	KOR	NB	SVM	GLM	KOR	NB	SVM	GLM	KOR	NB	SVM	GLM	KOR
Dokładność	1%	0,657	0,852	0,802	0,88	0,676	0,856	0,862	0,89	0,669	0,858	0,000	0,892	0,664	0,858	0,000	0,89	0,733	0,863	0,000	0,892
	2%	0,715	0,847	0,717	0,88	0,750	0,870	0,812	0,89	0,738	0,872	0,000	0,892	0,752	0,875	0,000	0,89	0,807	0,880	0,000	0,892
	3%	0,765	0,821	0,649	0,88	0,809	0,873	0,684	0,89	0,774	0,878	0,000	0,892	0,796	0,881	0,000	0,89	0,839	0,888	0,000	0,892
	5%	0,811	0,807	0,641	0,88	0,834	0,849	0,682	0,89	0,834	0,882	0,000	0,892	0,835	0,886	0,000	0,89	0,869	0,893	0,000	0,892
	7%	0,822	0,808	0,612	0,88	0,849	0,814	0,579	0,89	0,856	0,867	0,000	0,892	0,862	0,887	0,000	0,89	0,887	0,894	0,000	0,892
	10%	0,836	0,808	0,633	0,88	0,859	0,815	0,630	0,89	0,867	0,817	0,000	0,892	0,872	0,879	0,000	0,89	0,894	0,895	0,000	0,892
	13%	0,856	0,809	0,584	0,88	0,868	0,816	0,619	0,89	0,877	0,818	0,000	0,892	0,885	0,819	0,000	0,89	0,898	0,897	0,000	0,892
	20%	0,875	0,809	0,562	0,88	0,881	0,816	0,516	0,89	0,889	0,818	0,000	0,892	0,891	0,820	0,000	0,89	0,908	0,867	0,000	0,892
	30%	0,859	0,809	0,584	0,88	0,896	0,817	0,557	0,89	0,894	0,818	0,000	0,892	0,899	0,820	0,000	0,89	0,912	0,820	0,000	0,892
	50%	0,833	0,808	0,587	0,88	0,878	0,816	0,533	0,89	0,897	0,818	0,000	0,892	0,910	0,821	0,000	0,89	0,914	0,821	0,000	0,892
F1	1%	0,582	0,866	0,815	0,896	0,610	0,870	0,866	0,891	0,602	0,872	0,000	0,893	0,594	0,872	0,000	0,89	0,691	0,875	0,000	0,893
	2%	0,665	0,863	0,750	0,896	0,710	0,880	0,822	0,891	0,695	0,883	0,000	0,893	0,713	0,886	0,000	0,89	0,782	0,890	0,000	0,893
	3%	0,734	0,845	0,695	0,896	0,790	0,883	0,690	0,891	0,742	0,888	0,000	0,893	0,772	0,890	0,000	0,89	0,823	0,896	0,000	0,893
	5%	0,795	0,835	0,687	0,896	0,822	0,865	0,724	0,891	0,821	0,891	0,000	0,893	0,823	0,895	0,000	0,89	0,860	0,900	0,000	0,893
	7%	0,808	0,835	0,651	0,896	0,841	0,840	0,602	0,891	0,848	0,879	0,000	0,893	0,856	0,895	0,000	0,89	0,882	0,901	0,000	0,893
	10%	0,823	0,836	0,675	0,896	0,852	0,840	0,675	0,891	0,861	0,842	0,000	0,893	0,868	0,889	0,000	0,89	0,892	0,902	0,000	0,893
	13%	0,846	0,836	0,599	0,896	0,861	0,841	0,659	0,891	0,872	0,843	0,000	0,893	0,882	0,844	0,000	0,89	0,896	0,904	0,000	0,893
	20%	0,866	0,836	0,550	0,896	0,875	0,841	0,507	0,891	0,886	0,843	0,000	0,893	0,888	0,844	0,000	0,89	0,907	0,881	0,000	0,893
	30%	0,842	0,836	0,614	0,896	0,890	0,842	0,579	0,891	0,889	0,843	0,000	0,893	0,897	0,845	0,000	0,89	0,911	0,844	0,000	0,893
	50%	0,804	0,836	0,613	0,896	0,865	0,841	0,542	0,891	0,891	0,843	0,000	0,893	0,908	0,845	0,000	0,89	0,913	0,844	0,000	0,893

Źródło: obliczenia własne.

Tabela 11. Jakość klasyfikacji dla zbioru *esklepyzbil*, dla metody CHI w pierwszym etapie.

Miara	% terminów	Zbiór uczący 3%				Zbiór uczący 6%				Zbiór uczący 10%				Zbiór uczący 15%				Zbiór uczący 25%			
		NB	SVM	GLM	KOR	NB	SVM	GLM	KOR	NB	SVM	GLM	KOR	NB	SVM	GLM	KOR	NB	SVM	GLM	KOR
Dokładność	1%	0,668	0,852	0,794	0,88	0,689	0,857	0,863	0,89	0,681	0,858	0,000	0,892	0,673	0,858	0,000	0,89	0,743	0,864	0,000	0,892
	2%	0,733	0,846	0,715	0,88	0,770	0,869	0,815	0,89	0,764	0,872	0,000	0,892	0,779	0,875	0,000	0,89	0,810	0,880	0,000	0,892
	3%	0,766	0,822	0,656	0,88	0,818	0,873	0,760	0,89	0,808	0,879	0,000	0,892	0,811	0,882	0,000	0,89	0,843	0,888	0,000	0,892
	5%	0,809	0,807	0,642	0,88	0,837	0,861	0,676	0,89	0,844	0,881	0,000	0,892	0,845	0,885	0,000	0,89	0,879	0,893	0,000	0,892
	7%	0,819	0,808	0,638	0,88	0,852	0,814	0,606	0,89	0,861	0,867	0,000	0,892	0,866	0,886	0,000	0,89	0,894	0,895	0,000	0,892
	10%	0,832	0,808	0,624	0,88	0,859	0,815	0,639	0,89	0,873	0,817	0,000	0,892	0,875	0,879	0,000	0,89	0,895	0,896	0,000	0,892
	13%	0,846	0,809	0,596	0,88	0,868	0,816	0,608	0,89	0,875	0,817	0,000	0,892	0,880	0,819	0,000	0,89	0,902	0,897	0,000	0,892
	20%	0,851	0,809	0,570	0,88	0,871	0,816	0,591	0,89	0,885	0,818	0,000	0,892	0,893	0,820	0,000	0,89	0,909	0,875	0,000	0,892
	30%	0,835	0,809	0,611	0,88	0,875	0,816	0,564	0,89	0,890	0,818	0,000	0,892	0,898	0,820	0,000	0,89	0,911	0,820	0,000	0,892
	50%	0,829	0,809	0,592	0,88	0,866	0,816	0,569	0,89	0,885	0,818	0,000	0,892	0,899	0,820	0,000	0,89	0,912	0,821	0,000	0,892
F1	1%	0,601	0,866	0,804	0,896	0,630	0,870	0,867	0,891	0,617	0,872	0,000	0,893	0,606	0,872	0,000	0,89	0,702	0,876	0,000	0,893
	2%	0,690	0,862	0,751	0,896	0,738	0,879	0,824	0,891	0,728	0,882	0,000	0,893	0,749	0,886	0,000	0,89	0,785	0,889	0,000	0,893
	3%	0,733	0,845	0,699	0,896	0,802	0,884	0,782	0,891	0,786	0,888	0,000	0,893	0,791	0,891	0,000	0,89	0,828	0,896	0,000	0,893
	5%	0,792	0,835	0,690	0,896	0,827	0,874	0,720	0,891	0,834	0,890	0,000	0,893	0,835	0,894	0,000	0,89	0,872	0,900	0,000	0,893
	7%	0,803	0,835	0,687	0,896	0,844	0,840	0,637	0,891	0,854	0,880	0,000	0,893	0,860	0,895	0,000	0,89	0,891	0,902	0,000	0,893
	10%	0,817	0,836	0,663	0,896	0,852	0,840	0,683	0,891	0,869	0,842	0,000	0,893	0,871	0,889	0,000	0,89	0,892	0,902	0,000	0,893
	13%	0,832	0,836	0,608	0,896	0,862	0,841	0,639	0,891	0,870	0,842	0,000	0,893	0,876	0,844	0,000	0,89	0,900	0,904	0,000	0,893
	20%	0,834	0,836	0,587	0,896	0,863	0,841	0,618	0,891	0,881	0,843	0,000	0,893	0,891	0,844	0,000	0,89	0,908	0,886	0,000	0,893
	30%	0,808	0,836	0,646	0,896	0,864	0,841	0,588	0,891	0,885	0,843	0,000	0,893	0,895	0,845	0,000	0,89	0,910	0,844	0,000	0,893
	50%	0,798	0,836	0,616	0,896	0,850	0,841	0,593	0,891	0,875	0,843	0,000	0,893	0,894	0,845	0,000	0,89	0,911	0,844	0,000	0,893

Źródło: obliczenia własne.

Tabela 12. Jakość klasyfikacji dla zbioru *hotele*, dla metody IG w pierwszym etapie.

Miara	% terminów	Zbiór uczący 3%				Zbiór uczący 6%				Zbiór uczący 10%				Zbiór uczący 15%				Zbiór uczący 25%			
		NB	SVM	GLM	KOR	NB	SVM	GLM	KOR	NB	SVM	GLM	KOR	NB	SVM	GLM	KOR	NB	SVM	GLM	KOR
Dokładność	1%	0,783	0,760	0,694	0,848	0,805	0,779	0,738	0,852	0,821	0,804	0,778	0,851	0,843	0,826	0,808	0,85	0,850	0,836	0,831	0,851
	2%	0,794	0,764	0,640	0,848	0,815	0,780	0,704	0,852	0,832	0,807	0,735	0,851	0,849	0,829	0,760	0,85	0,858	0,839	0,810	0,851
	3%	0,798	0,765	0,623	0,848	0,823	0,785	0,645	0,852	0,837	0,800	0,709	0,851	0,855	0,827	0,755	0,85	0,860	0,839	0,788	0,851
	5%	0,794	0,764	0,616	0,848	0,826	0,786	0,612	0,852	0,842	0,804	0,607	0,851	0,857	0,820	0,685	0,85	0,865	0,837	0,749	0,851
	7%	0,789	0,751	0,616	0,848	0,820	0,785	0,571	0,852	0,845	0,802	0,583	0,851	0,861	0,821	0,612	0,85	0,865	0,832	0,727	0,851
	10%	0,788	0,751	0,613	0,848	0,814	0,780	0,571	0,852	0,840	0,794	0,521	0,851	0,860	0,820	0,558	0,85	0,867	0,831	0,652	0,851
	13%	0,791	0,753	0,616	0,848	0,811	0,781	0,619	0,852	0,834	0,785	0,559	0,851	0,858	0,812	0,479	0,85	0,869	0,830	0,564	0,851
	20%	0,801	0,754	0,610	0,848	0,814	0,781	0,579	0,852	0,831	0,790	0,548	0,851	0,851	0,807	0,556	0,85	0,865	0,820	0,528	0,851
	30%	0,801	0,769	0,610	0,848	0,821	0,781	0,568	0,852	0,832	0,790	0,519	0,851	0,849	0,806	0,467	0,85	0,862	0,820	0,540	0,851
	50%	0,790	0,769	0,610	0,848	0,812	0,790	0,568	0,852	0,833	0,812	0,532	0,851	0,852	0,807	0,459	0,85	0,861	0,820	0,556	0,851
F1	1%	0,865	0,852	0,799	0,93	0,876	0,862	0,830	0,931	0,885	0,877	0,859	0,938	0,899	0,890	0,878	0,94	0,903	0,895	0,892	0,94
	2%	0,873	0,856	0,747	0,93	0,883	0,864	0,803	0,931	0,893	0,879	0,827	0,938	0,902	0,892	0,839	0,94	0,908	0,897	0,878	0,94
	3%	0,876	0,858	0,733	0,93	0,888	0,868	0,751	0,931	0,896	0,876	0,805	0,938	0,906	0,892	0,840	0,94	0,909	0,898	0,863	0,94
	5%	0,876	0,859	0,727	0,93	0,891	0,870	0,721	0,931	0,900	0,879	0,718	0,938	0,908	0,888	0,785	0,94	0,912	0,897	0,834	0,94
	7%	0,874	0,849	0,727	0,93	0,889	0,870	0,660	0,931	0,902	0,878	0,695	0,938	0,911	0,889	0,722	0,94	0,913	0,895	0,819	0,94
	10%	0,875	0,850	0,723	0,93	0,887	0,868	0,662	0,931	0,901	0,874	0,615	0,938	0,911	0,889	0,664	0,94	0,914	0,894	0,757	0,94
	13%	0,876	0,851	0,725	0,93	0,886	0,869	0,728	0,931	0,898	0,870	0,662	0,938	0,910	0,885	0,566	0,94	0,916	0,894	0,672	0,94
	20%	0,881	0,851	0,720	0,93	0,888	0,868	0,683	0,931	0,897	0,873	0,651	0,938	0,907	0,882	0,658	0,94	0,914	0,889	0,624	0,94
	30%	0,881	0,858	0,720	0,93	0,891	0,867	0,669	0,931	0,897	0,873	0,612	0,938	0,906	0,881	0,521	0,94	0,913	0,889	0,629	0,94
	50%	0,876	0,858	0,720	0,93	0,887	0,873	0,669	0,931	0,898	0,884	0,624	0,938	0,908	0,881	0,505	0,94	0,912	0,888	0,660	0,94

Źródło: obliczenia własne.

Tabela 13. Jakość klasyfikacji dla zbioru *hotele*, dla metody CHI w pierwszym etapie.

Miara	% terminów	Zbiór uczący 3%				Zbiór uczący 6%				Zbiór uczący 10%				Zbiór uczący 15%				Zbiór uczący 25%			
		NB	SVM	GLM	KOR	NB	SVM	GLM	KOR	NB	SVM	GLM	KOR	NB	SVM	GLM	KOR	NB	SVM	GLM	KOR
Dokładność	1%	0,745	0,649	0,627	0,848	0,770	0,676	0,701	0,852	0,777	0,656	0,758	0,851	0,784	0,662	0,779	0,85	0,788	0,668	0,790	0,851
	2%	0,745	0,679	0,579	0,848	0,789	0,712	0,626	0,852	0,794	0,691	0,705	0,851	0,800	0,698	0,746	0,85	0,805	0,703	0,785	0,851
	3%	0,757	0,700	0,565	0,848	0,801	0,736	0,586	0,852	0,805	0,733	0,657	0,851	0,812	0,733	0,721	0,85	0,818	0,737	0,744	0,851
	5%	0,783	0,712	0,566	0,848	0,815	0,760	0,579	0,852	0,817	0,754	0,566	0,851	0,826	0,766	0,650	0,85	0,830	0,763	0,733	0,851
	7%	0,808	0,730	0,567	0,848	0,821	0,777	0,564	0,852	0,825	0,779	0,572	0,851	0,838	0,789	0,586	0,85	0,838	0,785	0,689	0,851
	10%	0,805	0,734	0,572	0,848	0,826	0,785	0,551	0,852	0,837	0,795	0,571	0,851	0,845	0,804	0,583	0,85	0,847	0,813	0,613	0,851
	13%	0,801	0,736	0,545	0,848	0,826	0,787	0,536	0,852	0,840	0,799	0,560	0,851	0,850	0,815	0,564	0,85	0,854	0,820	0,564	0,851
	20%	0,798	0,736	0,560	0,848	0,815	0,788	0,537	0,852	0,840	0,801	0,545	0,851	0,856	0,821	0,565	0,85	0,860	0,834	0,488	0,851
	30%	0,798	0,748	0,578	0,848	0,813	0,783	0,555	0,852	0,829	0,793	0,499	0,851	0,854	0,817	0,480	0,85	0,866	0,835	0,501	0,851
	50%	0,788	0,751	0,574	0,848	0,809	0,790	0,518	0,852	0,831	0,811	0,532	0,851	0,848	0,808	0,484	0,85	0,858	0,830	0,533	0,851
F1	1%	0,836	0,725	0,730	0,93	0,856	0,755	0,797	0,931	0,861	0,730	0,844	0,938	0,865	0,733	0,858	0,94	0,868	0,737	0,866	0,94
	2%	0,835	0,760	0,689	0,93	0,867	0,793	0,730	0,931	0,870	0,768	0,798	0,938	0,874	0,770	0,829	0,94	0,877	0,773	0,860	0,94
	3%	0,843	0,785	0,673	0,93	0,874	0,817	0,697	0,931	0,877	0,808	0,758	0,938	0,881	0,805	0,810	0,94	0,885	0,806	0,823	0,94
	5%	0,863	0,800	0,676	0,93	0,883	0,841	0,687	0,931	0,884	0,830	0,679	0,938	0,889	0,836	0,750	0,94	0,891	0,830	0,819	0,94
	7%	0,882	0,822	0,676	0,93	0,888	0,857	0,671	0,931	0,889	0,854	0,683	0,938	0,896	0,857	0,695	0,94	0,896	0,850	0,781	0,94
	10%	0,882	0,829	0,677	0,93	0,891	0,865	0,651	0,931	0,897	0,869	0,684	0,938	0,901	0,872	0,694	0,94	0,901	0,875	0,718	0,94
	13%	0,880	0,833	0,622	0,93	0,893	0,869	0,638	0,931	0,900	0,874	0,668	0,938	0,905	0,881	0,671	0,94	0,906	0,881	0,671	0,94
	20%	0,879	0,831	0,665	0,93	0,888	0,871	0,636	0,931	0,901	0,877	0,649	0,938	0,908	0,887	0,672	0,94	0,910	0,894	0,579	0,94
	30%	0,879	0,836	0,692	0,93	0,887	0,869	0,657	0,931	0,896	0,873	0,594	0,938	0,909	0,886	0,544	0,94	0,914	0,896	0,597	0,94
	50%	0,875	0,840	0,684	0,93	0,886	0,873	0,604	0,931	0,897	0,883	0,631	0,938	0,906	0,882	0,546	0,94	0,911	0,893	0,634	0,94

Źródło: obliczenia własne.

Tabela 14. Jakość klasyfikacji dla zbioru *hotelezbil*, dla metody IG w pierwszym etapie.

Miara	% terminów	Zbiór uczący 3%				Zbiór uczący 6%				Zbiór uczący 10%				Zbiór uczący 15%				Zbiór uczący 25%			
		NB	SVM	GLM	KOR	NB	SVM	GLM	KOR	NB	SVM	GLM	KOR	NB	SVM	GLM	KOR	NB	SVM	GLM	KOR
Dokładność	1%	0,710	0,653	0,546	0,845	0,757	0,718	0,627	0,846	0,782	0,716	0,669	0,847	0,796	0,813	0,721	0,849	0,802	0,820	0,747	0,848
	2%	0,723	0,658	0,527	0,845	0,777	0,712	0,515	0,846	0,804	0,724	0,611	0,847	0,818	0,791	0,681	0,849	0,827	0,828	0,721	0,848
	3%	0,727	0,665	0,522	0,845	0,782	0,723	0,509	0,846	0,817	0,738	0,556	0,847	0,828	0,770	0,630	0,849	0,843	0,810	0,723	0,848
	5%	0,731	0,673	0,525	0,845	0,784	0,714	0,517	0,846	0,824	0,736	0,501	0,847	0,836	0,773	0,522	0,849	0,852	0,802	0,650	0,848
	7%	0,720	0,663	0,521	0,845	0,782	0,718	0,508	0,846	0,826	0,736	0,497	0,847	0,841	0,776	0,486	0,849	0,857	0,796	0,573	0,848
	10%	0,711	0,660	0,521	0,845	0,777	0,709	0,507	0,846	0,827	0,741	0,497	0,847	0,841	0,777	0,500	0,849	0,862	0,800	0,503	0,848
	13%	0,691	0,662	0,521	0,845	0,772	0,705	0,510	0,846	0,821	0,739	0,502	0,847	0,840	0,781	0,494	0,849	0,861	0,805	0,508	0,848
	20%	0,672	0,662	0,521	0,845	0,757	0,700	0,510	0,846	0,807	0,731	0,503	0,847	0,830	0,777	0,491	0,849	0,862	0,808	0,500	0,848
	30%	0,652	0,652	0,521	0,845	0,729	0,698	0,510	0,846	0,792	0,729	0,499	0,847	0,806	0,773	0,505	0,849	0,848	0,803	0,507	0,848
	50%	0,611	0,650	0,521	0,845	0,673	0,695	0,510	0,846	0,755	0,727	0,497	0,847	0,780	0,772	0,493	0,849	0,821	0,807	0,524	0,848
F1	1%	0,700	0,611	0,559	0,939	0,762	0,723	0,640	0,938	0,795	0,710	0,679	0,94	0,816	0,846	0,737	0,937	0,834	0,859	0,777	0,938
	2%	0,719	0,620	0,555	0,939	0,787	0,706	0,539	0,938	0,820	0,727	0,631	0,94	0,838	0,815	0,712	0,937	0,858	0,863	0,757	0,938
	3%	0,723	0,628	0,552	0,939	0,792	0,726	0,528	0,938	0,832	0,745	0,587	0,94	0,848	0,802	0,667	0,937	0,872	0,845	0,770	0,938
	5%	0,727	0,642	0,558	0,939	0,793	0,715	0,544	0,938	0,839	0,739	0,533	0,94	0,855	0,814	0,564	0,937	0,880	0,840	0,702	0,938
	7%	0,706	0,626	0,556	0,939	0,790	0,725	0,534	0,938	0,842	0,739	0,524	0,94	0,860	0,818	0,532	0,937	0,884	0,834	0,633	0,938
	10%	0,690	0,625	0,556	0,939	0,781	0,710	0,533	0,938	0,842	0,747	0,529	0,94	0,860	0,822	0,547	0,937	0,888	0,839	0,567	0,938
	13%	0,654	0,629	0,556	0,939	0,772	0,702	0,537	0,938	0,835	0,745	0,536	0,94	0,858	0,828	0,542	0,937	0,887	0,845	0,574	0,938
	20%	0,610	0,625	0,556	0,939	0,751	0,691	0,539	0,938	0,817	0,742	0,534	0,94	0,847	0,820	0,540	0,937	0,888	0,847	0,564	0,938
	30%	0,564	0,603	0,556	0,939	0,711	0,687	0,539	0,938	0,799	0,737	0,527	0,94	0,822	0,815	0,552	0,937	0,875	0,846	0,575	0,938
	50%	0,466	0,598	0,556	0,939	0,614	0,679	0,539	0,938	0,748	0,732	0,524	0,94	0,792	0,814	0,537	0,937	0,850	0,851	0,600	0,938

Źródło: obliczenia własne.

Tabela 15. Jakość klasyfikacji dla zbioru *hotelezbil*, dla metody CHI w pierwszym etapie.

Miara	% terminów	Zbiór uczący 3%				Zbiór uczący 6%				Zbiór uczący 10%				Zbiór uczący 15%				Zbiór uczący 25%			
		NB	SVM	GLM	KOR	NB	SVM	GLM	KOR	NB	SVM	GLM	KOR	NB	SVM	GLM	KOR	NB	SVM	GLM	KOR
Dokładność	1%	0,693	0,625	0,549	0,845	0,742	0,655	0,618	0,846	0,769	0,696	0,684	0,847	0,793	0,772	0,710	0,849	0,797	0,773	0,730	0,848
	2%	0,710	0,632	0,518	0,845	0,761	0,668	0,534	0,846	0,791	0,702	0,598	0,847	0,813	0,750	0,674	0,849	0,820	0,767	0,705	0,848
	3%	0,717	0,631	0,521	0,845	0,765	0,675	0,525	0,846	0,800	0,703	0,553	0,847	0,821	0,755	0,627	0,849	0,835	0,762	0,673	0,848
	5%	0,716	0,631	0,516	0,845	0,775	0,696	0,517	0,846	0,813	0,717	0,513	0,847	0,830	0,763	0,522	0,849	0,846	0,768	0,637	0,848
	7%	0,718	0,647	0,507	0,845	0,773	0,697	0,515	0,846	0,816	0,725	0,506	0,847	0,832	0,764	0,489	0,849	0,852	0,776	0,546	0,848
	10%	0,700	0,648	0,507	0,845	0,776	0,707	0,514	0,846	0,814	0,725	0,506	0,847	0,833	0,770	0,496	0,849	0,856	0,779	0,505	0,848
	13%	0,685	0,650	0,508	0,845	0,769	0,708	0,513	0,846	0,816	0,723	0,504	0,847	0,833	0,771	0,502	0,849	0,856	0,790	0,505	0,848
	20%	0,669	0,651	0,504	0,845	0,745	0,706	0,513	0,846	0,808	0,727	0,494	0,847	0,829	0,776	0,505	0,849	0,855	0,791	0,509	0,848
	30%	0,643	0,653	0,504	0,845	0,724	0,706	0,512	0,846	0,788	0,726	0,494	0,847	0,813	0,776	0,506	0,849	0,848	0,802	0,501	0,848
	50%	0,608	0,649	0,504	0,845	0,669	0,693	0,516	0,846	0,750	0,727	0,496	0,847	0,780	0,779	0,508	0,849	0,834	0,801	0,508	0,848
F1	1%	0,676	0,560	0,556	0,939	0,744	0,618	0,624	0,938	0,779	0,683	0,692	0,94	0,812	0,792	0,722	0,937	0,830	0,806	0,759	0,938
	2%	0,698	0,574	0,544	0,939	0,767	0,634	0,550	0,938	0,804	0,692	0,604	0,94	0,832	0,767	0,693	0,937	0,851	0,799	0,740	0,938
	3%	0,707	0,571	0,546	0,939	0,770	0,646	0,552	0,938	0,814	0,691	0,577	0,94	0,841	0,777	0,646	0,937	0,864	0,796	0,712	0,938
	5%	0,705	0,572	0,534	0,939	0,781	0,678	0,546	0,938	0,827	0,710	0,542	0,94	0,849	0,790	0,564	0,937	0,875	0,806	0,680	0,938
	7%	0,702	0,597	0,533	0,939	0,777	0,681	0,542	0,938	0,830	0,725	0,536	0,94	0,851	0,792	0,489	0,937	0,880	0,814	0,599	0,938
	10%	0,671	0,599	0,533	0,939	0,779	0,694	0,540	0,938	0,827	0,725	0,539	0,94	0,852	0,803	0,531	0,937	0,883	0,817	0,568	0,938
	13%	0,642	0,601	0,533	0,939	0,768	0,695	0,540	0,938	0,828	0,722	0,533	0,94	0,851	0,804	0,541	0,937	0,883	0,831	0,570	0,938
	20%	0,609	0,601	0,529	0,939	0,733	0,691	0,539	0,938	0,818	0,730	0,521	0,94	0,846	0,809	0,554	0,937	0,881	0,832	0,575	0,938
	30%	0,543	0,606	0,529	0,939	0,698	0,688	0,538	0,938	0,794	0,729	0,521	0,94	0,828	0,809	0,557	0,937	0,875	0,844	0,566	0,938
	50%	0,459	0,599	0,529	0,939	0,604	0,676	0,551	0,938	0,741	0,732	0,519	0,94	0,789	0,816	0,562	0,937	0,860	0,844	0,579	0,938

Źródło: obliczenia własne.

Tabela 16. Jakość klasyfikacji dla zbioru *ksiazki*, dla metody IG w pierwszym etapie.

Miara	% terminów	Zbiór uczący 3%				Zbiór uczący 6%				Zbiór uczący 10%				Zbiór uczący 15%				Zbiór uczący 25%			
		NB	SVM	GLM	KOR	NB	SVM	GLM	KOR	NB	SVM	GLM	KOR	NB	SVM	GLM	KOR	NB	SVM	GLM	KOR
Dokładność	1%	0,688	0,689	0,686	0,831	0,725	0,715	0,712	0,829	0,736	0,731	0,727	0,835	0,754	0,748	0,000	0,834	0,758	0,749	0,000	0,835
	2%	0,707	0,698	0,607	0,831	0,740	0,729	0,618	0,829	0,753	0,747	0,623	0,835	0,765	0,764	0,000	0,834	0,773	0,766	0,000	0,835
	3%	0,713	0,703	0,581	0,831	0,746	0,731	0,587	0,829	0,761	0,751	0,590	0,835	0,774	0,773	0,000	0,834	0,778	0,775	0,000	0,835
	5%	0,731	0,703	0,565	0,831	0,756	0,745	0,569	0,829	0,767	0,755	0,571	0,835	0,781	0,780	0,000	0,834	0,781	0,781	0,000	0,835
	7%	0,730	0,705	0,560	0,831	0,763	0,735	0,562	0,829	0,773	0,764	0,563	0,835	0,785	0,778	0,000	0,834	0,784	0,785	0,000	0,835
	10%	0,734	0,714	0,542	0,831	0,769	0,741	0,544	0,829	0,778	0,760	0,545	0,835	0,787	0,781	0,000	0,834	0,788	0,784	0,000	0,835
	13%	0,739	0,706	0,540	0,831	0,769	0,749	0,542	0,829	0,782	0,760	0,542	0,835	0,791	0,776	0,000	0,834	0,791	0,788	0,000	0,835
	20%	0,745	0,695	0,539	0,831	0,775	0,735	0,540	0,829	0,785	0,766	0,540	0,835	0,795	0,775	0,000	0,834	0,799	0,784	0,000	0,835
	30%	0,728	0,669	0,536	0,831	0,779	0,723	0,537	0,829	0,790	0,755	0,537	0,835	0,798	0,779	0,000	0,834	0,800	0,789	0,000	0,835
	50%	0,722	0,638	0,535	0,831	0,770	0,660	0,536	0,829	0,787	0,751	0,536	0,835	0,803	0,763	0,000	0,834	0,805	0,782	0,000	0,835
F1	1%	0,750	0,777	0,728	0,86	0,784	0,797	0,744	0,859	0,793	0,811	0,752	0,86	0,808	0,822	0,000	0,854	0,812	0,825	0,000	0,855
	2%	0,762	0,781	0,651	0,86	0,791	0,803	0,659	0,859	0,803	0,818	0,663	0,86	0,813	0,829	0,000	0,854	0,821	0,833	0,000	0,855
	3%	0,765	0,779	0,633	0,86	0,795	0,803	0,638	0,859	0,808	0,818	0,639	0,86	0,818	0,834	0,000	0,854	0,823	0,838	0,000	0,855
	5%	0,785	0,785	0,609	0,86	0,800	0,811	0,613	0,859	0,813	0,820	0,614	0,86	0,823	0,837	0,000	0,854	0,824	0,842	0,000	0,855
	7%	0,778	0,779	0,602	0,86	0,809	0,806	0,604	0,859	0,816	0,825	0,605	0,86	0,826	0,835	0,000	0,854	0,827	0,845	0,000	0,855
	10%	0,783	0,788	0,577	0,86	0,813	0,809	0,579	0,859	0,822	0,824	0,580	0,86	0,827	0,837	0,000	0,854	0,830	0,844	0,000	0,855
	13%	0,793	0,786	0,573	0,86	0,812	0,813	0,574	0,859	0,825	0,823	0,575	0,86	0,832	0,834	0,000	0,854	0,832	0,846	0,000	0,855
	20%	0,806	0,783	0,577	0,86	0,821	0,806	0,578	0,859	0,826	0,825	0,578	0,86	0,834	0,832	0,000	0,854	0,839	0,844	0,000	0,855
	30%	0,773	0,775	0,575	0,86	0,830	0,800	0,575	0,859	0,833	0,820	0,576	0,86	0,836	0,834	0,000	0,854	0,839	0,846	0,000	0,855
	50%	0,764	0,748	0,574	0,86	0,812	0,776	0,574	0,859	0,827	0,816	0,575	0,86	0,844	0,826	0,000	0,854	0,845	0,843	0,000	0,855

Źródło: obliczenia własne.

Tabela 17. Jakość klasyfikacji dla zbioru *ksiazki*, dla metody CHI w pierwszym etapie.

Miara	% terminów	Zbiór uczący 3%				Zbiór uczący 6%				Zbiór uczący 10%				Zbiór uczący 15%				Zbiór uczący 25%			
		NB	SVM	GLM	KOR	NB	SVM	GLM	KOR	NB	SVM	GLM	KOR	NB	SVM	GLM	KOR	NB	SVM	GLM	KOR
Dokładność	1%	0,699	0,692	0,688	0,831	0,727	0,719	0,723	0,829	0,746	0,746	0,741	0,835	0,760	0,762	0,000	0,834	0,768	0,765	0,000	0,835
	2%	0,713	0,702	0,612	0,831	0,743	0,741	0,623	0,829	0,760	0,759	0,594	0,835	0,772	0,778	0,000	0,834	0,784	0,785	0,000	0,835
	3%	0,718	0,703	0,595	0,831	0,750	0,745	0,601	0,829	0,764	0,762	0,584	0,835	0,778	0,785	0,000	0,834	0,786	0,792	0,000	0,835
	5%	0,721	0,699	0,574	0,831	0,761	0,752	0,577	0,829	0,770	0,767	0,567	0,835	0,784	0,790	0,000	0,834	0,785	0,795	0,000	0,835
	7%	0,730	0,709	0,562	0,831	0,761	0,747	0,564	0,829	0,776	0,773	0,557	0,835	0,785	0,786	0,000	0,834	0,790	0,796	0,000	0,835
	10%	0,729	0,707	0,560	0,831	0,768	0,753	0,562	0,829	0,776	0,767	0,556	0,835	0,791	0,787	0,000	0,834	0,793	0,797	0,000	0,835
	13%	0,726	0,706	0,560	0,831	0,770	0,750	0,561	0,829	0,779	0,767	0,557	0,835	0,792	0,784	0,000	0,834	0,794	0,798	0,000	0,835
	20%	0,726	0,704	0,554	0,831	0,770	0,741	0,555	0,829	0,781	0,764	0,552	0,835	0,794	0,782	0,000	0,834	0,797	0,794	0,000	0,835
	30%	0,728	0,635	0,543	0,831	0,769	0,735	0,544	0,829	0,781	0,760	0,541	0,835	0,794	0,777	0,000	0,834	0,800	0,792	0,000	0,835
	50%	0,723	0,635	0,542	0,831	0,769	0,661	0,543	0,829	0,785	0,669	0,540	0,835	0,797	0,771	0,000	0,834	0,800	0,784	0,000	0,835
F1	1%	0,765	0,770	0,706	0,86	0,789	0,792	0,722	0,859	0,805	0,819	0,682	0,86	0,818	0,830	0,000	0,854	0,824	0,835	0,000	0,855
	2%	0,771	0,776	0,647	0,86	0,797	0,810	0,654	0,859	0,812	0,825	0,634	0,86	0,821	0,838	0,000	0,854	0,833	0,846	0,000	0,855
	3%	0,773	0,775	0,642	0,86	0,801	0,812	0,646	0,859	0,813	0,824	0,633	0,86	0,824	0,841	0,000	0,854	0,831	0,850	0,000	0,855
	5%	0,774	0,774	0,617	0,86	0,808	0,815	0,619	0,859	0,816	0,827	0,611	0,86	0,827	0,843	0,000	0,854	0,829	0,851	0,000	0,855
	7%	0,778	0,780	0,606	0,86	0,807	0,812	0,608	0,859	0,819	0,830	0,602	0,86	0,827	0,840	0,000	0,854	0,832	0,852	0,000	0,855
	10%	0,776	0,782	0,599	0,86	0,812	0,816	0,601	0,859	0,820	0,826	0,596	0,86	0,831	0,840	0,000	0,854	0,834	0,852	0,000	0,855
	13%	0,773	0,782	0,604	0,86	0,813	0,814	0,605	0,859	0,821	0,827	0,601	0,86	0,832	0,839	0,000	0,854	0,835	0,853	0,000	0,855
	20%	0,770	0,783	0,602	0,86	0,813	0,809	0,603	0,859	0,822	0,825	0,600	0,86	0,833	0,837	0,000	0,854	0,837	0,850	0,000	0,855
	30%	0,772	0,740	0,587	0,86	0,812	0,805	0,588	0,859	0,822	0,822	0,585	0,86	0,833	0,834	0,000	0,854	0,840	0,849	0,000	0,855
	50%	0,765	0,740	0,586	0,86	0,812	0,777	0,587	0,859	0,825	0,786	0,585	0,86	0,835	0,830	0,000	0,854	0,839	0,843	0,000	0,855

Źródło: obliczenia własne.

Tabela 18. Jakość klasyfikacji dla zbioru *ksiazkizbil*, dla metody IG w pierwszym etapie.

Miara	% terminów	Zbiór uczący 3%				Zbiór uczący 6%				Zbiór uczący 10%				Zbiór uczący 15%				Zbiór uczący 25%			
		NB	SVM	GLM	KOR	NB	SVM	GLM	KOR	NB	SVM	GLM	KOR	NB	SVM	GLM	KOR	NB	SVM	GLM	KOR
Dokładność	1%	0,680	0,544	0,623	0,82	0,718	0,586	0,678	0,833	0,741	0,661	0,726	0,831	0,752	0,725	0,000	0,835	0,760	0,730	0,000	0,835
	2%	0,696	0,596	0,566	0,82	0,739	0,585	0,630	0,833	0,760	0,594	0,697	0,831	0,773	0,595	0,000	0,835	0,784	0,737	0,000	0,835
	3%	0,704	0,596	0,550	0,82	0,748	0,584	0,613	0,833	0,771	0,582	0,673	0,831	0,785	0,581	0,000	0,835	0,796	0,692	0,000	0,835
	5%	0,714	0,594	0,541	0,82	0,754	0,586	0,543	0,833	0,781	0,589	0,621	0,831	0,795	0,585	0,000	0,835	0,807	0,625	0,000	0,835
	7%	0,719	0,602	0,539	0,82	0,758	0,588	0,543	0,833	0,786	0,590	0,576	0,831	0,799	0,590	0,000	0,835	0,813	0,591	0,000	0,835
	10%	0,719	0,605	0,525	0,82	0,763	0,594	0,532	0,833	0,788	0,592	0,530	0,831	0,805	0,585	0,000	0,835	0,818	0,594	0,000	0,835
	13%	0,720	0,605	0,525	0,82	0,762	0,599	0,528	0,833	0,791	0,597	0,514	0,831	0,806	0,584	0,000	0,835	0,819	0,596	0,000	0,835
	20%	0,719	0,610	0,525	0,82	0,762	0,603	0,519	0,833	0,791	0,604	0,512	0,831	0,808	0,582	0,000	0,835	0,819	0,600	0,000	0,835
	30%	0,717	0,611	0,524	0,82	0,762	0,603	0,524	0,833	0,790	0,605	0,509	0,831	0,806	0,585	0,000	0,835	0,820	0,605	0,000	0,835
	50%	0,713	0,611	0,524	0,82	0,762	0,604	0,524	0,833	0,789	0,605	0,516	0,831	0,804	0,585	0,000	0,835	0,817	0,606	0,000	0,835
F1	1%	0,733	0,540	0,676	0,849	0,777	0,729	0,743	0,853	0,790	0,762	0,777	0,854	0,797	0,789	0,000	0,86	0,803	0,795	0,000	0,852
	2%	0,744	0,730	0,615	0,849	0,787	0,728	0,684	0,853	0,802	0,734	0,740	0,854	0,812	0,733	0,000	0,86	0,820	0,801	0,000	0,852
	3%	0,750	0,727	0,604	0,849	0,793	0,728	0,669	0,853	0,809	0,728	0,715	0,854	0,821	0,726	0,000	0,86	0,828	0,781	0,000	0,852
	5%	0,755	0,729	0,586	0,849	0,796	0,728	0,591	0,853	0,817	0,731	0,671	0,854	0,828	0,728	0,000	0,86	0,837	0,748	0,000	0,852
	7%	0,758	0,730	0,582	0,849	0,800	0,729	0,586	0,853	0,819	0,731	0,625	0,854	0,831	0,730	0,000	0,86	0,842	0,731	0,000	0,852
	10%	0,756	0,732	0,560	0,849	0,802	0,731	0,581	0,853	0,821	0,731	0,563	0,854	0,835	0,690	0,000	0,86	0,845	0,732	0,000	0,852
	13%	0,757	0,732	0,558	0,849	0,801	0,733	0,547	0,853	0,823	0,733	0,551	0,854	0,837	0,688	0,000	0,86	0,845	0,734	0,000	0,852
	20%	0,754	0,734	0,563	0,849	0,802	0,735	0,554	0,853	0,821	0,736	0,542	0,854	0,838	0,669	0,000	0,86	0,846	0,735	0,000	0,852
	30%	0,749	0,734	0,562	0,849	0,804	0,735	0,565	0,853	0,822	0,736	0,542	0,854	0,836	0,671	0,000	0,86	0,846	0,736	0,000	0,852
	50%	0,739	0,734	0,562	0,849	0,803	0,735	0,568	0,853	0,819	0,736	0,555	0,854	0,836	0,672	0,000	0,86	0,844	0,736	0,000	0,852

Źródło: obliczenia własne.

Tabela 19. Jakość klasyfikacji dla zbioru *ksiazkizbil*, dla metody CHI w pierwszym etapie.

Miara	% terminów	Zbiór uczący 3%				Zbiór uczący 6%				Zbiór uczący 10%				Zbiór uczący 15%				Zbiór uczący 25%			
		NB	SVM	GLM	KOR	NB	SVM	GLM	KOR	NB	SVM	GLM	KOR	NB	SVM	GLM	KOR	NB	SVM	GLM	KOR
Dokładność	1%	0,676	0,545	0,602	0,82	0,711	0,585	0,666	0,833	0,743	0,633	0,719	0,831	0,759	0,734	0,000	0,835	0,763	0,759	0,000	0,835
	2%	0,693	0,587	0,561	0,82	0,731	0,585	0,633	0,833	0,767	0,620	0,697	0,831	0,784	0,650	0,000	0,835	0,796	0,767	0,000	0,835
	3%	0,698	0,586	0,558	0,82	0,738	0,591	0,591	0,833	0,775	0,592	0,666	0,831	0,793	0,581	0,000	0,835	0,804	0,716	0,000	0,835
	5%	0,703	0,597	0,545	0,82	0,749	0,592	0,561	0,833	0,783	0,593	0,577	0,831	0,802	0,582	0,000	0,835	0,814	0,606	0,000	0,835
	7%	0,712	0,587	0,539	0,82	0,750	0,592	0,555	0,833	0,786	0,594	0,582	0,831	0,807	0,582	0,000	0,835	0,819	0,608	0,000	0,835
	10%	0,712	0,588	0,540	0,82	0,762	0,593	0,520	0,833	0,788	0,595	0,531	0,831	0,811	0,582	0,000	0,835	0,820	0,609	0,000	0,835
	13%	0,711	0,588	0,543	0,82	0,760	0,592	0,520	0,833	0,791	0,598	0,543	0,831	0,812	0,583	0,000	0,835	0,822	0,609	0,000	0,835
	20%	0,712	0,587	0,539	0,82	0,760	0,592	0,529	0,833	0,791	0,599	0,517	0,831	0,814	0,582	0,000	0,835	0,825	0,610	0,000	0,835
	30%	0,716	0,613	0,530	0,82	0,761	0,594	0,529	0,833	0,792	0,599	0,526	0,831	0,814	0,582	0,000	0,835	0,826	0,609	0,000	0,835
	50%	0,711	0,612	0,530	0,82	0,762	0,603	0,519	0,833	0,791	0,605	0,508	0,831	0,809	0,584	0,000	0,835	0,826	0,609	0,000	0,835
F1	1%	0,733	0,521	0,642	0,849	0,775	0,726	0,728	0,853	0,793	0,698	0,769	0,854	0,808	0,790	0,000	0,86	0,811	0,812	0,000	0,852
	2%	0,744	0,642	0,604	0,849	0,785	0,711	0,677	0,853	0,809	0,742	0,751	0,854	0,823	0,754	0,000	0,86	0,833	0,820	0,000	0,852
	3%	0,747	0,657	0,608	0,849	0,789	0,730	0,635	0,853	0,814	0,729	0,703	0,854	0,829	0,664	0,000	0,86	0,837	0,793	0,000	0,852
	5%	0,749	0,702	0,590	0,849	0,795	0,730	0,611	0,853	0,820	0,730	0,622	0,854	0,835	0,668	0,000	0,86	0,844	0,732	0,000	0,852
	7%	0,754	0,700	0,584	0,849	0,796	0,730	0,605	0,853	0,821	0,731	0,618	0,854	0,839	0,668	0,000	0,86	0,847	0,735	0,000	0,852
	10%	0,752	0,701	0,580	0,849	0,803	0,731	0,563	0,853	0,823	0,731	0,580	0,854	0,841	0,670	0,000	0,86	0,848	0,736	0,000	0,852
	13%	0,750	0,702	0,587	0,849	0,802	0,731	0,552	0,853	0,824	0,733	0,589	0,854	0,843	0,670	0,000	0,86	0,849	0,737	0,000	0,852
	20%	0,749	0,699	0,586	0,849	0,803	0,731	0,567	0,853	0,824	0,733	0,542	0,854	0,843	0,670	0,000	0,86	0,852	0,737	0,000	0,852
	30%	0,746	0,734	0,574	0,849	0,804	0,731	0,570	0,853	0,825	0,733	0,556	0,854	0,844	0,670	0,000	0,86	0,851	0,737	0,000	0,852
	50%	0,736	0,734	0,574	0,849	0,804	0,735	0,559	0,853	0,822	0,736	0,527	0,854	0,840	0,671	0,000	0,86	0,852	0,737	0,000	0,852

Źródło: obliczenia własne.

Tabela 20. Jakość klasyfikacji dla zbioru *kurier*, dla metody IG w pierwszym etapie.

Miara	% terminów	Zbiór uczący 3%				Zbiór uczący 6%				Zbiór uczący 10%				Zbiór uczący 15%				Zbiór uczący 25%			
		NB	SVM	GLM	KOR	NB	SVM	GLM	KOR	NB	SVM	GLM	KOR	NB	SVM	GLM	KOR	NB	SVM	GLM	KOR
Dokładność	1%	0,894	0,909	0,880	0,897	0,893	0,914	0,887	0,9	0,900	0,924	0,900	0,901	0,914	0,930	0,000	0,901	0,923	0,936	0,000	0,902
	2%	0,905	0,908	0,881	0,897	0,912	0,910	0,881	0,9	0,921	0,923	0,884	0,901	0,928	0,927	0,000	0,901	0,940	0,938	0,000	0,902
	3%	0,906	0,905	0,881	0,897	0,918	0,911	0,882	0,9	0,920	0,920	0,883	0,901	0,936	0,928	0,000	0,901	0,945	0,936	0,000	0,902
	5%	0,905	0,901	0,882	0,897	0,919	0,906	0,886	0,9	0,933	0,919	0,868	0,901	0,940	0,926	0,000	0,901	0,948	0,937	0,000	0,902
	7%	0,913	0,902	0,882	0,897	0,919	0,903	0,882	0,9	0,931	0,915	0,875	0,901	0,942	0,922	0,000	0,901	0,949	0,935	0,000	0,902
	10%	0,923	0,903	0,877	0,897	0,922	0,903	0,886	0,9	0,929	0,908	0,878	0,901	0,941	0,919	0,000	0,901	0,950	0,932	0,000	0,902
	13%	0,928	0,906	0,877	0,897	0,927	0,906	0,887	0,9	0,927	0,907	0,880	0,901	0,940	0,914	0,000	0,901	0,950	0,926	0,000	0,902
	20%	0,928	0,909	0,882	0,897	0,937	0,912	0,887	0,9	0,940	0,910	0,888	0,901	0,940	0,913	0,000	0,901	0,948	0,920	0,000	0,902
	30%	0,927	0,908	0,882	0,897	0,937	0,913	0,887	0,9	0,943	0,916	0,888	0,901	0,948	0,917	0,000	0,901	0,953	0,920	0,000	0,902
	50%	0,927	0,908	0,882	0,897	0,935	0,913	0,887	0,9	0,943	0,916	0,888	0,901	0,947	0,917	0,000	0,901	0,955	0,920	0,000	0,902
F1	1%	0,943	0,949	0,933	0,93	0,942	0,952	0,937	0,932	0,945	0,957	0,944	0,931	0,952	0,960	0,000	0,932	0,957	0,963	0,000	0,933
	2%	0,948	0,949	0,933	0,93	0,952	0,950	0,933	0,932	0,956	0,957	0,935	0,931	0,960	0,958	0,000	0,932	0,966	0,965	0,000	0,933
	3%	0,948	0,947	0,934	0,93	0,955	0,951	0,934	0,932	0,956	0,955	0,934	0,931	0,964	0,959	0,000	0,932	0,969	0,964	0,000	0,933
	5%	0,948	0,946	0,934	0,93	0,955	0,948	0,936	0,932	0,962	0,955	0,925	0,931	0,966	0,958	0,000	0,932	0,970	0,964	0,000	0,933
	7%	0,952	0,946	0,934	0,93	0,955	0,946	0,934	0,932	0,961	0,953	0,930	0,931	0,967	0,956	0,000	0,932	0,971	0,963	0,000	0,933
	10%	0,957	0,947	0,931	0,93	0,957	0,947	0,936	0,932	0,960	0,949	0,931	0,931	0,967	0,955	0,000	0,932	0,972	0,962	0,000	0,933
	13%	0,959	0,948	0,931	0,93	0,959	0,948	0,936	0,932	0,959	0,949	0,932	0,931	0,966	0,952	0,000	0,932	0,972	0,959	0,000	0,933
	20%	0,959	0,950	0,934	0,93	0,964	0,951	0,937	0,932	0,966	0,950	0,937	0,931	0,966	0,952	0,000	0,932	0,971	0,955	0,000	0,933
	30%	0,959	0,950	0,934	0,93	0,964	0,952	0,937	0,932	0,968	0,953	0,937	0,931	0,970	0,954	0,000	0,932	0,973	0,955	0,000	0,933
	50%	0,959	0,949	0,934	0,93	0,964	0,952	0,937	0,932	0,967	0,953	0,937	0,931	0,970	0,954	0,000	0,932	0,975	0,955	0,000	0,933

Źródło: obliczenia własne.

Tabela 21. Jakość klasyfikacji dla zbioru *kurier*, dla metody CHI w pierwszym etapie.

Miara	% terminów	Zbiór uczący 3%				Zbiór uczący 6%				Zbiór uczący 10%				Zbiór uczący 15%				Zbiór uczący 25%			
		NB	SVM	GLM	KOR	NB	SVM	GLM	KOR	NB	SVM	GLM	KOR	NB	SVM	GLM	KOR	NB	SVM	GLM	KOR
Dokładność	1%	0,918	0,903	0,881	0,897	0,923	0,915	0,897	0,9	0,927	0,919	0,914	0,901	0,937	0,929	0,000	0,901	0,941	0,933	0,000	0,902
	2%	0,922	0,906	0,868	0,897	0,929	0,914	0,884	0,9	0,935	0,924	0,893	0,901	0,942	0,931	0,000	0,901	0,949	0,940	0,000	0,902
	3%	0,923	0,906	0,866	0,897	0,932	0,912	0,864	0,9	0,938	0,925	0,879	0,901	0,944	0,930	0,000	0,901	0,950	0,940	0,000	0,902
	5%	0,925	0,903	0,870	0,897	0,935	0,911	0,864	0,9	0,941	0,921	0,868	0,901	0,947	0,930	0,000	0,901	0,952	0,939	0,000	0,902
	7%	0,925	0,901	0,881	0,897	0,934	0,909	0,869	0,9	0,942	0,920	0,870	0,901	0,948	0,927	0,000	0,901	0,953	0,937	0,000	0,902
	10%	0,927	0,900	0,866	0,897	0,934	0,908	0,873	0,9	0,942	0,918	0,870	0,901	0,948	0,924	0,000	0,901	0,955	0,934	0,000	0,902
	13%	0,928	0,906	0,860	0,897	0,935	0,907	0,880	0,9	0,942	0,915	0,874	0,901	0,948	0,924	0,000	0,901	0,955	0,934	0,000	0,902
	20%	0,927	0,908	0,887	0,897	0,937	0,912	0,877	0,9	0,941	0,911	0,876	0,901	0,948	0,919	0,000	0,901	0,955	0,932	0,000	0,902
	30%	0,927	0,908	0,887	0,897	0,936	0,912	0,877	0,9	0,943	0,914	0,887	0,901	0,947	0,916	0,000	0,901	0,955	0,920	0,000	0,902
	50%	0,927	0,908	0,887	0,897	0,935	0,912	0,877	0,9	0,942	0,915	0,887	0,901	0,947	0,917	0,000	0,901	0,956	0,919	0,000	0,902
F1	1%	0,954	0,946	0,933	0,93	0,957	0,953	0,943	0,932	0,959	0,955	0,952	0,931	0,964	0,960	0,000	0,932	0,967	0,962	0,000	0,933
	2%	0,956	0,948	0,925	0,93	0,960	0,953	0,935	0,932	0,963	0,957	0,940	0,931	0,967	0,961	0,000	0,932	0,971	0,966	0,000	0,933
	3%	0,957	0,948	0,924	0,93	0,962	0,951	0,922	0,932	0,965	0,958	0,931	0,931	0,968	0,961	0,000	0,932	0,972	0,966	0,000	0,933
	5%	0,958	0,947	0,926	0,93	0,963	0,951	0,922	0,932	0,967	0,956	0,924	0,931	0,970	0,961	0,000	0,932	0,973	0,966	0,000	0,933
	7%	0,958	0,946	0,933	0,93	0,963	0,950	0,925	0,932	0,967	0,955	0,925	0,931	0,970	0,959	0,000	0,932	0,973	0,965	0,000	0,933
	10%	0,959	0,945	0,922	0,93	0,963	0,949	0,928	0,932	0,967	0,954	0,925	0,931	0,971	0,958	0,000	0,932	0,974	0,963	0,000	0,933
	13%	0,960	0,948	0,918	0,93	0,963	0,949	0,932	0,932	0,967	0,953	0,928	0,931	0,970	0,957	0,000	0,932	0,974	0,963	0,000	0,933
	20%	0,959	0,950	0,936	0,93	0,964	0,951	0,930	0,932	0,967	0,951	0,929	0,931	0,970	0,955	0,000	0,932	0,974	0,962	0,000	0,933
	30%	0,959	0,950	0,936	0,93	0,964	0,951	0,930	0,932	0,967	0,952	0,936	0,931	0,970	0,953	0,000	0,932	0,974	0,956	0,000	0,933
	50%	0,959	0,949	0,936	0,93	0,963	0,952	0,930	0,932	0,967	0,953	0,935	0,931	0,970	0,954	0,000	0,932	0,975	0,955	0,000	0,933

Źródło: obliczenia własne.

Tabela 22. Jakość klasyfikacji dla zbioru *kurierzbil*, dla metody IG w pierwszym etapie.

Miara	% terminów	Zbiór uczący 3%				Zbiór uczący 6%				Zbiór uczący 10%				Zbiór uczący 15%				Zbiór uczący 25%			
		NB	SVM	GLM	KOR	NB	SVM	GLM	KOR	NB	SVM	GLM	KOR	NB	SVM	GLM	KOR	NB	SVM	GLM	KOR
Dokładność	1%	0,623	0,725	0,698	0,88	0,710	0,804	0,761	0,887	0,715	0,800	0,819	0,886	0,702	0,814	0,834	0,888	0,791	0,817	0,000	0,889
	2%	0,689	0,731	0,690	0,88	0,746	0,736	0,689	0,887	0,744	0,797	0,741	0,886	0,750	0,821	0,784	0,888	0,811	0,828	0,000	0,889
	3%	0,729	0,733	0,690	0,88	0,756	0,741	0,689	0,887	0,759	0,760	0,695	0,886	0,763	0,827	0,755	0,888	0,824	0,839	0,000	0,889
	5%	0,759	0,734	0,690	0,88	0,774	0,742	0,695	0,887	0,778	0,747	0,680	0,886	0,772	0,785	0,695	0,888	0,834	0,843	0,000	0,889
	7%	0,779	0,735	0,692	0,88	0,802	0,742	0,692	0,887	0,785	0,749	0,679	0,886	0,776	0,750	0,666	0,888	0,832	0,816	0,000	0,889
	10%	0,776	0,734	0,694	0,88	0,832	0,741	0,693	0,887	0,809	0,749	0,676	0,886	0,790	0,751	0,653	0,888	0,822	0,764	0,000	0,889
	13%	0,775	0,732	0,693	0,88	0,846	0,741	0,696	0,887	0,811	0,750	0,678	0,886	0,818	0,752	0,601	0,888	0,828	0,755	0,000	0,889
	20%	0,777	0,730	0,693	0,88	0,843	0,741	0,698	0,887	0,864	0,750	0,646	0,886	0,835	0,751	0,612	0,888	0,857	0,755	0,000	0,889
	30%	0,768	0,730	0,693	0,88	0,836	0,739	0,698	0,887	0,860	0,749	0,670	0,886	0,877	0,751	0,556	0,888	0,888	0,755	0,000	0,889
	50%	0,761	0,728	0,693	0,88	0,818	0,739	0,688	0,887	0,861	0,749	0,672	0,886	0,864	0,751	0,552	0,888	0,894	0,756	0,000	0,889
F1	1%	0,533	0,790	0,756	0,917	0,635	0,839	0,796	0,92	0,648	0,838	0,842	0,919	0,633	0,849	0,857	0,919	0,769	0,853	0,000	0,92
	2%	0,647	0,794	0,747	0,917	0,713	0,798	0,744	0,92	0,700	0,836	0,786	0,919	0,710	0,854	0,818	0,919	0,797	0,861	0,000	0,92
	3%	0,700	0,795	0,749	0,917	0,728	0,800	0,748	0,92	0,729	0,813	0,757	0,919	0,735	0,859	0,800	0,919	0,815	0,869	0,000	0,92
	5%	0,740	0,796	0,748	0,917	0,758	0,801	0,753	0,92	0,764	0,804	0,743	0,919	0,754	0,832	0,755	0,919	0,831	0,872	0,000	0,92
	7%	0,761	0,796	0,749	0,917	0,794	0,801	0,750	0,92	0,772	0,805	0,742	0,919	0,763	0,809	0,708	0,919	0,832	0,855	0,000	0,92
	10%	0,746	0,796	0,754	0,917	0,831	0,801	0,749	0,92	0,801	0,805	0,737	0,919	0,781	0,809	0,690	0,919	0,823	0,820	0,000	0,92
	13%	0,742	0,795	0,753	0,917	0,845	0,800	0,750	0,92	0,802	0,806	0,743	0,919	0,817	0,810	0,617	0,919	0,829	0,814	0,000	0,92
	20%	0,743	0,793	0,754	0,917	0,838	0,801	0,753	0,92	0,864	0,806	0,695	0,919	0,835	0,809	0,647	0,919	0,863	0,814	0,000	0,92
	30%	0,724	0,793	0,753	0,917	0,828	0,800	0,753	0,92	0,857	0,806	0,726	0,919	0,881	0,809	0,574	0,919	0,897	0,814	0,000	0,92
	50%	0,711	0,792	0,753	0,917	0,798	0,800	0,741	0,92	0,856	0,805	0,730	0,919	0,865	0,809	0,556	0,919	0,899	0,814	0,000	0,92

Źródło: obliczenia własne.

Tabela 23. Jakość klasyfikacji dla zbioru *kurierzbil*, dla metody CHI w pierwszym etapie.

Miara	% terminów	Zbiór uczący 3%				Zbiór uczący 6%				Zbiór uczący 10%				Zbiór uczący 15%				Zbiór uczący 25%			
		NB	SVM	GLM	KOR	NB	SVM	GLM	KOR	NB	SVM	GLM	KOR	NB	SVM	GLM	KOR	NB	SVM	GLM	KOR
Dokładność	1%	0,632	0,722	0,703	0,88	0,738	0,803	0,763	0,887	0,744	0,813	0,820	0,886	0,736	0,813	0,000	0,888	0,805	0,812	0,000	0,889
	2%	0,674	0,729	0,688	0,88	0,760	0,736	0,666	0,887	0,771	0,804	0,743	0,886	0,764	0,821	0,000	0,888	0,819	0,828	0,000	0,889
	3%	0,714	0,732	0,687	0,88	0,766	0,739	0,678	0,887	0,774	0,760	0,704	0,886	0,773	0,828	0,000	0,888	0,834	0,838	0,000	0,889
	5%	0,747	0,732	0,689	0,88	0,780	0,741	0,679	0,887	0,779	0,746	0,685	0,886	0,789	0,778	0,000	0,888	0,835	0,845	0,000	0,889
	7%	0,759	0,733	0,689	0,88	0,799	0,741	0,680	0,887	0,787	0,748	0,653	0,886	0,801	0,751	0,000	0,888	0,832	0,807	0,000	0,889
	10%	0,763	0,733	0,689	0,88	0,821	0,741	0,686	0,887	0,810	0,749	0,670	0,886	0,819	0,751	0,000	0,888	0,832	0,764	0,000	0,889
	13%	0,767	0,733	0,689	0,88	0,826	0,741	0,653	0,887	0,814	0,749	0,642	0,886	0,825	0,752	0,000	0,888	0,849	0,754	0,000	0,889
	20%	0,769	0,732	0,687	0,88	0,827	0,741	0,644	0,887	0,840	0,750	0,654	0,886	0,837	0,751	0,000	0,888	0,867	0,755	0,000	0,889
	30%	0,761	0,730	0,688	0,88	0,827	0,740	0,680	0,887	0,840	0,750	0,684	0,886	0,850	0,751	0,000	0,888	0,885	0,755	0,000	0,889
	50%	0,757	0,729	0,687	0,88	0,817	0,739	0,650	0,887	0,856	0,749	0,641	0,886	0,848	0,751	0,000	0,888	0,893	0,755	0,000	0,889
F1	1%	0,542	0,788	0,757	0,917	0,681	0,839	0,799	0,92	0,690	0,847	0,843	0,919	0,685	0,849	0,000	0,919	0,786	0,851	0,000	0,92
	2%	0,619	0,793	0,746	0,917	0,732	0,798	0,725	0,92	0,743	0,841	0,787	0,919	0,729	0,854	0,000	0,919	0,806	0,861	0,000	0,92
	3%	0,677	0,795	0,745	0,917	0,746	0,799	0,740	0,92	0,752	0,813	0,762	0,919	0,746	0,859	0,000	0,919	0,827	0,868	0,000	0,92
	5%	0,716	0,795	0,746	0,917	0,766	0,801	0,741	0,92	0,766	0,804	0,745	0,919	0,777	0,827	0,000	0,919	0,832	0,873	0,000	0,92
	7%	0,728	0,796	0,747	0,917	0,788	0,801	0,742	0,92	0,774	0,805	0,691	0,919	0,795	0,809	0,000	0,919	0,831	0,849	0,000	0,92
	10%	0,726	0,795	0,746	0,917	0,816	0,800	0,745	0,92	0,801	0,806	0,728	0,919	0,818	0,809	0,000	0,919	0,834	0,820	0,000	0,92
	13%	0,729	0,796	0,747	0,917	0,818	0,801	0,691	0,92	0,805	0,806	0,687	0,919	0,825	0,810	0,000	0,919	0,855	0,814	0,000	0,92
	20%	0,730	0,794	0,748	0,917	0,815	0,801	0,686	0,92	0,833	0,806	0,697	0,919	0,837	0,809	0,000	0,919	0,876	0,814	0,000	0,92
	30%	0,714	0,794	0,748	0,917	0,814	0,800	0,740	0,92	0,831	0,806	0,746	0,919	0,847	0,809	0,000	0,919	0,893	0,814	0,000	0,92
	50%	0,704	0,793	0,748	0,917	0,797	0,800	0,689	0,92	0,850	0,806	0,694	0,919	0,842	0,809	0,000	0,919	0,897	0,814	0,000	0,92

Źródło: obliczenia własne.

Tabela 24. Jakość klasyfikacji dla zbioru *perfumy*, dla metody IG w pierwszym etapie.

Miara	% terminów	Zbiór uczący 3%				Zbiór uczący 6%				Zbiór uczący 10%				Zbiór uczący 15%				Zbiór uczący 25%			
		NB	SVM	GLM	KOR	NB	SVM	GLM	KOR	NB	SVM	GLM	KOR	NB	SVM	GLM	KOR	NB	SVM	GLM	KOR
Dokładność	1%	0,898	0,821	0,858	0,899	0,900	0,896	0,887	0,912	0,904	0,662	0,889	0,915	0,898	0,899	0,898	0,916	0,901	0,899	0,904	0,915
	2%	0,899	0,903	0,864	0,899	0,904	0,900	0,883	0,912	0,908	0,836	0,888	0,915	0,901	0,901	0,892	0,916	0,908	0,901	0,901	0,915
	3%	0,901	0,902	0,866	0,899	0,903	0,901	0,886	0,912	0,907	0,895	0,889	0,915	0,900	0,903	0,885	0,916	0,909	0,904	0,890	0,915
	5%	0,899	0,902	0,861	0,899	0,907	0,900	0,864	0,912	0,906	0,902	0,885	0,915	0,900	0,900	0,889	0,916	0,907	0,906	0,888	0,915
	7%	0,898	0,902	0,863	0,899	0,904	0,900	0,874	0,912	0,904	0,901	0,888	0,915	0,901	0,901	0,888	0,916	0,905	0,907	0,887	0,915
	10%	0,900	0,901	0,863	0,899	0,902	0,900	0,874	0,912	0,907	0,902	0,854	0,915	0,905	0,900	0,875	0,916	0,906	0,907	0,888	0,915
	13%	0,901	0,901	0,864	0,899	0,901	0,899	0,874	0,912	0,906	0,901	0,865	0,915	0,908	0,904	0,852	0,916	0,911	0,905	0,868	0,915
	20%	0,900	0,900	0,864	0,899	0,903	0,899	0,871	0,912	0,903	0,901	0,876	0,915	0,908	0,903	0,830	0,916	0,918	0,906	0,815	0,915
	30%	0,900	0,900	0,864	0,899	0,901	0,898	0,869	0,912	0,906	0,901	0,872	0,915	0,907	0,903	0,823	0,916	0,917	0,905	0,818	0,915
	50%	0,900	0,900	0,864	0,899	0,899	0,898	0,869	0,912	0,905	0,901	0,869	0,915	0,905	0,903	0,823	0,916	0,914	0,906	0,810	0,915
F1	1%	0,946	0,852	0,921	0,94	0,947	0,945	0,939	0,947	0,949	#N/D	0,940	0,948	0,946	0,945	0,945	0,948	0,948	0,944	0,948	0,947
	2%	0,946	0,948	0,925	0,94	0,949	0,947	0,936	0,947	0,951	0,880	0,939	0,948	0,948	0,947	0,941	0,948	0,951	0,946	0,946	0,947
	3%	0,947	0,948	0,926	0,94	0,949	0,947	0,938	0,947	0,950	0,943	0,939	0,948	0,947	0,948	0,937	0,948	0,952	0,948	0,940	0,947
	5%	0,946	0,948	0,923	0,94	0,951	0,947	0,924	0,947	0,950	0,948	0,937	0,948	0,947	0,947	0,940	0,948	0,951	0,949	0,939	0,947
	7%	0,946	0,948	0,924	0,94	0,949	0,947	0,930	0,947	0,949	0,948	0,939	0,948	0,947	0,947	0,939	0,948	0,949	0,949	0,938	0,947
	10%	0,947	0,948	0,924	0,94	0,948	0,947	0,930	0,947	0,950	0,948	0,918	0,948	0,949	0,947	0,931	0,948	0,950	0,950	0,938	0,947
	13%	0,947	0,947	0,925	0,94	0,948	0,947	0,930	0,947	0,950	0,948	0,925	0,948	0,951	0,949	0,917	0,948	0,953	0,948	0,927	0,947
	20%	0,947	0,947	0,925	0,94	0,948	0,946	0,928	0,947	0,948	0,948	0,931	0,948	0,951	0,949	0,902	0,948	0,956	0,950	0,893	0,947
	30%	0,947	0,947	0,925	0,94	0,948	0,946	0,926	0,947	0,950	0,947	0,928	0,948	0,951	0,949	0,898	0,948	0,955	0,950	0,895	0,947
	50%	0,947	0,947	0,925	0,94	0,947	0,946	0,926	0,947	0,949	0,948	0,926	0,948	0,950	0,949	0,898	0,948	0,954	0,950	0,889	0,947

Źródło: obliczenia własne.

Tabela 25. Jakość klasyfikacji dla zbioru *perfumy*, dla metody CHI w pierwszym etapie.

Miara	% terminów	Zbiór uczący 3%				Zbiór uczący 6%				Zbiór uczący 10%				Zbiór uczący 15%				Zbiór uczący 25%			
		NB	SVM	GLM	KOR	NB	SVM	GLM	KOR	NB	SVM	GLM	KOR	NB	SVM	GLM	KOR	NB	SVM	GLM	KOR
Dokładność	1%	0,901	0,837	0,872	0,899	0,908	0,899	0,892	0,912	0,912	0,809	0,898	0,915	0,897	0,676	0,885	0,916	0,897	0,705	0,894	0,915
	2%	0,902	0,897	0,868	0,899	0,911	0,900	0,873	0,912	0,914	0,870	0,890	0,915	0,894	0,734	0,864	0,916	0,900	0,791	0,885	0,915
	3%	0,902	0,902	0,875	0,899	0,910	0,900	0,880	0,912	0,915	0,870	0,889	0,915	0,895	0,796	0,809	0,916	0,903	0,843	0,869	0,915
	5%	0,902	0,902	0,874	0,899	0,909	0,900	0,880	0,912	0,914	0,899	0,876	0,915	0,902	0,890	0,760	0,916	0,905	0,808	0,814	0,915
	7%	0,902	0,901	0,874	0,899	0,907	0,900	0,880	0,912	0,912	0,900	0,869	0,915	0,905	0,901	0,746	0,916	0,909	0,883	0,795	0,915
	10%	0,903	0,901	0,872	0,899	0,906	0,899	0,880	0,912	0,911	0,901	0,853	0,915	0,905	0,901	0,727	0,916	0,913	0,895	0,777	0,915
	13%	0,903	0,901	0,874	0,899	0,905	0,899	0,880	0,912	0,910	0,901	0,847	0,915	0,904	0,901	0,729	0,916	0,915	0,904	0,782	0,915
	20%	0,901	0,900	0,871	0,899	0,905	0,899	0,863	0,912	0,909	0,901	0,845	0,915	0,903	0,901	0,748	0,916	0,913	0,905	0,775	0,915
	30%	0,900	0,901	0,871	0,899	0,901	0,899	0,875	0,912	0,909	0,900	0,839	0,915	0,906	0,902	0,775	0,916	0,910	0,905	0,727	0,915
	50%	0,900	0,901	0,871	0,899	0,900	0,899	0,875	0,912	0,905	0,901	0,842	0,915	0,905	0,902	0,779	0,916	0,913	0,905	0,780	0,915
F1	1%	0,947	0,890	0,929	0,94	0,951	0,947	0,941	0,947	0,953	0,879	0,945	0,948	0,945	0,781	0,938	0,948	0,946	0,810	0,943	0,947
	2%	0,948	0,945	0,927	0,94	0,952	0,947	0,930	0,947	0,954	0,924	0,940	0,948	0,944	0,830	0,925	0,948	0,947	0,875	0,937	0,947
	3%	0,948	0,948	0,931	0,94	0,952	0,947	0,934	0,947	0,954	0,924	0,939	0,948	0,944	0,878	0,890	0,948	0,948	0,909	0,928	0,947
	5%	0,947	0,948	0,931	0,94	0,951	0,947	0,934	0,947	0,954	0,946	0,931	0,948	0,948	0,941	0,857	0,948	0,949	0,845	0,892	0,947
	7%	0,947	0,948	0,931	0,94	0,951	0,947	0,934	0,947	0,953	0,947	0,928	0,948	0,949	0,947	0,847	0,948	0,951	0,935	0,879	0,947
	10%	0,948	0,948	0,930	0,94	0,950	0,947	0,935	0,947	0,952	0,947	0,917	0,948	0,950	0,947	0,834	0,948	0,953	0,943	0,867	0,947
	13%	0,948	0,947	0,931	0,94	0,950	0,947	0,935	0,947	0,952	0,947	0,913	0,948	0,949	0,947	0,835	0,948	0,954	0,949	0,871	0,947
	20%	0,948	0,947	0,929	0,94	0,950	0,946	0,922	0,947	0,951	0,947	0,912	0,948	0,948	0,948	0,846	0,948	0,953	0,949	0,866	0,947
	30%	0,947	0,947	0,929	0,94	0,948	0,946	0,931	0,947	0,951	0,947	0,906	0,948	0,950	0,948	0,863	0,948	0,952	0,949	0,824	0,947
	50%	0,947	0,947	0,929	0,94	0,947	0,946	0,931	0,947	0,949	0,947	0,908	0,948	0,949	0,948	0,866	0,948	0,953	0,949	0,868	0,947

Źródło: obliczenia własne.

Tabela 26. Jakość klasyfikacji dla zbioru *perfumyzbil*, dla metody IG w pierwszym etapie.

Miara	% terminów	Zbiór uczący 3%				Zbiór uczący 6%				Zbiór uczący 10%				Zbiór uczący 15%				Zbiór uczący 25%			
		NB	SVM	GLM	KOR	NB	SVM	GLM	KOR	NB	SVM	GLM	KOR	NB	SVM	GLM	KOR	NB	SVM	GLM	KOR
Dokładność	1%	0,519	0,575	0,619	0,82	0,607	0,562	0,671	0,878	0,626	0,606	0,693	0,883	0,665	0,674	0,734	0,885	0,740	0,741	0,786	0,89
	2%	0,559	0,574	0,612	0,82	0,645	0,590	0,653	0,878	0,651	0,602	0,682	0,883	0,680	0,598	0,706	0,885	0,770	0,657	0,744	0,89
	3%	0,581	0,584	0,600	0,82	0,664	0,587	0,644	0,878	0,674	0,605	0,679	0,883	0,699	0,599	0,681	0,885	0,778	0,607	0,707	0,89
	5%	0,598	0,600	0,551	0,82	0,658	0,585	0,627	0,878	0,698	0,599	0,673	0,883	0,719	0,599	0,665	0,885	0,795	0,604	0,702	0,89
	7%	0,601	0,599	0,558	0,82	0,683	0,616	0,564	0,878	0,694	0,599	0,674	0,883	0,740	0,593	0,665	0,885	0,816	0,600	0,675	0,89
	10%	0,596	0,595	0,559	0,82	0,681	0,606	0,588	0,878	0,705	0,609	0,623	0,883	0,733	0,592	0,656	0,885	0,813	0,627	0,671	0,89
	13%	0,598	0,608	0,560	0,82	0,676	0,606	0,600	0,878	0,713	0,608	0,605	0,883	0,752	0,604	0,641	0,885	0,817	0,625	0,669	0,89
	20%	0,619	0,609	0,540	0,82	0,663	0,605	0,600	0,878	0,707	0,604	0,610	0,883	0,751	0,602	0,627	0,885	0,801	0,619	0,641	0,89
	30%	0,617	0,600	0,539	0,82	0,680	0,612	0,601	0,878	0,702	0,601	0,623	0,883	0,742	0,600	0,644	0,885	0,827	0,618	0,600	0,89
	50%	0,615	0,595	0,539	0,82	0,671	0,610	0,616	0,878	0,717	0,600	0,604	0,883	0,757	0,599	0,629	0,885	0,799	0,615	0,615	0,89
F1	1%	0,500	0,663	0,682	0,84	0,631	#N/D	0,747	0,899	0,644	0,734	0,761	0,903	0,688	0,770	0,787	0,904	0,737	0,808	0,827	0,905
	2%	0,561	0,687	0,656	0,84	0,679	0,690	0,732	0,899	0,680	0,734	0,754	0,903	0,697	0,730	0,767	0,904	0,780	0,764	0,792	0,905
	3%	0,593	0,695	0,662	0,84	0,706	0,691	0,719	0,899	0,704	0,736	0,746	0,903	0,720	0,731	0,747	0,904	0,792	0,737	0,766	0,905
	5%	0,611	0,719	0,597	0,84	0,699	0,689	0,701	0,899	0,730	0,732	0,742	0,903	0,746	0,731	0,731	0,904	0,815	0,736	0,763	0,905
	7%	0,610	0,719	0,618	0,84	0,726	0,739	0,608	0,899	0,723	0,732	0,743	0,903	0,771	0,728	0,727	0,904	0,838	0,734	0,735	0,905
	10%	0,596	0,718	0,620	0,84	0,716	0,735	0,644	0,899	0,740	0,736	0,678	0,903	0,764	0,728	0,718	0,904	0,838	0,746	0,727	0,905
	13%	0,594	0,734	0,621	0,84	0,705	0,735	0,665	0,899	0,745	0,736	0,658	0,903	0,786	0,733	0,698	0,904	0,841	0,745	0,726	0,905
	20%	0,612	0,735	0,588	0,84	0,683	0,735	0,665	0,899	0,732	0,734	0,668	0,903	0,780	0,732	0,679	0,904	0,828	0,742	0,695	0,905
	30%	0,600	0,731	0,587	0,84	0,696	0,738	0,668	0,899	0,722	0,732	0,681	0,903	0,770	0,731	0,703	0,904	0,841	0,741	0,640	0,905
	50%	0,592	0,729	0,587	0,84	0,671	0,737	0,687	0,899	0,722	0,732	0,653	0,903	0,776	0,731	0,675	0,904	0,816	0,740	0,670	0,905

Źródło: obliczenia własne.

Tabela 27. Jakość klasyfikacji dla zbioru *perfumyzbil*, dla metody CHI w pierwszym etapie.

Miara	% terminów	Zbiór uczący 3%				Zbiór uczący 6%				Zbiór uczący 10%				Zbiór uczący 15%				Zbiór uczący 25%			
		NB	SVM	GLM	KOR	NB	SVM	GLM	KOR	NB	SVM	GLM	KOR	NB	SVM	GLM	KOR	NB	SVM	GLM	KOR
Dokładność	1%	0,560	0,562	0,589	0,82	0,651	0,552	0,643	0,878	0,662	0,624	0,691	0,883	0,716	0,658	0,731	0,885	0,731	0,759	0,776	0,89
	2%	0,580	0,606	0,591	0,82	0,677	0,593	0,669	0,878	0,698	0,603	0,661	0,883	0,734	0,600	0,696	0,885	0,762	0,677	0,741	0,89
	3%	0,607	0,610	0,622	0,82	0,677	0,594	0,656	0,878	0,714	0,604	0,665	0,883	0,744	0,600	0,682	0,885	0,778	0,606	0,712	0,89
	5%	0,618	0,604	0,550	0,82	0,685	0,601	0,656	0,878	0,716	0,596	0,661	0,883	0,745	0,601	0,678	0,885	0,775	0,603	0,709	0,89
	7%	0,625	0,605	0,549	0,82	0,689	0,605	0,635	0,878	0,719	0,593	0,651	0,883	0,760	0,606	0,672	0,885	0,804	0,601	0,703	0,89
	10%	0,629	0,594	0,551	0,82	0,694	0,604	0,628	0,878	0,716	0,595	0,625	0,883	0,757	0,604	0,663	0,885	0,806	0,604	0,683	0,89
	13%	0,632	0,599	0,543	0,82	0,689	0,603	0,640	0,878	0,727	0,597	0,626	0,883	0,769	0,602	0,650	0,885	0,804	0,603	0,680	0,89
	20%	0,619	0,599	0,527	0,82	0,695	0,605	0,632	0,878	0,724	0,594	0,628	0,883	0,766	0,600	0,641	0,885	0,802	0,622	0,663	0,89
	30%	0,622	0,591	0,542	0,82	0,684	0,612	0,620	0,878	0,730	0,594	0,638	0,883	0,769	0,597	0,648	0,885	0,806	0,619	0,663	0,89
	50%	0,622	0,587	0,541	0,82	0,673	0,609	0,632	0,878	0,721	0,602	0,595	0,883	0,760	0,598	0,626	0,885	0,804	0,616	0,657	0,89
F1	1%	0,573	#N/D	0,617	0,84	0,678	#N/D	0,706	0,899	0,689	0,742	0,752	0,903	0,737	0,759	0,786	0,904	0,725	0,815	0,818	0,905
	2%	0,603	0,731	0,627	0,84	0,711	0,683	0,745	0,899	0,734	0,734	0,731	0,903	0,754	0,731	0,761	0,904	0,772	0,774	0,789	0,905
	3%	0,629	0,735	0,706	0,84	0,716	0,714	0,728	0,899	0,750	0,735	0,732	0,903	0,765	0,732	0,744	0,904	0,792	0,737	0,770	0,905
	5%	0,635	0,732	0,562	0,84	0,724	0,725	0,724	0,899	0,753	0,731	0,731	0,903	0,769	0,732	0,738	0,904	0,792	0,735	0,767	0,905
	7%	0,638	0,733	0,563	0,84	0,730	0,730	0,690	0,899	0,753	0,729	0,724	0,903	0,787	0,734	0,731	0,904	0,824	0,734	0,758	0,905
	10%	0,644	0,728	0,563	0,84	0,725	0,730	0,688	0,899	0,748	0,730	0,690	0,903	0,783	0,733	0,723	0,904	0,830	0,735	0,741	0,905
	13%	0,646	0,731	0,557	0,84	0,714	0,733	0,705	0,899	0,752	0,731	0,688	0,903	0,798	0,732	0,704	0,904	0,828	0,735	0,739	0,905
	20%	0,616	0,731	0,516	0,84	0,717	0,734	0,688	0,899	0,743	0,729	0,689	0,903	0,787	0,731	0,700	0,904	0,828	0,744	0,724	0,905
	30%	0,611	0,728	0,559	0,84	0,699	0,738	0,671	0,899	0,745	0,729	0,704	0,903	0,787	0,730	0,707	0,904	0,825	0,742	0,718	0,905
	50%	0,603	0,726	0,559	0,84	0,674	0,736	0,690	0,899	0,725	0,733	0,642	0,903	0,777	0,730	0,673	0,904	0,819	0,740	0,713	0,905

Źródło: obliczenia własne.

Bibliografia

Abbasi A., Chen H., Salem A., (2008), *Sentiment Analysis in Multiple Languages: Feature Selection for Opinion Classification in Web Forums*, „ACM Transactions on Information Systems", vol. 26, issue 3, s. 1–34, https://doi.org/10.1145/1361684.1361685

Agarwal A., Xie B., Vovsha I., Rambow O., Passonneau R., (2011), *Sentiment Analysis of Twitter Data*, [w:] Proceedings of the Workshop on Language in Social Media (LSM), s. 30–38.

Agarwal B., Mittal N., (2012), *Categorical Probability Proportion Difference (CPPD): A Feature Selection Method for Sentiment Classification*, [w:] Proceedings of the 2nd Workshop on Sentiment Analysis Where AI Meets Psychology (SAAIP), Mumbai, India, s. 17–26.

Agarwal B., Mittal N., (2016), *Prominent Feature Extraction for Sentiment Analysis*, Springer International Publishing, Cham.

Aggarwal C. C., (2018), *Machine Learning for Text*, Springer International Publishing, Cham.

Agnihotri D., Verma K., Tripathi P., (2016), *Computing Correlative Association of Terms for Automatic Classification of Text Documents*, [w:] Proceedings of the International Symposium on Computer Vision and the Internet, Association for Computing Machinery, New York, https://doi.org/10.1145/2983402.2983424

Agnihotri D., Verma K., Tripathi P., Singh B., (2019), *Soft Voting Technique to Improve the Performance of Global Filter Based Feature Selection in Text Corpus*, „Applied Intelligence", vol. 49, issue 4, s. 1597–1619, https://doi.org/10.1007/s10489-018-1349-1

Bagheri A., Saraee M., Jong de F., (2013), *Sentiment Classification in Persian: Introducing a Mutual Information-Based Method for Feature Selection*, „21st Iranian Conference on Electrical Engineering (ICEE)", s. 1–6, https://doi.org/10.1109/IranianCEE.2013.6599671

Bahassine S., Madani A., Kissi M., (2016), *An Improved Chi-Square Feature Selection for Arabic Text Classification Using Decision Tree*, „11th International Conference on Intelligent Systems: Theories and Applications (SITA)", s. 1–5, https://doi.org/ 10.1109/SITA.2016.7772289

Bahassine S., Madani A., Al-Sarem M., Kissi M., (2018), *Feature Selection Using an Improved Chi-Square for Arabic Text Classification*, „Journal of King Saud University – Computer and Information Sciences", vol. 32, issue 2, s. 225–231, https://doi.org/10.1016/j.jksuci.2018.05.010

Bakus J., Kamel M., (2006), *Higher Order Feature Selection for Text Classification*, „Knowledge Information Systems", vol. 9, issue 4, s. 468–491, https://doi.org/10.1007/s10115-005-0209-6

Battiti R., (1994), *Using Mutual Information for Selecting Features in Supervised Neural Net Learning*, „IEEE Transactions on Neural Networks", vol. 5, issue 4, s. 537–550, https://doi.org/10.1109/72.298224

Blitzer J., Dredze M., Pereira F., (2007), *Biographies, Bollywood, Boom-Boxes and Blenders: Domain Adaptation for Sentiment Classification*, [w:] Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, Association of Computational Linguistics, Prague, s. 440–447, https://aclanthology.org/P07-1056.pdf (dostęp: 11.02.2022).

Boser B. E., Guyon I. M., Vapnik, V. N., (1992), *A Training Algorithm for Optimal Margin Classifiers*, [w:] Proceedings of the 5th Annual Workshop on Computational Learning Theory, Association for Computing Machinery, New York, s. 144–152, https://doi.org/10.1145/130385.130401

Cai J., Song F., (2008), *Maximum Entropy Modeling with Feature Selection for Text Categorization*, [w:] Proceedings of the 4th Asia Information Retrieval Conference on Information Retrieval Technology, s. 549–554.

Carvalho F., Guedes G. P., (2020), *TF-IDFC-RF: A Novel Supervised Term Weighting Scheme for Sentiment Analysis*, https://arxiv.org/pdf/2003.07193.pdf (dostęp: 20.07.2021).

Chen J., Huang H., Tian S., Qu Y., (2009), *Feature Selection for Text Classification with Naïve Bayes*, „Expert Systems with Applications", vol. 36, issue 3, https://doi.org/10.1016/j.eswa.2008.06.054

Chen X., Ma J., Lu Y., (2013), *Feature Selection for Chinese Online Reviews Sentiment Classification*, [w:] Proceedings of the Joint Conference of International Conference on Computational Problem-Solving and International High Speed Intelligent Communication Forum, s. 79–82, https://doi.org/10.1109/ICCPS.2013.6893490

Chen Y., Han B., Hou P., (2014), *New Feature Selection Methods Based on Context Similarity for Text Categorization*, [w:] Proceedings of the 11th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD), https://doi.org/10.1109/FSKD.2014.6980902

Combarro E., Montanes E., Diaz I., Ranilla J., Mones R., (2005), *Introducing a Family of Linear Measures for Feature Selection in Text Categorization*, „IEEE Transactions on Knowledge and Data Engineering", vol. 17, issue 9, s. 1223–1232, https://doi.org/10.1109/TKDE.2005.149

Cortes C., Vapnik V. N., (1995), *Support-Vector Networks*, „Machine Learning", vol. 20, no. 3, s. 273–297.

Dai L., Chen H., Li X., (2011), *Improving Sentiment Classification Using Feature Highlighting and Feature Bagging*, [w:] Proceedings of 11th IEEE International Conference on Data Mining Workshops, s. 61–66.

Dave K., Lawrence S., Pennock D. M., (2003), *Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews*, [w:] Proceedings of the 12th International Conference on World Wide Web (WWW–2003), Association for Computing Machinery, New York, s. 519–528, https://doi.org/10.1145/775152.775226

Davies A., Ghahramani Z., (2011), *Language-Independent Bayesian Sentiment Mining of Twitter*, [w:] Proceedings of the 5th Workshop on Social Network Mining and Analysis, s. 99–107.

Ding X., Tang Y., (2013), *Improved Mutual Information Method for Text Feature Selection*, [w:] Proceedings of the 8th International Conference on Computer Science and Education, s. 163–166, https://doi.org/10.1109/ICCSE.2013.6553903

Domański Cz., Pruska K., (2000), *Nieklasyczne metody statystyczne*, PWE, Warszawa.

Dunning T., (1993), *Accurate Methods for the Statistics of Surprise and Coincidence*, „Computational Linguistics", vol. 19, s. 61–74.

Elakkiya E., Selvakumar S., Velusamy R., (2020), *CIFAS: Community Inspired Firefly Algorithm with Fuzzy Cross-Entropy for Feature Selection in Twitter Spam Detection*, [w:] Proceedings of the 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT), s. 1–7, https://doi.org/10.1109/ICCCNT49239.2020.9225321

Eyheramendy S., Madigan D., (2007), *A Bayesian Feature Selection Score Based on Naïve Bayes Models*, [w:] H. Liu, H. Motoda (eds), *Computational Methods of Feature Selection*, Chapman and Hall, New York, s. 277–294.

Forman G., (2003), *An Extensive Empirical Study of Feature Selection Metrics for Text Classification*, „Journal of Machine Learning Research", vol. 3, s. 1289–1305.

Fragoudis D., Meretakis D., Likothanassis S., (2005), *Best Terms: an Efficient Feature-Selection Algorithm for Text Categorization*, „Knowledge and Information Systems", vol. 8, issue 1, s. 16–33.

Fukumoto F., Suzuki Y., (2015), *Temporal-based Feature Selection and Transfer Learning for Text Categorization*, [w:] Proceedings of the 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K), s. 17–26.

Galavotti L., Sebastiani F., Simi M., (2000), *Experiments on the Use of Feature Selection and Negative Evidence in Automated Text Categorization*, [w:] J. L. Borbinha, T. Baker (eds), *Proceedings of ECDL-00, 4th European Conference on Research and Advanced Technology for Digital Libraries*, Springer Verlag, Lisbon–Heidelberg, s. 59–68.

Gamon M., (2004), *Sentiment Classification on Customer Feedback Data: Noisy Data, Large Feature Vectors, and the Role of Linguistic Analysis*, [w:] Proceedings of the 20th International Conference on Computational Linguistics (COLING), Association for Computational Linguistics, Stroudsburg, PA, s. 841–847, https://doi.org/10.3115/1220355.1220476

Gao Z., Xu Y., Meng F., Qi F., Lin Z., (2014), *Improved Information Gain-Based Feature Selection for Text Categorization*, [w:] Proceedings of the 4th International Conference on Wireless Communications, Vehicular Technology, Information Theory and Aerospace and Electronic Systems (VITAE), s. 1–5, https://doi.org/10.1109/VITAE.2014.6934421

Garnes Ø. L., (2009), *Feature Selection for Text Categorisation*, Norwegian University of Science and Technology, Department of Computer and Information Science, https://ntnuopen.ntnu.no/ntnu-xmlui/bitstream/handle/11250/250768/347827_FULLTEXT01.pdf?sequence=1 (dostęp: 4.02.2022).

Genkin A., Lewis D., Madigan D., (2007), *Large-Scale Bayesian Logistic Regression for Text Categorization*, „Technometrics", vol. 49, no. 3, s. 291–304.

Ghareb A. S., Abu Bakara A., Al-Radaideh Q. A., Hamdan A. R., (2018), *Enhanced Filter Feature Selection Methods for Arabic Text Categorization*, „International Journal of Information Retrieval Research", vol. 8, issue 2, s. 1–24, https://doi.org/10.4018/IJIRR.2018040101

Govindarajan M., (2013), *Sentiment Analysis of Movie Reviews Using Hybrid Method of Naïve Bayes and Genetic Algorithm*, „International Journal of Advanced Computer Research", vol. 3, no. 4, s. 139–145.

Gündüz H., Çataltepe Z., (2015), *Borsa Istanbul (BIST) Daily Prediction Using Financial News and Balanced Feature Selection*, „Journal of Machine Learning", vol. 42, no. 22.

Guyon I., Elisseeff A., (2003), *An Introduction to Variable and Feature Selection*, „The Journal of Machine Learning Research", vol. 3, s. 1157–1182.

Hai N., Nghia N., Le H., Vu Thanh N., (2015), *A Hybrid Feature Selection Method for Vietnamese Text Classification*, Seventh International Conference on Knowledge and Systems Engineering (KSE), https://doi.org/10.1109/KSE.2015.25

Hand D., Mannila H., Smith P., (2005), *Eksploracja danych*, Wydawnictwo Naukowo-Techniczne, Warszawa.

Hatzivassiloglou V., Wiebe J., (2000), *Effects of Adjective Orientation and Gradability on Sentence Subjectivity*, [w:] Proceedings of the International Conference on Computational Linguistics (COLING), Association for Computational Linguistics, Stroudsburg, PA, s. 299–305, https://doi.org/10.3115/990820.990864

Hersh W., Buckley C., Leone T. J., Hickam D., (1994), *OHSUMED*: *An Interactive Retrieval Evaluation and New Large Test Collection for Research*, [w:] B. W. Croft, C. J. van Rijsbergen (eds), *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Dublin, s. 192–201.

Holland J., (1975), *Adaptation in Natural and Artificial Systems*, University of Michigan Press, Michigan, USA.

Hosmer D. W., Lemeshow S., Sturdivant R. X., (2013), *Applied Logistic Regression*, 3rd ed., John Wiley & Sons, New Jersey.

Internet Movie Database, www.imdb.com (dostęp: 15.10.2021).

Iqbal F., Hashmi J., Fung B., Batool R., Khattak A. M., Aleem S., Hung P., (2019), *A Hybrid Framework for Sentiment Analysis Using Genetic Algorithm Based Feature Reduction*, „IEEE Access", vol. 7, s. 14637–14652, https://doi.org/10.1109/ACCESS.2019.2892852

Jiang T., Yu H., (2015), *A Novel Feature Selection Based on Tibetan Grammar for Tibetan Text Classification*, [w:] Proceedings of the 6th IEEE International Conference on Software Engineering and Service Sciences (ICSESS), s. 445–448, https://doi.org/10.1109/ICSESS.2015.7339093

Jiang X.-Y., Shui J., (2013), *An Improved Mutual Information-Based Feature Selection Algorithm for Text Classification*, [w:] Proceedings of the 5th International Conference on Intelligent Human-Machine Systems and Cybernetics, s. 126–129, https://doi.org/10.1109/IHMSC.2013.37

Joachims T., (1998), *Text Categorization with Support Vector Machines: Learning with Many Relevant Features*, [w:] *Machine Learning: ECML-98. Lecture Notes in Computer Science (Lecture Notes in Artificial Intelligence)*, vol. 1398, Springer, Berlin, s. 137–142, https://doi.org/10.1007/BFb0026683

Joshi M., Penstein-Rosé C., (2009), *Generalizing Dependency Features for Opinion Mining*, [w:] Proceedings of the 47th ACL and the 4th IJCNLP Conference, Association for Computational Linguistics, ACL and AFNLP, Suntec, s. 313–316, https://aclanthology.org/P09-2079.pdf (dostęp: 11.02.2022).

Kwak N., Choi Ch.-H., (2002), *Input Feature Selection for Classification Problems*, „IEEE Transactions of Neural Works", vol. 13, no. 1, s. 143–159, https://doi.org/10.1109/72.977291

Lam W., Ho C. Y., (1998), *Using a Generalized Instance Set for Automatic Text Categorization*, [w:] Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, s. 81–89, https://doi.org/10.1145/290941.290961

Largeron C., Moulin C., Géry M., (2011), *Entropy Based Feature Selection for Text Categorization*, [w:] Proceedings of the 2011 ACM Symposium on Applied Computing, Taichung, s. 924–928, https://doi.org/10.1145/1982185.1982389

Lifang Y., Sijun Q., Huan Z., (2017), *Feature Selection Algorithm for Hierarchical Text Classification Using Kullback-Leibler Divergence*, [w:] Proceedings of the IEEE 2nd International Conference on Cloud Computing and Big Data Analysis (ICCCBDA), s. 421–424, https://doi.org/10.1109/ICCCBDA.2017.7951950

Lula P., (2018), *Statystyczne modelowanie zawartości dokumentów tekstowych*, Wydawnictwo Uniwersytetu Ekonomicznego w Krakowie, Kraków.

Lula P., Wójcik K., (2011), *Sentiment Analysis of Consumer Opinions Written in Polish*, „Economics and Management", vol. 16, s. 1286–1291.

Malik A., Novovicova J., (2005), *Information-Theoretic Feature Selection Algorithms for Text Classification*, [w:] Proceedings of the IEEE International Joint Conference on Neural Networks, Montreal, vol. 5, s. 3272–3278, https://doi.org/10.1109/IJCNN.2005.1556452

McCallum A., Nigam K., (1998), *A Comparison of Event Models for Naïve Bayes Text Classification*, [w:] Proceedings of the AAAI/ICML-98 Workshop on Learning for Text Categorization, AAAI Press, Madison, Wisconsin, s. 41–48.

Mladenic D., Grobelnik M., (1999), *Feature Selection for Unbalanced Class Distribution and Naive Bayes*, [w:] Proceedings of the 16th International Conference on Machine Learning (ICML), Bled, s. 258–267.

Mladenić D., Brank J., Grobelnik M., Milic-Frayling N., (2004), *Feature Selection Using Linear Classifier Weights: Interaction with Classification Models*, [w:] Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, s. 234–241, https://doi.org/10.1145/1008992.1009034

Mladenović M., Mitrović J., Krstev C., Vitas D., (2016), *Hybrid Sentiment Analysis Framework for a Morphologically Rich Language*, „Journal of Intelligent Information Systems", vol. 46, s. 599–620, https://doi.org/10.1007/s10844-015-0372-5

Movie Review Data, https://www.cs.cornell.edu/people/pabo/movie-review-data/ (dostęp: 15.10.2021).

Na J.-Ch., Khoo C., Wu P. H. J., (2005), *Use of Negation Phrases in Automatic Sentiment Classification of Product Reviews*, „Library Collections, Acquisitions & Technical Services", no. 29, s. 180–191, https://ccc.inaoep.mx/~villasen/bib/Use%20of%20negation%20phrases%20in%20automatic%20sentiment%20classification.pdf (dostęp: 11.02.2021).

Nasukawa T., Yi J., (2003), *Sentiment Analysis: Capturing Favorability Using Natural Language Processing*, [w:] Proceedings of the 2nd International Conference on Knowledge Capture, s. 70–77, https://doi.org/10.1145/945645.945658

Ng H. T., Goh W. B., Low K. L., (1997), *Feature Selection, Perceptron Learning, and a Usability Case Study for Text Categorization*, [w:] N. J. Belkin, A. D. Narasimhalu, P. Willett (eds), *Proceedings of the SIGIR-97, 20th ACM International Conference on Research and Development in Information Retrieval*, ACM Press, New York, vol. 31, s. 67–73, https://doi.org/10.1145/278459.258537

Nguyen T. H., Nghia N. H., Tuan D. L., Nguyen V. T., (2015), *A Hybrid Feature Selection Method for Vietnamese Text Classification*, [w:] Proceedings of the 7th International Conference on Knowledge and Systems Engineering (KSE), s. 91–96, https://doi.org/10.1109/KSE.2015.25

O'Keefe T., Koprinska I., (2009), *Feature Selection and Weighting Methods in Sentiment Analysis*, [w:] Proceedings of the 14th Australasian Document Computing Symposium, Sydney.

Ong B. Y., Goh S. W., Xu C., (2015), *Sparsity Adjusted Information Gain for Feature Selection in Sentiment Analysis*, [w:] Proceedings of the IEEE International Conference on Big Data, Santa Clara, USA, s. 2122–2128, https://doi.org/10.1109/BigData.2015.7363995

Ortega-Mendoza R. M., López-Monroy A., Franco-Arcega A., Montes-y-Gómez M., (2018), *Emphasizing Personal Information for Author Profiling: New Approaches for Term Selection and Weighting*, „Knowledge Based Systems", vol. 145, s. 169–181, https://doi.org/10.1016/J.KNOSYS.2018.01.014

Pakiet *e1071*, https://CRAN.R-project.org/package=e1071 (dostęp: 21.09.2021).

Pakiet *naivebayes*, https://CRAN.R-project.org/package=naivebayes (dostęp: 21.09.2021).

Pakiet *RWeka*, https://CRAN.R-project.org/package=RWeka (dostęp: 21.09.2021).

Pakiet *text2vec*, https://CRAN.R-project.org/package=text2vec (dostęp: 20.09.2021).

Pakiet *tm*, https://CRAN.R-project.org/package=tm (dostęp: 20.09.2021).

Paltoglou G., Thelwall M., (2010), *A Study of Information Retrieval Weighting Schemes for Sentiment Analysis*, [w:] Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL '10), Association for Computational Linguistics, Uppsala, s. 1386–1395, https://aclanthology.org/P10-1141.pdf (dostęp: 5.06.2021).

Pang B., Lee L., (2004), *A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts*, [w:] Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics (ACL '04), Association for Computational Linguistics, Stroudsburg, PA, s. 271–278, https://doi.org/10.3115/1218955.1218990

Pang B., Lee L., (2005), *Seeing Stars: Exploiting Class Relationships for Sentiment Categorization with Respect to Rating Scales*, [w:] Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics (ACL '05), Association for Computational Linguistics, Stroudsburg, PA, s. 115–124, https://doi.org/10.3115/1219840.1219855

Pang B., Lee L., Vaithyanathan S., (2002), *Thumbs up? Sentiment Classification Using Machine Learning Techniques*, [w:] Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002), Association for Computational Linguistics, Stroudsburg, PA, s. 79–86, https://www.cs.cornell.edu/home/llee/papers/sentiment.pdf (dostęp: 8.02.2021).

Patil L., Atique M., (2013), *A Novel Feature Selection Based on Information Gain Using WordNet*, [w:] Proceedings of the Science and Information Conference (SAI), London, s. 625–629.

Pintas J. T., Fernandes L. A. F., Garcia A. C. B., (2021), *Feature Selection Methods for Text Classification: A Systematic Literature Review*, „Artificial Intelligence Review", vol. 54, s. 6149–6200, https://doi.org/10.1007/s10462-021-09970-6

Rahate R. S., Emmanuel M., (2013), *Feature Selection for Sentiment Analysis by Using SVM*, „International Journal of Computer Applications", vol. 84, no. 5, s. 24–32, https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.402.3178&rep=rep1&type=pdf (dostęp: 4.02.2022).

Rastogi S., (2018), *Improving Classification Accuracy of Automated Text Classifiers*, [w:] Proceedings of the 7th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions), Noida, India, s. 1–7, https://doi.org/10.1109/ICRITO.2018.8748498

Ruiz M. E., Srinivasan P., (1999), *Hierarchical Neural Networks for Text Categorization*, [w:] Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, New York, s. 281–282, https://doi.org/10.1145/312624.312700

Saad S., Saberi B., (2017), *Sentiment Analysis or Opinion Mining*: *A Review*, „International Journal on Advanced Science, Engineering and Information Technology", vol. 7, no. 5, s. 1660–1666, https://doi.org/10.18517/ijaseit.7.5.2137

Saif H., He Y., Alani H., (2012), *Alleviating Data Sparsity for Twitter Sentiment Analysis*, [w:] *2nd Workshop on Making Sense of Microposts (#MSM2012): Big things come in small packages at the 21st International Conference on the World Wide Web (WWW'12)*, Lyon, CEUR Workshop Proceedings (CEUR-WS.org), s. 2–9, https://www2012.universite-lyon.fr/proceedings/nocompanion/MSM2012_paper_01.pdf (dostęp: 11.02.2022).

Salton G., Wong A., and Yang C. S., (1975), *A Vector Space Model for Automatic Indexing*, „Communication of the ACM", vol. 18, issue 11, s. 613–620, https://doi.org/10.1145/361219.361220

Sathic Ali P. U., Venkateswaran C. J., (2014), *A Dempster-Shafer Model for Feature Selection in Text Categorization*, „Research Journal of Applied Sciences, Engineering and Technology", vol. 7, no. 5, s. 981–985, http://dx.doi.org/10.19026/rjaset.7.347

Sebastiani F., (2002), *Machine Learning in Automated Text Categorization*, ACM Computing Surveys, vol. 34, issue 1, s. 1–47, https://doi.org/10.1145/505282.505283

Shan L.-L., Liu B.-Q., Sun C.-J., (2011), *Comparison and Improvement of Feature Selection Method for Text Categorization*, „Journal of Harbin Institute of Technology", vol. 43, no. 1, s. 319–324.

Shang W., Huang H., Zhu H., Lin Y., Qu Y., Wang Z., (2007), *A Novel Feature Selection Algorithm for Text Categorization*, „Expert Systems with Applications", vol. 33, issue 1, s. 1–5, https://doi.org/10.1016/j.eswa.2006.04.001

Shen K., Chen X., Ma J., Ke L., Lu Y., Zhang K., (2013), *A Blended Feature Selection Method in Text Classification*, [w:] Proceedings of the International Conference on Cyberspace Technology (CCT 2013), Beijing, China, s. 573–576, https://doi.org/10.1049/cp.2013.2077

Simeon M., Hilderman R., (2008), *Categorical Proportional Difference: A Feature Selection Method for Text Categorization*, [w:] J. F. Roddick, J. Li, P. Christen, P. Kennedy (eds), *The Seventh Australasian Data Mining Conference (AusDM 2008), Glenelg, South Australia. Conferences in Research and Practice in Information Technology (CRPIT)*, vol. 87, s. 201–208, https://dl.acm.org/doi/pdf/10.5555/2449288.2449320 (dostęp: 4.02.2022).

Stoplista, https://pl.wikipedia.org/wiki/Wikipedia:Stopwords (dostęp: 20.09.2021).

Subrahmanian V. S., Reforgiato D., (2008), *AVA: Adjective-Verb-Adverb Combinations for Sentiment Analysis*, „IEEE Intelligent Systems", vol. 23, no. 4, s. 43–50, https://doi.org/10.1109/MIS.2008.57

Sun J., Zhang X., Liao D., Chang V., (2017), *Efficient Method for Feature Selection in Text Classification*, [w:] Proceedings of International Conference on Engineering and Technology (ICET), Antalya, Turkey, s. 1–6, https://doi.org/10.1109/ICEngTechnol.2017.8308201

Tan S., Zhang J., (2008), *An Empirical Study of Sentiment Analysis for Chinese Documents*, „Expert Systems with Application", vol. 34, issue 4, s. 2622–2629, https://doi.org/10.1016/j.eswa.2007.05.028

Wang H., Bell D., (2004), *Extended k-Nearest Neighbours Based on Evidence Theory*, „The Computer Journal", no. 47, issue 6, s. 662–672, https://doi.org/10.1093/comjnl/47.6.662

Wang W., Kang Y., Wu X., (2008), *Study on Feature Selection in Text Categorization*, „Information Technology", no. 12, s. 29–31.

Wu G., Wang L., Zhao N., Lin H., (2015), *Improved Expected Cross Entropy Method for Text Feature Selection*, [w:] Proceedings of the International Conference on Computer Science and Mechanical Automation (CSMA), Massachusetts, USA, s. 49–54, https://doi.org/10.1109/CSMA.2015.17

Wu G., Xu J., (2015), *Optimized Approach of Feature Selection Based on Information Gain*, [w:] Proceedings of the International Conference on Computer Science and Mechanical Automation (CSMA), Massachusetts, USA, s. 157–161, https://doi.org/10.1109/CSMA.2015.38

Wu L., Wang Y., Zhang S., Zhang Y., (2017), *Fusing Gini Index and Term Frequency for Text Feature Selection*, [w:] Proceedings of the IEEE 3rd International Conference on Multimedia Big Data (BigMM), Laguna Hills, USA, s. 280–283, https://doi.org/10.1109/BigMM.2017.65

Xu H., Yu S., Chen J., Zuo X., (2018), *An Improved Firefly Algorithm for Feature Selection in Classification*, „Wireless Personal Communications: An International Journal", vol. 102, issue 4, s. 2823–2834, https://doi.org/10.1007/s11277-018-5309-1

Yang X.-S., (2009), *Firefly Algorithms for Multimodal Optimization*, [w:] Proceedings of the 5th International Symposium on Stochastic Algorithms: Foundations and Applications, Springer, Berlin–Heidelberg, s. 169–178.

Yang Y., Liu X., (1999), *A Re-Examination of Text Categorization Methods*, [w:] Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'99), Berkeley, USA, s. 42–49, https://doi.org/10.1145/312624.312647

Yazdani S. F., Murad M. A. A., Sharef N. M., Singh Y. P., Latiff A. R. A., (2017), *Sentiment Classification of Financial News Using Statistical Features*, „International Journal of Pattern Recognition and Artificial Intelligence", vol. 31, no. 3, s. 1–34, https://doi.org/10.1142/S0218001417500069

Yu L., Liu H., (2004), *Efficient Feature Selection Via Analysis of Relevance and Redundancy*, „Journal of Machine Learning Research", vol. 5, s. 1205–1224.

Zhang H., Ren Y., Yang X., (2013), *Research on Text Feature Selection Algorithm Based on Information Gain and Feature Relation Tree*, [w:] Proceedings of the 10th Web Information System and Application Conference, Yangzhou, China, s. 446–449, https://doi.org/10.1109/WISA.2013.90

Zhang T., Oles F. J., (2001), *Text Categorization Based on Regularized Linear Classification Methods*, „Information Retrieval", vol. 4, s. 5–31.

Zhen Z., Zeng X., Wang H., Han L., (2011), *A Global Evaluation Criterion for Feature Selection in Text Categorization Using Kullback-Leibler Divergence*, [w:] *International Conference of Soft Computing and Pattern Recognition (SoCPaR)*, Dalian, China, s. 440–445, https://doi.org/10.1109/SoCPaR.2011.6089284

Zhu L., Wang G., Zou X., (2017), *Improved Information Gain Feature Selection Method for Chinese Text Classification Based on Word Embedding*, [w:] Proceedings of the 6th International Conference of Software and Computer Applications, Bangkok, Thailand, s. 72–76, https://doi.org/10.1145/3056662.3056671