ZACHODNIOPOMORSKI UNIWERSYTET TECHNOLOGICZNY W SZCZECINIE

WYDZIAŁ INFORMATYKI

mgr inż. Marcin Gibert

Integracja metod eksploracji danych tekstowych i numerycznych w procesie podejmowania decyzji

Rozprawa doktorska

Promotor: dr hab. Bożena Śmiałkowska

Promotor pomocniczy: dr inż. Jarosław Jankowski

Szczecin 2016

Spis treści

1. Wprowadzenie

1.1. Eksploracja danych w procesie podejmowania decyzji

Proces podejmowania decyzji *PD* (inaczej proces decyzyjny) obejmuje różne czynności np. według literatury [94, s. 17] w tym procesie wyróżnia się następujące etapy:

1. Wyznaczenie przywództwa w procesie *PD* (wyznaczenie decydenta i jego roli w procesie *PD*),
2. Zdefiniowanie problemu decyzyjnego,
3. Opracowanie modelu (sposobu) oceny oraz sformułowanie wariantów decyzyjnych,
4. Zebranie znaczących i wiarygodnych danych,
5. Ocena wariantów decyzyjnych i podjęcie decyzji,
6. Opracowanie planu wdrożenia decyzji.

Proces *PD* wspierany jest różnymi metodami i technikami często bazującymi na komputerowym wspomaganiu. Szerokie zastosowanie mają tu komputerowe systemy wspomagania decyzji DSS (ang. Decision Suport Systems) [34, s. 24], często uzupełniane przez systemy odkrywania wiedzy z danych KDD (ang. Knowledge Discovery in Databases) [1, s. 1], w których wykorzystywane są metody eksploracji danych (ang. Data mining) [72, ss. 153–154]. Dotyczy to zwłaszcza tych systemów KDD, które oparte są na kolekcjach danych zgromadzonych w bazach i hurtowniach danych.

Rolą eksploracji danych w procesie *PD* jest wydobywanie użytecznych informacji zwłaszcza z dużych zbiorów danych [27, s. 2] poprzez automatyczne lub półautomatyczne przeszukiwanie i analizowanie danych w celu odkrywania znaczących wzorców i reguł [8, s. 7]. Wówczas za pomocą eksploracji danych można poszerzać wiedzę bazującą na wydobytych z danych informacjach - wiedzę stosowaną do realizacji zadań, czynności i rozwiązywania problemów w procesach *PD*.

Do najpopularniejszych i najczęściej wykorzystywanych metod eksploracji danych w procesach podejmowania decyzji zalicza się w literaturze przedmiotu [15, s. 21][61, s. 4] [55, s. 398]:

- klasyfikację,
- regresję,
- grupowanie,
- odkrywanie sekwencji,
- odkrywanie charakterystyk,

- analizę przebiegów czasowych,
- odkrywanie asocjacji,
- wykrywanie zmian i odchyleń.

Szczególnie przydatną i popularna, z praktycznego punktu widzenia procesów *PD* jest klasyfikacja [33, ss. 119–121] [102, s. 159]. Dlatego w pracy skoncentrowano się na tej właśnie metodzie.

Eksploracja danych, a zwłaszcza klasyfikacja, wspomagają ocenę oraz wybór wariantu decyzyjnego procesy *PD* (etap 5 zgodnie z [94,s.17]). Istotną częścią procesu *PD* na etapie definiowania problemu decyzyjnego (etap 2 procesu *PD*) po wyznaczeniu decydenta (etap 1 procesu *PD*) jest rozpoznanie kontekstu decyzyjnego [79, s. 24] [12, s. 35], który zgodnie z literaturą [56, s. 251] jest określony jako zespół współistniejących czynników wpływających na właściwe zrozumienie postawionego problemu. Kontekst ten bezpośrednio wpływa zarówno na określenie wariantów decyzyjnych (etap 3 procesu *PD*), jak też na możliwą procedurę postępowania w pozostałych etapach procesu *PD*. W zależności od kontekstu dobierane są również różne dane - istotne cechy opisujące analizowane obiekty (etap 4 procesu *PD*), na podstawie których dokonywana jest ocena i wybór wariantu decyzyjnego (etap 5 procesu *PD*) [79, s. 29]. Po wyznaczeniu decydenta (etap 1 procesu *PD*), etapy 2, 3, 4 oraz etap 6 procesu *PD* realizowane są przez decydenta, a nie przez metody eksploracji danych, które wspomagają decydenta. Dlatego z punktu widzenia możliwości wspomagania procesów decyzyjnych metodami eksploracyjnymi, zwłaszcza klasyfikacją, można zawęzić dalsze rozważania nad rolą tych metod w procesach *PD* do etapu 5.

Problem decyzyjny zdefiniowany w procesie *PD* (etap 2 procesu *PD*) rozważany jest w oparciu o dane (cechy i charakterystyki) opisujące obiekty analizowane w procesie decyzyjnym. Mogą one być wyrażone za pomocą danych numerycznych (za pomocą liczb), mogą wynikać z opisu sformułowanego za pomocą języka naturalnego, a także mogą być wyrażone innymi typami danych np. multimedialnymi, które w oryginalnej formie nie są ani danymi numerycznymi ani tekstowymi. Metody eksploracji danych wspomagające procesy *PD* w etapie 5 procesu *PD* operują na zbiorze wszystkich danych zgromadzonych w tym procesie. Zatem zbiór *Z*, będący przedmiotem eksploracji jest sumą mnogościową wszystkich opisanych typów danych zgodnie ze wzorem *(1)*.

$$Z = Z_N \cup Z_T \cup Z_P \qquad (1)$$

gdzie:

Z – zbiór dostępnych danych (zbiór danych w procesie *PD* poddanych eksploracji i uzyskanych w etapie 4 tego procesu),

Z_N – zbiór danych numerycznych,

Z_T – zbiór danych tekstowych,

Z_P – zbiór danych innych niż dane numeryczne i tekstowe, np. zbiór danych wizualnych (multimedialnych).

Zgodnie z literaturą [35] [64] [49] metody eksploracji danych koncentrują się przede wszystkim na danych Z_N oraz danych Z_T. Zaś eksploracja danych ze zbioru Z_P może być realizowana przez transformację tych danych do reprezentacji wyrażonych przez dane ustrukturyzowane ze zbioru Z_N [103, s. 20] lub dane opisane za pomocą danych ze zbioru Z_T [90, s. 108]. Z tego względu w eksploracji danych realizowanej w etapie 5 procesu *PD* skoncentrowano się na zbiorze danych:

$$Z_E = Z_N \cup Z_T \tag{2}$$

gdzie:

Z_E – zbiór dostępnych danych (zbiór danych w procesie *PD* poddanych eksploracji i uzyskanych w etapie 4 tego procesu, wyrażonych za pomocą danych tekstowych lub numerycznych),

Z_N – zbiór danych numerycznych,

Z_T – zbiór danych tekstowych.

Zarówno dane tekstowe jak i numeryczne mogą być wyrażone przez różne reprezentacje. Dobór odpowiedniej reprezentacji danych ma duży wpływ na wynik eksploracji [82, s. 8]. Jednak jest to szczególnie utrudnione zadanie w przypadku danych tekstowych, gdzie ze względu na charakter dobieranych cech (elementów reprezentacji) można wyróżnić dwa odmienne podejścia do klasyfikacji. Jedno z nich bazuje na cechach związanych z analizą struktury tekstu, natomiast drugie na cechach wynikających z precyzyjnego zrozumienia znaczenia tekstu [83, ss. 15–16][84, s. 22].

W zależności od rozpatrywanego problemu decyzyjnego w procesie *PD*, teksty mogą być klasyfikowane na podstawie porównania ich cech ilościowych np. liczby określonych wyrazów lub na podstawie precyzyjnie wyekstrahowanej z nich informacji znaczeniowej np. symptom choroby u pacjenta: *szybkie bicie serca* [92, s. 119]. Dane tekstowe

są zazwyczaj nieustrukturyzowane, czyli nie posiadają żadnej wewnętrznej struktury lub struktura ta jest określona częściowo (dane tzw. semistrukturalne) [36, s. 279]. Dlatego w eksploracji danych stosowane są zamienne reprezentacje dokumentów tekstowych, które umożliwiają stosowanie różnych technik eksploracji. Najczęściej do eksploracji danych tekstowych wykorzystywany jest model przestrzeni wektorowej (ang. Vector Space Model - VSM), w którym każdy dokument tekstowy jest reprezentowany za pomocą tzw. wektora cech, czyli zbioru elementów opisujących dokument tekstowy np. występujących w nim pojedynczych wyrazów, zwanych termami [14, s. 102]. Ogólnie reprezentację *R* dokumentu tekstowego w modelu przestrzeni wektorowej VSM można zdefiniować za pomocą wzoru *(3)* [25, s. 38]:

$$R = \{ r_1 \dots \ r_{mr} \} \tag{3}$$

gdzie:

R – reprezentacja dokumentu tekstowego,

mr – maksymalna liczba elementów reprezentacji dokumentu tekstowego,

r_j – element (cecha) reprezentacji dokumentu tekstowego, gdzie $j \in \{1,2, \dots mr\}$.

W literaturze wymienia się kilka propozycji reprezentacji danych tekstowych opartych o odpowiednie bazy przestrzeni cech, między innymi są to [24, s. 8]:

1. reprezentacja unigramowa – uwzględnia pojedyncze wyrazy,
2. reprezentacja n-gramowa – uwzględnia sekwencje o stałej n-wyrazowej długości,
3. reprezentacja γ-gramowa - uwzględnia sekwencje o zmiennej liczbie wyrazów.

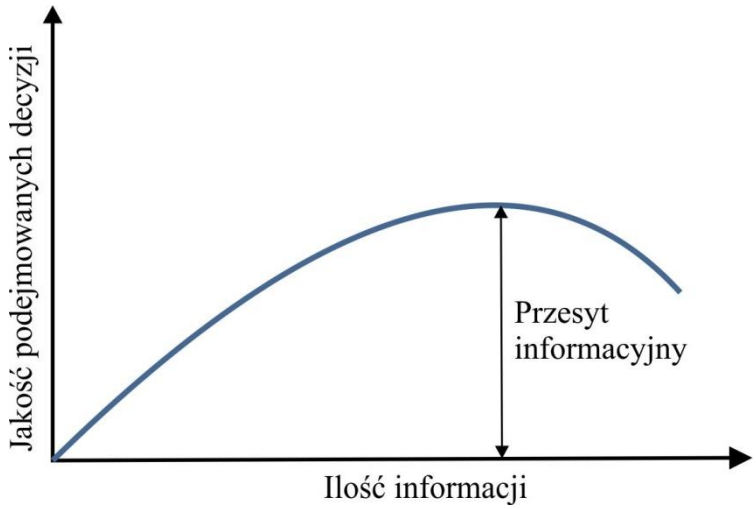
Rozwinięciem klasycznej eksploracji danych tekstowych w modelu przestrzeni wektorowej VSM jest niejawna analiza semantyczna (ang. Latent Semantic Analysis - LSA), której działanie opiera się na wykrywaniu niejawnych struktur semantycznych pomiędzy elementami reprezentacji tekstu [42, s. 10]. Wykryte struktury semantyczne są nowymi termami w zredukowanej przestrzeni, które lepiej odwzorowują niejawne zależności występujące pomiędzy pierwotnymi cechami (pojedynczymi wyrazami).

Na jakość eksploracji danych, a w konsekwencji jakość wyniku procesu *PD* (jakość decyzji) niewątpliwy wpływ ma reprezentacja wykorzystanych danych, zarówno danych tekstowych jak i numerycznych. Do wad reprezentacji danych zalicza się:

- nadmiarowość reprezentacji i wynikający z niej szum informacyjny, który komplikuje identyfikację informacji istotnych,
- utrudnione wydobycie informacji znaczeniowej tekstu,

- brak dostosowania reprezentacji danych (elementów reprezentacji tekstu) do rozważanego problemu decyzyjnego, co ogranicza możliwości interpretacyjne danych w procesie podejmowania decyzji. :
- źle przeprowadzoną dyskretyzację danych numerycznych o charakterze ciągłym,
- błędnie dobrane wartości nominalne.

W zależności od przyjętej strategii reprezentacji danych zmianie ulega ilość informacji, które przez te dane są przenoszone. Należy zauważyć, że zarówno niedobór istotnych danych w procesie *PD* jak również ich nadmiarowa ilość może powodować obniżenie jakości procesu podejmowania decyzji *PD* [26, s. 634]. Zjawisko to obrazuje tzw. krzywa przesytu informacyjnego o charakterze zgodnym z rysunkiem 1.



Rysunek 1. Krzywa przesytu informacyjnego.
Źródło: opracowanie własne na podstawie [26, s. 634]

Z przebiegu tej krzywej wynika, że wraz ze wzrostem ilości informacji, po osiągnięciu pewnego poziomu ilości informacji, pogorszeniu ulega jakość podejmowanych decyzji w procesie *PD*. Oznacza to, że różna jest nośność informacyjna danych będących podstawą eksploracji w procesie *PD*. Jakość decyzji uzależniona jest od redundancji reprezentacji danych i związanego z nią szumu informacyjnego, czyli nadmiaru danych, który może utrudnić wyodrębnienie informacji istotnych dla rozpatrywanego problemu decyzyjnego [91, s. 427]. W konsekwencji wpływa to na wynik eksploracji oraz na ostateczny wynik całego procesu *PD*. Dlatego bardzo ważne jest opracowanie odpowiedniej reprezentacji danych tekstowych jak i numerycznych wraz z całą procedurą eksploracji biorąc pod uwagę rozważany proces *PD*.

W związku z możliwością użycia podczas eksploracji w procesie *PD* różnych reprezentacji danych, dużego znaczenia nabiera określenie nośności informacyjnej danych [93, s. 166] [31].

Zakładając, że dostępne w procesie *PD* dane ze zbioru Z_E o reprezentacji *R* przenoszą istotne dla rozpatrywanego problemu decyzyjnego informacje ζ, to nośność informacyjna danych Z_E ze wzoru *(2)* jest zgodna ze wzorem *(4)*.

$$Ninf(Z_E, M) = g(W_M, R, \zeta) \qquad (4)$$

gdzie:

Ninf – nośność informacyjna danych Z_E,

Z_E – zbiór dostępnych danych (zbiór danych w procesie *PD* uzyskanych w etapie 4 tego procesu, wyrażonych za pomocą danych tekstowych lub numerycznych),

R – reprezentacja danych Z_E,

W_M – wynik procesu *PD* określony za pomocą miar jakości decyzji w tym procesie,

M – zbiór wybranych miar jakości decyzji w procesie *PD*,

ζ – niezidentyfikowany zbiór informacji mających wpływ na nośność informacyjną danych *Ninf*,

g – funkcja, której wartość określa nośność informacyjną *Ninf*.

W sytuacji, w której proces podejmowania decyzji *PD* w sensie proceduralnym pozostaje niezmienny, a zmianie ulega jedynie reprezentacja danych wejściowych, wówczas do oceny rangi nośności informacyjnej danych można wykorzystać wynik W_M procesu *PD* będący miarą jakości decyzji [74, s. 206]. Jest to analogia do metody testowania traktującego proces *PD* jako tzw. czarną skrzynkę (ang. black box), która w nieznany sposób realizuje wykonywane funkcje. W metodzie tej do procesu *PD* wprowadzane są dane wejściowe o określonej reprezentacji, a następnie analizie poddawane są wyniki otrzymywane na wyjściu. Dzięki takiemu podejściu, przy założeniu niezmienności procesu *PD*, można określić, która z różnych reprezentacji *R* danych Z_E umożliwia w procesie podejmowania decyzji *PD* uzyskanie wyższego wyniku W_M wyrażonego za pomocą miar jakości decyzji ze zbioru *M*. Wyższe wartości miar jakości decyzji oznaczają wykorzystanie większej ilości istotnych informacji przenoszonych przez dane Z_E reprezentowane przez *R* w procesie *PD*, które wpływają na poprawę wyniku W_M. Innymi słowy, przy powyższym założeniu, wyższy i wyrażony za pomocą miar jakości decyzji wynik W_M uzyskany dla określonej reprezentacji danych *R* w stosunku do pozostałych reprezentacji danych oznacza większą nośność informacyjną danych dla tej reprezentacji. Przyjęta miara nośności informacyjnej danych może być podstawową miarę procesu podejmowania decyzji *PD*.

Wyznaczenie nośności informacyjnej danych w procesie decyzyjnym *PD* ze wzoru *(4)* jest trudne ze względu na brak znajomości funkcji g oraz zbioru ζ. Dlatego do szacowania nośności informacyjnej danych dla różnych reprezentacji tych danych stosuje się metody porównawcze. Dla przykładu nośność informacyjna danych Z_E o reprezentacji R_1 jest wyższa od nośności informacyjnej danych Z_E o reprezentacji R_2 w przypadku osiągnięcia w procesie podejmowania decyzji *PD* korzystniejszego wyniku W_M dla danych reprezentowanych przez R_1 w stosunku do R_2. Taką zasadę pomiaru nośności informacyjnej przyjęto w niniejszej pracy.

Wynik W_M ze wzoru *(4)* w eksploracji danych jest określany przez zbiór wybranych miar jakości decyzji *M* [74, s. 206]. Przyjęte miary jakości decyzji uzależnione są od rozpatrywanego problemu decyzyjnego i przyjętej metody oceny jakości wyniku procesu *PD* (oceny jakości decyzji) [96, s. 88][45]. W literaturze wyróżnia się między innymi metody oceny jakości decyzji bazujące na wynikach klasyfikacji takie jak: metoda bazująca na miarach jakości klasyfikacji wynikających z macierzy pomyłek (ang. confusion matrix), krzywa ROC czy walidacja krzyżowa [105][45]. Jako podstawową metodę oceny jakości klasyfikacji w literaturze wskazuje się macierz pomyłek (ang. confusion matrix), która określa tendencje w klasyfikacji testowej w stosunku do rzeczywistych wyników [105]. Najczęściej wykorzystywanymi miarami jakości klasyfikacji wynikającymi z macierzy pomyłek są: współczynnik całkowitej dokładności ACC (ang. accuracy) oraz współczynnik całkowitego poziomu błędu – *ERR* (ang. error rate level) [60, s. 157][74, s. 144]. Miary jakości decyzji (elementy ze zbioru M) ze wzoru *(4)* takie jak całkowita dokładność – *ACC* oraz całkowity poziom błędu – ERR [63, ss. 181–182] określają jakość procesu *PD* opartego na eksploracji danych bazując na współczynnikach charakteryzujących klasyfikację, które zawierają się w macierzy pomyłek:

- współczynnik prawdziwy pozytywny (ang. True Positive - *TP*),
- współczynnik prawdziwy negatywny (ang. True Negative- *TN*),
- współczynnik fałszywy pozytywny (ang. False Positive - *FP*),
- współczynnik fałszywy negatywny (ang. True Positive - *FN*).

Za pomocą współczynników *ACC* oraz *ERR* można scharakteryzować jakość decyzji w procesie *PD* z uwzględnieniem wszystkich wskazań systemu klasyfikacji jednocześnie tj. z uwzględnieniem *TP, FP, TN* oraz *FN*. Dlatego zarówno miara *ACC* jak i *ERR* sprowadzają charakterystykę decyzji do jednej wartości liczbowej.

Pozostałe opisywane w literaturze miary jakości decyzji w procesie *PD* wynikające z macierzy pomyłek np. pozytywny współczynnik predykcji (ang. Positive

Predictive Value - *PPV*) czy współczynnik czułości (ang. Sensitivity - *SE*) definiowane są jedynie przez wybrane wskazania (współczynniki) systemu klasyfikacji np. jedynie współczynniki *TP* i *FP* w przypadku miary jakości *PPV* oraz współczynniki *TP* i *FN* w przypadku miary jakości *SE*. W związku z tym przy całościowej ocenie decyzji miary takie należy rozpatrywać zbiorczo np. jednocześnie miary *PPV* i *NPV*. Z tego powodu wybrane miary jakości decyzji (*ACC*, *ERR*), które wykorzystują wszystkie wskazania jakości klasyfikacji jednocześnie są najbardziej kompletnym wskaźnikiem charakteryzującym jakość klasyfikacji, na podstawie którego można wyznaczyć nośność informacyjną danych przy zadanej reprezentacji tych danych.

1.2. Problem badawczy, cel i hipoteza pracy

W procesie decyzyjnym *PD*, w zależności od typu analizowanych danych, wykorzystywane są różne dedykowane im metody i techniki eksploracji. Wśród opisanych w literaturze metod eksploracji danych znane są metody, które koncentrują się wyłącznie na zbiorze Z_N bądź wyłącznie na zbiorze Z_T. Jeśli proces *PD* oparty jest wyłącznie na jednym z tych zbiorów to w literaturze przedmiotu [102] [2, ss. 163–213] dostępnych jest wiele metod eksploracji danych do tych przypadków, przy czym w przypadku metod eksploracji danych tekstowych brak jest szerszych badań nad uwzględnieniem w ustrukturyzowanej reprezentacji danych tekstowych specyfiki języka polskiego (języka fleksyjnego), w którym, dzięki końcówkom fleksyjnym nadającym wyrazom właściwe znaczenie gramatyczne, istnieje możliwość zachowania znaczenia tekstu przy dowolnym, przestawnym szyku wyrazów w zadaniu [83, s. 8]. W badaniach nad eksploracją tekstów opisywanych w literaturze przyjmuje się metodę polegającą na sprowadzeniu różnych form fleksyjnych wyrazów do ich form podstawowych tzw. lematów lub ogranicza się długość wyrazów do części wspólnej we wszystkich formach fleksyjnych tzw. steamów. Jednak rozwiązanie takie jest pewnym uproszczeniem i powoduje znaczną utratę istotnych w procesie decyzyjnym informacji, z tego choćby powodu, że wyraz odmieniony niesie za sobą na ogół inną informację niż jego forma podstawowa [41, s. 46].

Głównym jednak problemem jest taka sytuacja, w której eksploracja realizowana jest jednocześnie w oparciu o dane ze zbiorów Z_N oraz Z_T. Wagę tego problemu podkreśla się również w literaturze [66, s. 974][5, s. 168]. P.Gawrysiak pisze, że przyszłe badania nad systemami eksploracji i kategoryzacji powinny skoncentrować się na hybrydowych rozwiązaniach uwzględniających zarówno zawartość tekstową dokumentów jak i dodatkowe atrybuty numeryczne [22, s. 100]. Również autorzy artykułu „A Roadmap for Web Mining:

From Web to Semantic Web" wskazują na główny problem do rozwiązania w przyszłych badaniach podkreślając, że ze względu na coraz częstsze występowanie informacji w sieci nie tylko w formie tekstowej, do klasyfikacji, grupowania, uczenia regułowego i sekwencyjnego - ogólnie ekstrakcji danych niezbędne jest połączenie metod dedykowanych tekstowi oraz danym numerycznym (liczbowym) [7, s. 19]. W literaturze [22, ss.99-100] zauważa się również, że integracja metod eksploracji danych może wpłynąć na osiąganie korzystniejszego wyniku w sensie kryterium nośności informacyjnej danych, wskaźników jakości eksploracji, a co za tym idzie jakości decyzji w procesie *PD* w stosunku do wyników osiąganych przez metody indywidualnie, dedykowane odrębnie danym numerycznym lub tekstowym [43, s. 2]. Dodatkowo w literaturze podkreślono istotne znaczenie problemu integracji metod eksploracji danych tekstowych i numerycznych w wielu dziedzinach, takich jak finanse, medycyna czy web mining [21, s. 310], [97, s. 151], [80, s. 18] oraz [59]. Ponadto w opracowaniach badawczych wskazano na możliwość pozyskania bardziej wartościowej wiedzy, gdy uwzględnia się jednocześnie w procesie *PD* eksplorację danych numerycznych i tekstowych [21, s. 314] [100, s. 368] [89, s. 4] oraz, że brakuje tu metody, która jednocześnie w sposób wieloaspektowy i systemowy umożliwiałaby eksplorację obu wyróżnionych typów danych w sposób adekwatny do tych typów.

W związku z powyższym sformułowano następujące pytania związane ze zidentyfikowanymi brakami metod eksploracji danych w procesie podejmowania decyzji *PD*, na które należy znaleźć odpowiedzi:

1. Czy możliwe jest opracowanie takiej metody eksploracji danych, która jednocześnie łącznie uwzględnia dane numeryczne i tekstowe?
2. Czy i w jaki sposób można zintegrować znane metody eksploracji danych tekstowych ze znanymi metodami eksploracji danych numerycznych, aby uzyskać lepszą nośność informacyjną eksplorowanych danych?
3. Jaki wpływ na wynik procesu *PD* ma integracja (łącznie) metod eksploracji danych tekstowych i numerycznych?
4. Czy i w jaki sposób można zwiększyć nośność informacyjną danych w eksploracji danych wspomagających proces *PD*?
5. Jaki wpływ ma wybór reprezentacji danych tekstowych i numerycznych na nośność informacyjna danych w rozpatrywanym problemie decyzyjnym?
6. W jaki sposób przy opracowywaniu reprezentacji danych tekstowych w oparciu o informacje znaczeniowe (informacje mające bezpośredni wpływ

na podejmowaną decyzję) można uwzględnić specyfikę języka naturalnego np. polskiego języka fleksyjnego?

Dlatego w kontekście odpowiedzi na pytania 1, 2, 3 i 6 za główny cel pracy przyjęto *opracowanie procedury integracji metod analizy fleksyjnej tekstu oraz metod eksploracji danych numerycznych*.

W odpowiedzi na postawione pytania 4 i 5 w pracy sformułowano następującą hipotezę: *integracja metod analizy fleksyjnej tekstu oraz eksploracji danych numerycznych zwiększy nośność informacyjną danych w wielokryterialnym procesie wspomagania decyzji.*

Aby osiągnąć cel pracy przyjęto następujące założenia:

- Proces decyzyjny *PD* jest oparty na eksploracji danych tekstowych i numerycznych,
- Dane poddawane eksploracji stanowią zbiór (zgodny ze wzorem *(2)*) wszystkich dostępnych danych w procesie *PD*,
- Podejmowana decyzja w procesie *PD* jest decyzją wielokryterialną (w szczególności jednokryterialną), a kryteria jej wyboru wynikają z dostępnych dla procesu eksploracji danych,
- Dane tekstowe w procesie *PD* są danymi opisanymi w języku fleksyjnym polskim, w którym ze względu na jego specyfikę możliwe jest występowania przestawnego szyku wyrazów w zdaniu zawartym w danych (dokumentach) tekstowych,
- Reprezentacja danych numerycznych wykorzystywana w eksploracji (wartości dyskretne oraz nominalne atrybutów) jest definiowana z uwzględnieniem struktury dziedziny wartości atrybutów, która odpowiada specyfice rozważanego problemu decyzyjnego,
- Elementami reprezentacji danych (dokumentów) tekstowych, która zgodnie ze wzorem *(3)* charakteryzuje dokument tekstowy, są rzeczowe informacje (ang. factual information), które mają bezpośredni wpływ na podejmowaną w procesie *PD* decyzję [17], przy czym rzeczowe informacje są tu rozumiane jako sekwencje wyrazów o zmiennej długości, które są ekstrahowane z dokumentów tekstowych na podstawie zdefiniowanych przez eksperta dziedzinowego wzorców informacyjnych,
- W pracy skoncentrowano się na metodzie eksploracji danych zwanej klasyfikacją, a to ze względu na mnogość oraz szeroki wachlarz zastosowań tych metod w procesach decyzyjnych *PD* [19, s. 55] [33, ss. 119–121] [102, s. 159],
- Ze względu na specyfikę wyniku klasyfikacji tj. występowanie zmiennych posiadających wyłącznie dwie kategorie (zmienne dychotomiczne), badanie prób zależnych danych (wyników klasyfikacji dla odpowiadających sobie przypadków

z różnych wariantach eksploracji z rysunku 2) oraz występowanie skali nominalnej zmiennych, do badań istotności i zgodności wyników klasyfikacji (weryfikacja statystyczna) zastosowano test McNemara [78, s. 197].

- Z powodu istnienia w procesach decyzyjnych *PD* luki informacyjnej wykazanej za pomocą wzoru *(4)*, brak jest możliwości dokładnego oszacowania nośności informacyjnej danych. Ponieważ nośność informacyjna nie tylko zależy od samych danych, ale również od wiedzy decydenta (możliwości interpretacyjnych tych danych) to wydaje się iż istnieje ścisła zależność iż im lepsza jest jakość eksploracji danych (np. klasyfikacji) tym nośność informacyjna danych na bazie których przeprowadzono tę eksplorację (klasyfikację) będzie wyższa. Dlatego w do weryfikacji hipotezy przyjęto założenie, że nośność informacyjna danych w procesie *PD* może być szacowana za pomocą wybranych miar jakości klasyfikacji takich jak: współczynnik ACC oraz współczynnik ERR [96, s. 7] [74, s. 206]. Miary jakości klasyfikacji (*ACC, ERR*) dla przykładowych procesów podejmowania decyzji *PD* w wariancie eksploracji A z rysunku 2 (integracja metod eksploracji danych tekstowych i numerycznych) będą wyższe w przypadku miary *ACC* oraz niższe w przypadku miary *ERR* od takich miar dla pozostałych wariantów B (eksploracja wyłącznie danych numerycznych), C (eksploracja wyłącznie danych tekstowych) oraz D (zintegrowanego wynik eksploracji z wariantów B i C).

W procedurze integracji opracowanej w ramach niniejszej pracy wykorzystano eksplorację danych tekstowych, która bazuje na modelu przestrzeni wektorowej VSM [58, ss. 45–54] oraz eksploracji danych numerycznych z wykorzystaniem metody Teorii Zbiorów Przybliżonych [70]. W eksploracji danych tekstowych wykorzystano również analizę fleksyjną języka polskiego po to by przy opracowaniu elementów reprezentacji tych danych zwiększyć ich możliwości interpretacyjne w kontekście postawionego problemu decyzyjnego oraz zweryfikować ich poprawność, co przekłada się na ostateczną jakość podejmowanych decyzji w procesie *PD*. Zaproponowano również metodę opracowywania γ-gramowej reprezentacji tekstu, której tzw. rzeczowe informacjach są ekstrahowane na podstawie wzorców informacyjnych. W pracy wykorzystano również takie metody jak:

- Analizę systemową,
- Analizę SWOT (mocne strony-szanse i zagrożenia),
- Analizę danych źródłowych ukierunkowaną na badanie istotności i wiarygodności tych danych (metoda Teoria Zbiorów Przybliżonych),
- Analizę przypadków użycia,

- Metodę statystyczną (test zgodności pomiędzy wynikami pomiarów McNemara).

Weryfikację hipotezy oparto na porównaniu nośności informacyjnej danych ze zbioru Z_E (wzór *(4)*), szacowaną współczynnikami *ACC* i *ERR*, w trzech przykładowych procesach podejmowania decyzji *PD* (dowód indukcyjny – studium przypadków). Rozważanymi trzema przypadkami weryfikującymi hipotezę badawczą były następujące przypadki użycia:

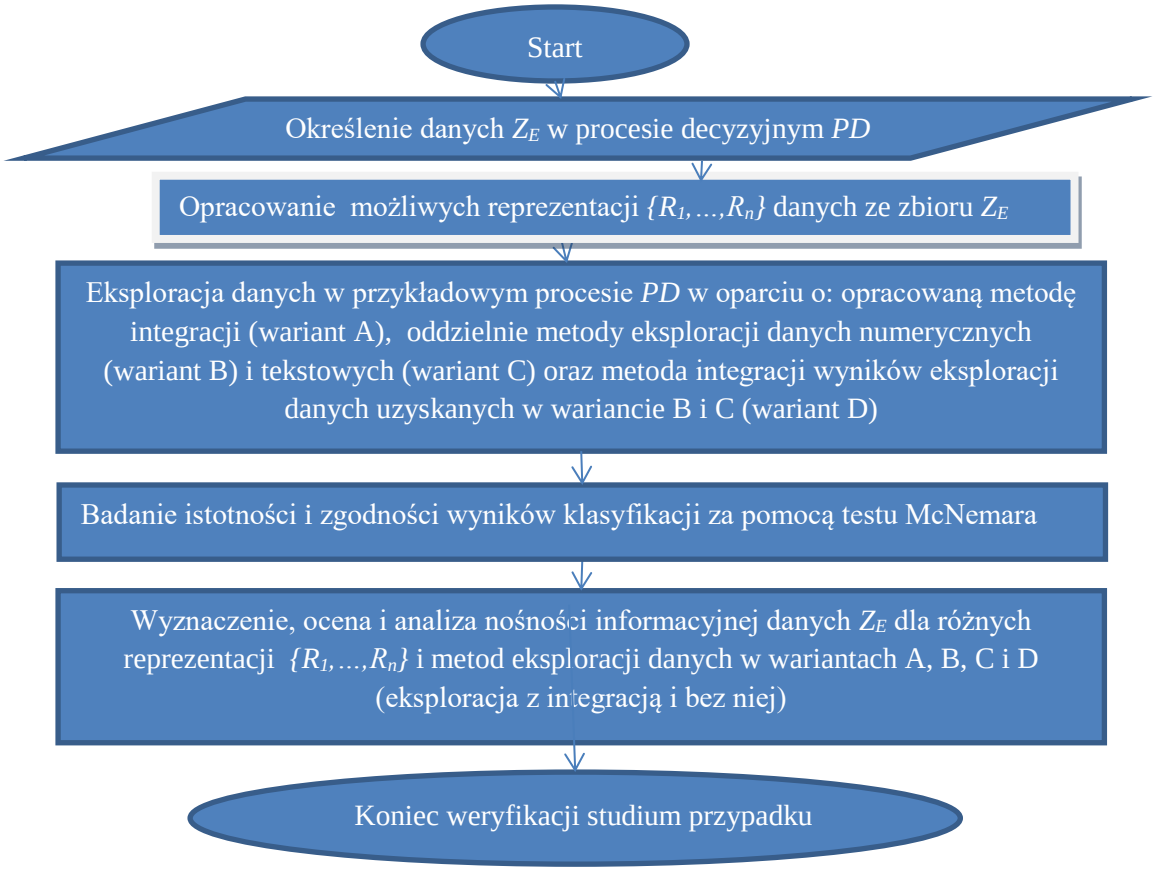
przypadek I. Problem decyzyjny dotyczący wyboru rentownych zamówień publicznych spośród zbioru takich zamówień,

przypadek II. Problem decyzyjny dotyczący sposobu inwestowania na Giełdzie Papierów Wartościowych,

przypadek III. Problem decyzyjny dotyczący wyszukiwania atrakcyjnych ofert pracy.

Dla każdego z przypadków I, II, i III dokonano weryfikacji hipotezy zgodnie z rysunkiem 2. W weryfikacji przypadków I, II i III przyjęto następujące założenia:

- W badaniach wykorzystano metodę eksploracji bazującą na klasyfikacji.
- Każdorazowo, do określenia rangi nośności informacyjnej danych w trzech przykładowych procesach decyzyjnych *PD*, dla wszystkich wariantów eksploracji (wariant A, B, C i D z rysunku 2) użyto miarę jakości decyzji *ACC* oraz *ERR* .
- Reprezentacja danych numerycznych została zdefiniowana z użyciem odpowiednio przeprowadzonej dyskretyzacji oraz doboru wartości nominalnych tych danych,
- Dane tekstowe zostały opracowane za pomocą trzech różnych reprezentacji, a mianowicie:
  - reprezentacji unigramowej - uwzględniającej pojedyncze wyrazy,
  - reprezentacji bigramowej – uwzględniającej sekwencje dwóch występujących po sobie wyrazów przy czym taka reprezentacja jest jednym z najczęściej wykorzystywanych typów reprezentacji n-gramowej [83, s. 20],
  - reprezentacji γ-gramowej - uwzględniająca sekwencje wyrazów o zmiennej długości ekstrahowane z tekstów za pomocą wzorców informacyjnych definiowanych przez eksperta oraz weryfikowane za pomocą analizy fleksyjnej języka polskiego.



Rysunek 2. Ogólny algorytm weryfikacji hipotezy dla każdego studium przypadków (przypadki I, II i III)
Źródło: opracowanie własne

1.3. Zakres pracy

Praca składa się z siedmiu rozdziałów. W rozdziale pierwszym opisano rolę eksploracji danych tekstowych i numerycznych w procesie podejmowania decyzji. Podkreślono znaczenie problemu integracji metod eksploracji danych tekstowych i numerycznych oraz określono cel pracy i hipotezę badawczą.

Rozdział drugi zawiera opis metod eksploracji danych tekstowych bazujących na klasyfikacji przeprowadzonej z wykorzystaniem modelu przestrzeni wektorowej VSM. Rozdział rozpoczyna się wprowadzeniem do analizy tekstu i omówieniem podstawowych zagadnień związanych eksploracją danych tekstowych. W szczególności scharakteryzowano tu różne podejścia do klasyfikacji dokumentów tekstowych, zarówno metody bazujące na uczeniu maszynowym jak i metody wykorzystujące wiedzę eksperta.

W rozdziale trzecim opisano metody klasyfikacji danych numerycznych. W rozdziale tym skoncentrowano się na Teorii Zbiorów Przybliżonych, która umożliwia budowanie wiedzy wykorzystywanej do podejmowania decyzji w procesie *PD* na bazie reguł decyzyjnych.

W szczególności opisano metody opracowania reprezentacji danych numerycznych (między innymi dyskretyzację danych numerycznych) oraz szczegółowo opisano metody wykorzystywane przy eliminacji szumu informacyjnego.

Przedmiotem rozdziału czwartego jest autorska procedura integracji metod klasyfikacji danych tekstowych i numerycznych w procesie podejmowania decyzji. W pierwszej kolejności został przedstawiony ogólny schemat procedury po czym szczegółowo opisano etapy tej procedury. Zaprezentowano tu metody budowy reprezentacji danych tekstowych oraz numerycznych. W szczególności opisano metodę opracowywania γ-gramowej reprezentacji danych tekstowych bazującą na wzorcach informacyjnych definiowanych przez eksperta dziedzinowego oraz analizie fleksyjnej rzeczowych informacji wyekstrahowanych z tekstu za pomocą wzorców. Kolejno opisano etap budowania systemu informacyjnego *SI*, na podstawie, którego generowana jest wiedza w procesie decyzyjnym *PD*.

W rozdziale piątym dokonano oceny różnych wariantów eksploracji (wariant A – z wykorzystaniem zintegrowanych metod eksploracji danych tekstowych i numerycznych, wariant B – z wykorzystaniem wyłącznie metody eksploracji danych tekstowych, wariant C – z wykorzystaniem wyłącznie metody eksploracji danych numerycznych, wariant D – z wykorzystaniem zintegrowanych wyników eksploracji z wariantów B i C), w oparciu wyniki badań testowych dotyczących przykładowych procesów podejmowania decyzji *PD*. W pierwszej części analizę poddano przykład, którego celem było wyszukiwanie rentownych zamówień publicznych w Biuletynie Zamówień Publicznych, kolejno przykład dotyczący inwestowania na Giełdzie Papierów Wartościowych oraz przykład związany z wyszukiwaniem atrakcyjnych ofert pracy.

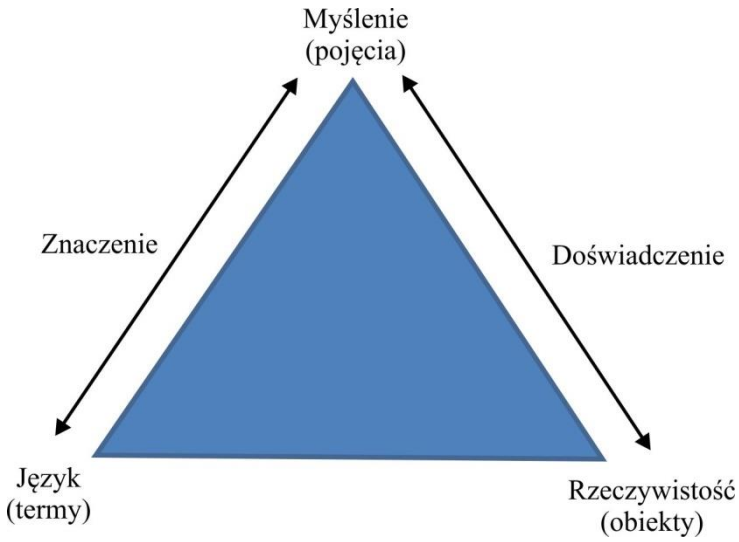
W rozdziale szóstym przeprowadzono dyskusję wyników uzyskanych w badaniach przypadków użycia.

Zakończenie pracy stanowi rozdział siódmy zawierający podsumowanie badań nad opracowaną procedurą integracyjną, sformułowano wnioski z realizacji celu pracy i weryfikacji postawionej hipotezy**.**

2. Klasyfikacja danych tekstowych

2.1. Wprowadzenie do eksploracji danych tekstowych

Dane tekstowe odnoszą się do zapisu tekstu, zazwyczaj w języku naturalnym. Dlatego w eksploracji danych tekstowych uwzględnia się reprezentację takich języków. Wówczas eksploracja danych tekstowych bazuje na strukturze zwanej trójkątem semiotycznym [76, ss. 253–254]. Jest to układ zależności zachodzący pomiędzy formą wyrażenia językowego tzw. *termem* (na przykład pojedynczy wyraz), *obiektem* (fragment rzeczywistości), na który wskazuje *term* oraz *pojęciem*, stanowiącym wyobrażenie (odwzorowanie) obiektu w umyśle człowieka. Schematycznie zaprezentowano to na rysunku 3.



Rysunek 3. Trójkąt semiotyczny.
Źródło: opracowanie własne na podstawie *[57, s. 12]*

Mając na uwadze powyższą zależność eksploracja tekstu powinna uwzględniać analizę na poziomie budowy tekstu oraz informacji znaczeniowej, której tekst jest nośnikiem. Z tego względu do pozyskania wiedzy w procesie eksploracji danych tekstowych można wyróżnić dwa zasadnicze podejścia [84, s. 22]:

- oparte o uczenie maszynowe,
- bazujące na wiedzy eksperta.

Pierwsze z nich - uczenie maszynowe - jest automatyczne i w głównej mierze opiera się na metodzie statystyczno-matematycznej. Podejście to polega na badaniu cech charakteryzujących strukturę dokumentu tekstowego np. zliczanie wyrazów czy rozkład występowania wyrazów w tekście. Drugie podejście większym stopniu koncentruje się na wiedzy eksperta, dotyczy technik zarządzania wiedzą i mocno związane jest z analizą

znaczeniową tekstu. Podejście to wykorzystuje reguły leksykalne i składniowe danego języka oraz bierze pod uwagę znaczenie analizowanych wyrazów i fraz. W tym podejściu istotna jest znajomość gramatyki analizowanego języka i specyfiki wypowiedzi związanej ze stosowanym słownictwem.

Popularniejsze, w związku z łatwością jego praktycznego zastosowania jest uczenie maszynowe. Wynika to głównie z jego funkcjonowania bez znaczącego udziału eksperta. Dlatego większość komercyjnych systemów bazuje na automatycznej analizie tekstu. Nie oznacza to jednak, że oparcie analizy tekstu wyłącznie na uczeniu maszynowym jest najkorzystniejsze. P.Gawrysiak pisze [22, s. 10]: „*Wydaje się raczej, że przyszłe systemy przetwarzania języka naturalnego (ang. Natural Language Processing, NLP), korzystać będą zarówno z wiedzy ekspertów-lingwistów, zapisanej w postaci bazy wiedzy, jak też i z systemów analizy automatycznej, dzięki której będą w stanie wiedzę tę modyfikować i uaktualniać”*. Z tego względu rozwiązaniem pełniejszym wydaje się uwzględnienie w eksploracji danych tekstowych tych dwóch podejść.

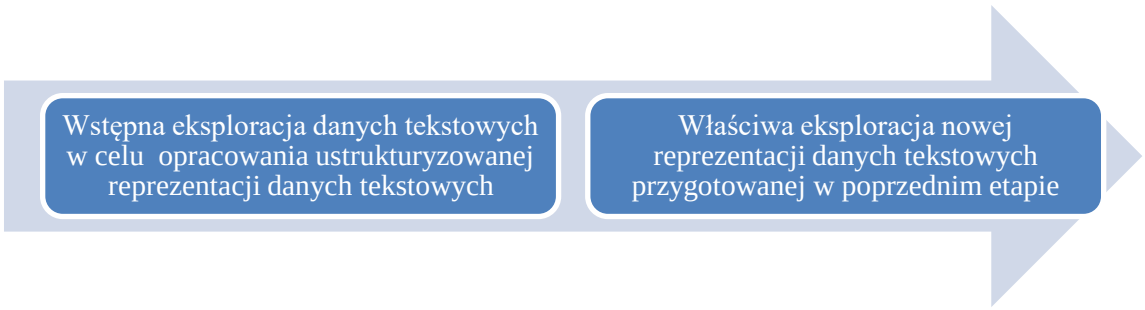
Wśród najważniejszych zadań eksploracji danych tekstowych wymienia się [50, ss. 71–74][71, s. 411][62, s. 4]:

- ranking dokumentów tekstowych,
- wyszukiwanie dokumentów tekstowych,
- klasyfikację dokumentów tekstowych,
- grupowanie dokumentów tekstowych,
- analizę powiązań dokumentów tekstowych,
- wizualizacja cech dokumentów tekstowych oraz wyników pozostałych zadań eksploracji dokumentów tekstowych.

W kolejnych częściach niniejszego rozdziału zostały omówione strategie eksploracji danych tekstowych bazującej na uczenie maszynowym oraz wiedzy eksperta, wykorzystywanej do zadania klasyfikacji (zgodnie z założeniami w rozdziale 1.2).

2.2. Klasyfikacja danych tekstowych z wykorzystaniem metod uczenia maszynowego

Eksploracja danych tekstowych zazwyczaj bazuje na modelu przestrzeni wektorowej (ang. Vector Space Model, VSM), który stanowi formalny sposób reprezentacji dokumentów tekstowych w wielowymiarowej przestrzeni euklidesowej [54, ss. 531–532]. Procedurę eksploracji danych tekstowych w tym modelu można podzielić na dwa główne etapy, przedstawione na rysunku 4.



Rysunek 4. Dwuetapowy proces eksploracji danych tekstowych.
Źródło: opracowanie własne

Nieodłącznym etapem poprzedzającym właściwą eksplorację danych tekstowych z wykorzystaniem modelu przestrzeni wektorowej oraz mającym olbrzymi wpływ na wynik eksploracji jest opracowanie ustrukturyzowanej reprezentacji dokumentów tekstowych. Rolą wstępnej eksploracji, na podstawie której opracowywana jest ustrukturyzowana reprezentacja danych tekstowych jest przede wszystkim eliminacja elementów zbędnych występujących w tekście (tzw. szumu informacyjnego), które mogłyby negatywnie wpłynąć na wynik właściwej eksploracji. W ramach wstępnej eksploracji dobierany jest zbiór odpowiednich cech reprezentujących dokumenty tekstowe.

Dokument tekstowy *t* w pierwotnej formie języka naturalnego jest to ciąg *mw* wyrazów rozdzielonych znakami, które dzielą tekst na zdania, co zdefiniowano wzorem *(5)* [24, s. 35].

$$t = (w_1, w_2, \dots, z_1, \dots, w_{mw}, z_{mz}), \forall iw \in \langle 1, mw \rangle, iz \in \langle 1, mz \rangle; w_{iw} \epsilon\, V, z_{iz} \epsilon Z \quad (5)$$

gdzie:

V – słownik wszystkich wyrazów, które mogą wystąpić w dokumencie tekstowym,
mw – maksymalna liczba wyrazów wydobytych z dokumentu tekstowego,
iw – indeks wyrazu, wydobytego z dokumentu tekstowego,
w_{iw} – wyraz w dokumencie tekstowym, kilka wyrazów składa się na zdanie, którego granice wytycza znak *z*,
Z – zbiór wszystkich znaków, które mogą kończyć zdanie,
mz – maksymalna liczba znaków wydobyta z dokumentu tekstowego,
iz – indeks wyrazu *z*, wydobytego z dokumentu tekstowego,
z_{iz} – znak rozdzielający zdania (np. kropka, wykrzyknik).

W etapie wstępnej eksploracji danych, dokumenty tekstowe w formie języka naturalnego zostają odwzorowane zastępczą reprezentacją wyrażoną wzorem *(3)*, w postaci wektora cech tj. zbioru elementów charakteryzujących dokumenty tekstowe np. występujących w nich

pojedynczych wyrazów, zwanych termami. Reprezentacja dokumentów tekstowych za pomocą wektorów pozwala na wykonywanie określonych formalnych przekształceń na danych tekstowych, co umożliwia wykorzystanie w ich analizie zaawansowanych metod i algorytmów właściwej eksploracji danych.

Dla przykładu niech zbiór sześciu cech (termów) będących elementami reprezentacji *R* jest zgodny z tabelą 1.

Tabela 1. Cechy *r1…r6* reprezentacji *R*.

Symbol cechy	Cechy w postaci pojedynczych wyrazów
r_1	bazy
r_2	SQL
r_3	indeks
r_4	regresja
r_5	wiarygodność
r_6	liniowa

Źródło: opracowanie własne na podstawie [62, s. 15]

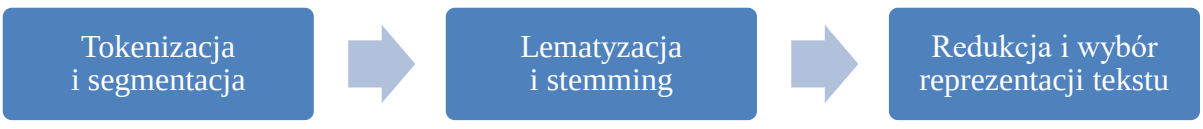
Zaś reprezentacja *R* dla dziewięciu przykładowych dokumentów tekstowych t_1, $t_2,\ldots,t_9$ w modelu przestrzeni wektorowej VSM jest zgodna z tabelą 2.

Tabela 2. Reprezentacja dokumentów tekstowych *t1…t9* w modelu przestrzeni wektorowej składająca się z cech *r1…r6*.

	r_1	r_2	r_3	r_4	r_5	r_6
t_1	24	21	9	0	0	3
t_2	32	10	5	0	3	0
t_3	12	16	5	0	0	0
t_4	6	7	2	0	0	0
t_5	43	31	20	0	3	0
t_6	2	0	0	18	7	16
t_7	0	0	1	32	12	0
t_8	1	0	0	34	27	25
t_9	6	0	0	17	4	23

Źródło: opracowanie własne na podstawie [62, s. 15]

Opracowanie ustrukturyzowanej reprezentacji danych tekstowych składa się zazwyczaj z trzech głównych części zaprezentowanych na rysunku 5 [75, s. 4537].



Rysunek 5. Przygotowanie danych tekstowych.

Źródło: opracowanie własne

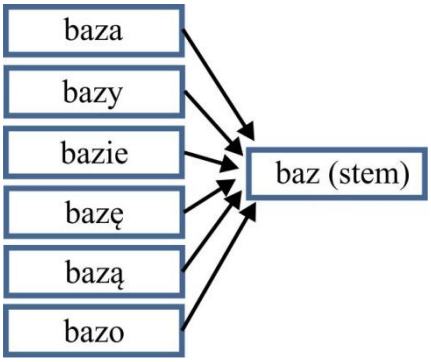
W pierwszej kolejności tekst zostaje przekształcony z formy ciągłej w zbiory zdań i pojedynczych wyrazów. Przekształcenie tekstów realizowane jest poprzez tokenizację i segmentację. Tokenizacja jest procesem, w którym monolityczny tekst zostaje podzielony na ciąg pojedynczych tokenów, zazwyczaj pojedynczych wyrazów [47]. W piśmie języków, w którym granice wyrazów nie są wyraźnie zaznaczone, tokenizacja jest rozumiana jako segmentacja [69, s. 1]. Segmentacja jest procesem podziału tekstu na językowe jednostki znaczeniowe np. wyrazy czy całe zdania [69, s. 1]. Czasem segmentacja dotyczy również podziału tekstu na większe jednostki - części tekstu dotyczące wyodrębnionych podtematów (ang. TextTiling), które mogą składać się z kilku zdań lub akapitów [28, s. 34]. W przypadku segmentacji tekstu napotyka się wiele zjawisk, takich jak haplologia kropki [73, s. 36] niejednoznaczne skróty, przeniesienie części wyrazu do następnej linii, które wymagają konstruowania bardzo precyzyjnych reguł segmentacji [73, ss. 36–38]. W tym celu wykorzystuje się reguły segmentacji tekstu zdefiniowane za pomocą wyrażeń regularnych, szerzej opisane w literaturze [40].

Kolejnym etapem przygotowania ustrukturyzowanej reprezentacji danych tekstowych jest proces lematyzacji, czyli sprowadzenia wyrazów do ich form podstawowych (lematu) np. *materia* - mianownik liczby pojedynczej dla rzeczownika, *materializować* - bezokolicznik dla czasownika itd. Dzięki powyższej operacji odmienne formy gramatyczne traktowane są jako jeden wyraz, co pozwala na zidentyfikowanie wystąpień tego samego wyrazu w różnych miejscach tekstu. Stemming w odróżnieniu od lematyzacji jest to proces polegający na wydobyciu z wybranego wyrazu tzw. rdzenia (ang. stem), a więc tej jego części wyrazu, która jest odporna na odmianę [98, ss. 21–23]. Przykład stemmingu dla wyrazu *baza* przedstawiono na rysunku 6.

Zarówno w przypadku lematyzacji, jak i stemmingu, można wyodrębnić dwa odmienne podejścia do ich realizacji [39, s. 16]:

1. Słownikowe, które polega na wykorzystaniu słownika np. słownika fleksyjnego języka polskiego, który zawiera zarówno formę podstawową wyrazu (lemat lub rdzeń) jak i jego różne formy gramatyczne. Wydobyty z tekstu wyraz jest wyszukiwany w słowniku, a następnie z bazy wyrazów jest pobierana jego forma podstawowa.
2. Algorytmiczne, które polega na wykorzystaniu zbioru reguł pozwalających wykryć i usunąć różnice pomiędzy poszczególnymi formami gramatycznymi

wyrazów. Przykładem takiego rozwiązania jest system generowania Słownika Fleksyjnego Języka Polskiego szerzej opisany w literaturze [49, ss. 47–67].



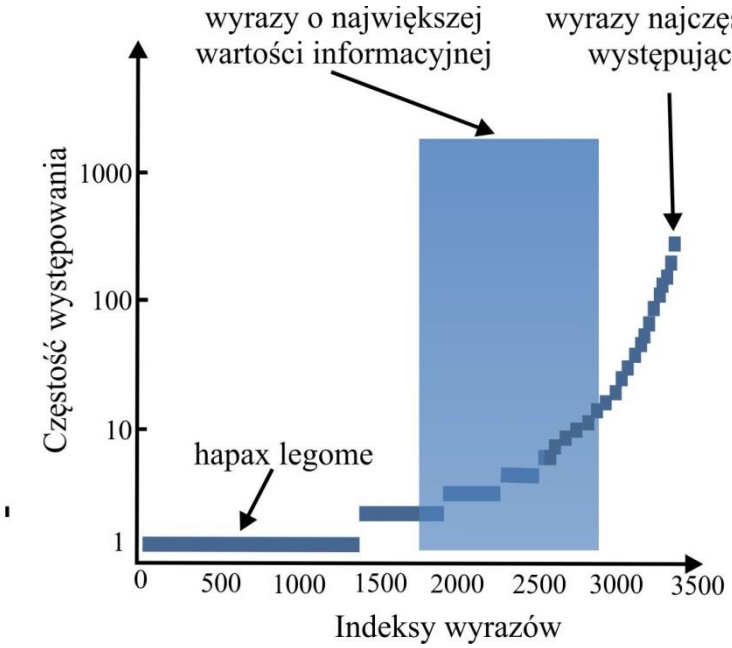
Rysunek 6. Proces stemmingu.
Źródło: Opracowanie własne na podstawie [75, s. 4537]

W praktyce istnieje możliwość równoczesnego użycia zarówno lematyzacji jak i stemmingu. W takim przypadku lematyzacji używa się w stosunku do wyrazów rozpoznanych w dokumencie tekstowym, natomiast stemmingu do wyrazów nierozpoznanych [49, s. 166].

W ostatnim etapie opracowania ustrukturyzowanej reprezentacji danych tekstowych wykonywana jest redukcja obszernego zbiór wyrazów wydobytych z tekstu, z których jedynie część ma istotne znaczenie. Ograniczenie wielkości zbioru wyrazów może zostać zrealizowane za pomocą pięciu różnych technik:

1. Użycie tzw. list stopujących (ang. stop list), czyli list zawierających wyrazy, które nie wpływają bezpośrednio na znaczenie tekstu, a jedynie kształtują tok wypowiedzi. Są to przeważnie wyrazy najczęściej używane w tekstach danego języka np. spójniki, zaimki, rodzajniki itp. Stop listy definiowane są zazwyczaj na podstawie analizy częstości występowania wyrazów w dużym, zróżnicowanym zbiorze tekstów danego języka. Wyrazy ze stop listy usuwane są ze zbioru wyrazów wydobytych z korpusu dokumentów tekstowych [39, s. 14][49, s. 166].
2. Ograniczenie zbioru wyrazów wydobytych z tekstu do zdefiniowanego zamkniętego katalogu wyrazów [100, s. 365]. Maksymalny rozmiar zbioru wydobytych i rozpoznanych wyrazów z korpusu może być równy liczbie wyrazów w zdefiniowanym katalogu.
3. Wykorzystanie reguł stopujących dla analizowanego korpusu tekstów, przeważnie bazujących na prawie Zipfa [13, ss. 94–95], zgodnie z którym mały zbiór wyrazów

występujące w tekstach najczęściej nie jest wartościowy informacyjnie. Z drugiej strony usunięcie wyrazów występujących najrzadziej również eliminuje pewien szum informacyjny. Jest to najczęściej związane z jednokrotnym wystąpieniem wyrazów w tekście (tzw. hapax legomena). Pośród tych wyrazów znajdują się również błędy typu literówki. Dlatego z pełnego zbioru wyrazów z danego korpusu tekstów wyodrębnia się wyrazy najrzadziej i najczęściej występujące, które są następnie usuwane. Ostatecznie do dalszej analizy brane są pod uwagę wyłącznie wyrazy o największej wartości informacyjnej. Graficznie problem ten zilustrowano na rysunku 7.



Rysunek 7. Eliminacja szumu informacyjnego.
Źródło: opracowanie własne na podstawie [23, s. 47]

4. Redukcja reprezentacji dokumentów tekstowych wykorzystująca automatyczne wykrywanie najlepiej powiązanych semantycznie ze sobą wyrazów w tekście [48, ss. 119–131]. Metoda redukcji reprezentacji tekstu wykorzystująca najlepiej powiązane semantycznie wyrazy ze sobą bazuje na generowaniu tzw. list skojarzeniowych wyrazów za pomocą metody statystyczno-matematycznej, zgodnie ze wzorem *(6)* [49, s. 150].

$$sk = \frac{cw}{lw} \qquad (6)$$

gdzie:

sk – wyrażona w procentach miara skojarzenia wyrazu definiującego z definiowanym,

cw – częstość względna wyrazu definiującego, tj. ilość wystąpień wyrazu definiującego w z daniach z wyrazem definiowanym,

lw – częstość bezwzględna wyrazu definiowanego, czyli ilość wystąpień wyrazu w korpusie tekstów, który posłużył do wygenerowania listy skojarzeniowej.

Za pomocą zdefiniowanej we wzorze *(6)* miary skojarzeniowej możliwe jest wygenerowanie list skojarzeniowych dla wybranych wyrazów występujących w korpusie tekstów, a następnie ograniczenie całego zbioru wyrazów w tekście do tych, które posiadają wystarczającą miarę skojarzeniową. W celu wyeliminowania wyrazów o najniższej wartości miary skojarzeniowej eksperymentalnie dopierany jest próg (wartość graniczna miary skojarzeniowej), który decyduje o uwzględnieniu lub odrzuceniu danych wyrazów. Nieco inną możliwą do zastosowania metodą czyszczenia list skojarzeniowych opisywaną w literaturze jest analiza odległości pomiędzy częstościami względnymi wyrazów na liście. W przypadku gdy odległość jest mniejsza niż arbitralnie ustalona wielkość następuje usunięcie wyrazów [49, s. 151].

5. W celu redukcji reprezentacji dokumentów tekstowych wykorzystywane są również tzw. tezaurusy, czyli słowniki wyrazów bliskoznacznych. Przykładem takich słowników jest WordNet lub Słownik Semantyczny Języka Polskiego [20][49, ss. 243–248]. Dzięki wykryciu wszystkich wyrazów bliskoznacznych w tekstach korpusu i zamianie ich na jeden, wspólny wyraz zwiększa się stopień podobieństwa dokumentów tekstowych. Wykorzystanie tezaurusów zazwyczaj poprawia jakość wyniku uzyskanego w procesie eksploracji danych. Czasami tezaurusy stosuje się zamiennie z niejawną indeksacją semantyczną (ang. Latent Semantic Indexing), która wykrywa związki semantyczne pomiędzy wyrazami odpowiadające synonimii za pomocą obliczeń algebraicznych.

Newralgicznym punktem opracowania reprezentacji danych w modelu VSM jest dobranie właściwych cech *r* opisujących dokument tekstowy *t* [22, ss. 18–19]. W literaturze opisano kilka najczęściej stosowanych reprezentacji dokumentów tekstowych tj. reprezentacje unigramową, n-gramową oraz γ-gramowa [24, s. 8].

Reprezentacja **unigramowa** (ang. unigram model) bazuje na zliczaniu częstości występowania pojedynczych wyrazów w dokumencie tekstowym, które następnie wykorzystywane są do zbudowania wektora reprezentującego dany dokument tekstowy. Zliczanie występowania wyrazów można zrealizować poprzez odnotowanie (lub nie) danego wyrazu (reprezentacja binarna) lub zliczanie liczby wystąpień wyrazu (unigramowa reprezentacja częstościowa) w treści dokumentu tekstowego. Unigramową reprezentację binarną dokumentu tekstowego *t* stanowi wektor *R* taki, że elementy tej reprezentacji (elementy wektora) obliczane są zgodnie ze wzorem *(7)* [22, s. 36]:

$$r_j = \begin{cases} 1\ gdy\ w_{iw} = v_{iv}, v_{iv} \in V \\ 0\ wpw. \end{cases} \quad (7)$$

gdzie:

r_j – element reprezentacji dokumentu tekstowego, gdzie $j \in \{1,2, \dots mr\}$,

mr – maksymalna liczba elementów reprezentacji dokumentu tekstowego,

V – słownik wyrazów,

mv – maksymalna liczba wyrazów w słowniku,

iv – indeks wyrazu w słowniku *V*, gdzie $iv \in \langle 1, mv \rangle$,

v_{iv} – wyraz w słowniku,

mw – maksymalna liczba wyrazów wydobytych z dokumentu tekstowego,

iw – indeks wyrazu wydobytego z dokumentu tekstowego, gdzie $iw \in \langle 1, mw \rangle$,

w_{iw} – wyraz występujący w dokumencie tekstowym *t*,

wpw. – skrót oznaczający „w przeciwnym wypadku”.

Elementy unigramowej reprezentacji częstościowa obliczane są zgodnie ze wzorem *(8)* [22, s. 36]:

$$r_j = \sum_{iw=1}^{mw} \begin{cases} 1\ gdy\ w_{iw} = v_{iv}, v_{iv} \in V \\ 0\ wpw. \end{cases} \quad (8)$$

gdzie:

r_j – element reprezentacji dokumentu tekstowego, gdzie $j \in \{1,2, \dots mr\}$,

mr – maksymalna liczba elementów reprezentacji dokumentu tekstowego,

V – słownik wyrazów,

mv – maksymalna liczba wyrazów w słowniku,

iv – indeks wyrazu w słowniku *V*, gdzie , $iv \in \langle 1, mv \rangle$,

v_{iv} – wyraz w słowniku,

mw – maksymalna liczba wyrazów wydobytych z dokumentu tekstowego,

iw – indeks wyrazu wydobytego z dokumentu tekstowego, gdzie , $iw \in \langle 1, mw \rangle$,

w_{iw} – wyraz występujący w dokumencie tekstowym *t*,

wpw. – skrót oznaczający „w przeciwnym wypadku”.

Ze względu na objętość dokumentu tekstowego, która może zmniejszać znaczenie tematyki tekstu, bezwzględną częstość wyrazów w wektorze *R* zastępuje się częstością względną zdefiniowaną wzorem *(9)* [22, s. 36].

$$r_j = \frac{\sum_{iw=1}^{mw} \begin{cases} 1 \; gdy \; w_{iw}=v_{iv}, v_{iv} \in V \\ 0 \; wpw. \end{cases}}{mw} \qquad (9)$$

gdzie:

r_j – element reprezentacji dokumentu tekstowego, gdzie $j \in \{1,2, \dots mr\}$,

mr – maksymalna liczba elementów reprezentacji dokumentu tekstowego,

V – słownik wyrazów,

mv – maksymalna liczba wyrazów w słowniku,

iv – indeks wyrazu w słowniku *V*, gdzie $iv \in \langle 1, mv \rangle$,

v_{iv} – wyraz w słowniku,

mw – maksymalna liczba wyrazów wydobytych z dokumentu tekstowego,

iw – indeks wyrazu wydobytego z dokumentu tekstowego, gdzie $iw \in \langle 1, mw \rangle$,

w_{iw} – wyraz występujący w dokumencie tekstowym *t*,

wpw. – skrót oznaczający „w przeciwnym wypadku”.

Kolejna reprezentacja **n-gramowa** bazuje na zliczaniu częstości występowania sekwencji obejmujących określoną liczbę występujących po sobie w tekście wyrazów ze słownika *V*. Elementy reprezentacji n-gramowej dokumentu tekstowego *t* obliczane są zgodnie ze wzorem *(10)* [22, s. 42].

$$r_j = \sum_{iw=1}^{mw-lw} \begin{cases} 1 \; gdy \; (w_{iw}, w_{iw+1}, \dots, w_{iw+lw-1}) = s_{is} \wedge w_{iw} = v_{iv}, v_{iv} \in V \\ 0 \; wpw. \end{cases} \qquad (10)$$

gdzie:

r_j – element reprezentacji dokumentu tekstowego, gdzie $j \in \{1,2, \dots mr\}$,

mr – maksymalna liczba elementów reprezentacji dokumentu tekstowego,

V – słownik wyrazów,

mv – maksymalna liczba wyrazów w słowniku,

iv – indeks wyrazu w słowniku *V*, gdzie $iv \in \langle 1, mv \rangle$,

v_{iv} – wyraz w słowniku,

mw – maksymalna liczba wyrazów wydobytych z dokumentu tekstowego,

iw – indeks wyrazu wydobytego z dokumentu tekstowego, gdzie $iw \in \langle 1, mw \rangle$,

w_{iw} – wyraz występujący w dokumencie tekstowym t,

lw – długość sekwencji wyrazów,

is – indeks sekwencji wyrazów, gdzie $is \in (1, (mw\text{-}lw))$,

s_{is} – sekwencje obejmująca lw wyrazów v ze słownika V,

wpw. – skrót oznaczający „w przeciwnym wypadku”.

Trzecim rodzajem reprezentacji jest reprezentacja ***γ*-gramowa**, której elementy stanowią sekwencje wyrazów o zmiennej długości. Reprezentacja γ-gramowa może być oparta na monotonicznej funkcji oceniającej $\gamma(w_1,\ldots,w_{lw})$, której wartości odpowiadają przydatności danej sekwencji wyrazów $w_1,\ldots,w_k$ w analizie dokumentu tekstowego [22, s. 45]. Wyłonienie najbardziej przydatnych sekwencji o zmiennej długości realizowane jest w ramach tzw. wstępnej eksploracji danych tekstowych za pomocą odpowiedniego algorytmu.

Pomimo tego, że powyższe reprezentacje dokumentów tekstowych należą do najczęściej stosowanych możliwe jest również wykorzystanie innych baz przestrzeni cech [39, s. 37]. Są to przeważnie propozycje reprezentacji dokumentów tekstowych bazujących na rozszerzeniach lub przekształceniach wymienionych reprezentacji. Przykładem może być **reprezentacja pozycyjna**, która przechowuje informację zarówno o częstości występowania wyrazów w tekście jak i o miejscach tych wystąpień. Dlatego jest to stosunkowo proste rozszerzenie reprezentacji unigramowej wyrażonej za pomocą funkcji gęstości wyrazów oraz wektora skalującego o wartościach takich jak wektor względnej unigramowej reprezentacji częstościowej [22, ss. 47–52].

Po etapie opracowania reprezentacji danych tekstowych realizowana jest właściwa eksploracja danych bazująca na odpowiednich względem zadania eksploracji technikach i metodach. W celu zwiększenia dokładności eksploracji tekstu, elementom reprezentacji (termom) nadaje się wagi wykorzystując do tego funkcje istotności. Podstawowa funkcja istotności nadaje elementom reprezentacji tekstu wagi binarne, czyli odnotowuje obecność danego termu w dokumencie t za pomocą wartości 0 lub 1. W literaturze wyróżnia się również bardziej złożone funkcje istotności nadające wagi lokalne, które określają wpływ termu w obrębie tekstu, wagi globalne w obrębie całego korpusu oraz wagi mieszane będące połączeniem dwóch poprzednich wag. Przykładem wagi lokalnej jest waga *tf* (ang. Term Frequency) odpowiadająca częstościowej reprezentacji unigramowej zdefiniowanej

za pomocą wzoru *(9)*, globalnej *idf* (ang. Inverse Document Frequency) oraz mieszanej *tfidf* [49, s. 168]. Wagę *tf* określa stosunek liczby danego elementu reprezentacji występującego w dokumencie tekstowych do sumy wystąpień wszystkich elementów reprezentacji dokumentu tekstowego i jest zdefiniowana wzorem *(11)*.

$$tf_j = \frac{\sum r_j}{\sum_{jj=1}^{mr} \sum r_{jj}} \tag{11}$$

gdzie:

tf_j – częstość termów (waga lokalna) w dokumencie tekstowym dla elementu r_j, gdzie $j \in \{1,2, \dots mr\}$,

r_j – element reprezentacji dokumentu tekstowego,

mr – maksymalna liczba elementów reprezentacji dokumentu tekstowego,

$\sum r_j$ – suma wystąpień elementu r_j w dokumencie tekstowym,

$\sum_{jj=1}^{mr} \sum r_{jj}$ – suma wystąpień wszystkich elementów w dokumencie tekstowym, gdzie $jj \in \{1,2, \dots mr\}$.

Waga *idf* jest wyrażona wzorem *(12)*.

$$idf = log\left(\frac{mt}{\sum t : r_j \in t}\right) + 1 \tag{12}$$

gdzie:

idf – odwrotna częstość termów (waga globalna),

t – dokument tekstowy, dla elementów którego obliczana jest waga *idf*,

mt – liczba dokumentów tekstowych w korpusie,

$\sum t : r_j \in t$ – liczba dokumentów tekstowych zawierających przynajmniej jedno wystąpienie termu r_j, , gdzie $j \in \{1,2, \dots mr\}$,

r_j – element reprezentacji dokumentu tekstowego,

mr – maksymalna liczba elementów reprezentacji dokumentu tekstowego.

Waga *tfidf* jest wyrażona wzorem *(13)*:

$$tfidf = tf * idf \tag{13}$$

gdzie:

tfidf – waga wynikająca z połączenia częstość termów z odwrotną częstością termów (waga mieszana),

tf – częstość termów (waga lokalna),

idf – odwrotna częstość termów (waga globalna).

Oprócz wyżej wymienionych funkcji istotności istnieją również takie, które uwzględniają w swoich obliczeniach bardziej zaawansowane techniki statystyczne bazujące np. na tzw. listach skojarzeniowych, czyli statystykach współwystępowaniu ze sobą w tekście poszczególnych wyrazów.

W eksploracji danych tekstowych istnieje również możliwość zastosowania niejawnej indeksacji semantycznej (ang. Latent Semantic Indexing - LSI), która jest matematyczno-statystyczną metodą eksploracji danych wykorzystywaną do wykrywania za pomocą obliczeń algebraicznych ukrytych struktur semantycznych istniejących w tekście pomiędzy pojedynczymi wyrazami. Poprzez redukcję ilości wymiarów reprezentacji tekstu dąży się do wykrycia wzorców współwystępowania ze sobą różnych wyrazów, które stanowią nowe elementy zredukowanej reprezentacji tekstu. Dzięki tej technice istnieje możliwość wykrycia niejawnych zależności pomiędzy poszczególnymi elementami reprezentacji tekstu (standardowo pojedynczymi wyrazami), które pozwalają go jeszcze lepiej odwzorować. Przy użyciu metody LSI dokonuje się redukcji wymiaru wektorów do określonej ilości wykrytych struktur semantycznych pomiędzy pojedynczymi wyrazami. Metoda LSI dobiera optymalne rzutowanie dla zadanej ilości struktur semantycznych, ponieważ odpowiadają one największemu zróżnicowaniu pierwotnych elementów reprezentacji. Redukcji wymiarowości reprezentacji dokumentów tekstowych dokonuje się dzięki rozkładowi według wartości osobliwych (ang. Singular Value Decomposition, SVD) dla macierzy termów i dokumentów tekstowych. Rzadka macierz termów i tekstów *Mrt* zostaje zdekomponowana zgodnie ze wzorem *(14)* [30, ss. 179–180]:

$$Mrt_{mr\ x\ mt} = Mr_{mr\ x\ o} * O_{o\ x\ o} * (Mt_{o\ x\ mt})^T \qquad (14)$$

gdzie:

Mrt – dekomponowana macierz termów i dokumentów tekstowych o wymiarze (mr x mr),

mr – liczba termów,

mt – liczba dokumentów tekstowych,

o – liczba wartości osobliwych, gdzie $o = min(mr, mt)$,

O – macierz wartości osobliwych,

Mr – macierz termów,

Mt – macierz dokumentów tekstowych,

T – transpozycja macierzy.

W celu uzyskania określonej liczby wykrytych struktur semantycznych pomiędzy termami macierz wartości osobliwych oraz macierz tekstów zostaje zredukowana do *zo* kolumn, gdzie $zo<o$. Następnie zredukowane macierze O_{zoxzo} i $Mt_{zoxmt}{}^{T}$ zostają przemnożone. W wyniku czego otrzymywana jest macierz, której elementami (termami) są struktury semantyczne będące nowymi elementami reprezentacji dokumentu tekstowego.

Wartość *zo* dobiera się na podstawie analizy wartości osobliwych za pomocą określonych reguł. W literaturze opisano metody wyboru ilości struktur semantycznych na podstawie analizy wartości osobliwych. Kilka przykładów z nich to [49]:

1) reguła ilości tekstów – taka ilość wartości osobliwych idąc od największej, że ich suma jest równa lub przekracza ilość tekstów,
2) reguła wartości większych od 1,
3) reguła proporcji – taka ilość wartości osobliwych idąc od największej, że proporcja ich sumy do sumy wszystkich jest równa lub przekracza zadaną proporcję.

Ostatecznie, w przypadku zadania eksploracji polegającego na wyszukiwaniu dokumentów tekstowych, zostaje obliczone podobieństwo dokumentu w stosunku do zapytania (tekstu wzorcowego, zapytania w postaci języka naturalnego lub wybranych słów kluczowych). Podobieństwo dokumentów tekstowych określa się poprzez obliczoną odległość pomiędzy wektorami, które je reprezentują. Najczęściej stosuje się miarę kosinusową, zgodnie ze wzorem *(15)* [9, s. 27]:

$$cos(t, tw) = \frac{\sum_{j=1}^{mr} rw_j r_j}{\sqrt{\sum_{j=1}^{mr} rw_j{}^2}\sqrt{\sum_{j=1}^{mr} r_j{}^2}} \qquad (15)$$

gdzie:

t – dokument tekstowy,

r – element reprezentacji dokumentu tekstowego,

tw – wzorcowy dokument tekstowy,

rw – element reprezentacji wzorcowego dokumentu tekstowego,

r_j – element reprezentacji dokumentu tekstowego, gdzie $j \in \{1,2,\dots mr\}$,

mr – maksymalna liczba elementów reprezentacji dokumentu tekstowego.

W wyniku obliczeń zgodnych ze wzorem *(15)* powstaje ranking wartości podobieństwa posortowany od najwyższego do najniższego podobieństwa. Zgodnie z rankingiem najbardziej

odpowiadające zapytaniu teksty znajdują się na szczycie listy rankingowej. Pozwala to na zidentyfikowanie wyszukiwanego dokument lub zbiór dokumentów tekstowych i jest wykorzystywane przy klasyfikacji.

Klasyfikacja danych tekstowych polega ona na przyporządkowaniu ocenianego dokumentu tekstowego do pewnej klasy (kategorii) zdefiniowanej przez eksperta i reprezentowanej przez zbiór wzorcowych (treningowych) dokumentów tekstowych, co zdefiniowano wzorem *(16)*.

$$f: t \rightarrow kl \tag{16}$$

gdzie:

f – funkcja przyporządkowująca dokumenty tekstowe *t* do klas *kl*,

t – dokument tekstowy,

kl – klasa reprezentowana przez zbiór wzorcowych (treningowych) dokumentów tekstowych $\{tw_{1,} \ldots tw_{mtw}\}$, zwanym zbiorem treningowym,

mtw – liczba dokumentów wzorcowych.

W modelu przestrzeni wektorowej klasyfikacja dokumentów tekstowych odbywa się za pośrednictwem formalnego modelu zwanego klasyfikatorem, który budowany jest na podstawie danych tekstowych w postaci zbioru treningowego. Klasyfikator służy do predykcji klasy, w oparciu o określony algorytm klasyfikacji, dla ocenianych dokumentów tekstowych, dla których przydział do odpowiedniej klasy nie jest znany. Najczęściej wykorzystywanymi algorytmami klasyfikacji są: Rocchio, k-Nearest Neighbors (kNN), klasyfikator Bayesa, Support Vector Machnies (SVM) oraz sieci neuronowe [4][22, s. 19]. Poniżej zaprezentowano kilka z wymienionych klasyfikatorów, które wyróżniają się szybkością działania i prostą implementacją.

Klasyfikator kNN tzw. *k*-najbliższych sąsiadów należy do najefektywniejszych algorytmów klasyfikacji dokumentów tekstowych [22, s. 110]. Dla ocenianego dokumentu tekstowego wyszukiwanych jest *k* najbardziej podobnych dokumentów tekstowych ze zbioru treningowego. W pośród *k* wyszukanych dokumentów znajdują się wzorcowe (treningowe) dokumenty tekstowe reprezentujące różne kategorie. Dla każdej kategorii obliczane jest średnie podobieństwo pomiędzy treningowymi dokumentami tekstowymi do niej należącymi, a klasyfikowanym dokumentem tekstowych. Podobieństwo jest oznaczone jako funkcją *sim()* i obliczane zgodnie ze wzorem *(17)* [83, s. 25].

$$sim(t, kl) = \frac{\sum_{itw=1}^{mtw} sim(t, tw_{itw})}{mtw} \tag{17}$$

gdzie:

t – dokument tekstowy,

kl – klasa dokumentów wzorcowych (treningowych),

tw – dokument wzorcowy należący do klasy *kl*,

itw – indeks dokumentu wzorcowego,

mtw – liczba dokumentów wzorcowych należących do klasy *kl*.

Ostatecznie dokument tekstowy jest sklasyfikowany do kategorii, dla której obliczona wartość średniego podobieństwa jest najwyższa. Na wynik klasyfikacji duży wpływ ma wybór odpowiedniego współczynnika *k*.

Naiwny klasyfikator Bayesa jest klasyfikatorem statystycznym opartym o twierdzenie Bayesa [51, s. 50]. Na jego podstawie dokument tekstowy *t* klasyfikowany jest do kategorii *kl* na podstawie największej wartości prawdopodobieństwa *P(kl|t)*, które określa wzór *(18)* [83, s. 24].

$$P(kl|t) = \frac{P(kl)*P(t|kl)}{P(t)} \tag{18}$$

gdzie:

$P(t)$ – prawdopodobieństwo wystąpienia dokumentu tekstowego *t*,

$P(kl)$ – prawdopodobieństwo wystąpienia kategorii *kl*,

$P(t|kl)$ – prawdopodobieństwo przynależności tekstu *t* do kategorii *kl*.

Ze względu na założenie, że prawdopodobieństwo występowania dokumentu tekstowego *P(t)* jest niezależne od kategorii, zatem jego pominięcie nie zmienia rozkładu prawdopodobieństwa przynależności dokumentu do kategorii. Z powodu tego podejścia nazwa klasyfikatora zawiera określenie „naiwny”. Ostateczne uproszenie problemu przedstawia wzór *(19)* [22, s. 111].

$$P(kl|t) = P(kl) * \prod P(t|kl) \tag{19}$$

gdzie:

$P(kl)$ – obliczane jako stosunek liczby dokumentów należących do klasy *kl* do liczby dokumentów tekstowych z wszystkich klas,

$$P(t|kl) = \frac{1+\sum_{itw=1}^{mtw} \sum r_{itw}}{mr+\sum_{j=1}^{mr} r_j} \tag{20}$$

gdzie:

mr – liczba elementów reprezentacji dokumentu tekstowego,

j – indeks elementu reprezentacji dokumentu tekstowego,

mtw – liczba wzorcowych dokumentów tekstowych należący do kategorii *kl*,

itw – indeks wzorcowych dokumentów tekstowych,

$\sum_{jtw=1}^{mtw} \sum r_{jtw}$- liczba wystąpień *ir*-tego elementu reprezentacji we wzorcowych dokumentach tekstowych należących do kategorii *kl*.

W przypadku klasyfikatora Rocchio dla pojedynczej klasy *kl* budowany jest zbiór *Rk*, zawierający wektory reprezentujące dokumenty tekstowe relewantne do tej kategorii oraz zbiór *Rp* zawierający wektory reprezentujące pozostałe dokumenty tekstowe. Każda klasa *kl* jest odwzorowana za pomocą wektora WR_{kl} określonego wzorem *(21)* [104, s. 433].

$$WR_{kl} = \alpha * \sum_{R \in kl} \frac{R}{|R|} - \beta * \sum_{R \notin kl} \frac{R}{|R|} \qquad (21)$$

gdzie:

WR_{kl} – wektor odwzorowujący kategorię *kl*,

$R \in kl$ – zbiór reprezentacji dokumentów relewantnych w stosunku do klasy *kl*,

$R \notin kl$ – zbiór reprezentacji dokumentów nierelewantnych w stosunku do klasy *kl*,

R – reprezentacja dokumentu tekstowego ze zbioru treningowego,

|R| – wielkość reprezentacji dokumentu tekstowego,

α, β – parametry dobierane eksperymentalnie określające istotność dokumentów relewantnych i nierelewantnych.

Przy przyjęciu $\alpha = 1, \beta = 0$ formuła zostaje zredukowana do postaci wyrażonej wzorem *(22)*.

$$WR_{kl} = \sum_{R \in kl} \frac{R}{|R|} \qquad (22)$$

gdzie:

WR_{kl} – wektor odwzorowujący kategorię *kl*,

$R \in kl$ – zbiór reprezentacji dokumentów relewantnych w stosunku do klasy *kl*,

R – reprezentacja dokumentu tekstowego ze zbioru treningowego,

|R| – wielkość reprezentacji dokumentu tekstowego.

Klasyfikacja ocenianego dokumentu tekstowego jest realizowana poprzez obliczenie jego podobieństwa względem poszczególnych wektorów odwzorowujących kategorie zdefiniowane na zbiorze treningowym dokumentów tekstowych. Ostatecznie dokument tekstowy zostaje

przyporządkowany do kategorii, w stosunku do której uzyskano największy stopień podobieństwa.

Odmienne podejście do klasyfikacji danych tekstowych w grupie metod wyszukiwania dokumentów tekstowych w stosunku do metod opartych na modelu przestrzeni wektorowej VSM (metod rozmytych) cechuje tzw. wyszukiwanie dokładne (ang. exact match) bazujące na algebrze Boole`a. W wyszukiwaniu tym zapytanie traktowane jest jako zdanie logiczne. Wynikiem wyszukiwania są dokumenty tekstowe, dla których zdanie logiczne jest prawdziwe. Zapytania zawierają operatory *and, or, not*. Przykładowe zapytanie przedstawia wzór *(23)*.

$$w_1 \text{ or } (w_2 \text{ and } w_3) \qquad (23)$$

gdzie:

w_1, w_2, w_3– wyszukiwane wyrazy.

Wyszukiwanie dokładne boolowskie obarczone jest wieloma ograniczeniami i wadami. Zalicza się do nich między innymi [3, s. 3]:

- trudność w wyszukiwaniu tekstów bardzo podobnych, ale dla których zdanie logiczne w zapytaniu jest nieprawdziwe,
- przypadkowa kolejność wyszukanych dokumentów tekstowych,
- trudność budowy zapytania logicznego,
- trudność ograniczenia wielkości odpowiedzi.

Pomimo swoich wad i ograniczeń wyszukiwanie dokładne boolowskie jest stosowane z uwagi na małą złożoność obliczeniową oraz prostą implementacje.

2.3. Klasyfikacja danych tekstowych bazująca na wiedzy eksperta

Alternatywną metodą w stosunku do podejścia bazującego na uczeniu maszynowym jest eksploracja wykorzystująca wiedzę eksperta. W tym podejściu wykorzystywane są wzorce informacyjne lub formularze i reguły dopasowania zdefiniowane przez eksperta. W przypadku wyszukiwania lub klasyfikacji, dla każdej kategorii dokumentów tekstowych zostają opracowane przez eksperta dziedzinowego wzorce informacyjne do wykrywania w tekstach tzw. rzeczowych informacji (ang. factual information), które wyrażone są za pomocą sekwencji kilku wyrazów. Wzorce modelują pewne relacje pomiędzy wyrazami, natomiast wydobyte na ich podstawie rzeczowe informacje w lepszym stopniu niż pojedyncze wyrazy odwzorowują znaczenie tekstu [17, s. 1]. W literaturze przedstawiono kilka propozycji formalizacji zapisu wzorców informacyjnych [87, ss. 345–358] [49, ss. 155–

164] przy czym wzorce te budowane są zazwyczaj w oparciu o sformalizowany język zapytań. Przykładową konstrukcję wzorca określa formuła *(24)* [33, s. 126]:

(dolar (&n (amerykański ! singapurski)) *(24)*

Zapytanie wykorzystujące powyższy wzorzec wykrywa w tekście wyraz „dolar", ale z pominięciem frazy „dolar amerykański" oraz „dolar singapurski". Regułę klasyfikującą wykorzystującą strukturę informacyjną wykrytą za pomocą wzorca informacyjnego przedstawia formuła *(25)* [33, s. 236].

(25)

(jeśli
test:
(or [struktury informacyjnej:dolar-australijski]
and [struktury informacyjnej:dolar]
[struktury informacyjnej:australia]
(not struktury informacyjnej:dolar-amerykański])
(not [struktury informacyjnej:dolar-singapurski])))
wtedy: (przydziel do kategorii:dolar-australijski))

W momencie wykrycia w tekście struktury informacyjnej "dolar-australijski" dokładnie dopasowanej do wzorca lub wykrycie struktury „dolar" łącznie ze strukturą „australia", ale bez wykrycia struktur „dolar-amerykański" i „dolar-singapurski" tekst zostanie sklasyfikowany do kategorii „dolar-australijski".

W praktyce wykorzystywane są również bardziej rozwinięte techniki klasyfikacji dokumentów tekstowych w oparciu o formularze składające się z kategorii wyrazów oraz reguł określających stopień dopasowania formularza do danej kategorii. Przykład formularza dopasowania wyrazów do kategorii przedstawia formuła *(26)* [49, s. 144]:

(26)

Napad, rachunek, kradzież => **Zdarzenie**
Nieznany, napastnik, mężczyzna, gangster => **Sprawca**
Pieniądze, gotówka => **Cel**
Bank, konwój, sklep, kasa => **Obiekt**
Szczecin, Żołnierska => **Miejsce**
Broń, uzbrojony, nóż, pistolet => **Narzędzie**
W południe, nad ranem => **Czas**

gdzie:

Zdarzenie, Sprawca, Cel, Obiekt, Miejsce, Narzędzie, Czas są kategoriami wyrazów.

Na podstawie powyższej zadeklarowanych kategorii wyrazów możliwe jest wydobycie z tekstu przykładowych elementów formularza dopasowania, który umieszczono w tabeli 3.

Tabela 3. Elementy wyekstrahowane z przykładowego tekstu za pomocą formularza dopasowania.

Kategoria	**Wyrazy wyekstrahowane z tekstu**
Zdarzenie	Rachunek
Sprawca	Mężczyzna, nieznany
Cel	Gotówka
Obiekt	Bank
Miejsce	Szczecin, Żołnierska
Narzędzie	Broń
Czas	?

Źródło: [49, s. 144]

Reguła dopasowania formularza do odpowiedniej kategorii dokumentów tekstowych może być zgodna z formułą *(27)*:

If** Sprawca **&** Cel **&** Obiekt **Then** dopasowanie **0.7 *(27)*

która oznacza, że w przypadku odnalezienia w tekście dowolnego wyrazu z kategorii „Sprawca" wraz z dowolnym wyrazem z kategorii „Cel" oraz „Obiekt", wskaźnik dopasowania dokumentu tekstowego do określonej kategorii wyniesie 0.7.

Jedną z bardziej rozwiniętych propozycji formalizacji wzorców informacyjnych jest ta, która bazuje na języku OWL (ang. Web Ontology Language), będącym standardem opisywania danych w postaci ontologii, czyli formalnej specyfikacji pewnego obszaru wiedzy w formie łatwej do wykorzystania przez komputery. Język OWL powstał na bazie standardu opisu zasobów sieci internetowej RDF (ang. Resource Description Framework) oraz języka reprezentacji wiedzy RDFS (ang. Resource Description Framework Schema) i jest wspierany przez World Wide Web Consortium. OWL jest wykorzystywany do budowania tzw. sieci semantycznej, czyli sieci internetowej, w której treści są opisywane w sposób umożliwiający ich przetwarzanie przez komputery odpowiednio do ich znaczenia.

Wzorce budowane przy użyciu OWL i udostępniane za pomocą adresów internetowych tworzą bazę wiedzy na wzór bazy opisanej w literaturze [11], która może być wykorzystywana do ekstrakcji z tekstu istotnych dla danego problemu informacji [17]. Definiowanie wzorców za pomocą języka opisu ontologii jest bardzo naturalne ze względu na zależności występujące pomiędzy obiektem, termem i pojęciem, które są wyrażone

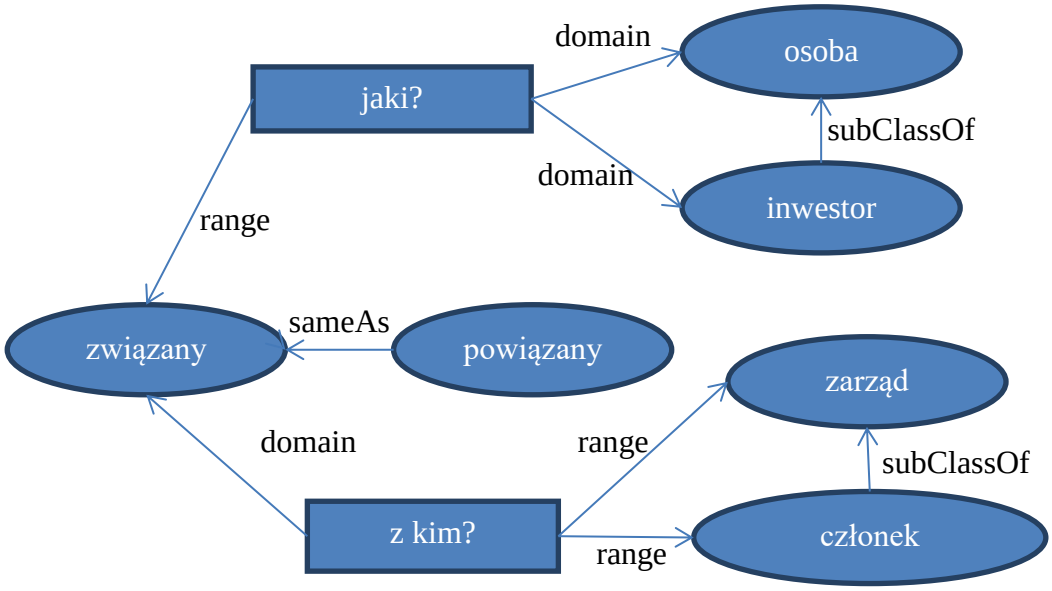
za pomocą trójkąta semiotycznego, opisanego w rozdziale 2.1 niniejszej pracy. Otóż każdemu z termów odpowiada pewne pojęcie, które jest tożsamy z pojęciem (konceptem) wykorzystywanym przy budowaniu ontologii. Również relacje pomiędzy pojęciami uwzględniane we wzorach informacyjnych są odwzorowywane w modelu ontologii. Różnica jaka istnieje pomiędzy wzorcami informacyjnymi, a ontologią jest taka, że każdy wzorzec dotyczy jedynie kilku konceptów, natomiast ontologia zawiera w sobie o wiele większą reprezentację danego obszaru wiedzy dziedzinowej.

Na bazie wybranych znaczników RDF, RDFS i OWL możliwych do wykorzystania przy konstruowaniu ontologii zdefiniowano poniższy przykład wzorca informacyjnego.

```
<rdf:RDF
        xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
        xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
        xmlns:owl="http://www.w3.org/2002/07/owl#"
<owl:Ontology rdf:about="http://localhost/wzorzec1">
<rdfs:comment>Wzorzec informacyjny: inwestor powiązany z zarządem </rdfs:comment>
<owl:versionInfo>1.0/09.05.2015</owl:versionInfo>
</owl:Ontology>
<owl:Class rdf:ID="osoba" />
<owl:Class rdf:ID="inwestor">
        <rdfs:subClassOf rdf:resource="#osoba"/>
</owl:Class>
<owl:Class rdf:ID="zarząd" />
<owl:Class rdf:ID="członek ">
        <rdfs:subClassOf rdf:resource="#zarząd"/>
</owl:Class>
<owl:Class rdf:ID="powiązany" />
<owl:Class rdf:ID="związany">
        <owl:sameAs rdf:resource="#powiązany>
</owl:Class>
<owl:ObjectProperty rdf:ID="jaki">
<rdf:Alt>
        <rdfs:domain rdf:resource="#osoba" />
        <rdfs:domain rdf:resource="#inwestor" />
</rdf:Alt>
        <rdfs:range rdf:resource="#powiązany" />
</owl:ObjectProperty>
```

```
<owl:ObjectProperty rdf:ID="z kim">
        <rdfs:domain rdf:resource="#powiazany" />
<rdf:Alt>
        <rdfs:range rdf:resource="#czlonek" />
        <rdfs:range rdf:resource="#zarząd" />
</rdf:Alt >
</owl:ObjectProperty>
</rdf:RDF>
```

Na podstawie powyższego zapisu istnieje możliwość zwizualizowania modelu wzorca informacyjnego za pomocą grafu ontologii, zaprezentowanego na rysunku 8.



Rysunek 8. Wzorzec informacyjny przedstawiony za pomocą grafu w języku OWL.
Źródło: opracowanie własne

Konstruowanie wzorców przy użycia języka OWL bazuje na strukturach złożonych z trójek:

- podmiot (*ang. subject*) – opisywany element,
- właściwość (*ang. property*) - właściwość elementu,
- obiekt lub orzeczenie (*ang. object*) – wartość właściwości.

W przypadku wzorca przedstawionego na rysunku 8 przykładową trójkę może tworzyć: podmiot – *inwestor*, właściwość – *jaki*, obiekt – *powiązany*. Podmioty oraz obiekty definiowane są za pomocą znaczników klas *owl:Class*. Każdą klasę można określić jako podklasę innej klasy z wykorzystaniem znacznika *rdfs:subClassOf*. Identyczne klasy identyfikuje znacznik *owl:sameAs*. Właściwość deklarowana jest przy użyciu znacznika *owl:ObjectProperty*. W

ramach właściwości określana jest relacja z podmiotem za pomocą znacznika *rdfs:domain* oraz relacja z obiektem z wykorzystaniem znacznika *rdfs:range.* Nazwy klas i właściwości określane są przy użyciu znacznika *rdf:ID*, natomiast odwołanie do danej nazwy realizowane jest za pomocą znacznika *rdf:resource="#...".* Dopełnieniem definicji wzorca jest określenie unikalnego URI, pod którym udostępniany jest wzorzec - *rdf:about=""* oraz komentarz w znaczniku *rdfs:comment* i wersja wzorca opisana za pomocą znacznika *owl:versionInfo.* W powyższym przykładzie zostały wykorzystane jedynie wybrane elementy jezyka OWL, natomiast jego kompletną specyfikację wraz ze szczegółowym opisem wszystkich znaczników w nim wykorzystywanych można znaleźć na stronie konsorcjum W3C [6].

Struktury informacyjne, które mogą zostać zbudowane i wyekstrahowane z tekstu na podstawie modelu wzorca informacyjnego zwizualizowane na rysunku 8 to:

- *inwestor - powiązany - członek,*
- *osoba - powiązany - członek,*
- *inwestor - związany - członek,*
- *osoba - związany - członek,*
- *inwestor - powiązany - zarząd,*
- *osoba - powiązany - zarząd,*
- *inwestor - związany - zarząd,*
- *osoba - związany - zarząd.*

Wszystkie podmioty i obiekty definiowane we wzorcu przyjmują nazwy odpowiadające formom podstawowym wyrazów. Dzięki temu istnieje możliwość łatwego powiązania elementów z modelu wzorca z wyrazami występującymi w dokumentach tekstowych, które zostały poddane lematyzacji. Na podstawie zdefiniowanych wzorców zostaje przeprowadzona ekstrakcja struktur informacyjnych z tekstu zgodnie z poniższym algorytmem:

1. wyszukiwanie w dokumencie tekstowych wyrazów zdefiniowanych we wzorcach jako nazwy podmiotów oraz obiektów,
2. identyfikacja wzorców informacyjnych,
3. generowanie oczekiwań wzorców informacyjnych,
4. weryfikacja oczekiwań.

Ostatecznie na podstawie wyekstrahowanych informacji dokonywana jest klasyfikacja dokumentu tekstowego np. z wykorzystaniem reguł klasyfikujących zgodnych ze wzorem *(27)*.

Klasyfikacja z wykorzystaniem technik bazujących na wiedzy eksperta może charakteryzować się uzyskiwaniem wyników o wysokiej jakości [33, s. 127]. Istotny wpływ na

jakość klasyfikacji wywierają jednak umiejętności konstruowania wzorców przez eksperta dziedzinowego oraz intuicyjne (nieprecyzyjne przy dużej ilości reguł) definiowanie reguł dopasowania. Równie ważnym elementem wpływającym negatywnie na użyteczność tej metody jest długi czas opracowywania pełnego modelu wzorców lub formularzy wraz z regułami dopasowania.

2.4. Porównanie metod eksploracji danych tekstowych

W celu dokonania oceny metod eksploracji danych tekstowych wykorzystywanych do zadania klasyfikacji, zaprezentowanych w rozdziale 2.2 i 2.3, przeprowadzono analizę SWOT (ang. Strengths Weaknesses Opportunities Threats) tych metod, a wyniki tej analizy zawarto w tabeli 4.

W przypadku wykorzystania metod eksploracji danych tekstowych w procesie podejmowania decyzji bardzo istotny staje się dobór odpowiednich cech reprezentujących dokumenty tekstowe. Szczególnie interesującym rozwiązaniem jest opracowanie reprezentacji w oparciu o tzw. rzeczowe informacje (sekwencje kilku wyrazów), które mają istotne znaczenie w kontekście podejmowanych decyzji. Z tego względu niezbędna w tym momencie jest baza wiedzy zdefiniowana przy udziale eksperta dziedzinowego, za pomocą której możliwe jest wydobycie z tekstu informacji istotnych dla rozważanego problemu decyzyjnego. Z drugiej strony biorąc pod uwagę długi czas opracowywania pełnego modelu wzorców lub formularzy i reguł dopasowania, dobrym rozwiązaniem wydaje się możliwość zautomatyzowania poszczególnych etapów eksploracji z wykorzystaniem metod uczenia maszynowego.

Z analizy SWOT zawartej w tabeli 4 wynika, że możliwym kierunkiem poprawy działania metod eksploracji danych tekstowych oraz ich dostosowania do rozważanego problemu decyzyjnego jest ich integracja. Specyfika metody automatycznej eksploracji danych tekstowych bazującej na modelu przestrzeni wektorowej VSM pozwala wykorzystać jako elementy reprezentacji dokumentów tekstowych rzeczowe informacje wyekstrahowane z wykorzystaniem bazy wiedzy np. w postaci wzorców informacyjnych zdefiniowanych przez eksperta dziedzinowego. Z tego względu w procedurze integracji metod eksploracji danych tekstowych i numerycznych opisanej w rozdziale 4 ujęto integrację zarówno metod opartych o automatyczną eksplorację (uczenie maszynowe) jak również metod eksploracji bazujących na wiedzy eksperta.

Tabela 4. Analiza SWOT metod klasyfikacji danych tekstowych.

Mocne strony (zalety)	Słabe strony (wady)

Metody automatycznej eksploracji tekstów (VSM): ▪ krótki czas opracowania modelu, ▪ automatyczna ocena podobieństwa, ▪ możliwość zastosowania funkcji istotności w stosunku do elementów reprezentacji tekstu, co pływa na poprawę jakości wyniku eksploracji. Metody automatycznej eksploracji tekstów (bez VSM): ▪ mała złożoność obliczeniowa, ▪ prosta implementacja. Metoda definiowania przez eksperta wzorców i reguł dopasowania (metody bazujące na wiedzy eksperta): ▪ dokładna analiza znaczeniowa tekstu, ▪ precyzyjne porównanie na poziomie rzeczowych informacji wyekstrahowany za pomocą wzorców informacyjnych, ▪ relatywnie mała ilość rzeczowych informacji i reguł dopasowania w stosunku do reprezentacji tekstu w metodach automatycznych.	Metody automatycznej eksploracji tekstów (VSM): ▪ niedokładna analiza znaczeniowa tekstu, ▪ nieprecyzyjne porównanie dokumentów tekstowych na poziomie pojedynczych wyrazów lub automatycznie wykrytych struktur semantycznych, ▪ duża wielkość reprezentacji i występowanie związanego z nią szumu informacyjnego wpływającego na obniżenie jakości wyniku eksploracji. Metody automatycznej eksploracji tekstów (bez VSM): ▪ ostre wyszukiwanie - nieodnajdywanie tekstów, które prawie pasują do zapytania, przypadkowa kolejność wyszukanych dokumentów tekstowych, trudność ograniczenia wielkości odpowiedzi, ▪ niska skuteczność przy długich tekstach Metoda definiowania przez eksperta wzorców i reguł dopasowania (metody bazujące na wiedzy eksperta): ▪ długi czas opracowania modelu wynikający z ręcznego definiowania reguł dopasowania, ▪ intuicyjne (nieprecyzyjne przy dużej ilości reguł) definiowanie reguł dopasowania, ▪ ze względu na ręczną definicję brak możliwości zastosowania funkcji istotności w stosunku do elementów wzorców.
Szanse	**Zagrożenia**
▪ możliwość integracji różnych rodzajów metod w celu zwiększenia ich użyteczności oraz jakości wyniku eksploracji	▪ uzyskanie wyniku eksploracji o niskiej jakości.

Źródło: opracowanie własne

3. Klasyfikacja danych numerycznych

3.1. Wprowadzenie do klasyfikacji danych numerycznych

Analogicznie jak w przypadku eksploracji danych tekstowych dla wybranego zadania decyzyjnego, w którym wykorzystywana jest eksploracja danych numerycznych, dobierana jest właściwa metoda eksploracji. Najpopularniejsze metody eksploracji danych numerycznych wymienianych w literaturze to [53][38, s. 10][107]:

- zbiory przybliżone,
- logika rozmyta,
- sieci neuronowe,
- metody statystyczne,
- metody ewolucyjne.

Szczególnie skuteczną metodą w zastosowaniach praktycznych klasyfikacji danych numerycznych jest metoda Teoria Zbiorów Przybliżonych (TZP), co potwierdzają liczne eksperymenty opisane w literaturze [46][37][85]. Metoda TZP umożliwia również ocenę jakości wiedzy zdefiniowanej za pomocą reprezentacji danych numerycznych oraz ocenę istotności wybranych danych (atrybutów) numerycznych. Ułatwia to wybór odpowiedniej reprezentacji danych numerycznych i ograniczenie tzw. szumu informacyjnego. Z tego względu w pracy skoncentrowano się na tej właśnie metodzie eksploracji danych numerycznych.

W zależności od przyjętej metody eksploracji danych numerycznych uzyskuje się różne formy reprezentacji wiedzy odkrytej na podstawie analizowanych danych. Jedną z najbardziej popularnych form reprezentacji wiedzy są reguły decyzyjne [81, s. 22]. Reguły stanowią reprezentację najbardziej zbliżoną do sposobu zapisu wiedzy przez człowieka. Dlatego uznawane są za najprostszy do interpretacji język reprezentacji hipotez będących wynikiem działania algorytmu uczącego się i najczęściej stosowane są w tych wszystkich zadaniach eksploracji danych, w których jedną z ważnych cech modelu danych jest jego czytelność. W ogólny sposób reguły decyzyjne (zamiennie reguły klasyfikujące) wyrażone są, zgodnie z literaturą [81, s. 22], wzorem *(34)*:

$$\phi \longrightarrow \psi \qquad (28)$$

gdzie:

ϕ – przesłanka, pewna formuła logiczna,

ψ – konkluzja, działanie podjęte po spełnieniu przesłanki.

Reguły decyzyjne są wyrażeniami logiki matematycznej I rzędu i reprezentują zależności pomiędzy przesłankami, a konkluzjami w celu opisania określonych obiektów. Mogą to być opisy dokładne lub często spotykane w rzeczywistych sytuacjach opisy przybliżone (aproksymacyjne), co wynika z niepełności lub niedokładności w opisie tej rzeczywistości. Jednym z podejść, które rozważa te dwa rodzaje opisów oraz definiuje sposób formalny jest Teoria Zbiorów Przybliżonych (TZP), w której rozważa się wiedzę i jej reprezentację jako elementy wynikające ze zdolności do klasyfikacji. Metoda TZP jest zatem metodą wykorzystywaną do konstruowania algorytmów klasyfikujących, które bazują na regułach decyzyjnych. Proces eksploracji danych numerycznych bazujący na Teorii Zbiorów Przybliżonych można podzielić na trzy główne etapy, przedstawione na rysunku 9.

Rysunek 9. Trzyetapowy proces eksploracji danych numerycznych
Źródło: opracowanie własne

Wstępem do eksploracji danych numerycznych z wykorzystaniem metody TZP jest dyskretyzacja danych oraz przypisanie wartości nominalnych poszczególnym atrybutom opisującym badane przypadki (obiekty). W procesie dyskretyzacji ciągłe dane numeryczne zostają zamienione na atrybuty typu porządkowego o skończonej liczbie wartości. Polega to na podziale oryginalnej dziedziny atrybutu na określoną liczbę przedziałów i przypisaniu tym przedziałom wartości dyskretnych. Ze względu na wykorzystywane techniki podziału atrybutów ciągłych metody dyskretyzacji można podzielić na [68, s. 5]:

- prymitywne i zaawansowane – pierwsze w odróżnieniu od drugich są metodami, które w podziale atrybutu ciągłego nie biorą pod uwagę jego rozkładu oraz rozkładu klas decyzyjnych,
- metody globalne i lokalne – metody globalne biorą pod uwagę wyłącznie wartości atrybutu, który jest poddawany podziałowi, natomiast metody lokalne uwzględniają w podziale zależności wynikające z pozostałych atrybutów,

- metody z nadzorem i bez nadzoru – w przypadku pierwszej grupy metod przy podziale atrybutów brane są pod uwagę klasy decyzyjne, czego się nie uwzględnia w metodach dyskretyzacji bez nadzoru.
- metody wstępujące i zstępujące – metody wstępujące dzielą atrybut na określone przedziały na podstawie wyznaczonych wartości podziału, natomiast w przypadku metod zstępujących końcowy podział atrybutów jest efektem łącznia poszczególnych przedziałów, które na początku wyznaczone są na podstawie wszystkich wartości atrybutów tzn. każda wartość atrybutu stanowi jeden przedział.

Z praktycznego punktu widzenia najpopularniejsze techniki dyskretyzacji danych numerycznych to [68, ss. 5–7]:

- dyskretyzacja naiwna,
- dyskretyzacja według równej szerokości,
- dyskretyzacja według równej częstości,
- dyskretyzacja z wykorzystaniem wiedzy eksperta.

Dyskretyzacja naiwna polega na zastąpieniu rzeczywistych wartości atrybutu kolejnymi wartościami całkowitymi wynikającymi z ilości różnych wartości pierwotnych atrybutu np. wartości *{70,20,30,20}* są zastępowane wartościami *{3,1,2,1}*. Wadami tej metody jest utrata informacji o wartościach pierwotnych oraz brak uwzględnienia specyfiki danych – dane o podobnej wartości mogą uzyskać skrajnie różne wartości całkowite.

Dyskretyzacja według równej szerokości – polega na podziale dziedziny na *h* równych przedziałów. Ważnym elementem tej metody jest stała wartość przedziału w przypadku zwiększania się liczby opisanych obiektów. W metodzie tej dokonywane jest grupowanie pierwotnych wartości atrybutów, dlatego jedna wartość dyskretna odpowiada całej grupie wartości rzeczywistych.

W literaturze przedmiotu wymienia się kilka metod ustalenia liczby *h* przedziałów atrybutów np. za pomocą formuły Struges`a, Friedman-Diaconis`a czy Scott`sa. Jedną z prostszych metod jest formuła Sturges'a [32, s. 1], określona wzorem *(29)*.

$$h = 1 + log_2 mu \tag{29}$$

gdzie

h – liczba przedziałów,

m*u* – liczba opisanych obiektów.

Zgodnie z metodą dyskretyzacji dziedzina atrybutu ciągłego według równej częstości dzielona jest na *h* przedziałów na podstawie ustalonej liczby opisanych obiektów, które zawierają się w tych przedziałach. Dodanie opisu dodatkowego obiektu wiąże się z koniecznością przeprowadzenia ponownego podziału dziedziny atrybutu.

Wymienione metody dyskretyzacji mogą zostać doprecyzowana przez eksperta znającego dziedzinę eksplorowanych danych. Podział dziedziny wartości atrybutu wykonywany jest wówczas przez eksperta, który dzięki wiedzy i doświadczeniu jest w stanie z większą precyzją określić parametry dyskretyzacji (liczbę i szerokość przedziałów). Tak utworzona struktura dziedziny wartości atrybutu dużo lepiej odpowiada prawdziwemu charakterowi danych. Przykład dyskretyzacji atrybutu „*prawdopodobieństwo"* przedstawiono w tabeli 5.

Tabela 5. Dyskretyzacja wartości atrybutu *prawdopodobieństw*o.

Prawdopodobieństwo		
Wartość ciągła	**Wartość lingwistyczna**	**Forma zakodowana**
0 – 0,33	małe	1
0,34 – 0,67	średnie	2
0,68 - 1	duże	3

Źródło: opracowanie własne

Jak można zaobserwować na podstawie danych umieszczonych w tabeli 5 efektem dyskretyzacji jest redukcja rozmiaru danych, która jest niezbędna przy wykorzystaniu Teorii Zbiorów Przybliżonych w procesie podejmowania decyzji. Dyskretyzacja jest również istotnym etapem poprzedzającym eksplorację danych, od której w dużej mierze zależy jej wynik.

Równie ważnym elementem przygotowania reprezentacji danych numerycznych jest wybór wartości nominalnych atrybutów. Dotyczy to przede wszystkim atrybutów, które nie są wyrażone w formie ciągłej, natomiast dziedzina ich wartości może być określona przez eksperta za pomocą wartości nominalnych. W tym przypadku ekspert dziedzinowy musi podjąć decyzję, które wartości nominalne są istotne dla danego atrybutu ze względu na rozważany problem decyzyjny.

W wyniku dyskretyzacji danych numerycznych w formie ciągłej oraz doboru wartości nominalnych dla pozostałych atrybutów, wszystkie wartości atrybutów zostają poddane kodowaniu, czyli przypisaniu do nich odpowiednich wartości numerycznych, co przedstawiono w tabeli 5.

Etap drugi trzyetapowego procesu eksploracji danych numerycznych metodą TZP (Rysunek 9) polega na eksploracji danych numerycznych, których wartości zdefiniowane są za pomocą nowej reprezentacji w celu wygenerowania na ich podstawie reguł decyzyjnych. W etapie tym określa się współczynniki charakteryzujące eksplorację danych numerycznych. Są to współczynniki: jakość i dokładność przybliżenia konceptów decyzyjnych oraz istotność poszczególnych atrybutów.

Etap trzeci z rysunku 9 polega na ostatecznej klasyfikacji danych testowych za pomocą wygenerowanych w poprzednim etapie reguł decyzyjnych.

3.2. Zastosowanie Teorii Zbiorów Przybliżonych do klasyfikacji danych numerycznych

W Teorii Zbiorów Przybliżonych zbiór przypadków (obiektów) określa się nazwą *uniwersum* i oznacza symbolem *U*. Obiekty należące do uniwersum są opisywane za pomocą wybranych atrybutów (cech). Wartości atrybutów charakteryzujące obiekty należące do uniwersum przechowywane są w systemie informacyjnym *SI*, zdefiniowanym wzorami od *(30)* do *(32)* [67, s. 1]:

$$SI = (U, A, V, f) \tag{30}$$

$$V = \bigcup_{a \in A} V_a \tag{31}$$

$$f: U \times A \to V \ \textit{taka, że} \bigvee_{u \in U, a \in A} f(u, a) \in V_a \tag{32}$$

gdzie:

ma – maksymalna liczba atrybutów,

mu – maksymalna liczba obiektów,

U – uniwersum - skończony, niepusty zbiór obiektów $U=\{u_1,u_2 \ldots u_{mu}\}$,

A – skończony, niepusty zbiór atrybutów $A=\{a_1,a_2 \ldots a_{ma}\}$,

V_a – wartość atrybutu *a*,

V – zbiór wartości atrybutów ze zbioru *A*,

f – funkcja informacyjna.

W systemie informacyjnym *SI* przedstawionym w tabeli 6 można wyróżnić klasy abstrakcji (równoważności) zwane również zbiorami elementarnymi. Klasa równoważności dotyczy tych obiektów należących do *uniwersum*, które są nierozróżnialne ze względu na określony podzbiór atrybutów *B*. Klasy powstają z podziału zbioru obiektów ze względu na występowanie relacji równoważności określnej jako [86, s. 25]:

$$IND_{SI}(B) = \{(u_{iu}, u_{ju}) \in U \times U : \bigvee_{a \in B} f(u_{iu}, a) = f(u_{ju}, a)\}, B \subseteq A, 1 < iu, ju < mu, iu \neq ju \quad (33)$$

gdzie:

$IND_{SI}(B)$ – relacja równoważności określona na zbiorze *B*,

B – podzbiór zbioru wszystkich atrybutów *A*,

mu –liczba obiektów z uniwersum,

U – skończony, niepusty zbiór obiektów $u_1, u_2 \ldots u_{mu}$,

A – skończony, niepusty zbiór atrybutów $a_1, a_2 \ldots a_{ma}$,

iu, *ju* – indeksy atrybutów.

Przykład systemu informacyjnego *SI* zawierającego informacje kadrowo-płacowe pięciu pracowników składające się ze zbioru czterech atrybutów *A= {wykształcenie, doświadczenie zawodowe, ocena okresowa, zarobki}* opisanych w tabeli 6.

Tabela 6. Dane do systemu informacyjny zawierający informacje kadrowo-płacowe

Pracownik	**Wykształcenie (*w*)**	**Doświadczenie zawodowe (*d*)**	**Ocena okresowa (*o*)**	**Zarobki (*z*)**
1	wyższe	średnie	dobry	wysokie
2	średnie	duże	bardzo dobry	wysokie
3	średnie	średnie	dostateczny	niskie
4	wyższe	małe	dobry	przeciętne
5	podstawowe	średnie	dobry	niskie

Źródło: Opracowanie własne

można opisać następująco:

$$U = \{1,2,3,4,5\} \quad (34)$$

$$A = \{wykształcenie, doświadczenie\ zawodowe, ocena\ okresowa, zarobki\} \quad (35)$$

$$V = V_{wykształcenie} \cup V_{doświadczenie\ zawodowe} \cup V_{ocena\ okresowa} \cup V_{zarobki} \quad (36)$$

gdzie:

$$V_{\text{wykształcenie}} = \{podstawowe, średnie, wyższe\} \quad (37)$$

$$V_{\text{doświadczenie zawodowe}} = \{małe, średnie, duże\} \quad (38)$$

$$V_{\text{ocena okresowa}} = \{dostateczny, dobry, bardzo\ dobry\} \quad (39)$$

$$V_{\text{zarobki}} = \{niskie, przeciętne, wysokie\} \quad (40)$$

$$f: U \times A \rightarrow V\ np.\ f(1, ocena\ okresowa) = dobry; f(4, wykształcenie) = wyższe \quad (41)$$

Przykładami zbiorów relacji nierozróżnialności dla systemu informacyjnego *SI* z tabeli 6 są:

Dla zbioru atrybutów B_1={*doświadczenie*, *ocena*}:

$$IND_{SI}(B_1) = \{(1,1),(2,2),(3,3),(4,4),(5,5),(1,5),(5,1)\} \quad (42)$$

Dla zbioru atrybutów B_2={*doświadczenie*, *zarobki*}:

$$IND_{SI}(B_2) = \{(1,1),(2,2),(3,3),(4,4),(5,5),(3,5),(5,3)\} \quad (43)$$

Biorąc pod uwagę powyżej określone zbiory relacji nierozróżnialności ($IND_{SI}(B1)$ oraz $IND_{SI}(B2)$), wyróżnia się klasy abstrakcji tych relacji oznaczone jako $U/IND_{SI}(B)$. Dla przykładu z tabeli *6:*

$$U/IND_{SI}(B1) = \{\{1,5\},\{2\},\{3\},\{4\}\} \quad (44)$$

$$U/IND_{SI}(B2) = \{\{1\},\{2\},\{3,5\},\{4\}\} \quad (45)$$

Na podstawie wyznaczonych klas abstrakcji ($U/IND_{SI}(B1)$ oraz $U/IND_{SI}(B2)$) można określić klasę abstrakcji dotyczącą poszczególnych obiektów np. dla zbioru atrybutów B_1:

$$IND_{SI,B1}(1) = \{1,5\} \quad (46)$$

$$IND_{SI,B1}(2) = \{2\} \quad (47)$$

$$IND_{SI,B1}(3) = \{3\} \quad (48)$$

$$IND_{SI,B1}(4) = \{4\} \quad (49)$$

Ze względu na możliwość niejednoznacznego sklasyfikowania obiektu na podstawie wartości wybranych atrybutów do określonej klasy obiektów wyróżnia się określenia:

- zbiory ostre (dokładne, precyzyjne),
- zbiory nieostre (nieścisłe, nieprecyzyjne).

Zbiory ostre dotyczą tych wszystkich obiektów, które na podstawie wydobytej z systemu informacyjnego *SI* wiedzy mogą być jednoznacznie sklasyfikowane do danej klasy obiektów. Z kolei pojęcie zbioru nieostrego związane jest z obiektami, które na podstawie dostępnej w systemie informacyjnym *SI* wiedzy nie mogą być jednoznacznie (z pełną pewnością) sklasyfikowane do określonej klasy obiektów.

W Teorii Zbiorów Przybliżonych do formalizacji pojęcia nieostrego (nieprecyzyjnego) zostały wykorzystana para pojęć precyzyjnych, zwanych dolnym i górnym przybliżeniem. Dolne przybliżenie jest określone wzorem *(50)* [101, s. 4].

$$\underline{B}X = \{u \in U : IND_{SI,B}(u) \subseteq X\} \quad (50)$$

gdzie:

$\underline{B}X$ – przybliżenie dolne,

X – aproksymowany zbiór obiektów,

$IND_{SI,B}(u)$ – klasa abstrakcji obiektu u,

U – uniwersum, skończony, niepusty zbiór obiektów $u_1, u_2 \ldots u_{mu}$,

mu –liczba obiektów z uniwersum.

Do dolnego przybliżenie zbioru X należą wszystkie te obiekty z U, co do przynależności których do określonej klasy obiektów nie ma wątpliwości, biorąc pod uwagę wiedzę wynikającą ze zbioru wartości atrybutów B. Na podstawie dolnego przybliżenia można określić pozytywny obszar (ang. positive area) zbioru X w systemie informacyjnym, zgodny ze wzorem *(51)* [101, s. 5].

$$POS_B(X) = \underline{B}X \tag{51}$$

Górne przybliżenie zdefiniowane jest wzorem *(52)* [101, s. 4]:

$$\overline{B}X = \{u \in U : IND_{SI,B}(u) \cap X \neq \emptyset\} \tag{52}$$

gdzie:

$\overline{B}X$ – przybliżenie górne,

X – aproksymowany zbiór obiektów,

$IND_{SI,B}(u)$ – klasa abstrakcji obiektu u,

U – uniwersum, skończony, niepusty zbiór obiektów $u_1, u_2 \ldots u_{mu}$,

mu –liczba obiektów z uniwersum.

Do górnego przybliżenia należą wszystkie obiekty, które mogą (nie można wykluczyć, że są) być reprezentantami określonej klasy obiektów. Wykorzystując górne przybliżenie można określić negatywny obszar (ang. negative area) zbioru X w systemie informacyjnym SI, za pomocą wzoru *(53)* [101, s. 5]:

$$NEG_B(X) = U - \overline{B}X \tag{53}$$

gdzie:

X – aproksymowany zbiór obiektów,

$NEG_B(X)$ – negatywny obszar zbioru X,

U – uniwersum, skończony, niepusty zbiór obiektów $u_1, u_2 \ldots u_{mu}$,

mu –liczba obiektów z uniwersum,

$\overline{B}X$ – przybliżenie górne.

W celu doprecyzowania elementów występujących przy wykorzystywaniu pojęć dolnego i górnego przybliżenia określono również zbiory obiektów (obszar brzegowy), co do których nie ma pewności czy są czy nie są reprezentantami danego zbioru. Obszar brzegowy określono wzorem *(54)* [101, s. 5]:

$$BN_B(X) = \overline{B}X - \underline{B}X \qquad (54)$$

gdzie:

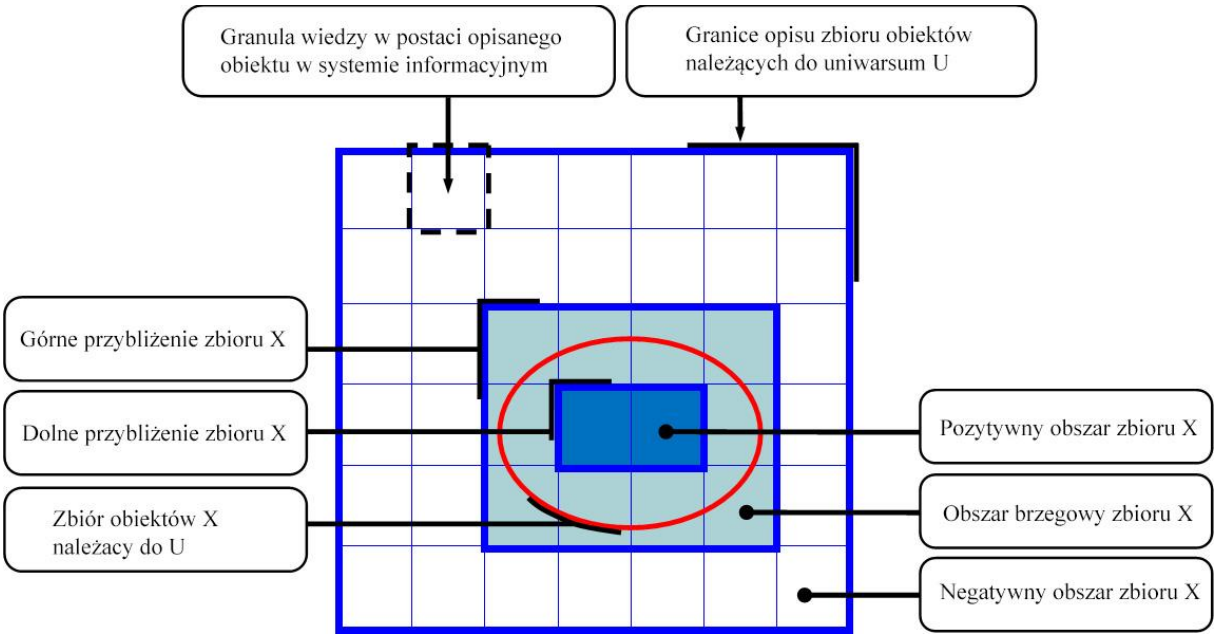
X – aproksymowany zbiór obiektów,

$BN_B(X)$ – obszar brzegowy,

$\underline{B}X$ – przybliżenie dolne,

$\overline{B}X$ – przybliżenie górne.

Formalny opis pojęcia nieostrego w Teorii Zbiorów przybliżonych zwizualizowano na rysunku 10.



Rysunek 10. Wizualizacja dolnego i górnego przybliżenia w Teorii Zbiorów Przybliżonych. *Źródło: Opracowanie własne na podstawie* [52]

Zbiory przybliżone można scharakteryzować za pomocą dwóch miar ilościowych:

- współczynnika jakości przybliżenia,
- współczynnika dokładności przybliżenia (aproksymacji).

Współczynnik jakości przybliżenia konceptów decyzyjnych de facto związany jest z pozytywnym obszarem zbioru X, a jego wartość zawiera się w przedziale $\langle 0,1 \rangle$. Współczynnik

jakości przybliżenia konceptów decyzyjnych zdefiniowany jest za pomocą wzoru *(55)* [16, s. 20].

$$\gamma_B(X) = \frac{card(POS_B(X))}{card(U)} \qquad (55)$$

gdzie:

X – aproksymowany zbiór obiektów,

U – uniwersum, skończony, niepusty zbiór obiektów $u_1, u_2 \ldots u_{mu}$,

mu –liczba obiektów z uniwersum,

$card(POS_B(X))$ – liczba obiektów (obiektów) w pozytywnym obszarze zbioru *X*,

card(U) – liczba obiektów w uniwersum *U*.

Współczynnik jakości przybliżenia konceptów decyzyjnych wyraża w jakim stopniu wiedza wydobyta z opisów przypadków z uniwersum *U* za pomocą atrybutów ze zbioru *B* w sposób jednoznaczny określa klasyfikację obiektów.

Współczynnik dokładności przybliżenia zbioru *X* względem zbioru atrybutów *B* przyjmuje wartości z przedziału $\langle 0,1 \rangle$ i określony jest za pomocą wzoru *(56)* [16, s. 20].

$$\beta_B(X) = \frac{card(POS_B(X))}{card(\overline{B}X)} = \frac{card(\underline{B}X)}{card(\overline{B}X)} \qquad (56)$$

gdzie:

X – aproksymowany zbiór obiektów,

$POS_B(X)$ – pozytywny obszar brzegowy,

$\underline{B}X$- dolne przybliżenie zbioru *X*,

$\overline{B}X$- górne przybliżenie zbioru *X*,

$card(\underline{B}X)$ – liczba przypadków należących do dolnego przybliżenia zbioru *X*,

$card(\overline{B}X)$ – liczba przypadków należących do górnego przybliżenia, znajdujących się w pozytywnym oraz brzegowym obszarze zbioru *X*.

Współczynnik dokładności przybliżenia zbioru *X* określa stopień dokładności klasyfikacji obiektów za pomocą wiedzy wyrażonej przez zbiór atrybutów *B*. Z wartości tego współczynnika jakości wynikają następujące wnioski:

- Jeżeli $\beta_B(X) = 1$ to *X* jest zbiorem dokładnym,
- Jeżeli $\beta_B(X) < 1$ to *X* jest zbiorem przybliżonym.

Szczególną reprezentacją systemu informacyjnego *SI* jest tablica decyzyjna TD sformułowana za pomocą wzoru *(57)* [67, s. 13].

$$TD = (U, Aw, ad, V, f), gdzie\ Aw, ad \in A; Aw \neq \emptyset; Aw \cup ad = A;\ Aw \cap Ad = \emptyset \qquad (57)$$

gdzie:

A – skończony, niepusty zbiór atrybutów,

Aw – zbiór atrybutów warunkowych,

ad – atrybut decyzyjny, którego wartości wyznaczają klasy decyzyjne – zbiór decyzji oznaczony jako *{d}*,

f – funkcja decyzyjna,

U – skończony, niepusty zbiór obiektów,

V – zbiór wartości atrybutów ze zbioru *A*,

∅ – zbiór pusty.

W systemie informacyjnym *SI* reprezentowanym za pomocą tabeli decyzyjnej *TD* występuje podział atrybutów na atrybuty warunkowe *Aw*. W przypadku *SI* z tabeli 6 są to *wykształcenie, doświadczenie zawodowe, ocena okresowa*, oraz zbiór decyzji (klas decyzyjnych) *{d}*, które wyznaczają wartości atrybutu decyzyjnego *ad - zarobki*.

Relację nierozróżnialności względem decyzji *d* na zbiorze *U* określoną za pomocą zbioru atrybutów *Bw*, gdzie $Bw \subseteq Aw$, zdefiniowana jest wzorem *(58)*.

$$IND_{TD}(Bw, d) = \{(u_{iu}, u_{ju}) \in U \times U: (u_{iu}, u_{ju}) \in IND_{SI}(Bw) \vee f(u_{iu}, d) = f(u_{ju}, d)\},$$
$$1 < iu, ju < mu, iu \neq ju \qquad (58)$$

gdzie:

u – obiekt z uniwersum,

U - skończony, niepusty zbiór obiektów $u_1, u_2 \ldots u_{mu}$,

mu – liczba obiektów w uniwersum,

iu, ju – indeksy obiektów z uniwersum,

d – klasa decyzyjna, wartość atrybutu decyzyjnego *ad*,

f() – funkcja określająca wartość atrybutu dla danego obiektu,

Bw – zbiór atrybutów warunkowych.

Istotną różnicą pomiędzy relacją nierozróżnialności zdefiniowanej za pomocą wzoru *(33)*, a relacją nierozróżnialności względem decyzji określoną za pomocą wzoru *(58)* jest to,

że obiekty mające takie same wartości decyzji nie są rozróżniane, również w przypadku kiedy opisy tych obiektów różnią się wartościami atrybutów warunkowych ze zbioru *Bw*.

W przypadku eksploracji danych numerycznych, analogicznie do eksploracji danych tekstowych, pojawia się problem szumu informacyjnego, który obniża nośność informacyjną danych oraz negatywnie wpływa na jakość i dokładność przybliżenia. W tym przypadku szum informacyjny związany jest z:

- opisami obiektów, które zawierają błędy (są przekłamane) wynikające z przyczyn technicznych lub merytorycznych,
- niewłaściwie przeprowadzoną dyskretyzacja atrybutów,
- nadmierną ilością lub niepoprawnie dobranymi atrybutami charakteryzującymi obiekty.

W celu identyfikacji obiektów, których opisy generują szum informacyjny wykorzystuje się metody wyznaczające współczynnik wiarygodności obiektu. Przykładami takich metod, które zostały opisane w literaturze [16, s. 86], są:

- metoda statystyczno-częstotliwościowa,
- metoda oparta na przybliżeniach klas decyzyjnych,
- metoda hybrydowa.

Metoda statystyczno-częstotliwościowa bazuje na analizie ilościowej wartości atrybutów opisujących obiekty. Współczynnik wiarygodności obiektu jest wyższy, gdy istnieje więcej obiektów w tej samej klasie decyzyjnej z identycznymi wartościami opisujących je atrybutów. Formalny zapis współczynnika wiarygodności obiektu *Wo* dla tej metody stanowi wzór *(59)* oraz wzory *(60) i (61)* [16, s. 86].

$$Wo_{TD}(u,d) = \frac{\sum_{a \in Aw} \frac{card(Wo_{u,d,a})}{card(Kw_{u,d})}}{card(Aw)} \tag{59}$$

$$Kw_{u,d} = \{u_{ju} \in U : f(u_{ju}, d) = f(u_{iu}, d)\}, 1 < iu, ju < mu, iu \neq ju \tag{60}$$

$$Wo_{u,d,a} = \{u_{ju} \in U : f(u_{ju}, d) = f(u_{iu}, d) \wedge f(u_{ju}, a) = f(u_{iu}, a)\}, 1 < iu, ju < mu, iu \neq ju \tag{61}$$

gdzie:

Wo – współczynnik wiarygodności obiektu,

u – obiekt z uniwersum,

U – skończony, niepusty zbiór obiektów u_1,u_2…u_{mu},

mu – liczba obiektów w uniwersum,

iu, ju – indeksy obiektów z uniwersum,

a – atrybuty warunkowy, jeśli $a \in Aw$,

d – klasa decyzyjna, wartość atrybutu decyzyjnego ad,

$f()$ – funkcja określająca wartość atrybutu dla danego obiektu,

$Kw_{u,d}$ oraz $Wo_{u,d,a}$ – współczynniki pomocnicze,

Aw – zbiór atrybutów warunkowych.

Zaletą metody statystyczno-częstotliwościowej jest niewątpliwie jej prostota i mała złożoność obliczeniowa.

Metoda bazująca na przybliżeniach klas decyzyjnych oparta jest bezpośrednio na pojęciach zdefiniowanych w Teorii Zbiorów Przybliżonych takich jak: obszar pozytywny, negatywny oraz brzegowy. W metodzie tej wprowadza się pojęcie obiektu niekonfliktowego oraz konfliktowego względem klasy decyzyjnej d wyznaczonej przez wartość atrybutu decyzyjnego, występujących jeżeli zachodzą warunki określone we wzorach *(62)* oraz *(63)* [16, s. 87].

Obiekt jest niekonfliktowy względem klasy decyzyjnej d, jeśli:

$$u \in POS_{Aw}(X) \cup NEG_{Aw}(X), gdzie\ X \coloneqq \{u_{ju} \in U: f(u_{ju}, ad) = d\}, 1 < ju < mu \qquad (62)$$

Obiekt jest konfliktowy względem klasy decyzyjnej, jeśli:

$$u \in BN_{Aw}(X), gdzie\ X \coloneqq \{u_{ju} \in U: f(u_{ju}, ad) = d\}, 1 < ju < mu \qquad (63)$$

gdzie:

X – aproksymowany zbiór obiektów,

$POS_{Aw}(X)$ – pozytywny obszar zbioru X,

$NEG_{Aw}(X)$ – negatywny obszar zbioru X,

$BN_{Aw}(X)$ – obszar brzegowy zbioru X,

u – obiekt z uniwersum U,

ju – indeks obiektu z uniwersum,

mu – liczba obiektów w uniwersum,

ad – atrybut decyzyjny,

d – klasa decyzyjna,

Aw – zbiór atrybutów warunkowych.

W metodzie bazującej na przybliżeniach klasy decyzyjnej wyznacza się współczynnik wiarygodności obiektu w oparciu o to czy obiekt względem każdej klasy decyzyjnej jest niekonfliktowy czy konfliktowy. Formalny zapis współczynnika wiarygodności klasyfikacji w tej metodzie stanowią wzory od *(64)* do *(71)* [16, s. 89].

$$Wo_{TD}(u_{iu}) = \frac{\sum def(u_{iu},d)}{card(\{d\})} \tag{64}$$

$$def(u_{iu},d) = \begin{cases} 1\ dla\ u \in \{u_{iu} \in POS_{Aw}(X) \cup NEG_{Aw}(X)\} \\ \frac{\sum_{a\in A} pom(u_{iu},d,a)}{card(Aw)}\ dla\ u_{iu} \in BN_{Aw}(X) \end{cases} \tag{65}$$

$$pom(u_{iu},d,a) = \begin{cases} \frac{card(Wpos_{u_{iu},a})}{card(Kpos)}\ dla\ f(u_{iu},ad) = d \\ \frac{card(Wneg_{u_{iu},a})}{card(Kneg)}\ dla\ f(u_{iu},ad) \neq d \end{cases} \tag{66}$$

$$Wpos_{u_{iu},a} = \{u_{ju} \in POS_{Aw}(X) \wedge f(u_{ju},a) = f(u_{iu},a)\}, 1 < iu, ju < mu, iu \neq ju \tag{67}$$

$$Wneg_{u_{iu},a} = \{u_{ju} \in NEG_{Aw}(X) \wedge f(u_{ju},a) = f(u_{iu},a)\} \tag{68}$$

$$Kpos = \{u_{iu} \in U : u_{iu} \in POS_{Aw}(X)\} \tag{69}$$

$$Kneg = \{u_{iu} \in U : u_{iu} \in NEG_{Aw}(X)\} \tag{70}$$

$$1 < iu, ju < mu, iu \neq ju \tag{71}$$

gdzie:

Wo- współczynnik wiarygodności obiektu,

X – aproksymowany zbiór obiektów,

$POS_{Aw}(X)$ – pozytywny obszar zbioru *X*,

$NEG_{Aw}(X)$ – negatywny obszar zbioru *X*,

$BN_{Aw}(X)$ – obszar brzegowy zbioru *X*,

u – obiekt z uniwersum *U*,

iu, *ju* – indeksy obiektu z uniwersum,

mu – liczba obiektów w uniwersum,

ad – atrybut decyzyjny,

d – klasa decyzyjna, gdzie $d = v_{ad} \in \mathrm{V_{ad}}$,

Aw – zbiór atrybutów warunkowych.

Jeśli obiekt względem danej klasy jest niekonfliktowy wówczas współczynnik wiarygodności obiektu jest zwiększany o wartość 1. W przeciwnym wypadku, gdy obiekt jest konfliktowy,

wtedy współczynnik wiarygodności obiektu jest zwiększany o wartość współczynnika pomocniczego, który wyznaczany jest następująco:

1. Jeżeli obiekt należy do klasy decyzyjnej *d* to określany jest zbiór obiektów *Kpos* należących do obszaru pozytywnego klasy decyzyjnej *d*. W przeciwnym wypadku, jeżeli obiekt nie należy do klasy decyzyjnej *d* określany jest zbiór obiektów *Kneg* należących do obszaru negatywnego klasy decyzyjnej *d*.

2. Określane są zbiory obiektów *Wpos* i *Wneg*, które mają identyczną wartość atrybutu *a* co obiekt u_{ui}. W kolejnym kroku obliczany jest współczynnik pomocniczy na podstawie zbiorów *Wpos, Kpos, Wneg, Kneg*. Ostatecznie współczynnik pomocniczy jest normalizowany względem ilości wszystkich atrybutów.

3. W ostatnim kroku współczynnik wiarygodności klasyfikacji jest zwiększany o wartość współczynnika pomocniczego oraz normalizowany względem ilości wszystkich obiektów.

Wadą metody opartej na przybliżeniach klas decyzyjnych jest możliwość otrzymywania złych rezultatów klasyfikacji w sytuacji gdy większość obiektów należy do obszaru brzegowego klas decyzyjnych.

Metoda hybrydowa wyznaczenia współczynnika wiarygodności obiektu łączy dwie wcześniej opisane metody (statystyczno-częstotliwościową oraz metodę opartą na przybliżeniach klas decyzyjnych) i eliminuje wadę metody opartej o przybliżenia klas decyzyjnych poprzez obliczanie współczynnika wiarygodności dla obiektów konfliktowych zgodnie z metodą statystyczno-częstotliwościową. Z kolei dla obiektów niekonfliktowych współczynnik wiarygodności obliczany jest tak jak w metodzie opartej o przybliżenia klas decyzyjnych. Formalną definicję współczynnika wiarygodności obiektu w metodzie hybrydowej zdefiniowano za pomocą wzorów od (72) do (76) [16, s. 92].

$$Wo_{TD}(u,d) = \frac{\sum def(u,d)}{card(\{d\})} \tag{72}$$

$$def(u_{iu},d) = \begin{cases} 1 \; dla \; u \in \{u_{iu} \in POS_{Aw}(X) \cup NEG_{Aw}(X)\} \\ \frac{\sum_{a \in A} \frac{card(W_{wu_{iu},d,a})}{card(Kw_{u_{iu}})}}{card(Aw)} \; dla \; u_{iu} \in BN_{Aw}(X) \end{cases} \tag{73}$$

$$Kw_{u_{iu}} = \{u_{ju} \in U : f(u_{ju}, ad) = f(u_{iu}, ad)\}, 1 < iu, ju < mu, iu \neq ju \tag{74}$$

$$Kw_{u_{iu}} = \{u_{ju} \in U : f(u_{ju}, ad) = f(u_{iu}, ad)\}, 1 < iu, ju < mu, iu \neq ju \tag{75}$$

$$1 < iu, ju < mu, iu \neq ju \tag{76}$$

gdzie:

Wo – współczynnik wiarygodności obiektu,

X – aproksymowany zbiór obiektów,

$POS_{Aw}(X)$ – pozytywny obszar zbioru X,

$NEG_{Aw}(X)$ – negatywny obszar zbioru X,

BN_{Aw} – obszar brzegowy zbioru X,

u – obiekt z uniwersum U,

iu, ju – indeksy obiektu z uniwersum,

mu – liczba obiektów w uniwersum,

ad – atrybut decyzyjny,

d – klasa decyzyjna, gdzie $d = v_{ad} \in \mathrm{V_{Ad}}$,

Aw – zbiór atrybutów warunkowych.

W celu wyznaczenia błędnych opisów obiektów w pierwszej kolejności określa się dopuszczalny próg dla współczynnika wiarygodności obiektów. Jedną z możliwości jest tu określenie progu na podstawie wyznaczonego minimalnego współczynnika wiarygodności dla poprawnie zweryfikowanych opisów obiektów. Określony w ten sposób próg jest wykorzystywany do wyznaczenia błędnych opisów nowych obiektów dodawanych do uniwersum. Jeżeli współczynnik wiarygodności nowego obiektu jest niższy od określonego progu, wówczas należy zastosować techniki eliminacji szumu informacyjnego (powstałych niespójności w tablicy decyzyjnej), wśród których stosuje się następujące rozważania:

- opis obiektu o niskim współczynniku wiarygodności zostaje zweryfikowany przez eksperta, który podejmuję decyzję o modyfikacji opisu na zgodny z prawidłową klasyfikacją obiektu. W niektórych sytuacja ekspert niestety nie jest w stanie jednoznacznie wskazać prawidłowej klasyfikacji.
- całkowite usunięcie opisu obiektu z tablicy decyzyjnej.

W związku doborem odpowiedniej ilości atrybutów oraz ich dyskretyzacją i doborem wartości nominalnych, w eksploracji danych numerycznych pojawiają się trzy główne przyczyny generowania szumu informacyjnego, do których należą:

- błędnie dobrana reprezentacja danych - dyskretyzacja lub wybór wartości nominalnych danych numerycznych,

- dobór atrybutów które przenoszą błędne informacje, co powoduje obniżenie jakości i dokładności przybliżenia,
- dobór nadmiernej ilości atrybutów, co zmniejsza czytelność wydobytej wiedzy poprzez nadmierną złożoność reguł decyzyjnych.

Wszystkie wymienione powyżej elementy mogą generować szum informacyjny w tablicy decyzyjnej w postaci niespójności. Poprawność wyboru atrybutów, dyskretyzacji wartości tych atrybutów lub doboru ich wartości nominalnych może być weryfikowana za pomocą współczynnika istotności $\sigma_{\mathrm{Aw}}(\mathrm{a_{ia}})$, obliczanego zgodnie ze wzorem (77) [95, s. 158]:

$$\sigma_{Aw}(a_{ia}) = 1 - \frac{\gamma_{Aw}(X) - \gamma_{Aw-\{a_{ja}\}}(X)}{\gamma_{Aw}(X)}\ , o\ ile\ 1 < ia, ja < card(Aw), ia \neq ja \qquad (77)$$

gdzie:

$\gamma_{\mathrm{Aw}}(\mathrm{X})$ – współczynnik jakości atrybutów ze zbioru Aw,

card(Aw) – liczba atrybutów warunkowych *Aw*,

a – atrybut warunkowy, jeśli $a \in Aw$,

X – aproksymowany zbiór obiektów,

ia, ja – indeksy atrybutów.

Współczynnik istotności określa wpływ atrybutu na klasyfikację, która jest uzyskana dla pełnego zbioru atrybutów względem klasyfikacji z pominięciem atrybutu, dla którego wyznaczana jest istotność. Atrybuty, których istotność jest równa 0 lub jej wartość jest poniżej określonego progu nie są uwzględniane w systemie informacyjnym *SI* podczas klasyfikacji.

Biorąc pod uwagę współczynnik istotności atrybutów można wyznaczyć tzw. redukt aproksymacyjny *RED(Bw)*, czyli zredukowany, minimalny zbiór atrybutów $Bw \subseteq Aw$, przy których zapewniona jest ta sama klasyfikacja jak w przypadku pełnego zbiorów atrybutów warunkowych *Aw*. Zbiór wszystkich niezbędnych do zachowania pierwotnej klasyfikacji atrybutów, które występują we wszystkich reduktach nazwany jest rdzeniem (jądrem) i oznaczony symbolem CORE() jest zgodne ze wzorem *(78)*:

$$CORE(Aw) = \bigcap RED(Aw) \qquad (78)$$

gdzie:

$\cap$ - suma mnogościowa.

Aby uzyskać ostateczny opis klas decyzyjnych w celu wykorzystania go do klasyfikacji obiektów należy wyznaczyć reguły decyzyjne, które składają się z minimalnej ilości atrybutów niezbędnych do zróżnicowania klas decyzyjnych. Reguły decyzyjne w postaci wzoru (28), wyznaczone na podstawie tablicy decyzyjnej (wzór 57) wyrażone są wzorem *(79)*.

$$v_{aw_1} \wedge v_{aw_2} \wedge \ldots \wedge v_{aw_{min}} \rightarrow v_{ad},\ o\ ile\ \{aw_1, \ldots, aw_{maw}\} \in Aw, v_{aw_{iaw}} \in V_{aw_{iaw}}, iaw = 1,2,..,awmin \quad (79)$$

gdzie:

aw – atrybut warunkowy,

v_{aw} – wartość atrybutu warunkowego, ze zbioru V_{aw},

v_{ad} – wartość atrybutu decyzyjnego, ze zbioru V_{aa},

V_{aw} – zbiór wartości atrybutów warunkowych,

maw – liczba atrybutów warunkowych,

V_{ad} – zbiór wartości atrybutu decyzyjnego,

awmin – minimalna ilość atrybutów warunkowych w jądrze,

iaw – indeks atrybutów warunkowych.

Do wyznaczenia reguł złożonych z minimalnej ilości atrybutów można wykorzystać tzw. macierz odróżnialności (lub tzw. macierz rozróżnialności) zdefiniowana za pomocą wzoru *(80)* [67, s. 9]:

$$Mo(SI)_{iu,ju} = \{a \in A: f(u_{iu}, a) \neq f(u_{ju}, a)\}, dla\ iu, ju = 1, \ldots, mu \quad (80)$$

gdzie:

U – uniwersum, skończony, niepusty zbiór obiektów $u_1, u_2 \ldots u_{iu}$,

u – obiekt z uniwersum *U*,

mu – liczba obiektów w uniwersum,

iu, ju – indeks obiektów z uniwersum,

a – atrybut,

A – zbiór atrybutów.

W danej komórce macierzy *Mo* zawierają się te atrybuty, dzięki którym obiekty z uniwersum są rozróżnialne. Przykład macierzy rozróżnialności systemu informacyjnego *SI* z tabeli 6 przedstawiono w tabeli 7.

Tabela 7. Macierz rozróżnialności dla systemu informacyjnego.

Pracownik	**1**	**2**	**3**	**4**	**5**

1					
2	w,d,o				
3	w,o	d,o			
4	d	w,d,o	w,d,o		
5	w	w,d,o	w,o	w,d	

gdzie:

w – atrybut *wykształcenie*.

d – atrybut *doświadczenie zawodowe,*

o – atrybut *ocena okresowa*.

Źródło: opracowanie własne na podstawie [67, s. 9]

Zgodnie ze wzorem *(81)* atrybuty, które występują w macierzy pojedynczo są składowymi jądra [67, s. 12].

$$CORE(A) = \left\{a \subseteq A : mo_{iu,ju} = \{a\}\right\}, dla\ 0 < iu, ju \leq mu \qquad (81)$$

gdzie:

iu, ju – indeks obiektów z uniwersum,

mu – liczba obiektów w uniwersum,

mo – element macierzy rozróżnialności,

Aw – zbiór atrybutów warunkowych,

aw –atrybut warunkowy.

Z kolei redukty są elementami macierzy, które z każdym innym elementem macierzy, który jest niepusty, posiadają wspólną część (poszczególne atrybuty).

Wiedzą zawartą w macierzy rozróżnialności można również przedstawić za pomocą funkcji rozróżnialności zdefiniowanej za pomocą wzoru *(82)* [67, ss. 10–11].

$$f_{SI}(a_1^*, \dots, a_{ma}^*) = \bigcap\left\{\bigcup\left(X_{iu,ju} : 1 \leq iu, ju \leq mu \wedge mo_{iu,ju} \neq \emptyset\right)\right\} \qquad (82)$$

gdzie:

f_{SI} – funkcja rozróżnialności,

ma – maksymalny indeks atrybutu,

*a** – zmienna,

iu, ju – indeks obiektów z uniwersum,

$\emptyset$ – zbiór pusty,

mu – liczba obiektów w uniwersum,

mo – element macierzy rozróżnialności,

X – aproksymowany zbiór obiektów.

Funkcja rozróżnialności jest funkcją boolowską f_{SI} zmiennych $a_1^*, \dots, a_{ma}^*$ odpowiadającym atrybutom $a_1,\dots,a_{ma}$. Przykład funkcji rozróżnialności dla systemu informacyjnego zawartego w tabeli 6 określono zgodnie ze wzorem:

$$f_{SI}(w^*, d^*, o^*) = (w^* \vee d^* \vee o^*) \wedge (w^* \vee o^*) \wedge (d^*) \wedge (w^*) \wedge (d^* \vee o^*) \wedge (w^* \vee d^* \vee o^*) \wedge (w^* \vee d^* \vee o^*) \wedge (w^* \vee d^* \vee o^*) \wedge (w^* \vee o^*) \wedge (w^* \vee d^*) \quad (83)$$

gdzie, zgodnie z tabelą 6:

w* – atrybut *wykształcenie*.

d* – atrybut *doświadczenie zawodowe*,

o *– atrybut *ocena okresowa*.

Stosując prawa rachunku zdań powyższą funkcję można uprościć (zminimalizować) do postaci wzoru *(84)*.

$$f_{TD}(w^*, d^*, o^*) = w^* \wedge d^* \quad (84)$$

Oznacza to, że reguły minimalne powinny składać się z atrybutów *w* (*wykształcenie*) i *o* (*ocena*). Przykład reguły minimalnej wyznaczonej na podstawie tablicy decyzyjnej 6 zdefiniowano za pomocą wzoru *(85)*.

$$jeżeli\ o = "średnie"\ i\ w = "bardzo\ dobry"\ to\ z = "wysokie" \quad (85)$$

Jeżeli identyczne przesłanki (wartości atrybutów warunkowych) w regułach decyzyjnych implikują różne konkluzje (wartości atrybutu decyzyjnego) wówczas reguła określana jest jako niepewna (aproksymacyjna). W przeciwnym wypadu jest to reguła pewna (dokładna). Dla określenia poziomu niepewności reguł obliczany jest współczynnik pewności, zdefiniowany wzorami *(86) i (87)* [18, s. 12]:

$$\phi \rightarrow \psi: cer(\phi, \psi) = \frac{sup_{Regi}(\phi,\psi)}{card(\|\phi\|)} \quad (86)$$

$$\phi \rightarrow \psi: sup_{Regi}(\phi, \psi) = card(\| \phi \wedge \psi \|) \quad (87)$$

gdzie:

cer (ϕ,ψ) – prewność reguły,

ϕ – przesłanka i-tej reguły,

ψ – konkluzja i-tej reguły,

$card(\|\phi\|)$ – liczebność reguł o przesłankach ϕ

Regi – indeks reguły,

$sup_{Regi}(\phi,\psi)$ – wsparcie reguły o indeksie *Regi*.

Przy ocenie jakości eksploracji oprócz współczynników jakości zdefiniowanych w metodzie TZP możliwe jest również badanie wyniku eksploracji (w tym przypadku klasyfikacji) przy wykorzystaniu wielu miar jakości, które definiowane są na podstawie macierzy pomyłek (ang. confusion matrix). Miary jakości określane są wówczas za pomocą współczynników systemu klasyfikacji takich jak: współczynnik prawdziwy pozytywny (ang. True Positive - TP), prawdziwy negatywny (ang. True Negative - TN), fałszywy pozytywny (ang. False Positive - FP), fałszywy negatywny (ang. False Negative - FN). Ogólną specyfikę systemu takiej klasyfikacji przedstawiono w tabeli 8.

Tabela 8. Specyfika system klasyfikacji z wykorzystaniem wielu miar jakości

Kategoria	**Stan rzeczywisty: TAK**	**Stan rzeczywisty: NIE**
Klasyfikacja systemowa: TAK	TP	FP
Klasyfikacja systemowa: NIE	FN	TN

Źródło: [65, ss. 21–22]

gdzie:

TP – współczynnik prawdziwy pozytywny,

FP – współczynnik fałszywy pozytywny,

TN – współczynnik prawdziwy negatywny,

FN – współczynnik fałszywy negatywny.

Do najczęściej wykorzystywanych miar jakości klasyfikacji należy współczynnik całkowitej dokładności (*ACC*) zgodny ze wzorem (88) oraz współczynnik całkowitego poziomu błędu (*ERR*) odpowiednio zgodny ze wzorem (89) [60, s. 157]. Za pomocą tych dwóch współczynników można scharakteryzować jakość decyzji procesie *PD* z uwzględnieniem wszystkich wskazań systemu klasyfikacji jednocześnie tj. z uwzględnieniem TP, FP, TN oraz FN. Z tego względu zarówno miara ACC jak i ERR sprowadzają charakterystykę decyzji do jednej liczby, która w tym przypadku jest opisem całościowym wyniku klasyfikacji. Pozostałe miary jakości klasyfikacji zdefiniowane za pomocą wzorów od *(90)* do *(94)* muszą być sparowane, tzn. przy ocenie decyzji trzeba

jednocześnie wziąć pod uwagę kilka miar np. PPV i NPV, aby charakterystyka klasyfikacji była opisem całościowym [110].

$$ACC = \frac{TP+TN}{TP+TN+FP+FN} \quad (88)$$

$$ERR = \frac{FP+FN}{FP+FN+TP+TN} \quad (89)$$

*Współczynnik ACC (*całkowitej dokładności) określa prawdopodobieństwo prawidłowego wskazania systemu klasyfikacji zarówno dla przypadków pozytywnych jak i negatywnych. Współczynnik całkowitej dokładności powinien być jak najwyższy. Współczynnik ERR (współczynnik całkowitego poziomu błędu) określa prawdopodobieństwo błędnego wskazania systemu klasyfikacji zarówno dla przypadków pozytywnych jak i negatywnych. Współczynnik całkowitego poziomu błędu powinien być jak najniższy.

Pozostałymi miarami jakości charakteryzujące jakość klasyfikacji są:

- współczynnik czułości oznaczony symbolem *SE*, określający zdolność wykrywania przypadków prawdziwych pozytywnych. W przypadku Współczynnik czułości powinien być jak najwyższy. W przypadku analizy danych tekstowych współczynnik czułości określany jest jako zwrot (ang. recall),
- współczynnik specyficzności *SP* określający zdolność wykrywania przypadków negatywnych pozytywnych,
- pozytywny współczynnik predykcji *PPV* - określa prawdopodobieństwo, że przypadek, który system sklasyfikował jako pozytywny jest faktycznie przypadkiem pozytywnym. Pozytywny współczynnik predykcji powinien być jak najwyższy. W przypadku analizy danych tekstowych pozytywny współczynnik predykcji określany jest jako precyzja (ang. precision).
- negatywny współczynnik predykcji *NPV* określa prawdopodobieństwo, że przypadek, który system sklasyfikował jako negatywny jest faktycznie przypadkiem negatywnym. Negatywny współczynnik predykcji powinien być jak najwyższy.

- *F* – współczynnik, który jest ogólnym wskaźnikiem jakości,

zdefiniowane odpowiednio za pomocą wzorów od *(90)* do (*94).*

$$SE = \frac{TP}{TP+FN} \quad (90)$$

$$SP = \frac{TN}{TN+FP} \quad (91)$$

$$PPV = \frac{TP}{TP+FP} \quad (92)$$

$$NPV = \frac{TN}{TN+FN} \quad (93)$$

$$F = \frac{2*PPV*SE}{PPV+SE} \quad (94)$$

W podsumowaniu niniejszego rozdziału pracy należy stwierdzić, że wybór atrybutów, które mają istotny wpływ na definicję wiedzy w postaci reguł decyzyjnych oraz właściwe opracowanie ich reprezentacji (wybór wartości dyskretnych i nominalnych) jest szczególne ważne dla eliminacji tzw. szumu informacyjnego występującego w eksploracji danych. Dzięki zastosowaniu metody TZP możliwe jest odpowiednie dobranie tych elementów między innymi na podstawie współczynnika istotności atrybutów.

Drugi kluczowy etap eksploracji danych numerycznych polega na generowaniu przy wykorzystaniu metody TZP reguł decyzyjnych, na podstawie których podejmowane są decyzje w procesie *PD*. Szczególnego znaczenia posiadają tutaj współczynniki charakteryzujące reguły tj. współczynnik pewności reguł oraz wsparcie, które pozwalają na eliminacje reguł o niskiej pewności.

Ze względu na istotny wpływ na wynik eksploracji danych, dwa powyżej wymienione elementy eksploracji zostały uwzględnione w procedurze integracji metod eksploracji danych tekstowych i numerycznych w procesie podejmowania decyzji, która została opisana w rozdziale 4 niniejszej pracy.

3.3. Badanie istotności i zgodności wyników klasyfikacji

Badanie istotności i zgodności wyników klasyfikacji danych numerycznych ma na celu uwierzytelnienie wyników poprzez weryfikację postawionej hipotezy statystycznej dotyczącej różnic w wynikach poszczególnych prób. Za pomocą wybranego testu statystycznego stwierdza się czy porównywane wyniki różnią się w sposób istotny, czy przypadkowy.

Ze względu na specyfikę wyniku eksploracji danych otrzymanych w badaniach do weryfikacji statystycznej zastosowano test McNemara [78, s. 197]. Jest to test nieparametryczny, który pozwala na badanie prób zależnych z uwzględnieniem zmiennych dychotomicznych tj. zmiennych posiadających tylko dwie kategorie.

Test McNemara można przeprowadzić na podstawie tabeli kontyngencji (tabela 9).

Tabela 9. Tabela kontyngencji dla testu McNemara

		Wyniki drugiej metody eksploracji		
		Kategoria II	Kategoria I	Suma
Wyniki pierwszej metody eksploracji	Kategoria I	*Mc1*	*Mc2*	*Mc1+Mc2*
	Kategoria II	*Mc3*	*Mc4*	*Mc3+Mc4*
	Suma	*Mc1+Mc3*	*Mc2+Mc4*	*Mc1+Mc2+Mc3+Mc4*

Źródło: opracowanie na podstawie [78, s. 197]

gdzie:

Mc1 – liczba przypadków, w których doszło do zmiany wyniku klasyfikacji z kategorii I na II,

Mc2 i *Mc3* – liczba przypadków, u których nie stwierdzono zmiany w klasyfikacji,

Mc4 – liczba przypadków, w których doszło do zmiany wyniku klasyfikacji z kategorii II na I.

Dla testu McNemara formułuje się następujące hipotezy statystyczne zgodne ze wzorami *(95)* i *(96)*.

$$H_0: Mc1 = Mc4 \qquad (95)$$

oraz

$$H_1: Mc1 \neq Mc4 \qquad (96)$$

gdzie:

H_0, H_1 – hipotezy statystyczne,

Mc1 – liczba przypadków, w których doszło do zmiany wyniku klasyfikacji z kategorii I na kategorię II, współczynnik *Mc1* określony jest w tabeli 9,

Mc4 – liczba przypadków, w których doszło do zmiany wyniku klasyfikacji z kategorii II na kategorię I, współczynnik *Mc4* określony jest w tabeli 9.

Hipoteza H_0 oznacza, że w związku z zastosowaniem innej metody eksploracji nie doszło do zmiany wyniku (zmiany klasyfikacji), czyli współczynnik *Mc1* i *Mc4* z tabeli 9 są identyczne, natomiast hipoteza H_1, że wynik eksploracji (klasyfikacji) istotnie się zmienił tzn. wartości współczynników *Mc1* i *Mc4* z tabeli 9 różnią się.

W metodzie McNemara wykorzystywana jest statystyka testowa χ^2, o rozkładzie chi-kwadrat z jednym stopniem swobody. Dla określonego poziomu istotności α z tablic tego

rozkładu odczytuje się prawdopodobieństwo p. Wartość p porównywana jest z wartością p_{ob} obliczoną według wzoru *(97)* [78, s. 197].

$$p_{ob} = \frac{(|Mc1 - Mc4| - 1)^2}{Mc1 + Mc4} \qquad (97)$$

gdzie:

Mc1 – liczba przypadków, w których doszło do zmiany wyniku klasyfikacji z kategorii I na II, współczynnik *Mc1* określony jest w tabeli 9,

Mc4 – liczba przypadków, w których doszło do zmiany wyniku klasyfikacji z kategorii II na I, współczynnik *Mc4* określony jest w tabeli 9,

Jeżeli wartości p oraz p_{ob} spełniają warunek *(98)*

$$p \leq p_{ob} \qquad (98)$$

to należy odrzucić hipotezę H_0 określoną wzorem *(95)*. Wówczas stawia się wniosek, że prawdziwa jest hipoteza H_1 ze wzoru *(96)*.

Gdy

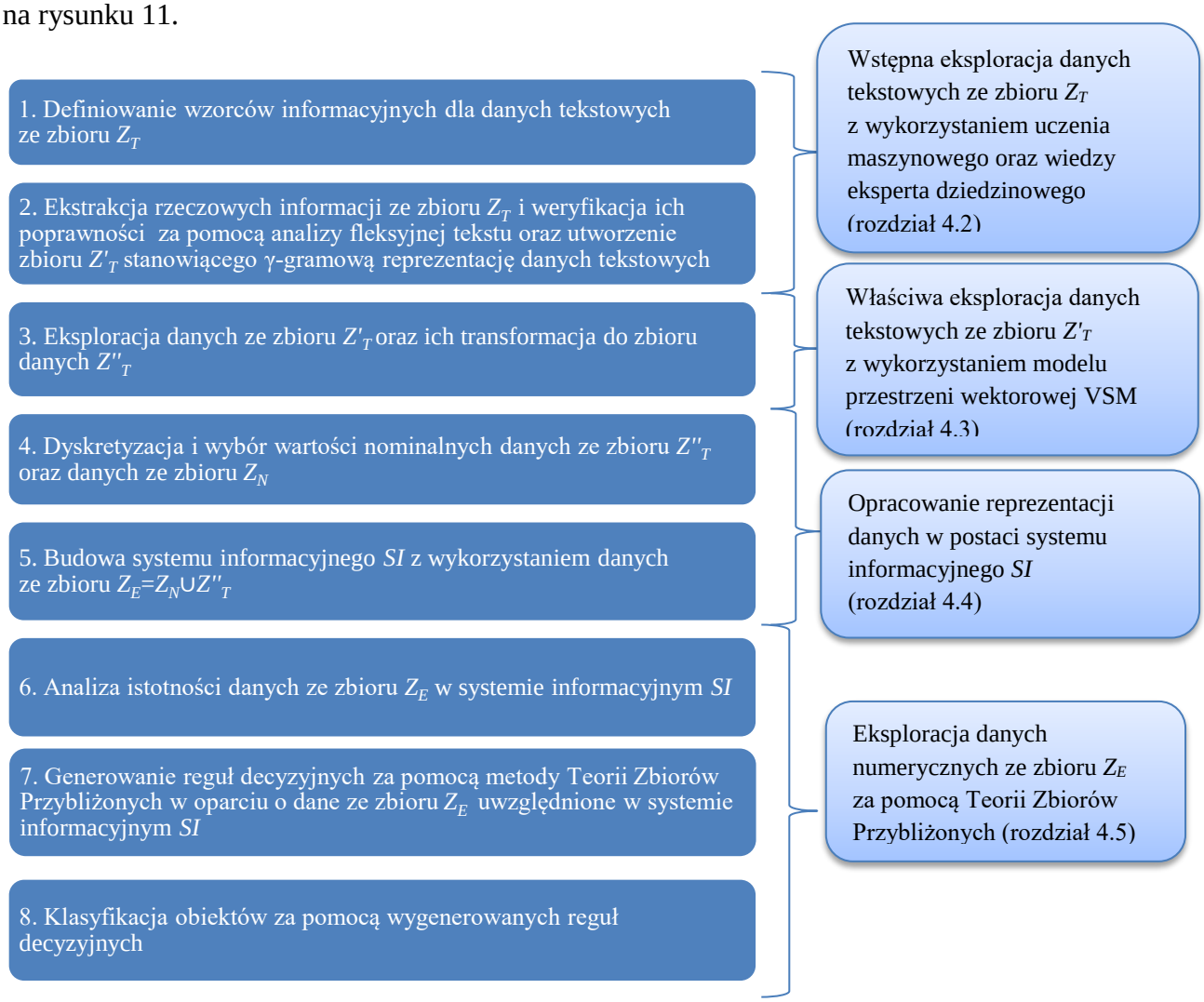
$$p \geq p_{ob} \qquad (99)$$

to nie ma podstaw do odrzucenia hipotezy H_0.

4. Procedura integracji metod klasyfikacji danych tekstowych i numerycznych w procesie podejmowania decyzji

4.1. Ogólny schemat procedury

Procedura integracji metod klasyfikacji danych tekstowych i numerycznych w procesie podejmowania decyzji składa się z 8 podstawowych etapów, które zaprezentowano na rysunku 11.



Rysunek 11. Etapy procedury integracji metod eksploracji danych tekstowych i numerycznych w procesie podejmowania decyzji.
Źródło: opracowanie własne

Na wynik eksploracji duży wpływ ma zastosowanie odpowiedniej reprezentacji danych. W przypadku eksploracji danych tekstowych bazującej na modelu przestrzeni wektorowej VSM kluczowe jest ustrukturyzowanie danych polegające na wyborze odpowiedniej ich reprezentacji. Z tego względu etapy 1 i 2 z rysunku 11, realizowane w ramach tzw. wstępnej eksploracji danych tekstowych ze zbioru Z_T, są wykorzystywane

do opracowania właściwej względem rozpatrywanego problemu decyzyjnego reprezentacji danych tekstowych. W przygotowaniu ustrukturyzowanej reprezentacji danych tekstowych w modelu przestrzeni wektorowej VSM przeważnie wykorzystywane są automatyczne metody uczenia maszynowego. Ze względu na to, że w literaturze podkreśla się korzyści wynikające również z zastosowania przy eksploracji danych tekstowych metod bazujących na wiedzy eksperta np. wzorców informacyjnych [33, s. 127], co potwierdzają również wyniki analizy SWOT z rozdziału 2.4, w procedurze zgodnej z rysunkiem 11 wykorzystano reprezentację γ-gramową, która jest opracowywana z wykorzystaniem zarówno metod bazujących na wiedzy eksperta jak i metod uczenia maszynowego. Standardowo elementy reprezentacji γ-gramowej budowane są w oparciu o funkcję oceniającą $\gamma(w_1,\ldots,w_{lw})$, której wartości odpowiadają przydatności danej sekwencji wyrazów $w_1,\ldots,w_{lw}$ w analizie dokumentu tekstowego, natomiast w pracy elementy reprezentacji budowane są z rzeczowych informacji (sekwencji wyrazów o zmiennej długości) wyekstrahowanych z tekstów za pomocą wzorców informacyjnych. W opracowanej procedurze z rysunku 11 wykorzystano zatem metody dotyczące budowy wzorców informacyjnych oraz ekstrakcji na ich podstawie tzw. rzeczowych informacji, które stanowią elementy nowej reprezentacji danych tekstowych w modelu przestrzeni wektorowej VSM [17, s. 1].

Rzeczowe informacje wyekstrahowane za pomocą wzorców informacyjnych stanowią informacje znaczeniowe przenoszone przez dane tekstowe, które są szczególnie istotne w kontekście rozpatrywanego problemu decyzyjnego. W celu uwzględnienia we wstępnej eksploracji danych tekstowych specyfiki języka polskiego do wyboru poprawnych elementów (rzeczowych informacji) reprezentujących dane tekstowe, w etapie 2 procedury z rysunku 11 wykorzystano analizę fleksyjną tekstu [88, ss. 240–242]. W wyniku realizacji tego etapu procedury zostaje ostatecznie opracowana nowa, ustrukturyzowana reprezentacja danych tekstowych (zbiór Z'_T).

Etap 3 procedury z rysunku 11 jest właściwą eksploracją danych tekstowych, w ramach której wykonywana jest transformacja zbioru danych tekstowych Z'_T do zbioru danych numerycznych Z''_T charakteryzujących dane tekstowe. Transformacja taka jest niezbędna do zbudowania w etapie 5 tej procedury systemu informacyjnego *SI* z uwzględnieniem wszystkich eksplorowanych danych w procesie podejmowania decyzji *PD*. W celu przeprowadzenia właściwej eksploracji danych tekstowych ze zbioru Z'_T wykorzystywane są wybrane metody i techniki eksploracji danych tekstowych. Wynik eksploracji (dane ze zbioru Z''_T) poddawane są odpowiedniej dyskretyzacji w 4 etapie procedury z rysunku 11.

Dyskretyzacja danych numerycznych oraz wybór wartości nominalnych, realizowane w oparciu o zbiory danych Z''_T oraz Z_N w etapie 4 procedury z rysunku 11, ma istotny wpływ na przyszłą definicję wiedzy generowanej na podstawie systemu informacyjnego *SI*, w którym uwzględniane są zakodowane wartości nominalne oraz wartości dyskretne danych. Z tego względu ważne jest aby dyskretyzacja była przeprowadzana, zgodnie z przyjętą metodą opisaną w rozdziale 3.1, na rzeczywistych wartościach (dziedzinie) danych pomiarowych, a w szczególności na rzeczywistych i poprawnych wartościach danych stanowiących wynik eksploracji danych tekstowych.

W etapie 5 procedury z rysunku 11 na podstawie wybranych danych numerycznych ze zbioru Z_E budowany jest system informacyjny *SI* w postaci tablicy decyzyjnej *TD*, który w późniejszym etapie procedury jest wykorzystywany do generowania wiedzy w postaci reguł decyzyjnych.

Dane uwzględnione w systemie informacyjnym *SI* w etapie 6 z rysunku 11 poddawane są ocenie istotności. Ze względu na wykorzystanie w eksploracji danych numerycznych metody Teorii Zbiorów Przybliżonych do oceny istotności danych wykorzystano miary jakości i dokładności wynikające z tej metody tj. współczynnik jakości i dokładności rodziny konceptów decyzyjnych. Za pomocą tych współczynników określany jest wpływ reprezentacji danych opracowanej w etapie 4 i 5 procedury z rysunku 11 oraz uwzględnionej w systemie informacyjnych *SI* na definiowanie wiedzy wykorzystywanej w procesie podejmowania decyzji *PD*. Z kolei ocena istotności poszczególnych atrybutów (danych numerycznych) realizowana jest na podstawie współczynnika istotności.

W etapie 7 procedury z rysunku 11 na podstawie systemu informacyjnego *SI*, w którym uwzględniono najistotniejsze dane (atrybuty) ze zbioru Z_E generowany jest zbiór reguł decyzyjnych wykorzystywanych do podejmowania ostatecznych decyzji (klasyfikacji) w etapie 8 procedury.

Poszczególne etapy procedury z rysunku 11 są zależne od swoich poprzedników, a wynik działania każdego z nich ma wpływ na poprawność integracji metod eksploracji danych tekstowych i numerycznych.

W dalszej części pracy zaprezentowano przykładową implementację procedury z rysunku 11 z uwzględnieniem szczegółowego opisu poszczególnych etapów tej procedury.

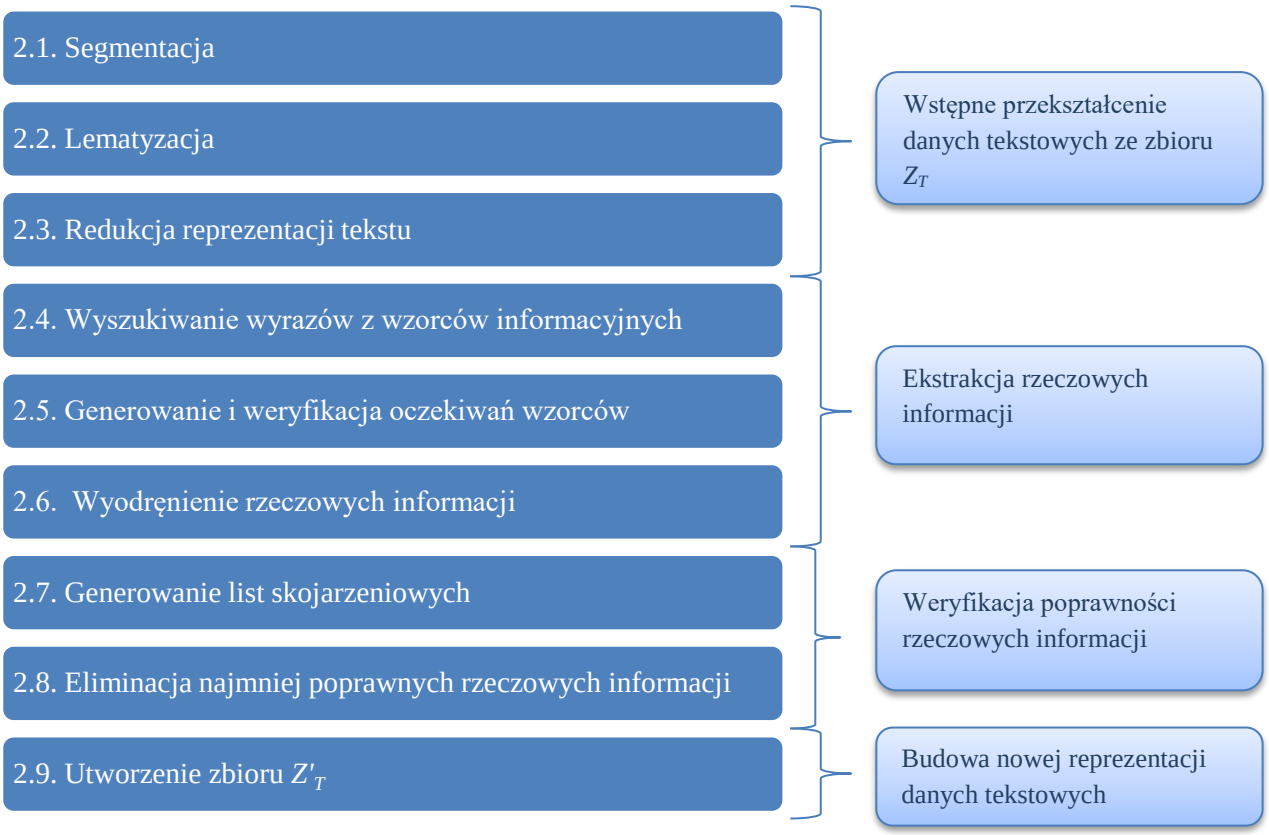
4.2. Szczegółowy opis wstępnej eksploracji danych tekstowych

W pierwszej kolejności dla danych ze zbioru Z_T, uwzględniając kontekst decyzyjny, definiowane są przez eksperta dziedzinowego wzorce informacyjne. Wzorce są ogólnym

modelem najbardziej istotnej informacji semantycznej (rzeczowych informacji) przenoszonej przez tekst względem rozpatrywanego problemu decyzyjnego, która traktowana jest jako układ wybranych wyrazów [49, s. 137]. Formalny zapis wzorców stanowi mechanizm, który automatyzuje i ułatwia definiowanie różnych sekwencji wyrazów, na podstawie których budowana jest reprezentacja tekstu. W literaturze przedstawiono kilka propozycji formalizacji zapisu wzorców informacyjnych [87, ss. 345–358] [49, ss. 155–164], które opisano w rozdziale 2.3. W niniejszej pracy wykorzystano do tego celu języka OWL (ang. Web Ontology Language). Definiowanie wzorców za pomocą języka OWL wraz z udostępnianiem ich za pomocą adresów internetowych ułatwia i przyśpiesza wykorzystanie właściwych wzorców do określonego problemu decyzyjnego. Szczególnie istotne jest szybkie rozbudowywanie poszczególnych wzorców o nowe elementy, co znaczenie usprawnia proces konstruowanie pełnego modelu wiedzy niezbędnego do ekstrakcji wszystkich rzeczowych informacji, które są istotne w kontekście danego problemu decyzyjnego. Definiowanie wzorców informacyjnych w oparciu o język OWL realizowane jest zgodnie opisem zamieszczonym w rozdziale 2.3.

Po zdefiniowaniu wzorców informacyjnych przez eksperta dziedzinowego realizowany jest etap 2 procedury z rysunku 11, który schematycznie został przedstawiony na rysunku 12. W pierwszej kolejności dane tekstowe w oryginalnej postaci dokumentów tekstowych ze zbioru Z_T zostają przekształcone, zgodnie z trzema pierwszymi etapami (etapy 2.1, 2.2, 2.3) z rysunku 12. W niniejszej pracy segmentacja polegająca na przekształceniu tekstów z formy ciągłej na zdania oraz pojedyncze wyrazy wykonywana jest na podstawie zdefiniowanych wyrażeń regularnych. Z połączonych segmentów będących pojedynczymi wyrazami w późniejszym etapie budowane są elementy reprezentacji γ-gramowej. Z kolei segmentacja na całe zdania pozwala na precyzyjne określenie granic połączeń pomiędzy pojedynczymi wyrazami.

W pracy wykorzystano łatwiejszą do realizacji lematyzację bazującą na słowniku fleksyjnym, która polega na wyszukiwaniu kolejnych wyrazów z tekstu w bazie wyrazów Słownika Języka Polskiego - SJP.PL [109].



Rysunek 12. Części składowe etapu 2 procedury z rysunku 11
Źródło: opracowanie własne

Po odszukaniu wyrazu w oryginalnej formie fleksyjnej pobierana jest jego forma podstawowa. Za formę podstawową wyrazów będących odmiennymi częściami mowy w słowniku SJP.PL uznaje się:

1) dla rzeczownika: mianownik liczby pojedynczej lub mianownik liczby mnogiej dla występujących tylko w liczbie mnogiej, a dla rzeczownika odczasownikowego: bezokolicznik,
2) dla czasownika: bezokolicznik,
3) dla przymiotnika: stopień równy przymiotnika w mianowniku,
4) dla liczebnika: mianownik,
5) dla zaimka mianownik odpowiedniej liczby, a dla zaimka zwrotnego tj. dla form siebie, sobie, sobą, się, się – „się”.

W rozbudowanej bazie Słownika Języka Polskiego – SJP.PL uwzględniono wyrazy jedno i wielosegmentowe, które są wyrazami pospolitymi oraz wybranymi nazwami własnymi.

Po przeprowadzeniu lematyzacji budowany jest zbiór wszystkich wyrazów występujących w dokumentach tekstowych przy jednoczesnym zachowaniu ich podstawowych (po lematyzacji) i oryginalnych form fleksyjnych oraz informacji

o granicach zdań. Następnie, w etapie 2.3 z rysunku 12, utworzony zbiór wyrazów redukowany jest wyłącznie do wyrazów występujących we wzorcach. Pozostałe wyrazy zostają uznane za szum informacyjny i zostają pominięte.

W etapie 2.4 z rysunku 12 w utworzonym zbiorze wyszukiwane są wyrazy odpowiadające nazwom klas określonym we wzorcach zdefiniowanych przy użyciu języka OWL. Odbywa się to na podstawie listy wszystkich unikalnych nazw klas występujących we wzorcach. Po zidentyfikowaniu danej nazwy w zbiorze wyrazów reprezentujących dokument tekstowy wyszukiwane są zawierające ją wzorce.

W etapie 2.5 z rysunku 12 dla wzorców generowane są oczekiwania, czyli kolejne elementy, które znajdują się w ich pełnej definicji. Przykładowo dla wcześniej omawianego w rozdziale 2.3 wzorca informacyjnego przedstawionego na rysunek 8, po zidentyfikowaniu w zdaniu wyrazu *inwestor*, generowane będzie oczekiwanie wyrazu *powiązany*, który wynika z relacji z właściwością *jaki*. Z listy wyszukanych wzorców usuwane są te, dla których nie odnaleziono pełnej listy oczekiwań zgodnej z ich całkowitą definicją.

W etapie 2.6 z rysunku 12, na podstawie zweryfikowanej listy wzorców, zostają wyekstrahowane z poszczególnych zdań wszystkie możliwe rzeczowe informacje (sekwencje wyrazów zdefiniowane za pomocą wzorców) z uwzględnieniem podstawowych oraz oryginalnych form fleksyjnych wyrazów.

W tekście napisanym w języku polskim, ze względu na końcówki fleksyjne nadające wyrazom odpowiednią formę gramatyczną kolejność wyrazów w zdaniu może być zmienna, co negatywnie wpływa na poprawność budowy reprezentacji bazujących na sekwencjach kilku wyrazów. W przypadku reprezentacji γ-gramowej, w której elementami reprezentacji są sekwencje wybranych wyrazów tzw. rzeczowe informacje (ang. factual information) [17] dużym wyzwaniem jest ekstrakcja sekwencji wyrazów z uwzględnieniem ich poprawnej odmiany fleksyjnej (właściwych oryginalnych końcówek fleksyjnych wyrazów). Jeśli ekstrakcja sekwencji wyrazów jest przeprowadzana jedynie w oparciu o formy podstawowe wyrazów, wówczas może łączyć ze sobą wyrazy o niespójnej odmianie fleksyjne, które nie powinny budować danej rzeczowej informacji. Na przykład dla dwóch różnych zdań:

1) *Kandydat jest dobry, ale nie potrafi obsługiwać rejestratora Psion.*
2) *Stażysta jest dobry w obsługiwaniu się rejestratorem Psion XT 15.*

Po sprowadzeniu wyrazów do form podstawowych może nastąpić ekstrakcja zdefiniowanej za pomocą wzorca informacyjnego sekwencji wyrazów *dobry – obsługiwać – rejestrator*, co oznacza, że kandydat lub stażysta wykazuje się dobrą obsługą rejestratora. Jednak w przypadku zdania pierwszego jest to nieprawda, ponieważ pierwotna forma wyrazów zawiera

końcówki fleksyjne, które sobie nie odpowiadają tj. *dobry – obsługiwać – rejestrator*, zamiast np. *dobrze – obsługuje – rejestrator*. Poza skojarzeniem wyrazów o poprawnej odmianie istotnym problemem jest wybór wyrazów o prawidłowym znaczeniu, tak aby budowały logiczną strukturę informacyjną (rzeczową informację). Dlatego przy użyciu technik uczenia maszynowego dla wydobytych z tekstu rzeczowych informacji określona zostaje poprawności dopasowania do siebie form fleksyjnych poszczególnych wyrazów wchodzących w ich skład. W tym celu w etapie 2.7 z rysunku 12 zostaje przeprowadzona analiza fleksyjna wyekstrahowanych rzeczowych informacji, która polega na automatycznym generowaniu list skojarzeniowych dla wyekstrahowanych z tekstu oryginalnych form fleksyjnych wyrazów. Podstawą metody generowania list skojarzeniowych jest statystyczno-matematyczne wyliczenie miary skojarzeniowej dla form fleksyjnych elementów z kolejnych trójek (podmiot, orzeczenie/właściwość, obiekt) zawierających się w poszczególnych wzorcach, zgodnie ze wzorem (6), gdzie *cw* w tym przypadku oznacza częstość względna określonej formy fleksyjnej nazwy podmiotu i obiektu występujących razem w zdaniach, natomiast *lw* oznacza częstość bezwzględna uwzględniająca wszystkie formy fleksyjne nazwy obiektu, które występują w zadaniach z określoną formą fleksyjną nazwy podmiotu. Przykładem listy skojarzeniowej dla trójki: podmiot – *inwestor,* orzeczenie – *jaki,* obiekt – *powiązany,* jest lista zawarta w tabeli 10.

Tabela 10. Lista skojarzeniowa form fleksyjnych wyrazów.

Podmiot	Orzeczenie/Właściwość	Obiekt	cw	lw	Miara skojarzeniowa [%]
inwestor	*jaki?*	*związany*	24	27	88
		związana	2		7
		związanego	1		4
inwestora		*związanego*	5	7	71
		związana	1		14
		związany	1		14

Źródło: opracowanie własne

Innymi słowy miara skojarzeniowa uwzględnia liczbę par wyrazów odpowiadającym nazwom podmiotów i obiektów w konkretnych formach fleksyjnych występujących w zdaniach. Jest to analogia do list skojarzeniowych budowanych automatycznie w celu wykrycia najlepiej powiązanych semantycznie ze sobą wyrazów w tekście [48, ss. 119–131]. Jednak

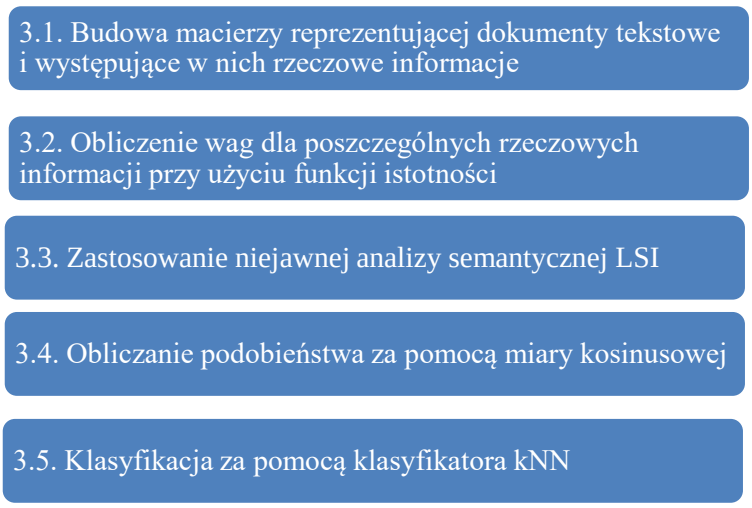
w tym przypadku dla wyrazów wyekstrahowanych z tekstu na podstawie wzorców informacyjnych obliczana jest wartość wynikająca ze skojarzenia ze sobą odpowiednich form fleksyjnych. Obliczona miara skojarzeniowa jest wskazaniem na najbardziej poprawne formy fleksyjne wyrazów dla danej rzeczowej informacji.

W etapie 2.8 z rysunku 12 w celu wyeliminowania najmniej poprawnych form fleksyjnych wyrazów wyekstrahowanych według poszczególnych wzorców informacyjnych eksperymentalnie dopierany jest próg (wartość graniczna miary skojarzeniowej), który decyduje o uwzględnieniu lub odrzuceniu rzeczowej informacji w dalszej części procedury eksploracji. W przypadku listy skojarzeniowej z tabeli 10 oraz progu miary skojarzeniowej wynoszącego 15%, w dalszej analizie zostałyby uwzględnione jedynie struktury: *inwestor związany* oraz *inwestora związanego*.

Rezultatem działania etapu 9 z rysunku 12 jest γ-gramowa reprezentacja danych tekstowych tworząca zbiór Z'_T.

4.3. Szczegółowy opis właściwej eksploracji danych tekstowych

W celu transformacji danych ze zbioru Z'_T do postaci danych numerycznych (stanowiących zbiór Z''_T), które mogą zostać uwzględnione w systemie informacyjnym *SI* wykorzystywanym w dalszych etapach procedury, zgodnie z etapem 3 procedury z rysunku 11, realizowana jest właściwa eksploracja danych tekstowych ze zbioru Z'_T. Dane ze zbioru Z''_T stanowią zatem wynik eksploracji danych tekstowych ze zbioru Z'_T. Schematycznie kolejne podetapy etapu 3 procedury z rysunku 11 zaprezentowano na rysunku 13.



Rysunek 13. Części składowe etapu 3 procedury z rysunku 11
Źródło: opracowanie własne

W celu przeprowadzenia właściwej eksploracji danych tekstowych w pierwszej kolejności (etap 3.1 z rysunku 13) opracowywana jest macierz reprezentująca wyekstrahowane i zweryfikowane za pomocą analizy fleksyjnej w poprzednim etapie 2 procedury z rysunku 11 rzeczowe informacyjne oraz dokumenty tekstowe. Wielkość macierzy określa liczba dokumentów tekstowych oraz liczba wyekstrahowanych z dokumentów tekstowych rzeczowych informacji, które uwzględniają w swojej budowie różne formy fleksyjnej wyrazów. Przykładowa, fragmentaryczna macierz rzeczowych informacji i dokumentów tekstowych, w której elementy macierzy wyrażone są za pomocą wagi binarnej przedstawiono w tabeli 11.

Tabela 11. Fragmentaryczna macierz rzeczowych informacji i tekstów.

Lp.	Rzeczowe informacje	Tekst 1	Tekst2	Tekst 3	Tekst 4	…
1	*inwestor związany z zarządem*	1	0	0	1	…
2	*inwestora związanego z zarządem*	0	0	0	0	…
3	*inwestor związany z osobą*	0	1	1	0	
4	*inwestora związanego z osobą*	1	1	0	0	
…	…	…	…	…		…

Źródło: opracowanie własne

Kolejno dobierane są odpowiednie metody i techniki eksploracji, które umożliwią uzyskanie wyniku o jak najwyższej jakości. Na wynik uzyskany za pomocą określonych metody eksploracji danych tekstowych mają wpływ poszczególne techniki wykorzystywane w ramach tych metod. Jest to szczególnie istotne w przypadku eksploracji danych tekstowych z wykorzystaniem modelu przestrzeni wektorowej VSM, w której istnieje możliwość wykorzystania wielu różnych technik wspomagających eksplorację. Wybór poszczególnych technik może być zrealizowany za pomocą metody eksperymentalnej (na podstawie wyniku eksploracji) lub poprzez wskazanie eksperta wynikające z określonych zależności, wymagań lub ograniczeń np. ograniczeń wydajnościowych jednostki obliczeniowej. Przykładowe techniki, które mogą być użyte w procesie eksploracji danych tekstowych zostały szczegółowo opisane w rozdziale 2.2.

Jedną z podstawowych technik wykorzystywanych w eksploracji danych tekstowych w modelu przestrzeni wektorowej VSM, którą wykorzystano w etapie 3.2 z rysunku 13, jest funkcji istotności, która elementom (termom) opracowanej reprezentacji nadaje odpowiednie wagi. Mogą to być wagi lokalne, globalne lub mieszane, co opisano w rozdziale 2.2 niniejszej pracy. Zadaniem wag jest podkreślenie znaczenia poszczególnych elementów reprezentacji (podkreślenie znaczenia właściwych rzeczowych informacji wyekstrahowanym

z dokumentów tekstowych), które są istotne w kontekście postawionego problemu decyzyjnego.

Kolejną techniką zastosowaną w etapie 3.3 z rysunku 13 jest niejawna analiza semantyczna (ang. Latent Semantic Indexing - LSI). W standardowym rozwiązaniu niejawną indeksację semantyczną stosuje się do wykrycia ukrytych struktur semantycznych istniejących w tekście pomiędzy pojedynczymi wyrazami (elementami reprezentacji). W pracy analogicznie wykorzystano niejawną analizę semantyczną do wykrycia ukrytych struktur semantycznych pomiędzy elementami reprezentacji dokumentu tekstowego. W tym przypadku elementami reprezentacji są jednak rzeczowe informacje wydobyte za pomocą zdefiniowanych przez eksperta wzorców. Przy użyciu metody LSI dokonuje się redukcji wymiaru reprezentacji do określonej liczby wykrytych struktur semantycznych pomiędzy rzeczowymi informacyjnymi. W celu uzyskania określonej liczby wykrytych struktur semantycznych pomiędzy rzeczowymi informacjami, macierz wartości osobliwych oraz macierz rzeczowych informacji zostaje zredukowana do *oz* kolumn, zgodnie ze wzorem *(14)*. Do dobrania wartości *oz* w badaniach testowych przyjęto regułę wyznaczoną przez autora pracy, która została zweryfikowana eksperymentalnie. Jest to taka liczba wartości osobliwych idąc od największej, że ich suma przekracza połowę sumy wszystkich wartości osobliwych. Ostatecznie, ilość obliczona w powyższy sposób jest zwiększana o jedną, kolejną wartość osobliwą. Dzięki zastosowaniu metody LSI teksty, których reprezentacje zawierają rzeczowe informacje o podobnym znaczeniu, ale wyekstrahowane za pomocą odmiennych konstrukcji zdefiniowanych we wzorcach lub zbudowane z odmiennych form fleksyjnych wyrazów, uzyskają wysoką miarę podobieństwa.

W etapie 3.4 z rysunku 13 obliczane jest podobieństwo klasyfikowanych dokumentów tekstowych w stosunku do treningowych (wzorcowych) dokumentów tekstowych reprezentujących poszczególne kategorii za pomocą miary kosinusowej, zgodnej ze wzorem *(15)*.

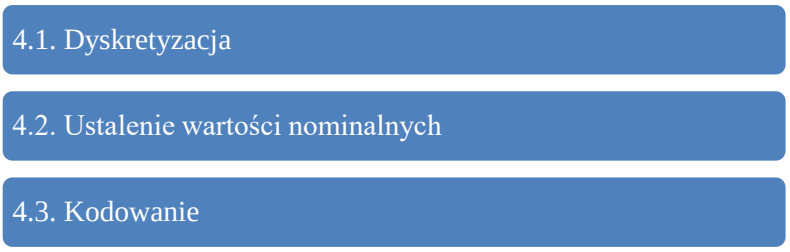
W końcowym etapie 3.5 z rysunku 13 dokonywana jest klasyfikacja danych tekstowych za pomocą klasyfikatora kNN (ang. k-Nearest Neighbor), który został opisany w rozdziale 2.2. Ze względu na integrację metod eksploracji danych tekstowych i numerycznych wynik eksploracji danych tekstowych nie jest jednak przedstawiany w standardowej formie jako ostateczna decyzja przypisania klasyfikowanego dokumentu tekstowego do wybranej kategorii w formie binarnej (0 lub 1), ale jako znormalizowana wartość średniego podobieństwa pomiędzy klasyfikowanym dokumentem, a treningowymi dokumentami tekstowymi reprezentującymi daną kategorię. Dzięki przedstawieniu wyniku

eksploracji tekstowej w formie ciągłych danych numerycznych z przedziału <0,1> możliwe jest lepsze dostosowanie reprezentacji danych numerycznych za pomocą dyskretyzacji do rozważanego problemu decyzyjnego. Standardowo metoda kNN klasyfikuje tekst do kategorii, dla której osiągnięta została wyższa wartość podobieństwa. Również w przypadku gdy różnica pomiędzy wartościami podobieństwa do poszczególnych kategorii jest minimalna. Może to powodować błędną klasyfikację i wpływać niekorzystnie na wynik procesu *PD*. W przypadku dyskretyzacji danych numerycznych w etapie 4 z rysunku 11, będących wynikiem eksploracji danych tekstowych możliwe, jest dobranie odpowiednich przedziałów, które pozwolą na sklasyfikowaniu dokumentu tekstowego do danej kategorii z większą pewnością np. w przypadku osiągnięcia znormalizowanej średniej wartości podobieństwa do dokumentów tekstowych reprezentujących daną kategorię powyżej wartości 0,6.

Ostatecznie, po przeprowadzeniu klasyfikacji danych tekstowych, otrzymywany jest zbiór danych Z''_T stanowiący wynik eksploracji danych tekstowych ze zbioru Z'_T, który podlega dyskretyzacji w kolejnym etapie procedury z rysunku 11.

4.4. Opracowanie reprezentacji danych numerycznych poprzez dyskretyzację i wybór wartości nominalnych danych

Opracowanie reprezentacji przetransformowanych danych tekstowych (dane ze zbioru Z''_T) oraz danych numerycznych ze zbioru Z_N w etapie 4 procedury z rysunku 11 przebiega zgodnie ze schematem z rysunku 14.



Rysunek 14. Części składowe etapu 4 procedury z rysunku 11
Źródło: opracowanie własne

W pierwszej kolejności, w podetapie 4.1 z rysunku 14, dane numeryczne w formie ciągłej zostają poddane dyskretyzacji. W opracowaniu przyjęto metodę dyskretyzacji opartą o podział o równej częstości, czyli podział dziedziny atrybutów (danych numerycznych) na przedziały zawierające równą liczbę przypadków (obiektów), która w niektórych sytuacjach została doprecyzowana przez eksperta. Przykład dyskretyzacji danych numerycznych

odnoszących się do liczby mieszkańców danej miejscowości przedstawiono w tabeli 12.

Tabela 12. Zamiana wartości ciągłych w formę zakodowaną dla atrybutu *liczba mieszkańców*.

Liczba mieszkańców		
Wartość ciągła [mieszkaniec]	**Wartość lingwistyczna**	**Forma zakodowana**
Poniżej 461531	mała	1
461531-1111704	średnia	2
powyżej 1726581	duża	3

Źródło: opracowanie własne

Wynik eksploracji danych tekstowych stanowiący zbiór Z''_T również poddawany jest dyskretyzacji w podetapie 4.1 procedury z rysunku 11, co przedstawiono na przykładzie zawartym w tabeli 13.

Tabela 13. Zamiana wartości ciągłych w formę zakodowaną dla atrybutu *wynik eksploracji danych tekstowych.*

Wynik eksploracji danych tekstowych		
Wartość ciągła [podobieństwo]	**Wartość lingwistyczna**	**Forma zakodowana**
Poniżej 0,6	Klasa I	1
0,6 i powyżej	Klasa II	2

Źródło: opracowanie własne

W podetapie 4.2 z rysunku 14, w przypadku atrybutów, które nie podlegają dyskretyzacji przypisywane są wartości nominalne. Następnie w podetapie 4.3 z rysunku 14 kolejnym wartościom dyskretnym i nominalnym przypisywana jest numeryczna forma kodowa, co przedstawiono w tabeli 14.

Tabela 14. Zamiana wartości lingwistycznych w formę kodową dla atrybutu *decyzyjnego atrakcyjność.*

Atrakcyjność	
Wartość lingwistyczna	**Forma zakodowana**
TAK	1
NIE	2

Źródło: opracowanie własne

Na podstawie nowej reprezentacji danych w postaci zakodowanych wartości nominalnych i dyskretnych (dane ze zbiorów Z''_T oraz Z_N) wszystkich atrybutów poddawanych eksploracji zostaje utworzony zbiór Z_E.

W etapie 5 procedury z rysunku 11, dane ze zbioru Z_E zostają uwzględnione w systemie informacyjnym *SI* w postaci tablicy decyzyjnej *TD*. Tablica decyzyjna *TD* jest opracowywana zgodnie ze wzorem (57) oraz opisem w rozdziale 3.2.

4.5. Klasyfikacja danych w procesie decyzyjnym poprzez eksplorację danych numerycznych

Po etapie opracowania reprezentacji danych numerycznych, na bazie tablicy decyzyjnej *TD* oraz współczynników wynikających z metody Teorii Zbiorów Przybliżonych, obliczana jest istotność danych uwzględnionych w systemie informacyjnym *SI*, która obliczana jest zgodnie z etapami:

1. obliczanie ogólnej istotności danych,
2. obliczanie istotności poszczególnych atrybutów.

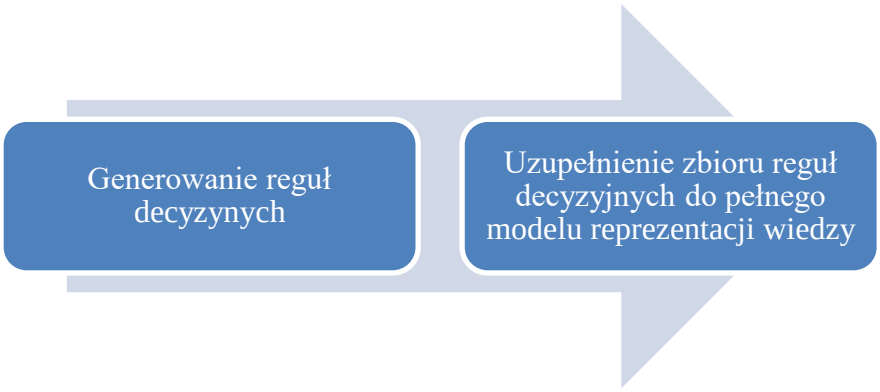
Do obliczenia ogólnej istotności danych ze względu na jakość wiedzy zawierającej się w systemie informacyjnym *SI* wykorzystywane są współczynniki jakości i dokładności rodziny konceptów decyzyjnych, natomiast do obliczenia istotności poszczególnych atrybutów, ze szczególnym uwzględnieniem atrybutu, który wyraża wynik eksploracji danych tekstowych, używany jest współczynnik istotności. Współczynniki te zostały dokładnie opisane w rozdziale 3.2 niniejszej pracy. Dzięki wykorzystaniu współczynnika istotności poszczególnych atrybutów można stwierdzić zasadność zastosowania integracji metod eksploracji danych tekstowych i numerycznych w procesie podejmowania decyzji.
W przypadku błędnie dobranych atrybutów oraz ich reprezentacji, wartość współczynnika istotności atrybutu *wynik eksploracji danych tekstowych* w granicznym przypadku może być równy zeru, wówczas atrybut taki podlega eliminacji, a integracja metod eksploracji danych tekstowych i numerycznych nie przyniesie korzyści w postaci poprawy jakości wyniku w procesie *PD*. W takim przypadku istotność atrybutu reprezentującego wynik eksploracji danych tekstowych może być zwiększona poprzez:

- modyfikację reprezentacji danych (dyskretyzację oraz dobranie wartości nominalnych), które zostały opracowane w etapie 4 z rysunku 11,
- zastosowanie metod eliminacji szumu informacyjnego, które zostały opisane w rozdziale 3.2 niniejszej pracy,

- odpowiedni dobór pozostałych atrybutów reprezentujących dane numeryczne (poza atrybutem reprezentującym wynik eksploracji danych tekstowych).

Dobór pozostałych atrybutów reprezentujących dane numeryczne w celu uzyskania systemie informacyjnym *SI* wiedzy wykorzystywanej do podejmowania decyzji w procesie *PD* o jak najwyższej jakości możliwy jest na podstawie współczynników istotność atrybutów oraz na podstawie miar ogólnej jakości wiedzy wyrażonej za pomocą współczynników jakości i dokładności przybliżenia konceptów decyzyjnych.

Po ostatecznym opracowaniu systemu informacyjnego *SI* przy użyciu metody TZP w etapie 7 procedury z rysunku 11 generowane są reguły decyzyjne. Zdefiniowanie zbioru reguł decyzyjnych pokrywających pełną dziedzinę analizowanych zależności odbywa się zgodnie ze schematem przedstawionym na rysunku 15.



Rysunek 15. Etapy definiowania pełnego zbioru reguł decyzyjnych
Źródło: opracowanie własne

Do automatycznego generowania reguł z danych eksperymentalnych wykorzystano Teorię Zbiorów Przybliżonych. Za pomocą tej metody, w pierwszym etapie przedstawionym na rysunku 15, generowane, a następnie upraszczane są reguły decyzyjne, które wykorzystywane są w dalszej eksploracji danych. Model wnioskowania zbudowany na bazie reguł wyekstrahowanych z danych eksperymentalnych posiada dużą dokładność i wymaga relatywnie niewielkich nakładów pracy w stosunku do reguł pozyskiwanych z wiedzy eksperta. W literaturze opisano również metody integracji tych dwóch źródeł wiedzy (dane eksperymentalne oraz wiedza eksperta), jednak do realizacji założeń pracy ekstrakcję reguł oparto głównie o dane eksperymentalne [77].

W przypadku generowania reguł z ograniczonej ilości danych eksperymentalnych istnieje ryzyko braku pokrycia pełnej dziedziny analizowanych zależności. Dlatego w celu uzupełnienia zbioru reguł decyzyjnych w pełni pokrywających dziedzinę analizowanych

zależności, dla obiektów, dla których nie istnieje właściwa reguła wyekstrahowana z danych eksperymentalnych zostaje utworzona reguła wyznaczona przez eksperta dziedzinowego. W tym przypadku dla wszystkich kombinacji wartości atrybutów (warunkowych i decyzyjnego), które nie są uwzględnione (pokryte) przez reguły decyzyjne wygenerowane z danych treningowych, ekspert dziedzinowy buduje reguły, które klasyfikują obiekt do wybranej kategorii, np. do kategorii, która oznacza decyzję (wariant) negatywną w procesie *PD*.

Wynikiem operacji przeprowadzonych w etapie 7 procedury z rysunku 11 jest wiedza zapisana w postaci reguł decyzyjnych, na podstawie której możliwe jest podejmowanie ostatecznych decyzji w procesie podejmowania decyzji *PD*.

Ostatecznie w etapie 8 procedury z rysunku 11, za pomocą reguł wygenerowanych przy użyciu Teorii Zbiorów Przybliżonych, obiekty zostają poddane klasyfikacji.

5. Badania testowe

5.1. Opis badań testowych

Badania testowe mają na celu potwierdzenie hipotezy badawczej postawionej w pracy. W związku z tym w ramach badań testowych zostały porównane cztery warianty eksploracji, zgodne z rysunkiem 2, tj.:

Wariant A. Zintegrowana eksploracja danych tekstowych i numerycznych,

Wariant B. Eksploracja wyłącznie danych numerycznych,

Wariant C. Eksploracja wyłącznie danych tekstowych.

Wariant D. Zintegrowany wynik eksploracji z wariantów B i C.

W wariantach A, C oraz D eksploracji, w których użyto eksploracji danych tekstowych wykorzystano trzy różne reprezentacje danych tekstowych, tj.:

- Reprezentację unigramową - uwzględniającą pojedyncze wyrazy,
- Reprezentację n-gramową (bigramową) – uwzględniającą sekwencje dwóch wyrazów,
- Reprezentację γ-gramową - uwzględniającą sekwencje wyrazów o zmiennej długości, opracowaną za pomocą wzorców informacyjnych definiowanych przez eksperta dziedzinowego oraz metod analizy fleksyjnej tekstu.

W badaniach wykorzystano trzy różne przykładowe procesy podejmowania decyzji (trzy przypadki użycia), w których możliwe było zastosowanie zintegrowanej eksploracji danych tekstowych i numerycznych, tj.:

- przykład I. proces decyzyjny dotyczący wyboru rentownych zamówień publicznych spośród zbioru takich zamówień,
- przykład II. proces decyzyjny dotyczący sposobu inwestowania na Giełdzie Papierów Wartościowych,
- przykład III. proces decyzyjny dotyczący wyszukiwania atrakcyjnych ofert pracy.

Zintegrowana eksploracja danych tekstowych i numerycznych (Wariant A) została przeprowadzona w 8 etapach, zgodnych ze schematem procedury integracji przedstawionym na rysunku 11. W ramach integracji metod eksploracji danych tekstowych i numerycznych zostały dobrane przez eksperta następujące techniki wykorzystywane w eksploracji danych tekstowych:

- funkcja istotności nadająca wagi termom – dla przykładu I została wybrana funkcja nadająca wagę binarną, dla przykładu II wagę tfidf, natomiast, a dla przykładu III wagę idf, zdefiniowane wzorami *(11), (12), (13).*

- wykorzystanie niejawnej indeksacji semantycznej LSI do wykrywania ukrytych struktur semantycznych pomiędzy elementami reprezentacji dokumentu tekstowego, która została opisana w rozdziale 2.2,
- podobieństwo dokumentów tekstowych obliczane na podstawie miary kosinusowej obliczanej zgodnie ze wzorem *(15)*.
- Jako klasyfikator dokumentów tekstowych wykorzystano klasyfikator kNN, opisany w rozdziale 2.2 niniejszej pracy.

W etapie 7 z rysunku 11 uzupełnienie zbioru reguł decyzyjnych w celu w pełnego pokrycia dziedziny analizowanych zależności, dla obiektów, dla których nie istnieje właściwa reguła wyekstrahowana z danych eksperymentalnych została utworzona reguła wyznaczona przez eksperta.

Do zbadania istotności i wiarygodności wyników różnych wariantów eksploracji z rysunku 2, wykorzystano test statystyczny MyNemara.

Po określeniu czy wyniki dla porównywanych wariantów eksploracji z rysunku 2 różnią się w sposób istotny statystycznie została przeprowadzana weryfikacja hipotezy badawczej polegająca na ocenie nośności informacyjnej danych. Do oceny nośności informacyjnej danych wykorzystano współczynnik całkowitej dokładności – ACC oraz współczynnik całkowitego poziomu błędu – ERR, zdefiniowane za pomocą wzorów *(88)* i *(89)*. Są to miary określające jakość klasyfikacji.

W celu realizacji badań testowych zostało opracowane autorskie oprogramowanie realizujące poszczególne etapy procedury. Oprogramowanie bazuje na serwerze Apache (serwer HTTP) w połączeniu z interpreterem języka skryptowego PHP oraz bazą danych MySQL [99, ss. 1–15]. Do przeprowadzenia zasobochłonnych obliczeń wykonywanych na dużych macierzach w oprogramowaniu zastosowano protokół sieciowy oparty o XML - PHP / Java Bridge [10]. Jest to realizacja transmisji strumieniowej pomiędzy PHP, a maszyną wirtualną Java [44]. Rozwiązanie to wymaga mniej zasobów po stronie serwera WWW i umożliwia w łatwy sposób wywoływanie metod Java z poziomu PHP. Do przeprowadzania rozkładu macierzy według wartości osobliwych (ang. Singular Values Decomposition) w oprogramowaniu zaadoptowano bibliotekę JAMA [29] (obliczenia numeryczne w algebrze liniowej). Wykorzystywana w badaniach baza słownika fleksyjnego języka polskiego „Słownik Języka Polskiego – SJP.PL" została przekonwertowana z formy pliku tekstowego dostępnego pod adresem: http://sjp.pl/slownik/growy/ do postaci rekordów w tabeli bazy danych.

5.2. Przykład I: Wyszukiwanie rentownych zamówień publicznych

Pierwszy przypadek procesu *PD* wykorzystany w badaniach testowych dotyczy analizy ogłoszeń opublikowanych w Biuletynie Zamówień Publicznych (BZP) [106]. Każdego dnia w BZP publikowanych jest kilkaset nowych ogłoszeń o zamówieniu. Miesięcznie jest to nawet kilkanaście tysięcy ogłoszeń. Każdemu zamówieniu przydzielane są odpowiednie kody CPV [108], które teoretycznie powinny ułatwić wykonawcom właściwą identyfikację przedmiotu zamówienia. W praktyce jednak zamówieniom często nadawane są jedynie kody CPV najwyższego rzędu, przez co nieuwzględnione jest bardziej szczegółowe kodowanie. W konsekwencji utrudnia to sprawne wyszukiwanie właściwych zamówień publicznych przez wykonawców. Z drugiej strony, przy tak dużej ilości zamówień, dokładna analiza opisów przedmiotu zamówienia oraz pozostałych kryteriów wpływających na decyzję o przystąpieniu do procedury udzielania zamówienia publicznego jest bardzo czasochłonna. Z tego względu alternatywnym rozwiązaniem jest zastosowanie opisywanej w pracy procedury integracji metod eksploracji danych tekstowych i numerycznych do rozwiązania przykładowego problemu decyzyjnego.

Zadanie decyzyjne polega wytypowaniu zamówień publicznych, w których opis przedmiotu zamówienia jest tożsamy z zakresem prac wykonywanych przez podmiot. W rozważanym przykładzie proce *PD* jest realizowany przez firmę, która wykonuje usługi w zakresie mechanicznego koszenia traw, głównie koszenia w pasach drogowych przy czym usługi takie firma wykonuje za pomocą kosiarek bijakowych lub rotacyjnych doczepianych do ciągników. W procesie wyboru zamówień dostępnych w BZP dużą trudnością jest odróżnienie opisów przedmiotu zamówienia dotyczących ręcznego koszenia traw od opisów dotyczących mechanicznego koszenia trawy wyłącznie z ręcznym obkaszaniem słupków, barier i pozostałych tego typu elementów. Dodatkową trudnością jest wybór takich zamówień z BZP, które według wybranych kryteriów wskazują na rentowność.

Badanie testowo przeprowadzano na zbiorze 200 testowych przypadków zamówień publicznych z wykorzystaniem 11 przypadków treningowych dla kategorii rentownych.

Po przeprowadzeniu klasyfikacji danych testowych, w pierwszej kolejności, za pomocą testu McNemara dokonano sprawdzenia istotności i zgodności otrzymanych wynikach dla różnych wariantów eksploracji danych z rysunku 2. Fragmentaryczne wyniki sprawdzenia zaprezentowano w tabelach od 15 do 17.

Tabela 15. Weryfikacja statystyczna za pomocą testu McNemara wyniku eksploracji dla wariantu C z wykorzystaniem reprezentacji unigramowej oraz wariantu A z wykorzystaniem reprezentacji unigramowej

Próba 1	Próba 2	Poziom zwrotu	p-value
Wariant C – reprezentacja unigramowa	Wariant A – reprezentacja unigramowa	1	<1E−4
		2	<1E−4
		3	<1E−4
		4	<1E−4
		5	<1E−4
		6	<1E−4
		7	<1E−4
		8	<1E−4
		9	<1E−4
		10	<1E−4
		11	<1E−4

Źródło: opracowanie własne

Tabela 16. Weryfikacja statystyczna za pomocą testu McNemara wyniku eksploracji dla Wariantu C z wykorzystaniem reprezentacji n-gramowej oraz Wariantu A z wykorzystaniem reprezentacji n-gramowej

Próba 1	Próba 2	Poziom zwrotu	p-value
Wariant C – reprezentacja n-gramowa	Wariant A – reprezentacja n-gramowa	1	<1E−4
		2	<1E−4
		3	<1E−4
		4	<1E−4
		5	<1E−4
		6	<1E−4
		7	<1E−4
		8	<1E−4
		9	<1E−4
		10	<1E−4
		11	<1E−4

Źródło: opracowanie własne

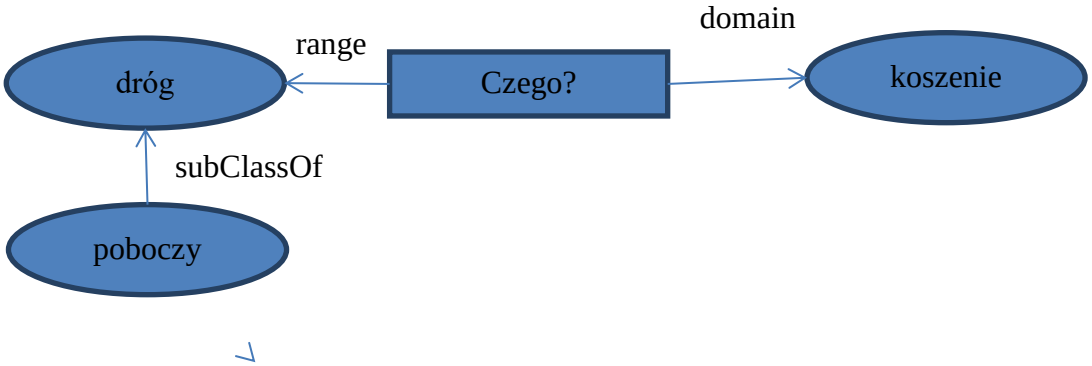
Tabela 17. Weryfikacja statystyczna za pomocą testu McNemara wyniku eksploracji dla Wariantu C z wykorzystaniem reprezentacji γ-gramowej oraz Wariantu A z wykorzystaniem reprezentacji γ -gramowej

Próba 1	Próba 2	Poziom zwrotu	p-value
Wariant C – reprezentacja γ -gramowa	Wariant A – reprezentacja γ - gramowa	1	<1E−4
		2	<1E−4
		3	<1E−4
		4	<1E−4
		5	<1E−4
		6	<1E−4
		7	<1E−4
		8	<1E−4
		9	<1E−4
		10	<1E−4
		11	<1E−4

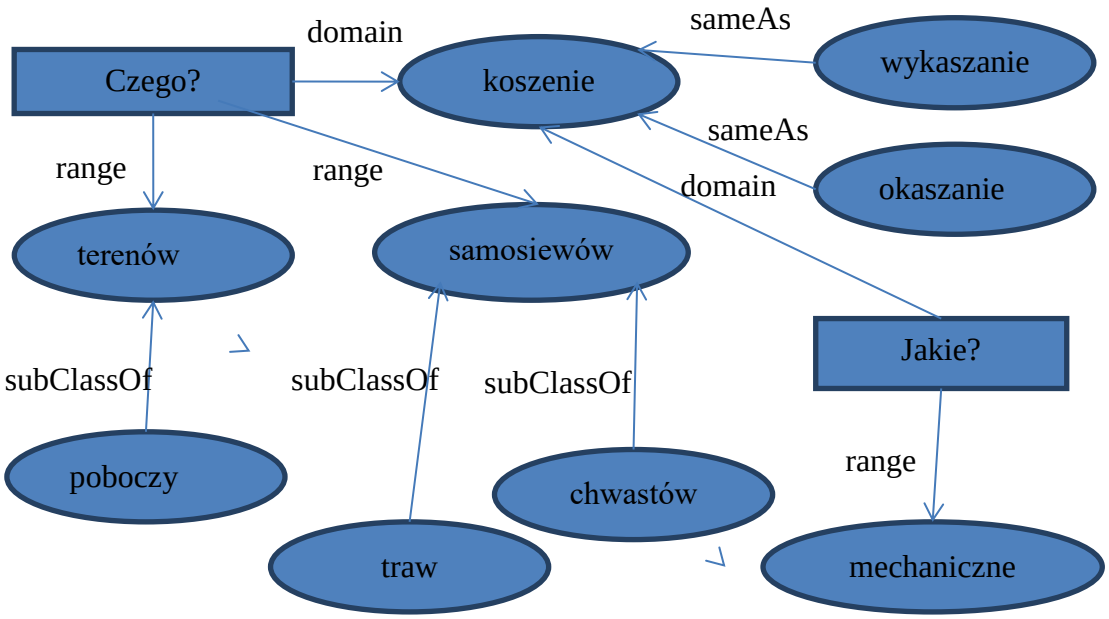
Źródło: opracowanie własne

Dla określonego poziomu istotności $\alpha=0,05$ przeważająca większości wyników badań za pomocą testu McNemara posiada wartość *p-value* poniżej wartości α, przez co można potwierdzić, że różnice pomiędzy wynikami porównywanych wariantów są istotne (biorąc pod uwagę przyjęty poziom istotności). Ze względu na wykazaną istotność statystyczną w większości porównywanych wyników różnych wariantów eksploracji danych, w dalszej części badań testowych, na podstawie wyników klasyfikacji, zostały obliczone wybrane miary jakości decyzji (ACC, ERR).

W pierwszej kolejności został zrealizowany wariant A z rysunku 2 eksploracji (autorska metoda z rzozdziału 4) bazującej na opracowanej procedurze integracji metod eksploracji zarówno danych tekstowych jak i danych numerycznych. W tym celu na wstępie została opracowana reprezentacja danych tekstowych Z_T. W badaniach uwzględniono reprezentację unigramową tekstu oraz reprezentację bigramową (n-gramową). Reprezentację unigramową opracowano w oparciu o pojedyncze wyrazy, natomiast bigramową bazując na parach występujących po sobie wyrazów w dokumencie tekstowym. Reprezentacje te zostały opracowane zgodnie z opisem treścią rozdziału 2.2. Do opracowania reprezentacji γ-gramowej zastosowano zdefiniowane przez eksperta dziedzinowego trzy zwizualizowane za pomocą grafów ontologii wzorce informacyjne, zgodne z rysunkami 16, 17 i 18.



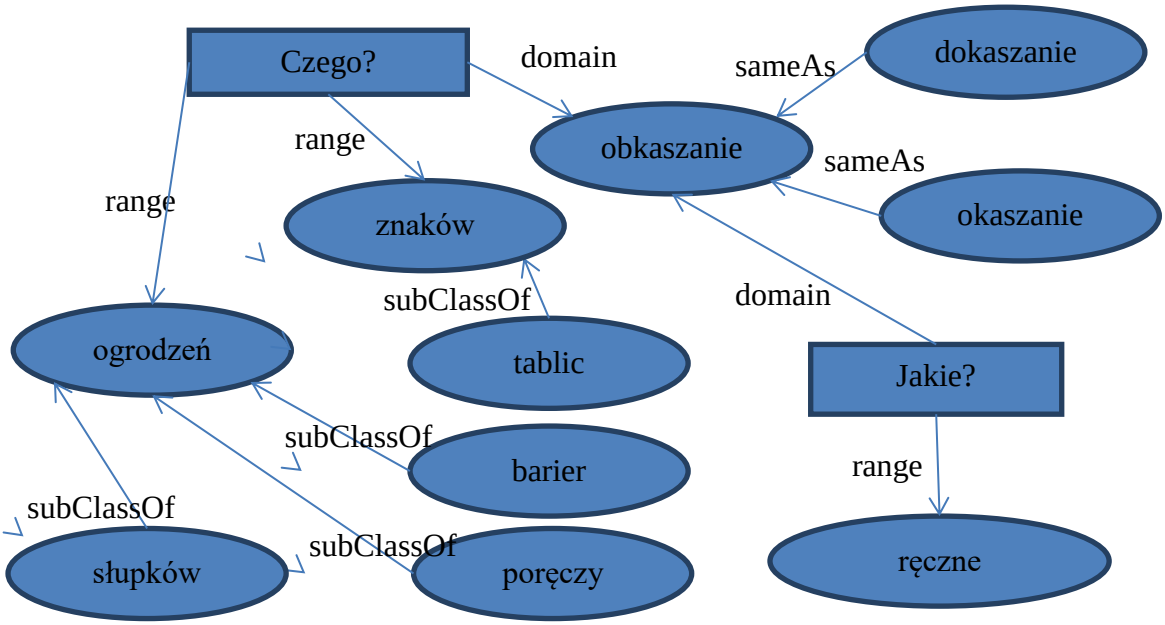
Rysunek 16. Wzorzec informacyjny numer 1 przykładowego procesu *PD* w języku OWL
Źródło: opracowanie własne



Rysunek 17. Wzorzec informacyjny numer 2 przykładowego procesu *PD* w języku OWL
Źródło: opracowanie własne

Przykładowe rzeczowe informacje wyekstrahowane na podstawie trzech zdefiniowanych wzorców (rysunki 16, 17 i 18) to:

- koszenie poboczy,
- koszenie dróg,
- mechaniczne koszenie trawy,
- mechaniczne wykaszanie samosiewów,
- ręczne obkaszanie znaków,
- ręczne okaszanie barier.



Rysunek 18. Wzorzec informacyjny numer 3 przykładowego procesu *PD* w języku OWL
Źródło: opracowanie własne

W drugim etapie autorskiej procedury integracji (rysunek 11 – rozdział 4) za pomocą trzech wzorców łącznie z tekstów wyekstrahowano 168 rzeczowych informacji z 200 zamówień z BZP zawierających różne formy fleksyjne wyrazów. Następnie rzeczowe informacje zgodnie z rysunkami 11 i 12 poddano analizie fleksyjnej. Na podstawie obliczonych miar skojarzeniowych wyeliminowano rzeczowe informacje zawierające najmniej poprawne powiązania form fleksyjne wyrazów. W tym celu eksperymentalnie został dobrany próg, czyli wartość graniczną miary skojarzeniowej, która umożliwiła uwzględnienie lub odrzucenie rzeczowych informacji w dalszej części procedury eksploracji. Przykładowy fragment listy skojarzeniowej dla pierwszego wzorca informacyjnego z rysunku 16 przedstawiono w tabeli 18.

Tabela 18. Fragment listy skojarzeniowej dla trójki: podmiot – *kosić*, właściwość – *czego*, obiekt – *droga*.

Lp.	Podmiot	Obiekt	cw	lw	sk [%]
1	*koszenie*	*dróg*	121	174	0,695
2		*drogi*	23		0,132
3		*(przy) drogach*	19		0,109
4		*droga*	9		0,051
5		*drogami*	2		0,011

gdzie:

sk – wyrażona w procentach miara skojarzenia podmiotu z obiektem,

cw – częstość względna wyrazu definiującego, tj. ilość współwystąpień podmiotu w zdaniach z obiektem,

lw – częstość bezwzględna podmiotu, czyli ilość wystąpień podmiotu w korpusie tekstów, który posłużył do wygenerowania listy skojarzeniowej.

Źródło: opracowanie własne

Po przeprowadzeniu eliminacji najmniej poprawnych powiązań form fleksyjnych wyrazów pozostało 91 rzeczowych informacji będących elementami γ-gramowej reprezentacji dokumentów tekstowych w modelu przestrzeni wektorowej, stanowiącej zbiór $Z'_{T.}$

W etapie 3 procedury z rysunku 11 dokonano transformacji danych ze zbioru Z'_T do Z''_T za pomocą właściwej eksploracji danych tekstowych w modelu przestrzeni wektorowej VSM. W celu realizacji tego etapu przez eksperta dziedzinowego dobrane zostały odpowiednie techniki eksploracji danych tekstowych. Są to techniki opisane w rozdziale 5.1. W efekcie działania tego etapu został utworzony zbiór danych Z''_T, który w kolejnym etapie procedury integracji został poddany dyskretyzacji.

W ramach etapu czwartego procedury integracji (Rysunek 11 – *Etap 4*) dane ze zbiorów Z_N oraz Z''_T poddano dyskretyzacji według równej ilości obiektów w przedziałach oraz doborowi wartości nominalnych, co zaprezentowano w tabelach 19, 20 oraz 21.

Tabela 19. Zamiana wartości ciągłych w dyskretne wartości lingwistyczne oraz formę kodową dla atrybutu warunkowego – *odległość od siedziby firmy.*

Odległość od siedziby firmy		
Wartość ciągła [km]	**Wartość lingwistyczna**	**Forma zakodowana**
Poniżej 362	mała	1
362-562	średnia	2
powyżej 562	duża	3

Źródło: opracowanie własne

Tabela 20. Zamiana wartości ciągłych w dyskretne wartości lingwistyczne oraz formę kodową dla atrybutu warunkowego – *obszar do koszenia.*

Obszar do koszenia		
Wartość ciągła [ha]	**Wartość lingwistyczna**	**Forma zakodowana**
Poniżej 67750	mały	1
67750- 455000	średni	2
powyżej 455000	duży	3

Źródło: opracowanie własne

Tabela 21. Zamiana wartości lingwistycznych w formę kodową dla atrybutu decyzyjnego warunkowego – *rentowność zamówienia*.

Rentowność zamówienia	
Wartość lingwistyczna	**Forma zakodowana**
TAK	1
NIE	2

Źródło: opracowanie własne

W przypadku wyniku eksploracji danych tekstowych przedziały zostały zdefiniowane przez eksperta (tabela 22).

Tabela 22. Zamiana ciągłych w dyskretne wartości lingwistyczne oraz formę kodową dla atrybutu warunkowego – *wynik eksploracji danych tekstowych.*

Wynik eksploracji danych tekstowych		
Wartość ciągła [podobieństwo]	**Wartość lingwistyczna**	**Forma zakodowana**
Poniżej 0,5	Klasa I	1
0,5 i powyżej	Klasa II	2

Źródło: opracowanie własne

Ostatecznie w etapie 5 procedury z rysunku 11 opracowano reprezentację danych numerycznych (zbiór Z_E) w postaci systemu informacyjnego – tablicy decyzyjnej, której fragment zawiera się w tabeli 23.

Tabela 23. Fragmentaryczna reprezentacja danych numerycznych w postaci tablicy decyzyjnej

Lp.	**Atrybuty warunkowe**			**Atrybut decyzyjny**
	Obszar do koszenia (q_1)	**Odległość od siedziby firmy (q_2)**	**Wynik eksploracji danych tekstowych (q_3)**	**Rentowność zamówienia**
1	1	1	1	1
2	2	1	1	1
3	2	2	1	1
4	3	1	1	1
5	3	2	1	1
6	3	3	1	1
…	…	…	…	…

Źródło: opracowanie własne

W etapie 6 procedury integracji z rysunku 11 przeprowadzono analizę istotności danych ze zbioru Z_E przy użyciu współczynników wynikających z metody Teorii Zbiorów Przybliżonych. W systemie informacyjnym *SI* uwzględniono 22-elementowy zbiór przypadków (obiektów). Bazując na wartościach atrybutów poszczególnych przypadków otrzymano tabele

informacyjną w formie zakodowanej (tabela 24), w której wyodrębniono elementarne zbiory warunkowe E_i oraz zbiory decyzyjne (koncepty) X_i.

Tabela 24. Tabela informacyjna z przekształconymi wartościami atrybutów do formy zakodowanej.

Lp.	**Atrybuty warunkowe**			**Zbiory elementarne**	**Atrybut decyzyjny**	**Koncept**
	Obszar do koszenia (q_1)	**Odległość od siedziby firmy (q_2)**	**Wynik eksploracji danych tekstowych (q_3)**	***Eie***	**Rentowność zamówienia**	***Xix***
1	1	1	1	E_1	1	X_1
2	2	1	1	E_2	1	X_1
3	2	2	1	E_3	1	X_1
4	3	1	1	E_4	1	X_1
5	3	2	1	E_5	1	X_1
6	3	3	1	E_6	1	X_1
7	1	1	1	E_1	1	X_1
8	1	2	1	E_7	1	X_1
9	3	1	1	E_4	1	X_1
10	3	2	1	E_5	1	X_1
11	2	2	1	E_3	1	X_1
12	1	2	1	E_7	2	X_2
13	1	2	2	E_8	2	X_2
14	1	3	2	E_9	2	X_2
15	2	1	2	E_{10}	2	X_2
16	2	3	1	E_{11}	2	X_2
17	2	3	1	E_{11}	2	X_2
18	2	2	1	E_3	2	X_2
19	1	3	2	E_9	2	X_2
20	1	2	2	E_8	2	X_2
21	1	2	1	E_7	2	X_2
22	1	1	1	E_1	2	X_2

Źródło: opracowanie własne

gdzie:

q_1, q_2, q_3 – atrybuty warunkowe,
E_{ie} – elementarne zbiory warunkowe odpowiadające klasom abstrakcji,
ie – indeks zbioru warunkowego,
X – koncept decyzyjny,
ix – indeks konceptu decyzyjnego.

Poszczególne przypadki zostały przyporządkowane do odpowiednich zbiorów elementarnych E_{ie} zaprezentowanych w *(100)*.

(100)

E_1={1, 7, 22}, E_2={2}, E_3={3, 11, 18}, E_4={4, 9}, E_5={5, 10}, E_6={6}, E_7={8, 12, 21}, E_8={13, 20}, E_9={14, 19}, E_{10}={15}, E_{11}={16, 17},

Kolejno przeprowadzono obliczenia, w celu wyznaczenia dolnych przybliżeń konceptów decyzyjnych $\underline{B}X_{ix}$ oraz liczby przypadków $card(\underline{B}X_{ix})$ należących do dolnych przybliżeń, wzory *(101)* oraz *(102).*

(101)

X_1 = {1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11}

$\underline{B}X_1 = E_2 \cup E_4 \cup E_5 \cup E_6$ = {2, 4, 5, 6, 9, 10}

$card(\underline{B}X_1)$ = 6

(102)

X_2 = {12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22}

$\underline{B}X_2 = E_8 \cup E_9 \cup E_{10} \cup E_{11}$ = {13, 14, 15, 16, 17, 19, 20}

$card(\underline{B}X_2)$ = 7

Na podstawie liczby przypadków należących do dolnych przybliżeń wyznaczono pozytywny obszar rodziny konceptów decyzyjnych $POS_B(X)$, zgodny ze wzorem (103).

(103)

$POS_B(X) = E_2 \cup E_4 \cup E_5 \cup E_6 \cup E_8 \cup E_9 \cup E_{10} \cup E_{11}$ = {2, 4, 5, 6, 9, 10, 13, 14, 15, 16, 17, 19, 20}

Ilość wszystkich przypadków w pozytywnym obszarze card($POS_B(X)$) jest równa 13. Współczynnik jakości przybliżenia rodziny konceptów decyzyjnych $\gamma_B(X)$ jest zatem równy wartości 0,59 co oznacza, że 59% przypadków wykorzystanych w obliczeniach pozwala na generowanie reguł pewnych. Następnie przeprowadzono obliczenia, które zostały wykorzystane do określenia górnych przybliżeń konceptów $\overline{B}X_{ix}$ oraz liczby przypadków $card(\overline{B}X_{ix})$ należące do górnych przybliżeń, wzory *(104)* i *(105).*

(104)

X_1 = {1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11}

$\overline{B}X_1 = E_1 \cup E_2 \cup E_3 \cup E_4 \cup E_5 \cup E_6 \cup E_7$ = {1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 18, 21, 22}

$card(\overline{B}X_1)$ = 15

(105)

X_2 = {12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22}

$\overline{B}X_2$ = $E_1 \cup E_3 \cup E_7 \cup E_8 \cup E_9 \cup E_{10} \cup E_{11}$ = {1, 3, 7, 8, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22}

$card(\overline{B}X_2)$ = 16

Na podstawie obliczeń otrzymano dokładności przybliżeń kolejnych konceptów decyzyjnych równe $\mu(X_1)$ = 0,4, $\mu(X_2)$ = 0,44. Dokładność przybliżenia rodziny konceptów decyzyjnych $\beta_B(X)$, czyli przeciętny stopień zrozumienia decyzji wyniósł 0,42.

W dalszej części badań testowych wyliczono istotność kolejnych atrybutów warunkowych q_1,q_2,q_3, zgodnie ze wzorem *(77)*. Wyniki obliczeń przedstawiono w tabeli 25.

Tabela 25. Jakość i dokładność przybliżenia konceptów decyzyjnych X oraz istotności usuwanego atrybutu warunkowego.

Atrybut warunkowy	**Istotność atrybutu**	**Dokładność przybliżenia rodziny konceptów $\beta_B(X)$**	**Jakość przybliżenia rodziny konceptów $\gamma_B(X)$**
q_1	0,61	0,13	0,23
q_2	0,24	0,29	0,45
q_3	0,31	0,26	0,41

Źródło: opracowanie własne

Na podstawie wyznaczonych istotności poszczególnych atrybutów warunkowych stwierdzono, że każdy z atrybutów ma znaczący wpływ na definicję wiedzy wynikającą z systemu informacyjnego *SI*. Istotność szczególnie ważnego ze względu na integrację atrybutu q_3 – *wynik eksploracji danych tekstowych*, wynosi 0,31. Oznacza to, że znacząco wpływa on na wiedzę w postaci zbioru reguł wygenerowanych na podstawie danych reprezentowanych przez system informacyjny *SI* i potwierdza zasadność uwzględnienie w systemie *SI* atrybutu wynikającego z eksploracji danych tekstowych.

W wyniku realizacji etapu 7 procedury z rysunku 11 został zbudowany algorytm decyzyjny, który opiera się na wyodrębnionych regułach decyzyjnych. Dla reguły pewnych oraz niepewnych obliczono wsparcie i współczynnik pewności, które przedstawiono tabelach 26 i 27.

Formalny zapis zbioru reguł pewnych określono wzorem *(106)*.

(106)

Reg$_1$: IF ((q_1 = 2) AND (q_2 = 1) AND (q_3 = 1)) THEN d = 1

Reg$_2$: IF ((q_1 = 3) AND (q_2 = 1) AND (q_3 = 1)) THEN d = 1

Reg$_3$: IF ((q_1 = 3) AND (q_2 = 2) AND (q_3 = 1)) THEN d = 1

Reg$_4$: IF ((q_1 = 3) AND (q_2 = 3) AND (q_3 = 1)) THEN d = 1

Reg_5: IF ((q_1 = 1) AND (q_2 = 2) AND (q_3 = 2)) THEN d = 2

Reg_6: IF ((q_1 = 1) AND (q_2 = 3) AND (q_3 = 2)) THEN d = 2

Reg_7: IF ((q_1 = 2) AND (q_2 = 1) AND (q_3 = 2)) THEN d = 2

Reg_8: IF ((q_1 = 2) AND (q_2 = 3) AND (q_3 = 1)) THEN d = 2

Tabela 26. Reguły pewne algorytmu decyzyjnego w formie tabeli informacyjnej.

Reguła (Reg_{ir})	**Atrybuty warunkowe**			**Atrybut decyzyjny**	**Pewność**	**Wsparcie**
	Obszar do koszenia (q_1)	**Odległość od siedziby firmy (q_2)**	**Wynik eksploracji danych tekstowych (q_3)**	**Rentowność zamówienia**		
Reg_1	2	1	1	1	1	1
Reg_2	3	1	1	1	1	2
Reg_3	3	2	1	1	1	2
Reg_4	3	3	1	1	1	1
Reg_5	1	2	2	2	1	2
Reg_6	1	3	2	2	1	2
Reg_7	2	1	2	2	1	1
Reg_8	2	3	1	2	1	2

Źródło: opracowanie własne

gdzie:

q_1, q_2, q_3 – atrybuty warunkowe,
Reg_{ir} – reguła decyzyjna,
ir – indeks reguły decyzyjnej.

Tabela 27. Reguły niepewne algorytmu decyzyjnego w formie tabeli informacyjnej.

Reguła (Reg_{ir})	**Atrybuty warunkowe**			**Atrybut decyzyjny**	**Pewność**	**Wsparcie**
	Obszar do koszenia (q_1)	**Odległość od siedziby firmy (q_2)**	**Wynik eksploracji danych tekstowych (q_3)**	**Rentowność zamówienia**		
Reg_1	1	1	1	1	0,67	3
Reg_2	2	2	1	1	0,67	3
Reg_3	1	2	1	2	0,67	3

Źródło: opracowanie własne

gdzie:

q_1, q_2, q_3 – atrybuty warunkowe,
Reg_{ir} – reguła decyzyjna,
ir – indeks reguły decyzyjnej.

Formalny zapis zbioru reguł niepewnych określono wzorem *(107)*.

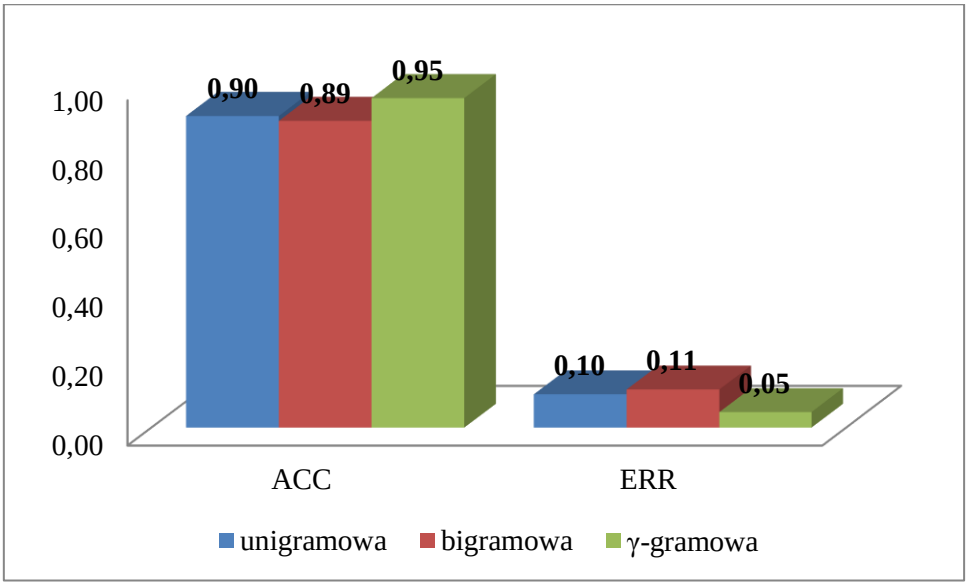
(107)

Reg_1: IF ((q_1 = 1) AND (q_2 = 1) AND (q_3 = 1)) THEN d = 1

Reg_2: IF ((q_1 = 2) AND (q_2 = 2) AND (q_3 = 1)) THEN d = 1

Reg_3: IF ((q_1 = 1) AND (q_2 = 2) AND (q_3 = 1)) THEN d = 2

Ostatecznie zbiór 200 przypadków zamówień publicznych poddano klasyfikacji w etapie 8 procedury integracji z rysunku 11. Następnie obliczono miary jakości decyzji (ACC oraz ERR) dotyczących sklasyfikowania przypadków do dwóch kategorii tj. do kategorii reprezentującej rentowne zamówienia publiczne oraz do kategorii z pozostałymi zamówieniami. Ostateczny wynik obliczeń dla wariantu A z rysunku 2 w postaci średniej arytmetycznej poszczególnych miar jakości (ACC, ERR) dla 11 wartości współczynnika *k* klasyfikatora kNN, zaprezentowano na rysunku 19.



Rysunek 19. Wartości miar jakości decyzji (ACC oraz ERR) dla trzech reprezentacji danych tekstowych (unigramowej, bigramowej oraz γ-gramowej) w przypadku eksploracji danych tekstowych i numerycznych (wariant A z rysunku 2)
Źródło: opracowanie własne

W celu realizacji wariantu B z rysunku 2 eksploracji z zastosowaniem wyłącznie metod eksploracji danych numerycznych pominięte zostały etapy 1, 2, 3 procedury integracji z rysunku 11. Do eksploracji danych numerycznych wykorzystano reprezentację danych opracowaną, w etapach 4 i 5 procedury z rysunku 2, dla wariantu A badań testowych, które przedstawiono w tabeli 23. Na podstawie tej reprezentacji danych numerycznych, jednak bez

wykorzystania atrybutu stanowiącego wynik eksploracji danych tekstowych, opracowano tablicę decyzyjną zawierającą się w tabeli 28.

Tabela 28. Tabela informacyjna z przekształconymi wartościami atrybutów do formy zakodowanej.

Lp.	Atrybuty warunkowe		Elementarne zbiory warunkowe	Atrybut decyzyjny	Koncept
	Obszar do koszenia (q_1)	**Odległość od siedziby firmy (q_2)**	***Eie***	**Rentowność zamówienia**	***Xix***
1	1	1	E_1	1	X_1
2	2	1	E_2	1	X_1
3	2	2	E_3	1	X_1
4	3	1	E_4	1	X_1
5	3	2	E_5	1	X_1
6	3	3	E_6	1	X_1
7	1	1	E_1	1	X_1
8	1	2	E_7	1	X_1
9	3	1	E_4	1	X_1
10	3	2	E_5	1	X_1
11	2	2	E_3	1	X_1
12	1	2	E_7	2	X_2
13	1	2	E_7	2	X_2
14	1	3	E_8	2	X_2
15	2	1	E_2	2	X_2
16	2	3	E_9	2	X_2
17	2	3	E_9	2	X_2
18	2	2	E_3	2	X_2
19	1	3	E_8	2	X_2
20	1	2	E_7	2	X_2
21	1	2	E_7	2	X_2
22	1	1	E_1	2	X_2

Źródło: opracowanie własne

gdzie:

q_1, q_2, q_3 – atrybuty warunkowe,
E_{ie} – elementarne zbiory warunkowe odpowiadające klasom abstrakcji,
ie – indeks zbioru warunkowego,
X – koncept decyzyjny,
ix – indeks konceptu decyzyjnego.

W etapie 6 procedury integracji z rysunku 11 przeprowadzono analizę istotności danych ze zbioru Z_E przy użyciu współczynników wynikających z metody Teorii Zbiorów

Przybliżonych. W tym celu Poszczególne przypadki zostały przyporządkowane do odpowiednich elementarnych zbiorów warunkowych E_{ie} zaprezentowanych w *(108)*.

(108)

E_1={1, 7, 22}, E_2={2, 15}, E_3={3, 11, 18}, E_4={4, 9}, E_5={5, 10}, E_6={6}, E_7={8, 12, 13, 20, 21 }, E_8={14, 19}, E_9={16, 17},

Kolejno przeprowadzono obliczenia, które zostały wykorzystane do określenia dolnych przybliżeń konceptów decyzyjnych $\underline{B}X_{ix}$ oraz ilości przypadków $card(\underline{B}X_{ix})$ należących do dolnych przybliżeń, wzory *(109)* oraz *(110)*.

(109)

X_1 = {1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11}

$\underline{B}X_1 = E_4 \cup E_5 \cup E_6$ = {4, 5, 6, 9, 10}

$card(\underline{B}X_1)$= 5

(110)

X_2 = {12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22}

$\underline{B}X_2 = E_8 \cup E_9$ = {14, 16, 17, 19}

$card(\underline{B}X_2)$= 4

Na podstawie ilości przypadków należących do dolnych przybliżeń wyznaczono pozytywny obszar rodziny konceptów decyzyjnych $POS_B(X)$, zgodny ze wzorem *(111)*.

(111)

$POS_B(X) = E_4 \cup E_5 \cup E_6 \cup E_8 \cup E_9$ = {4, 5, 6, 9, 10, 14, 16, 17, 19}

Ilość wszystkich przypadków w pozytywnym obszarze card($POS_B(X)$) jest równa 9. Współczynnik jakości przybliżenia rodziny konceptów decyzyjnych $\gamma_B(X)$ jest zatem równy wartości 0,41 co oznacza, że 41% przypadków wykorzystanych w obliczeniach pozwala na generowanie reguł pewnych. Następnie przeprowadzono obliczenia, które zostały wykorzystane do określenia górnych przybliżeń konceptów $\overline{B}X_{ix}$ oraz ilości przypadków $card(\overline{B}X_{ix})$ należące do górnych przybliżeń, wzory *(112)* oraz *(113)*.

(112)

X_1 = {1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11}

$\overline{B}X_1 = E_1 \cup E_2 \cup E_3 \cup E_4 \cup E_5 \cup E_6 \cup E_7$ = {1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 15, 18, 20, 21, 22}

$card(\overline{B}X_1)$ = 18

(113)

X_2 = {12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22}

$\overline{B}X_2 = E_1 \cup E_2 \cup E_3 \cup E_7 \cup E_8 \cup E_9 = \{1, 2, 3, 7, 8, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22\}$

$card(\overline{B}X_2) = 17$

Na podstawie obliczeń otrzymano dokładności przybliżeń kolejnych konceptów decyzyjnych równe $\mu(X_1) = 0{,}28$, $\mu(X_2) = 0{,}24$. Dokładność przybliżenia rodziny konceptów decyzyjnych $\beta_B(X)$, czyli przeciętny stopień zrozumienia decyzji wyniósł 0,26.

W dalszej części badań testowych wyliczono istotność kolejnych atrybutów warunkowych q_1, q_2, q_3, zgodnie ze wzorem *(77)*. Wyniki obliczeń przedstawiono w tabeli 29.

Tabela 29. Jakość i dokładność przybliżenia konceptów decyzyjnych X oraz istotności usuwanego atrybutu warunkowego.

Atrybut warunkowy	**Istotność atrybutu**	**Dokładność przybliżenia rodziny konceptów $\beta_B(X)$**	**Jakość przybliżenia rodziny konceptów $\gamma_B(X)$**
q_1	0,68	0,13	0,13
q_2	1	0	0

Źródło: opracowanie własne

Na podstawie wyznaczonych istotności poszczególnych atrybutów warunkowych stwierdzono, że każdy z atrybutów ma wpływ na definicję wiedzy na podstawie informacji zawartych w systemie informacyjnym – tablicy decyzyjnej zaprezentowanej w tabeli 28.

W wyniku realizacji etapu 7 z rysunku 11 został zbudowany algorytm decyzyjny, który opiera się na wyodrębnionych regułach decyzyjnych. Dla reguły pewnych oraz niepewnych obliczono wsparcie i współczynnik pewności, które przedstawiono tabelach 30 oraz 31.

Tabela 30. Reguły pewne algorytmu decyzyjnego w formie tabeli informacyjnej.

Reguła (Reg_{ir})	**Atrybuty warunkowe**		**Atrybut decyzyjny**	**Pewność**	**Wsparcie**
	Obszar do koszenia (q_1)	**Odległość od siedziby firmy (q_2)**	**Rentowność zamówienia**		
Reg_1	3	1	1	1	2
Reg_2	3	2	1	1	2
Reg_3	3	3	1	1	1
Reg_4	1	3	2	1	2
Reg_5	2	3	2	1	2

Źródło: opracowanie własne

gdzie:

q_1, q_2 – atrybuty warunkowe,

Reg_{ir} – reguła decyzyjna,
ir – indeks reguły decyzyjnej.

Formalny zapis zbioru reguł pewnych określono wzorem *(114)*.

(114)

Reg_1: IF ((q_1 = 3) AND (q_2 = 1)) THEN d = 1

Reg_2: IF ((q_1 = 3) AND (q_2 = 2)) THEN d = 1

Reg_3: IF ((q_1 = 3) AND (q_2 = 3)) THEN d = 1

Reg_4: IF ((q_1 = 1) AND (q_2 = 3)) THEN d = 2

Reg_5: IF ((q_1 = 2) AND (q_2 = 3)) THEN d = 2

Tabela 31. Reguły niepewne algorytmu decyzyjnego w formie tabeli informacyjnej.

Reguła (Reg_{ir})	**Atrybuty warunkowe**		**Atrybut decyzyjny**	**Pewność**	**Wsparcie**
	Obszar do koszenia (q_1)	**Odległość od siedziby firmy (q_2)**	**Rentowność zamówienia**		
Reg_1	1	1	1	0,66	3
Reg_2	2	1	2	0,5	2
Reg_3	2	2	1	0,66	3
Reg_4	1	2	2	0,8	5

Źródło: opracowanie własne

gdzie:

q_1, q_2 – atrybuty warunkowe,
Reg_{ir} – reguła decyzyjna,
ir – indeks reguły decyzyjnej.

Formalny zapis zbioru reguł niepewnych określono wzorem *(115)*.

(115)

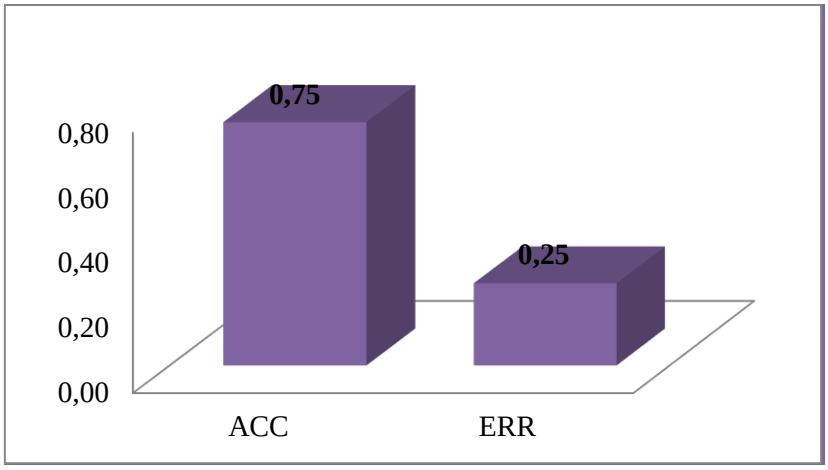
Reg_1: IF ((q_1 = 1) AND (q_2 = 1)) THEN d = 1

Reg_2: IF ((q_1 = 1) AND (q_2 = 2)) THEN d = 2

Reg_3: IF ((q_1 = 2) AND (q_2 = 1)) THEN d = 1

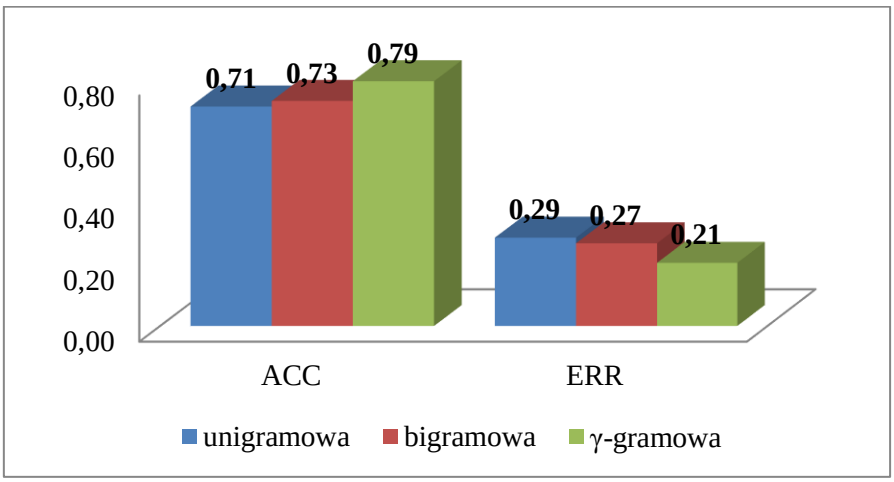
Reg_4: IF ((q_1 = 2) AND (q_2 = 2)) THEN d = 2

Ostatecznie zbiór 200 przypadków zamówień publicznych poddano klasyfikacji w etapie 8 procedury integracji z rysunku 11. Ostateczny wynik klasyfikacji danych numerycznych w postaci miar jakości decyzji (ACC, ERR) dla wariantu B z rysunku 2 przedstawiono na rysunku 20.



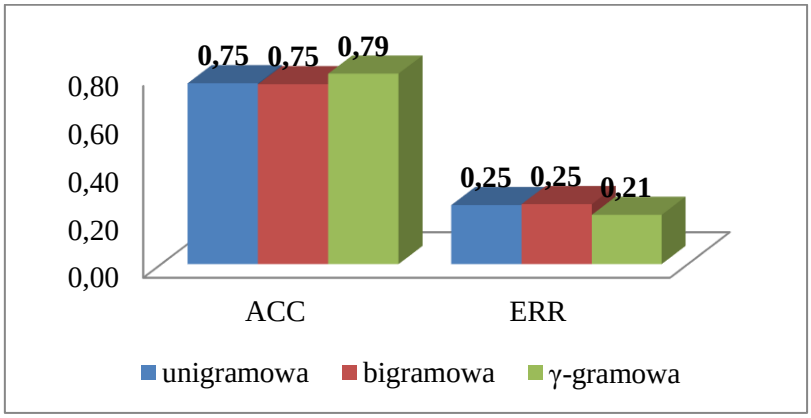
Rysunek 20. Średnie wartości miar jakości decyzji (ACC, ERR) w przypadku eksploracji danych numerycznych (wariant B z rysunku 2)
Źródło: opracowanie własne

W celu realizacji wariantu C z rysunku 2 eksploracji z zastosowaniem wyłącznie metod eksploracji danych tekstowych pominięte zostały etapy 4, 5, 6, 7 i 8 procedury integracji z rysunku 11. Do eksploracji wyłącznie danych tekstowych wykorzystano reprezentacje danych opracowane w pierwszym wariancie eksploracji (wariant A z rysunku 2). Zgodnie z wybranymi technikami eksploracji danych tekstowych w wariancie z wykorzystaniem zintegrowanych metod eksploracji danych tekstowych i numerycznych (wariant A z rysunku 2) dokonano klasyfikacji danych tekstowych, a wyniki w postaci miar jakości decyzji przedstawiono na rysunku 21.



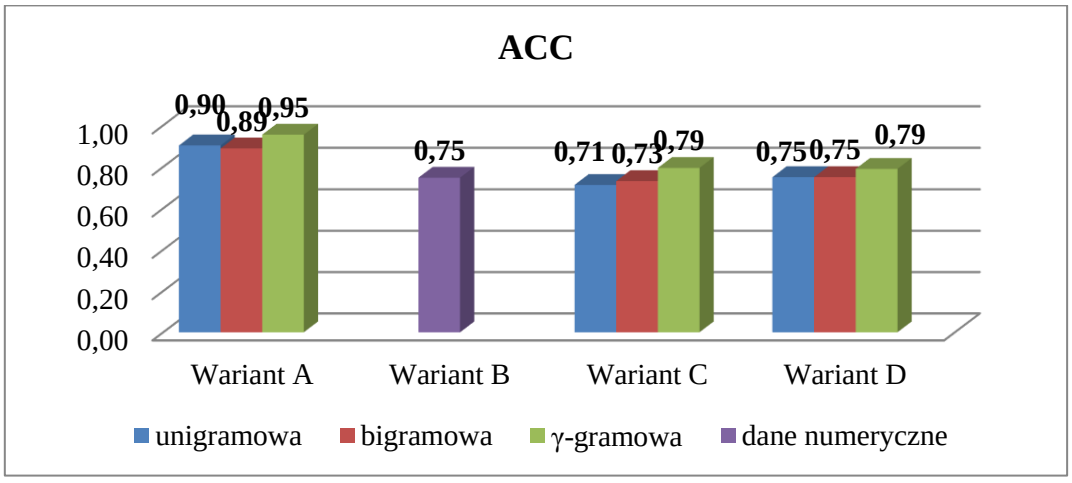
Rysunek 21. Wartości miar jakości (ACC oraz ERR) dla trzech reprezentacji danych tekstowych (unigramowej, bigramowej oraz γ-gramowej) dla wariantu eksploracji danych tekstowych (wariant C z rysunku 2)
Źródło: opracowanie własne

Wariant D z rysunku 2 dotyczy metody integracji wyników oddzielnych eksploracji danych numerycznych (wariant B z rysunku 2) oraz eksploracji danych tekstowych (wariant C z rysunku 2). Integracja wyników eksploracji w tym wariancie polega na wyborze korzystniejszego (z wyższą miarą ACC i niższą miarą ERR) wariantu eksploracji z pośród wariantów B i C dla poszczególnych poziomów zwrotu (od 1 do 11). Następnie miary jakości decyzji zostały uśrednione dla wszystkich 11 poziomów zwrotu, tak jak to przebiegało w poprzednich wariantach A, B oraz C. Wyniki w postaci miar jakości decyzji przedstawiono na rysunkach 22.

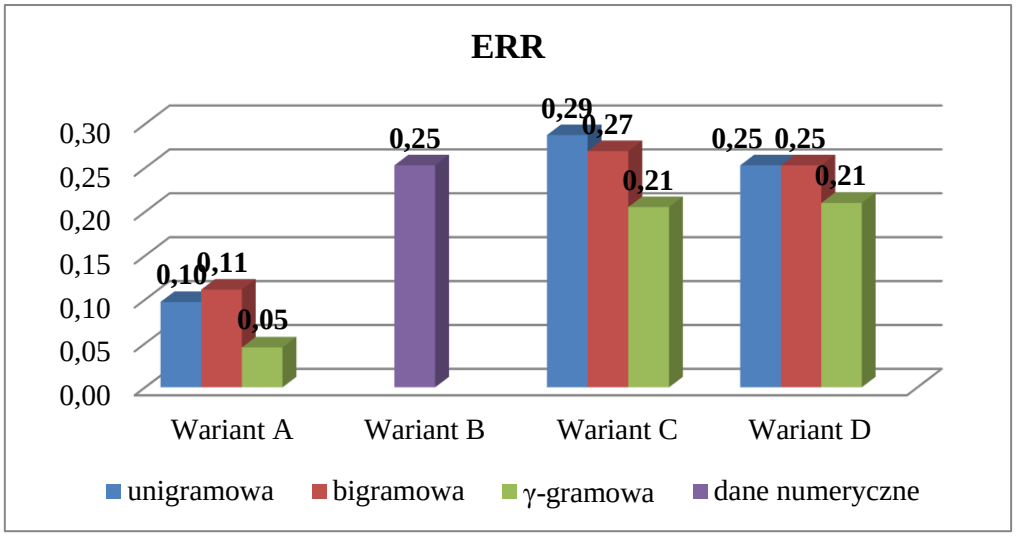


Rysunek 22. Wartości miar jakości (ACC oraz ERR) dla trzech reprezentacji danych tekstowych (unigramowej, bigramowej oraz γ-gramowej) dla metody integracji wyników eksploracji danych uzyskanych w wariancie B i C (wariant D z rysunku 2)
Źródło: opracowanie własne

Reasumując w wyniku obliczeń parametrów jakości kwalifikacji w rozpatrywanym przykładzie uzyskano wyniki zgodne z rysunkami 23 i 24.



Rysunek 23. Średnie wartości miar jakości decyzji ACC osiągnięte dla I przypadku procesu *PD*
Źródło: opracowanie własne



Rysunek 24. Średnie wartości miar jakości decyzji ERR osiągnięte dla I przypadku procesu *PD*
Źródło: opracowanie własne

5.3. Przykład II: Inwestowanie na Giełdzie Papierów Wartościowych

Drugi przypadek procesu *PD* wykorzystany w badaniach testowych dotyczy inwestowania na Giełdzie Papierów Wartościowych w tzw. trybie daytrading, czyli dokonywania transakcji kupna i sprzedaży akcji w czasie jednej sesji giełdowej. Co kilka dni, a czasami codziennie spółki giełdowe emitują komunikaty przeznaczone dla swoich akcjonariuszy. W zależności od charakteru komunikatu nastroje akcjonariuszy ulegają zmianie, co wpływa na rodzaj i ilość transakcji dokonywanych na akcjach lub kontraktach terminowych spółki. Komunikatami, które mogą wpływać na zmianę nastojów akcjonariuszy jest zakup lub sprzedaż papierów wartościowych spółki przez osoby, które mają dostęp do informacji poufnych lub osoby blisko związanej z członkami zarządu spółki.

Oprócz emisji przez spółkę określonych komunikatów giełdowych bardzo ważnym sygnałem do realizacji transakcji przez akcjonariuszy jest również wynik analizy technicznej. Analiza techniczna polega na badaniu wartości różnego rodzaju wskaźników wyznaczających najlepszy moment dokonywania transakcji kupna lub sprzedaży akcji spółki.

Połączenie zarówno analizy komunikatów giełdowych oraz wyników analizy technicznej daje pewną ilość historycznych obserwacji (przypadków), które umożliwiają zbudowanie reguł wyznaczających najkorzystniejszy moment dokonywania określonych transakcji na Giełdzie Papierów Wartościowych.

W niniejszym przypadku procedurę integracji metod eksploracji danych tekstowych i numerycznych zastosowano do analizy komunikatów spółek oraz analizy wskaźników takich jak: wolumen obrotu akcjami przez osobę związaną z zarządem spółki oraz wysokość cen kupna i sprzedaży tych akcji. Celem wykorzystania integracji analizy danych tekstowych oraz danych numerycznych w tym przypadku jest pozyskanie informacji o momencie spadku cen akcji spółki. Dzięki tej informacji możliwe jest wykonanie operacji w trybie daytrading na instrumentach finansowych uwzględniających spadki cen akcji spółki. Badania testowe przeprowadzono analizując historyczne notowania akcji spółki PKN Orlen SA.

Zadanie decyzyjne polega na wytypowaniu komunikatów spółki, w których podano informacje o tym, że osoba blisko związana z zarządem dokonała transakcji na akcjach spółki. W tym przypadku jedną z większych trudności jest odróżnienie komunikatów o dokonaniu transakcji na akcjach oraz na kontraktach terminowych, jak również dokonaniu tych transakcji przez inne podmioty. Dodatkowo ocenie podlegają dane numeryczne tj. wielkości sprzedaży akcji oraz cena sprzedaży akcji opublikowane w komunikacie giełdowym. Atrybutem decyzyjnym jest w tym przypadku cena akcji w dniu emisji komunikatu giełdowego. Na podstawie eksploracji danych tekstowych oraz numerycznych możliwe jest np. odkrycie reguł stwierdzających, że w przypadku znaczącej ilości sprzedaży akcji po cenie niższej od zakupu przez osobę blisko związaną z zarządem spółki rynek zazwyczaj będzie reagował negatywie, czyli ceny akcje spółki po emisji takiego komunikatu będą spadały.

Badanie testowo przeprowadzano na zbiorze 200 testowych przypadków transakcji z wykorzystaniem 11 przypadków treningowych dla kategorii oznaczającej spadek kursu akcji.

Po przeprowadzeniu klasyfikacji danych testowych, w pierwszej kolejności, za pomocą testu McNemara dokonano sprawdzenia istotności i zgodności otrzymanych wynikach dla różnych wariantów eksploracji danych z rysunku 2. Sumaryczne wyniki sprawdzenia zaprezentowano w tabelach od 32 do 34.

Dla określonego poziomu istotności $\alpha=0,05$ przeważająca większości wyników badań za pomocą testu McNemara posiada wartość *p-value* poniżej wartości α, przez co można potwierdzić, że różnice pomiędzy wynikami porównywanych wariantów są istotnie (biorąc pod uwagę przyjęty poziom istotności). Ze względu na wykazaną istotność statystyczną w większości porównywanych wyników różnych wariantów eksploracji danych, w dalszej części badań testowych, na podstawie wyników klasyfikacji, zostały obliczone wybrane miary jakości decyzji (ACC, ERR).

Tabela 32. Weryfikacja statystyczna za pomocą testu McNemara wyniku eksploracji Wariantu C z wykorzystaniem reprezentacji unigramowej oraz Wariantu A z wykorzystaniem reprezentacji unigramowej

Próba 1	Próba 2	Poziom zwrotu	p-value
Wariant C – reprezentacja unigramowa	Wariant A - reprezentacja unigramowa	1	<1E−4
		2	<1E−4
		3	<1E−4
		4	<1E−4
		5	377E-4
		6	1042E-4
		7	<1E−4
		8	<1E−4
		9	<1E−4
		10	<1E−4
		11	<1E−4

Źródło: opracowanie własne

Tabela 33. Weryfikacja statystyczna za pomocą testu McNemara wyniku eksploracji Wariantu C z wykorzystaniem reprezentacji *n*-gramowej oraz Wariantu A z wykorzystaniem reprezentacji n-gramowej
Źródło: opracowanie własne

Próba 1	Próba 2	Poziom zwrotu	p-value
Wariant C – reprezentacja n-gramowa	Wariant A - reprezentacja n-gramowa	1	<1E−4
		2	<1E−4
		3	<1E−4
		4	<1E−4
		5	<1E−4
		6	<1E−4
		7	<1E−4
		8	<1E−4
		9	<1E−4
		10	<1E−4
		11	<1E−4

Źródło: opracowanie własne

Tabela 34. Weryfikacja statystyczna za pomocą testu McNemara wyniku eksploracji Wariantu C z wykorzystaniem reprezentacji γ-gramowej oraz Wariantu A z wykorzystaniem reprezentacji γ-gramowej

Próba 1	Próba 2	Poziom zwrotu	p-value
Wariant C– reprezentacja γ -gramowa	Wariant A – reprezentacja γ -gramowa	1	35E-4
		2	49E-4
		3	49E-4
		4	49E-4
		5	49E-4
		6	49E-4
		7	49E-4
		8	49E-4
		9	<1E−4
		10	<1E−4
		11	<1E−4

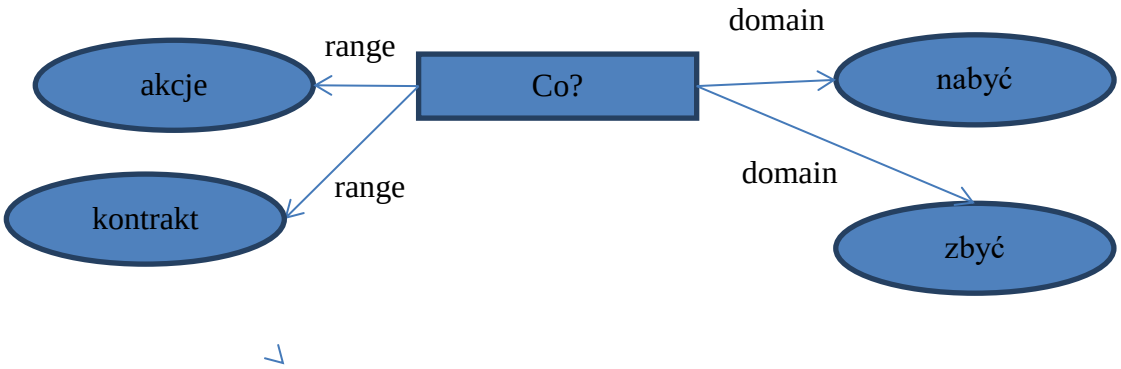
Źródło: opracowanie własne

W pierwszej kolejności został zrealizowany wariant A z rysunku 2 eksploracji (autorska metoda z rzozdziału 4) bazującej na opracowanej procedurze integracji metod eksploracji zarówno danych tekstowych jak i danych numerycznych. W tym celu na wstępie została opracowana reprezentacja danych tekstowych Z_T. W badaniach uwzględniono reprezentację unigramową tekstu oraz reprezentację bigramową (n-gramową). Reprezentację unigramową opracowano w oparciu o pojedyncze wyrazy, natomiast bigramową bazując na parach występujących po sobie wyrazów w dokumencie tekstowym. Reprezentacje te zostały opracowane zgodnie z opisem z treścią rozdziału 2.2. Do opracowania reprezentacji γ-gramowej zastosowano zdefiniowane przez eksperta dziedzinowego trzy zwizualizowane za pomocą grafów ontologii wzorce informacyjne, zgodne z rysunkami 25, 26 i 27.

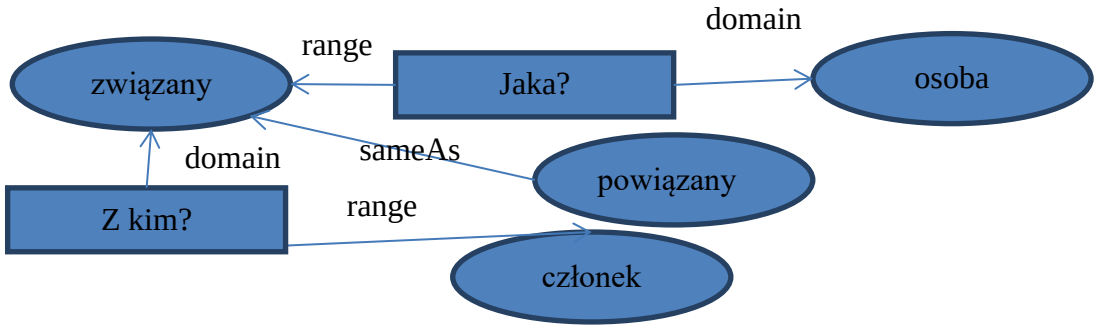
Przykładowe rzeczowe informacyjne wyekstrahowane na podstawie trzech zdefiniowanych wzorców (rysunki 25,26 i 27) to:

- nabycie akcji,
- zbycie akcji,
- nabycie kontraktów,
- osoba związana (z) członkiem,
- transakcja (na) instrumentach spółki.

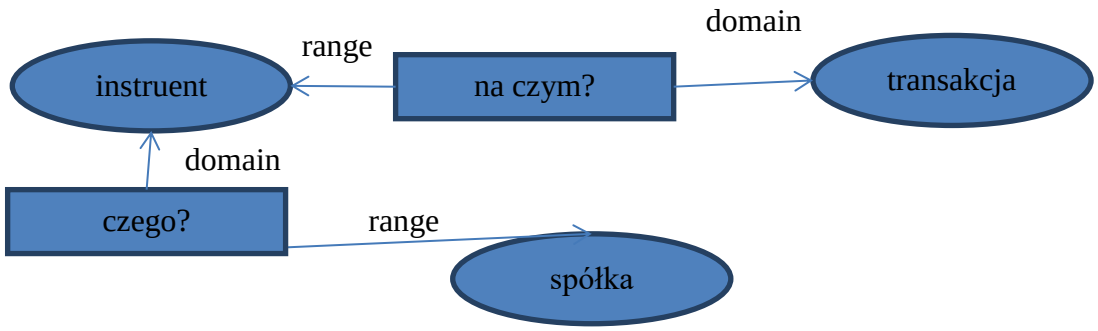
() – elementy rzeczowych informacji umieszczone w nawiasach zostały pominięte we wzorcach.



Rysunek 25. Wzorzec informacyjny numer 1 przykładowego procesu *PD* w języku OWL
Źródło: opracowanie własne



Rysunek 26. Wzorzec informacyjny numer 2 przykładowego procesu *PD* w języku OWL
Źródło: opracowanie własne



Rysunek 27. Wzorzec informacyjny numer 3 przykładowego procesu *PD* w języku OWL
Źródło: opracowanie własne

W drugim etapie autorskiej procedury integracji (rysunek 11 – rozdział 4) za pomocą trzech wzorców łącznie z tekstów wyekstrahowano 32 rzeczowych informacji z 200 zamówień z BZP zawierających różne formy fleksyjne wyrazów. Następnie rzeczowe informacje zgodnie z rysunkami 11 i 12 poddano analizie fleksyjnej. Na podstawie obliczonych miar skojarzeniowych wyeliminowano rzeczowe informacje zawierające najmniej poprawne

powiązania form fleksyjne wyrazów. W tym celu eksperymentalnie został dobrany próg, czyli wartość graniczną miary skojarzeniowej, która umożliwiła uwzględnienie lub odrzucenie rzeczowych informacji w dalszej części procedury eksploracji. Przykładowy fragment listy skojarzeniowej dla pierwszego wzorca informacyjnego z rysunku 25 przedstawiono w tabeli 35.

Tabela 35. Lista skojarzeniowa dla trójki: podmiot – *zbyć*, właściwość – *co*, obiekt – *kontrakt*.

Lp.	Podmiot	Obiekt	cw	lw	Sk [%]
1	*zbyła*	*kontrakt*	48	107	0,449
2		*(ilość np. 70) kontraktów*	38		0,355
3		*kontrakty*	21		0,20

gdzie:

sk – wyrażona w procentach miara skojarzenia podmiotu z obiektem,

cw – częstość względna wyrazu definiującego, tj. ilość współwystąpień podmiotu w zdaniach z obiektem,

lw – częstość bezwzględna podmiotu, czyli ilość wystąpień podmiotu w korpusie tekstów, który posłużył do wygenerowania listy skojarzeniowej.

Źródło: opracowanie własne

Po przeprowadzeniu eliminacji najmniej poprawnych powiązań form fleksyjnych wyrazów pozostało 24 rzeczowych informacji będących elementami γ-gramowej reprezentacji dokumentów tekstowych w modelu przestrzeni wektorowej, stanowiącej zbiór $Z'_{T.}$

W etapie 3 procedury z rysunku 11 dokonano transformacji danych ze zbioru Z'_T do Z''_T za pomocą właściwej eksploracji danych tekstowych w modelu przestrzeni wektorowej VSM. W celu realizacji tego etapu przez eksperta dziedzinowego dobrane zostały odpowiednie techniki eksploracji danych tekstowych. Są to techniki opisane w rozdziale 4.2. W efekcie działania tego etapu został utworzony zbiór danych Z''_T, który w kolejnym etapie procedury integracji został poddany dyskretyzacji.

W ramach etapu czwartego procedury integracji (Rysunek 11 – *Etap 4*) dane ze zbiorów Z_N oraz Z''_T poddano dyskretyzacji według równej ilości obiektów w przedziałach oraz doborowi wartości nominalnych, co zaprezentowano w tabelach od 36 do 40.

Tabela 36. Zamiana wartości ciągłych w dyskretne wartości lingwistyczne oraz formę kodową dla atrybutu warunkowego – *cena sprzedaży.*

Cena sprzedaży		
Wartość ciągła	**Wartość lingwistyczna**	**Forma zakodowana**
Niższa cena sprzedaży od ceny kupna	niższa	1
Wyższa cena sprzedaży od ceny kupna	wyższa	2
Cena sprzedaży i kupna są równe	równa	3

Źródło: opracowanie własne

Tabela 37. Zamiana wartości ciągłych w dyskretne wartości lingwistyczne oraz formę kodową dla atrybutu warunkowego – *wielkość sprzedaży.*

Wielkość sprzedaży		
Wartość ciągła	**Wartość lingwistyczna**	**Forma zakodowana**
Mniej sprzedanych akcji niż kupionych	mniej	1
Więcej sprzedanych niż kupionych	więcej	2
Tyle samo sprzedanych co kupionych	tyle samo	3

Źródło: opracowanie własne

Tabela 38. Zamiana wartości lingwistycznych w formę kodową dla atrybutu decyzyjnego warunkowego – *spadek cen akcji.*

Spadek cen akcji	
Wartość lingwistyczna	**Forma zakodowana**
TAK	1
NIE	2

Źródło: opracowanie własne

W przypadku wyniku eksploracji danych tekstowych przedziały zostały zdefiniowane przez eksperta (Tabela 45).

Tabela 39. Zamiana wartości lingwistycznych w formę kodową dla atrybutu warunkowego – *wynik eksploracji danych tekstowych.*

Wynik eksploracji danych tekstowych		
Wartość ciągła [podobieństwo]	**Wartość lingwistyczna**	**Forma zakodowana**
Poniżej 0,5	Klasa I	1
0,5 i powyżej	Klasa II	2

Źródło: opracowanie własne

Ostatecznie w etapie 5 procedury z rysunku 11 opracowano reprezentację danych numerycznych (zbiór Z_E) w postaci systemu informacyjnego – tablicy decyzyjnej, której fragment zawiera się w tabeli 40.

Tabela 40. Fragmentaryczna reprezentacja danych numerycznych w postaci tablicy decyzyjnej

Lp.	Atrybuty warunkowe			Atrybut decyzyjny
	Cena sprzedaży (q_1)	**Wielkość sprzedaży (q_2)**	**Klasyfikacja (q_3)**	**Spadek cen akcji**
1	1	3	1	1
2	2	3	2	1
3	1	2	1	1
4	2	2	1	1
5	1	2	1	1
6	2	1	1	1
…	…	…	…	…

Źródło: opracowanie własne

W etapie 6 procedury integracji z rysunku 11 przeprowadzono analizę istotności danych ze zbioru Z_E przy użyciu współczynników wynikających z metody Teorii Zbiorów Przybliżonych. W systemie informacyjnym *SI* uwzględniono 22-elementowy zbiór przypadków (obiektów). Bazując na wartościach atrybutów poszczególnych przypadków otrzymano tabele informacyjną w formie zakodowanej (tabela 41), w której wyodrębniono elementarne zbiory warunkowe E_i oraz zbiory decyzyjne (koncepty) X_i.

Poszczególne przypadki zostały przyporządkowane do odpowiednich elementarnych zbiorów warunkowych zaprezentowanych w *(116)*.

(116)

$E_1=\{1\}, E_2=\{2, 13\}, E_3=\{3, 5, 9\}, E_4=\{4, 7, 8, 16\}, E_5=\{6, 15, 17\}, E_6=\{10, 11, 22\}, E_7=\{12, 19, 20\}, E_8=\{14\}, E_9=\{18\}, E_{10}=\{21\},$

Kolejno przeprowadzono obliczenia, w celu wyznaczenia dolnych przybliżeń konceptów decyzyjnych $\underline{B}X_{ix}$ oraz liczby przypadków $card(\underline{B}X_{ix})$ należących do dolnych przybliżeń, wzory *(117)* oraz *(118)*.

(117)

$X_1 = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11\}$

$\underline{B}X_1 = E_1 \cup E_3 = \{1, 3, 5, 9\}$

$card(\underline{B}X_1) = 4$

(118)

$X_2 = \{12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22\}$

$\underline{B}X_2 = E_7 \cup E_8 \cup E_9 \cup E_{10} = \{12, 14, 18, 19, 20, 21\}$

$card(\underline{B}X_2) = 6$

Tabela 41. Tabela informacyjna z przekształconymi wartościami atrybutów do formy zakodowanej.

Lp.	Atrybuty warunkowe			Zbiory elementarne	Atrybut decyzyjny	Koncept
	Cena sprzedaży (q_1)	Wielkość sprzedaży (q_2)	Klasyfikacja (q_3)	*Eie*	Spadek cen akcji	*Xix*
1	1	3	1	E_1	1	X_1
2	2	3	2	E_2	1	X_1
3	1	2	1	E_3	1	X_1
4	2	2	1	E_4	1	X_1
5	1	2	1	E_3	1	X_1
6	2	1	1	E_5	1	X_1
7	2	2	1	E_4	1	X_1
8	2	2	1	E_4	1	X_1
9	1	2	1	E_3	1	X_1
10	2	1	2	E_6	1	X_1
11	2	1	2	E_6	1	X_1
12	2	2	2	E_7	2	X_2
13	2	3	2	E_2	2	X_2
14	1	1	1	E_8	2	X_2
15	2	1	1	E_5	2	X_2
16	2	2	1	E_4	2	X_2
17	2	1	1	E_5	2	X_2
18	1	1	2	E_9	2	X_2
19	2	2	2	E_7	2	X_2
20	2	2	2	E_7	2	X_2
21	1	2	2	E_{10}	2	X_2
22	2	1	2	E_6	2	X_2

Źródło: opracowanie własne

gdzie:

q_1, q_2, q_3 – atrybuty warunkowe,
E_{ie} – elementarne zbiory warunkowe odpowiadające klasom abstrakcji,
ie – indeks zbioru warunkowego,
X – koncept decyzyjny,
ix – indeks konceptu decyzyjnego.

Na podstawie liczby przypadków należących do dolnych przybliżeń wyznaczono pozytywny obszar rodziny konceptów decyzyjnych $POS_B(X)$, zgodny ze wzorem *(119)*.

(119)

$$POS_B(X) = E_1 \cup E_3 \cup E_7 \cup E_8 \cup E_9 \cup E_{10}\, E_{12} = \{1, 4, 5, 9, 12, 14, 18, 19, 20, 21\}$$

Ilość wszystkich przypadków w pozytywnym obszarze card($POS_B(X)$) jest równa 10. Jakość przybliżenia rodziny konceptów decyzyjnych $\gamma_B(X)$ jest zatem równa wartości 0,45 co oznacza, że 45% przypadków wykorzystanych w obliczeniach pozwala na generowanie reguł pewnych. Następnie przeprowadzono obliczenia, które zostały wykorzystane do określenia górnych przybliżeń konceptów $\overline{B}X_{ix}$ oraz ilości przypadków $card(\overline{B}X_{ix})$ należące do górnych przybliżeń, wzory *(120)* i *(121)*.

(120)

$X_1 = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11\}$

$\overline{B}X_1 = E_1 \cup E_2 \cup E_3 \cup E_4 \cup E_5 \cup E_6 = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 13, 15, 16, 17, 22\}$

$card(\overline{B}X_1) = 16$

(121)

$X_2 = \{12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22\}$

$\overline{B}X_2 = E_2 \cup E_4 \cup E_5 \cup E_6 \cup E_7 \cup E_8 \cup E_9 \cup E_{10} = \{2, 4, 6, 7, 8, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 21, 20, 22\}$

$card(\overline{B}X_2) = 18$

Na podstawie obliczeń otrzymano dokładności przybliżeń kolejnych konceptów decyzyjnych równe $\mu(X_1) = 0{,}25$, $\mu(X_2) = 0{,}33$. Dokładność przybliżenia rodziny konceptów decyzyjnych $\beta_B(X)$, czyli przeciętny stopień zrozumienia decyzji wyniósł 0,29.

W dalszej części badań testowych wyliczono istotność kolejnych atrybutów warunkowych q_1, q_2, q_3, zgodnie ze wzorem *(77)*. Wyniki obliczeń przedstawiono w tabeli 42.

Tabela 42. Jakość i dokładność przybliżenia konceptów decyzyjnych *X* oraz istotności usuwanego atrybutu warunkowego.

Atrybut warunkowy	**Istotność atrybutu**	**Dokładność przybliżenia rodziny konceptów $\beta_B(X)$**	**Jakość przybliżenia rodziny konceptów $\gamma_B(X)$**
q_1	0,6	0,1	0,18
q_2	0,8	0,05	0,09
q_3	1	0	0

Źródło: opracowanie własne

Na podstawie wyznaczonych istotności poszczególnych atrybutów warunkowych stwierdzono, że każdy z atrybutów ma znaczący wpływ na definicję wiedzy wynikającą z systemu informacyjnego *SI*. Istotność szczególnie ważnego ze względu na integrację atrybutu q_3 – *wynik eksploracji danych tekstowych,* wynosi 1. Oznacza to, że znacząco wpływa on na wiedzę w postaci zbioru reguł wygenerowanych na podstawie danych reprezentowanych przez

system informacyjny *SI* i potwierdza zasadność uwzględnienie w systemie SI atrybutu wynikającego z eksploracji danych tekstowych.

W wyniku realizacji etapu 7 procedury z rysunku 11 został zbudowany algorytm decyzyjny, który opiera się na wyodrębnionych regułach decyzyjnych. Dla reguły pewnych oraz niepewnych obliczono wsparcie i współczynnik pewności, które przedstawiono tabelach 43 i 44.

Tabela 43. Reguły pewne *Reg* algorytmu decyzyjnego w formie tabeli informacyjnej.

Reguła (Reg_{ir})	**Atrybuty warunkowe**			**Atrybut decyzyjny**	**Pewność**	**Wsparcie**
	Cena sprzedaży (q_1)	**Wielkość sprzedaży (q_2)**	**Klasyfikacja (q_3)**	**Spadek cen akcji**		
Reg_1	1	3	1	1	1	1
Reg_2	1	2	1	1	1	3
Reg_3	2	2	2	2	1	3
Reg_4	1	1	1	2	1	1
Reg_5	1	1	2	2	1	1
Reg_6	1	2	2	2	1	1

Źródło: opracowanie własne

gdzie:

q_1, q_2, q_3 – atrybuty warunkowe,
Reg_{ir} – reguła decyzyjna,
ir – indeks reguły decyzyjnej.

Formalny zapis zbioru reguł pewnych określono wzorem *(122).*

(122)

Reg_1: IF ((q_1 = 1) AND (q_2 = 3) AND (q_3 = 1)) THEN d = 1

Reg_2: IF ((q_1 = 2) AND (q_2 = 2) AND (q_3 = 1)) THEN d = 1

Reg_3: IF ((q_1 = 2) AND (q_2 = 2) AND (q_3 = 2)) THEN d = 2

Reg_4: IF ((q_1 = 1) AND (q_2 = 1) AND (q_3 = 1)) THEN d = 2

Reg_5: IF ((q_1 = 3) AND (q_2 = 1) AND (q_3 = 2)) THEN d = 2

Reg_6: IF ((q_1 = 1) AND (q_2 = 2) AND (q_3 = 2)) THEN d = 2

W przypadku reguły niepewnej Reg_3, ze względu na pewność równą 0,5 oraz atrybut *klasyfikacja* równy 2, ekspercko stwierdzono, że prawidłowy atrybut decyzyjny to 2.

Tabela 44. Reguły niepewne *Reg* algorytmu decyzyjnego w formie tabeli informacyjnej.

Reguła (Reg_{ir})	**Atrybuty warunkowe**			**Atrybut decyzyjny**	**Pewność**	**Wsparcie**
	Cena sprzedaży (q_1)	**Wielkość sprzedaży (q_2)**	**Klasyfikacja (q_3)**	**Spadek cen akcji**		
Reg_1	2	2	1	1	0,75	4
Reg_2	2	1	1	1	0,66	3
Reg_3	2	3	2	2	0,5	2
Reg_4	2	2	1	2	0,66	3

Źródło: opracowanie własne

gdzie:

q_1, q_2, q_3 – atrybuty warunkowe,
Reg_{ir} – reguła decyzyjna,
ir – indeks reguły decyzyjnej.

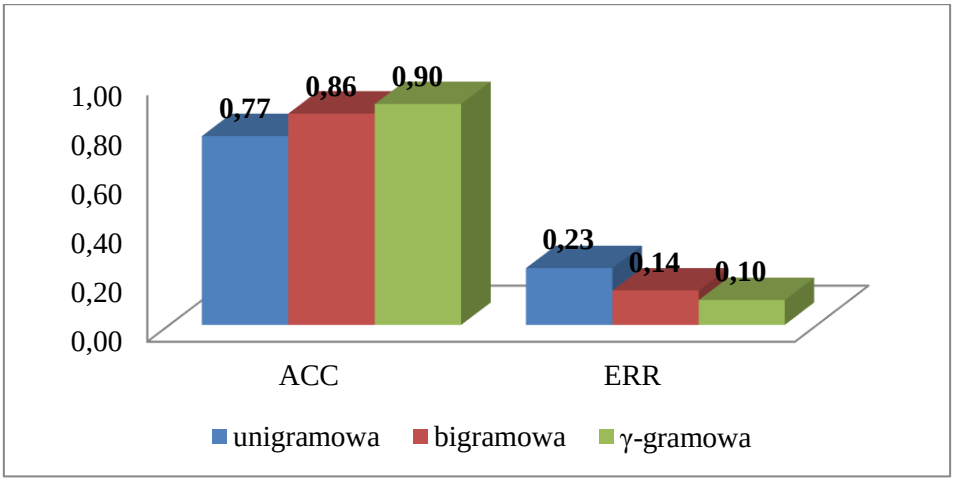
Formalny zapis zbioru reguł niepewnych określono wzorem *(123)*.

(123)

Reg_1: IF ((q_1 = 2) AND (q_2 = 2) AND (q_3 = 1)) THEN d = 1
Reg_2: IF ((q_1 = 2) AND (q_2 = 1) AND (q_3 = 1)) THEN d = 1
Reg_3: IF ((q_1 = 2) AND (q_2 = 3) AND (q_3 = 2)) THEN d = 2
Reg_4: IF ((q_1 = 2) AND (q_2 = 2) AND (q_3 = 1)) THEN d = 2

Ostatecznie zbiór 200 przypadków zamówień publicznych poddano klasyfikacji w etapie 8 procedury integracji z rysunku 11. Następnie obliczono miary jakości decyzji (ACC oraz ERR) dotyczących sklasyfikowania 200 przypadków do dwóch kategorii tj. do kategorii reprezentującej rentowne zamówienia publiczne oraz do kategorii z pozostałymi zamówieniami. Wyniki obliczeń w postaci średniej arytmetycznej poszczególnych miar jakości (ACC, ERR) dla 11 wartości współczynnika k klasyfikatora kNN, zaprezentowano na rysunku 28.

W celu realizacji wariantu B z rysunku 2 eksploracji z zastosowaniem wyłącznie metod eksploracji danych numerycznych pominięte zostały etapy 1, 2, 3 procedury integracji z rysunku 11. Do eksploracji danych numerycznych wykorzystano reprezentację danych opracowaną, w etapach 4 i 5 procedury z rysunku 2, dla wariantu A badań testowych, które przedstawiono w tabeli 40. Na podstawie tej reprezentacji danych numerycznych, jednak bez wykorzystania atrybutu stanowiącego wynik eksploracji danych tekstowych, opracowano tablicę decyzyjną zawierającą się w tabeli 45.



Rysunek 28. Średnie wartości miar jakości decyzji (ACC oraz ERR) dla trzech reprezentacji danych tekstowych (unigramowej, bigramowej oraz γ-gramowej) w przypadku eksploracji danych tekstowych i numerycznych (wariant A z rysunku 2)
Źródło: opracowanie własne

Tabela 45. Tabela informacyjna z przekształconymi wartościami atrybutów do formy zakodowanej.

Lp.	Atrybuty warunkowe		Zbiory elementarne	Atrybut decyzyjny	Koncept
	Cena sprzedaży (q_1)	Wielkość sprzedaży (q_2)	*Eie*	Spadek cen akcji	*Xix*
1	1	3	E_1	1	X_1
2	2	3	E_2	1	X_1
3	1	2	E_3	1	X_1
4	2	2	E_4	1	X_1
5	1	2	E_3	1	X_1
6	2	1	E_5	1	X_1
7	2	2	E_4	1	X_1
8	2	2	E_4	1	X_1
9	1	2	E_3	1	X_1
10	2	1	E_5	1	X_1
11	2	1	E_5	2	X_1
12	2	2	E_4	2	X_2
13	2	3	E_2	2	X_2
14	1	1	E_6	2	X_2
15	2	1	E_5	2	X_2
16	2	2	E_4	2	X_2
17	2	1	E_5	2	X_2
18	1	1	E_6	2	X_2
19	2	2	E_4	2	X_2
20	2	2	E_4	2	X_2
21	1	2	E_3	2	X_2
22	2	1	E_5	2	X_2

Źródło: opracowanie własne

gdzie:

q_1, q_2, q_3 – atrybuty warunkowe,
E_{ie} – elementarne zbiory warunkowe odpowiadające klasom abstrakcji,
ie – indeks zbioru warunkowego,
X – koncept decyzyjny,
ix – indeks konceptu decyzyjnego.

W etapie 6 procedury integracji z rysunku 11 przeprowadzono analizę istotności danych ze zbioru Z_E przy użyciu współczynników wynikających z metody Teorii Zbiorów Przybliżonych. W tym celu Poszczególne przypadki zostały przyporządkowane do odpowiednich elementarnych zbiorów warunkowych zaprezentowanych w *(124)*.

(124)

E_1={1}, E_2={2, 13}, E_3={3, 5, 9, 21}, E_4={4, 7, 8, 12, 16, 19, 20}, E_5={6, 10, 11, 15, 17, 22}, E_6={14, 18},

Kolejno przeprowadzono obliczenia, które zostały wykorzystane do określenia dolnych przybliżeń konceptów decyzyjnych $\underline{B}X_{ix}$ oraz ilości przypadków $card(\underline{B}X_{ix})$ należących do dolnych przybliżeń, wzory *(125)* oraz *(126)*.

(125)

X_1 = {1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11}
$\underline{B}X_1 = E_1$ = {1}
card(DP(X_1)) = 1

(126)

X_2 = {12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22}
DP(X_2) = E_6= {14, 18}
$card(\underline{B}X_2)$= 2

Na podstawie ilości przypadków należących do dolnych przybliżeń wyznaczono pozytywny obszar rodziny konceptów decyzyjnych $POS_B(X)$, zgodny ze wzorem *(127)*.

(127)

$POS_B(X)$ = $E_1 \cup E_6$ = {1, 14, 18}

Ilość wszystkich przypadków w pozytywnym obszarze card($POS_B(X)$) jest równa 3. Jakość przybliżenia rodziny konceptów decyzyjnych $\gamma_B(X)$ jest zatem równa wartości 0,14 co oznacza, że 14% przypadków wykorzystanych w obliczeniach pozwala na generowanie

reguł pewnych. Następnie przeprowadzono obliczenia, które zostały wykorzystane do określenia górnych przybliżeń konceptów $\overline{B}X_{ix}$ oraz ilości przypadków $card(\overline{B}X_{ix})$ należące do górnych przybliżeń, wzory *(128)* oraz *(129).*

(128)

X_1 = {1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11}

$\overline{B}X_1 = E_1 \cup E_2 \cup E_3 \cup E_4 \cup E_5 \cup E_6$ = {1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 15, 16, 17, 19, 21, 22}

$card(\overline{B}X_1)$ = 20

(129)

X_2 = {12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22}

$\overline{B}X_2 = E_2 \cup E_3 \cup E_4 \cup E_5 \cup E_6$ = {2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 21, 22}

$card(\overline{B}X_2)$ = 21

Na podstawie obliczeń otrzymano dokładności przybliżeń kolejnych konceptów decyzyjnych równe $\mu(X_1)$ = 0,05, $\mu(X_2)$ = 0,1. Dokładność przybliżenia rodziny konceptów decyzyjnych β (F), czyli przeciętny stopień zrozumienia decyzji wyniósł 0,07.

W dalszej części badań testowych wyliczono istotność kolejnych atrybutów warunkowych q_1, q_2, q_3, zgodnie ze wzorem *(77)*. Wyniki obliczeń przedstawiono w tabeli 46.

Tabela 46. Jakość i dokładność przybliżenia konceptów decyzyjnych *X* oraz istotności usuwanego atrybutu warunkowego.

Atrybut warunkowy	**Istotność atrybutu**	**Dokładność przybliżenia rodziny konceptów $\boldsymbol{\beta_B(X)}$**	**Jakość przybliżenia rodziny konceptów $\boldsymbol{\gamma_B(X)}$**
q_1	1	0	0
q_2	1	0	0

Źródło: opracowanie własne

Na podstawie wyznaczonych istotności poszczególnych atrybutów warunkowych stwierdzono, że każdy z atrybutów ma wpływ na definicję wiedzy na podstawie informacji zawartych w systemie informacyjnych – tablicy decyzyjnej zaprezentowanej w tabeli 45.

W wyniku realizacji etapu 7 z rysunku 11 został zbudowany algorytm decyzyjny, który opiera się na wyodrębnionych regułach decyzyjnych.. Dla reguły pewnych oraz niepewnych obliczono wsparcie i współczynnik pewności, które przedstawiono tabelach 47 oraz 48.

Tabela 47. Reguły pewne R*eg* algorytmu decyzyjnego w formie tabeli informacyjnej.

Reguła (Reg_{ir})	**Atrybuty warunkowe**		**Atrybut decyzyjny**	**Pewność**	**Wsparcie**
	Cena sprzedaży (q_1)	**Wielkość sprzedaży (q_2)**	**Spadek cen akcji (d)**		
Reg_1	1	3	1	1	1
Reg_2	1	1	2	1	2

Źródło: opracowanie własne

gdzie:

q_1, q_2 – atrybuty warunkowe,
Reg_{ir} – reguła decyzyjna,
ir – indeks reguły decyzyjnej.

Formalny zapis zbioru reguł pewnych określono wzorem *(130)*.

(130)

Reg_1: IF ((q_1 = 1) AND (q_2 = 3)) THEN d = 1
Reg_2: IF ((q_1 = 1) AND (q_2 = 1)) THEN d = 1

Tabela 48. Reguły niepewne *Reg* algorytmu decyzyjnego w formie tabeli informacyjnej.

Reguła (Reg_{ir})	**Atrybuty warunkowe**		**Atrybut decyzyjny**	**Pewność**	**Wsparcie**
	Cena sprzedaży (q_1)	**Wielkość sprzedaży (q_2)**	**Spadek cen akcji (d)**		
Reg_1	2	3	2	0,5	2
Reg_2	1	2	1	0,75	4
Reg_3	2	2	2	0,57	7
Reg_4	2	1	2	0,5	6

Źródło: opracowanie własne

gdzie:

q_1, q_2 – atrybuty warunkowe,
Reg_{ir} – reguła decyzyjna,
ir – indeks reguły decyzyjnej.

Formalny zapis zbioru reguł niepewnych określono wzorem *(131)*.

(131)

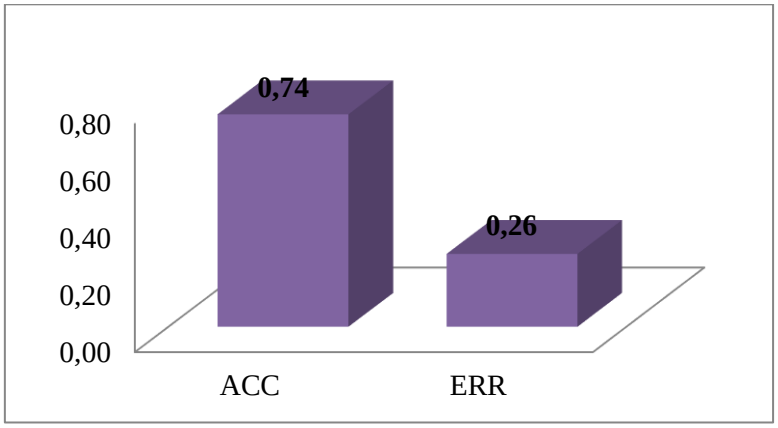
Reg_1: IF ((q_1 = 2) AND (q_2 = 3)) THEN d = 2

Reg_2: IF ((q_1 = 1) AND (q_2 = 2)) THEN d = 1

Reg_3: IF ((q_1 = 2) AND (q_2 = 2)) THEN d = 2

Reg_4: IF ((q_1 = 2) AND (q_2 = 1)) THEN d = 2

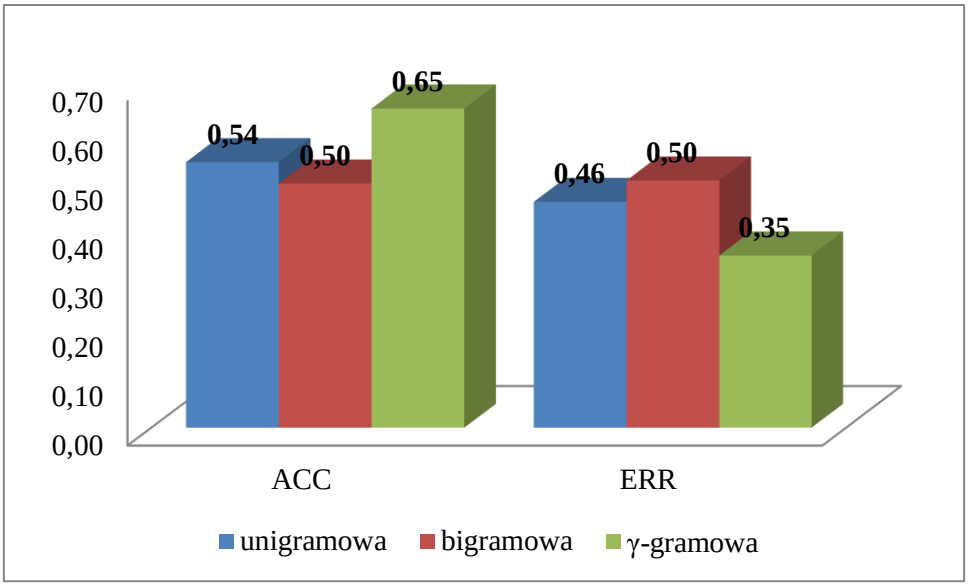
Ostatecznie zbiór 200 przypadków zamówień publicznych poddano klasyfikacji w etapie 8 procedury integracji z rysunku 11. Wyniki eksploracji danych numerycznych w postaci miar jakości decyzji (ACC, ERR) przedstawiono na rysunku 29.



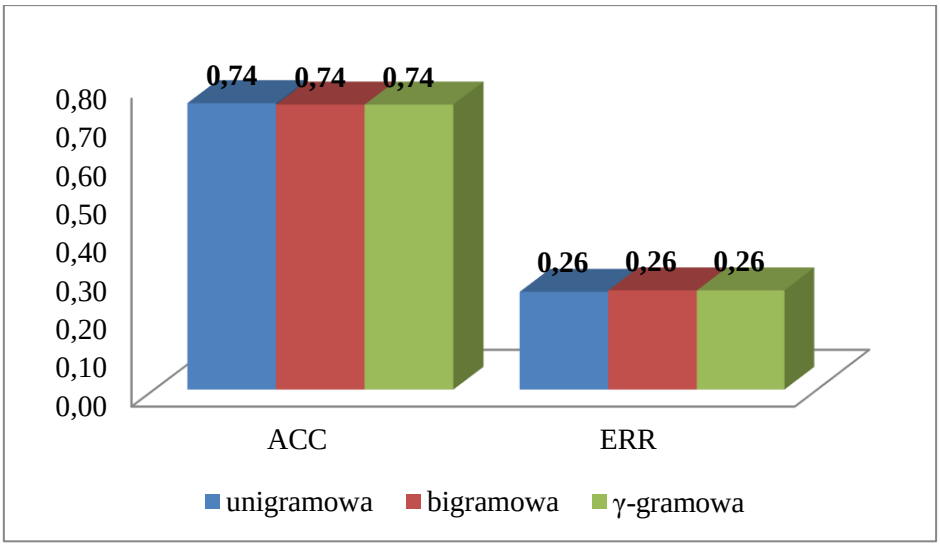
Rysunek 29. Średnie wartości miar jakości decyzji (F1 score, ACC, ERR) w przypadku eksploracji danych numerycznych (wariant B z rysunku 2)
Źródło: opracowanie własne

W celu realizacji wariantu C z rysunku 2 eksploracji z zastosowaniem wyłącznie metod eksploracji danych tekstowych pominięte zostały etapy 4, 5, 6, 7 i 8 procedury integracji z rysunku 11. Do eksploracji wyłącznie danych tekstowych wykorzystano reprezentacje danych opracowane w pierwszym wariancie eksploracji (wariant A z rysunku 2). Zgodnie z wybranymi technikami eksploracji danych tekstowych w wariancie z wykorzystaniem zintegrowanych metod eksploracji danych tekstowych i numerycznych (wariant A z rysunku 2) dokonano eksploracji danych tekstowych, a wyniki w postaci miar jakości decyzji przedstawiono na rysunkach 30.

Wariant D z rysunku 2 dotyczy metody integracji wyników oddzielnych eksploracji danych numerycznych (wariant B z rysunku 2) oraz eksploracji danych tekstowych (wariant C z rysunku 2). Integracja wyników eksploracji w tym wariancie polega na wyborze korzystniejszego (z wyższą miarą ACC i niższą miarą ERR) wariantu eksploracji z pośród wariantów B i C dla poszczególnych poziomów zwrotu (od 1 do 11). Następnie miary jakości decyzji zostały uśrednione dla wszystkich 11 poziomów zwrotu, tak jak to przebiegało w poprzednich wariantach A, B oraz C. Wyniki w postaci miar jakości decyzji przedstawiono na rysunkach 31.

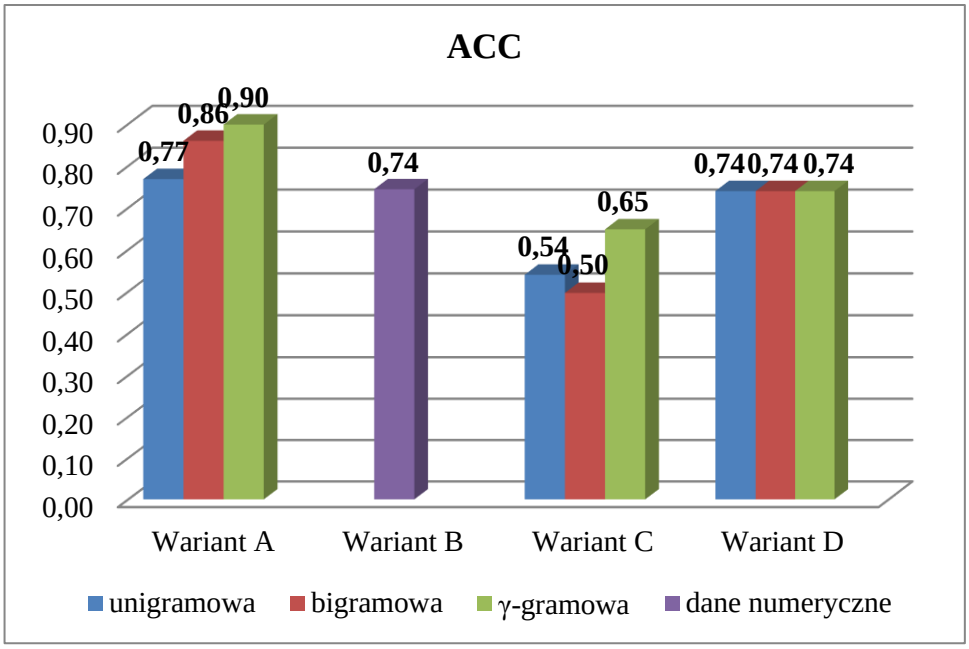


Rysunek 30. Wartości miar jakości (ACC oraz ERR) dla trzech reprezentacji danych tekstowych (unigramowej, bigramowej oraz γ-gramowej) dla wariantu eksploracji z wykorzystaniem integracji metod eksploracji danych tekstowych (wariant C z rysunku 2)
Źródło: opracowanie własne

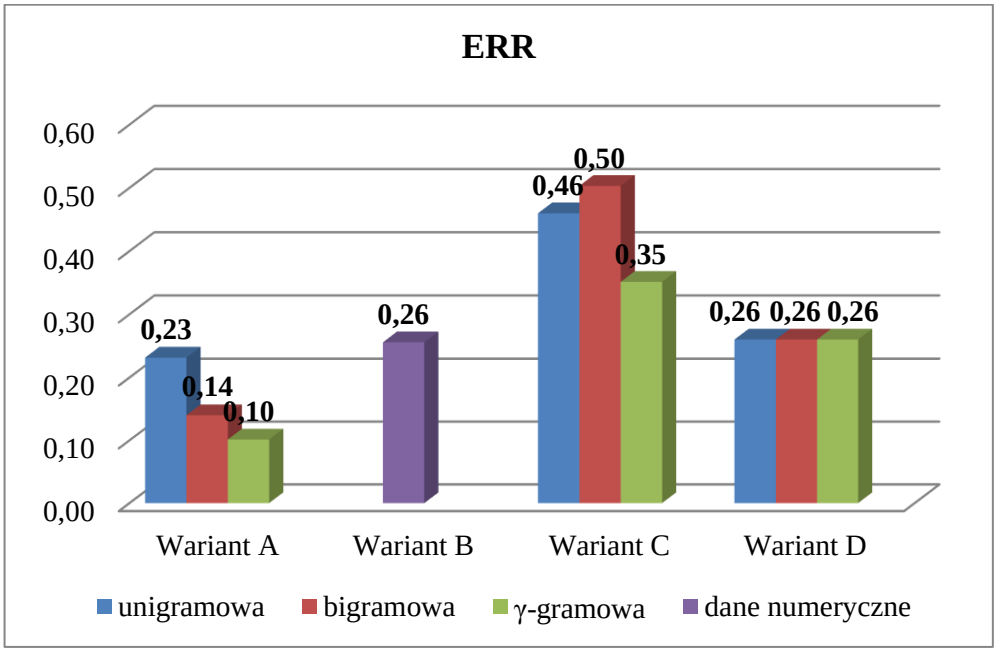


Rysunek 31. Wartości miar jakości (ACC oraz ERR) dla trzech reprezentacji danych tekstowych (unigramowej, bigramowej oraz γ-gramowej) dla metody integracji wyników eksploracji danych uzyskanych w wariancie B i C (wariant D z rysunku 2)
Źródło: opracowanie własne

Reasumując w wyniku obliczeń parametrów jakości kwalifikacji w rozpatrywanym przykładzie uzyskano wyniki zgodne z rysunkami 32 i 33.



Rysunek 32. Średnie wartości miar jakości decyzji ACC osiągnięte dla II przykładu procesu *PD*
Źródło: opracowanie własne



Rysunek 33. Średnie wartości miar jakości decyzji ERR osiągnięte dla II przykładu procesu *PD*
Źródło: opracowanie własne

5.4. Przykład III: Wyszukiwanie atrakcyjnych ofert pracy

Trzeci przykładowy proces *PD* wykorzystany w badaniach testowych dotyczy wyszukiwania atrakcyjnych ofert pracy umieszczonych w serwisach www.pracuj.pl i www.praca.pl. Zadanie decyzyjne polega na wytypowaniu ofert pracy, w których opis jest zgodny z charakterystyką pracy pracownika ds. rekrutacji. O atrakcyjności oferty oprócz zgodności opisu tekstowego świadczą dane numeryczne w postaci liczby mieszkańców miejscowości, w której znajduje się miejsce pracy oraz poziom hierarchii stanowiska (*stażysta, specjalista, kierownik*). Biorąc pod uwagę wartości wymienionych kryteriów badana jest atrakcyjność pracy względem historycznych wskazań kandydata.

Badanie testowo przeprowadzano na zbiorze 200 przypadków ofert pracy z wykorzystaniem 11 przypadków treningowych dla kategorii oznaczającej atrakcyjne oferty.

Po przeprowadzeniu klasyfikacji danych testowych, w pierwszej kolejności, za pomocą testu McNemara dokonano sprawdzenia sądów o otrzymanych wynikach dla różnych wariantów eksploracji danych z rysunku 2. Sumaryczne wyniki sprawdzenia zaprezentowano w tabelach od 49 do 51.

Tabela 49. Weryfikacja statystyczna za pomocą testu McNemara wyniku eksploracji danych trzeciego przypadku użycia dla Wariantu C z wykorzystaniem reprezentacji unigramowej oraz Wariantu A z wykorzystaniem reprezentacji unigramowej

Próba 1	Próba 2	Poziom zwrotu	p-value
Wariant C – reprezentacja unigramowa	Wariant A – reprezentacja unigramowa	1	<1E-4
		2	<1E-4
		3	<1E-4
		4	<1E-4
		5	<1E-4
		6	<1E-4
		7	<1E-4
		8	<1E-4
		9	<1E-4
		10	<1E-4
		11	<1E-4

Źródło: opracowanie własne

Tabela 50. Weryfikacja statystyczna za pomocą testu McNemara wyniku eksploracji danych trzeciego przypadku użycia dla Wariantu C z wykorzystaniem reprezentacji n-gramowej oraz Wariantu A z wykorzystaniem reprezentacji n-gramowej

Próba 1	**Próba 2**	**Poziom zwrotu**	**p-value**
Wariant C – reprezentacja n-gramowa	Wariant A - reprezentacja n-gramowa	1	<1E-4
		2	<1E-4
		3	<1E-4
		4	<1E-4
		5	2E-4
		6	3E-4
		7	2E-4
		8	<1E-4
		9	<1E-4
		10	5E-4
		11	77E-4

Źródło: opracowanie własne

Tabela 51. Weryfikacja statystyczna za pomocą testu McNemara wyniku eksploracji danych trzeciego przypadku użycia dla Wariantu C z wykorzystaniem reprezentacji γ-gramowej oraz Wariantu A z wykorzystaniem reprezentacji γ-gramowej

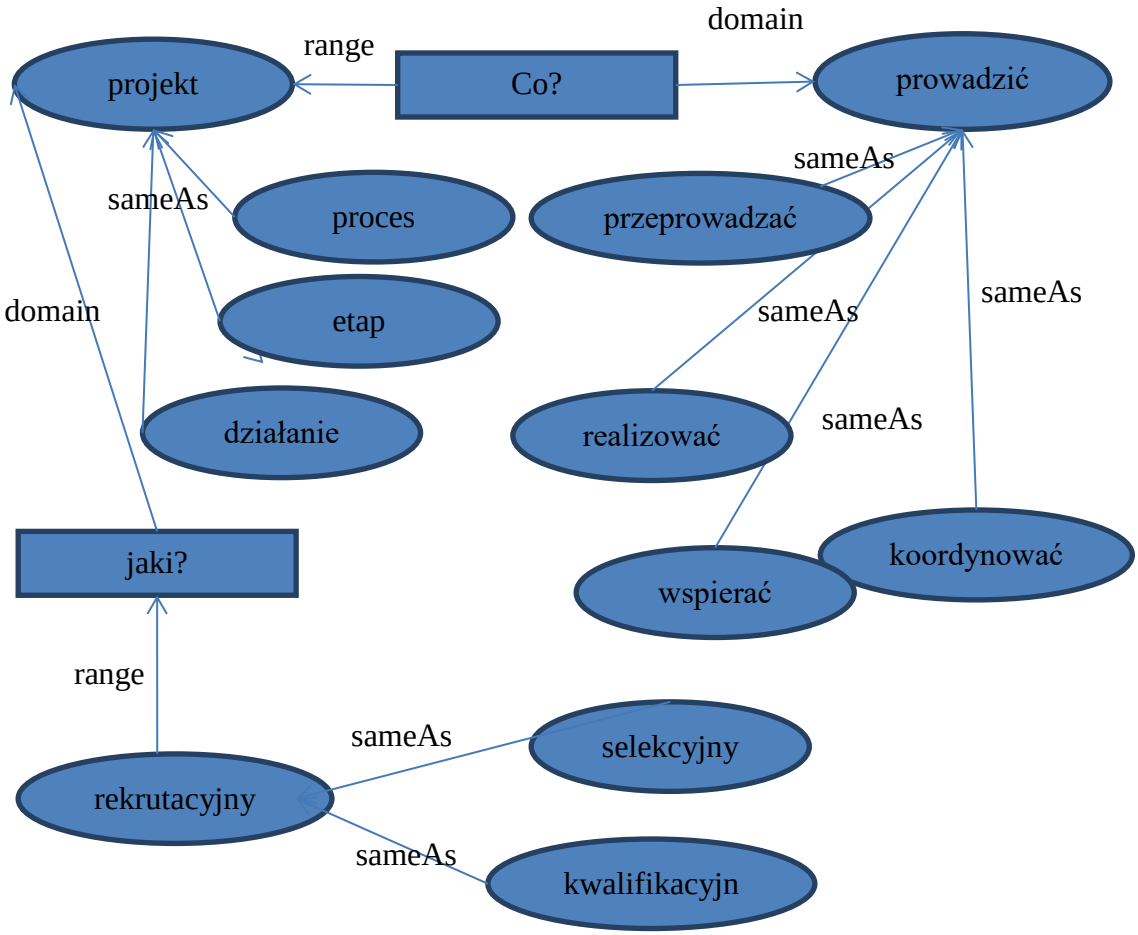
Próba 1	**Próba 2**	**Poziom zwrotu**	**p-value**
Wariant C reprezentacja γ -gramowa	Wariant A – reprezentacja γ -gramowa	1	<1E-4
		2	<1E-4
		3	<1E-4
		4	<1E-4
		5	<1E-4
		6	2E-4
		7	<1E-4
		8	<1E-4
		9	<1E-4
		10	<1E-4
		11	<1E-4

Źródło: opracowanie własne

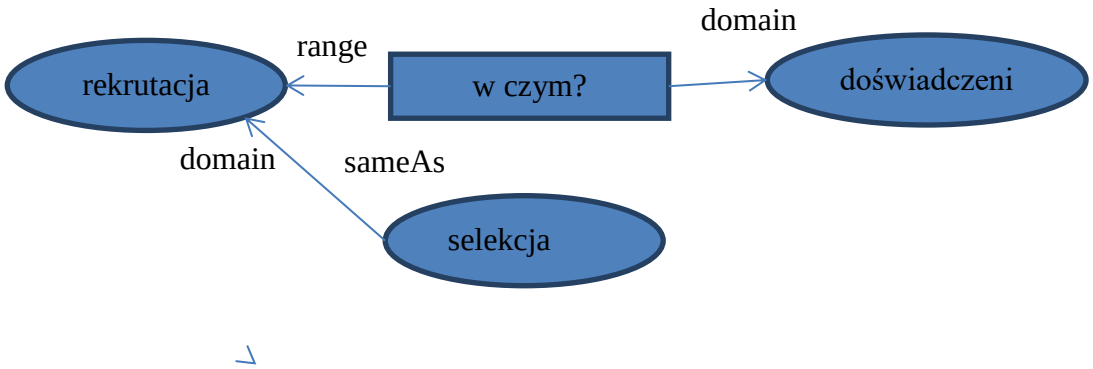
Dla określonego poziomu istotności α=0,05 przeważająca większości wyników badań za pomocą testu McNemara posiada wartość *p-value* poniżej wartości α,

przez co można potwierdzić, że różnice pomiędzy wynikami porównywanych wariantów są istotnie (biorąc pod uwagę przyjęty poziom istotności). Ze względu na wykazaną istotność statystyczną w większości porównywanych wyników różnych wariantów eksploracji danych, w dalszej części badań testowych, na podstawie wyników klasyfikacji, zostały obliczone wybrane miary jakości decyzji (ACC, ERR).

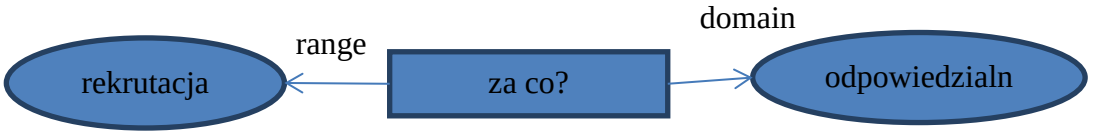
W pierwszej kolejności został zrealizowany wariant A z rysunku 2 eksploracji bazującej na opracowanej procedurze integracji metod eksploracji zarówno danych tekstowych jak i danych numerycznych. W tym celu na wstępie została opracowana reprezentacja danych tekstowych Z_T. Reprezentację unigramową opracowano w oparciu o pojedyncze wyrazy, natomiast bigramową bazując na parach występujących po sobie wyrazów w dokumencie tekstowym. Reprezentacje te zostały opracowane zgodnie z opisem z rozdziału 2.2. Do opracowania reprezentacji γ-gramowej zastosowano zdefiniowane przez eksperta dziedzinowego trzy zwizualizowane za pomocą grafów ontologii wzorce informacyjne, zgodne z rysunkami 34, 35 i 36.



Rysunek 34. Wzorzec informacyjny numer 1 przykładowego procesu *PD* w języku OWL
Źródło: opracowanie własne



Rysunek 35. Wzorzec informacyjny numer 2 przykładowego procesu *PD* w języku OWL
Źródło: opracowanie własne



Rysunek 36. Wzorzec informacyjny numer 3 przykładowego procesu *PD* w języku OWL
Źródło: opracowanie własne

Przykładowe rzeczowe informacje wyekstrahowane na podstawie trzech powyżej zdefiniowanych wzorców to:

- realizacja projektu rekrutacyjnego,
- prowadzenie działań selekcyjnych,
- doświadczenie (w) rekrutacji,
- doświadczenie (w) selekcji,
- odpowiedzialny (za) rekrutację.

() – elementy rzeczowych informacji w nawiasach zostały pominięte we wzorcach.

W drugim etapie autorskiej procedury integracji (rysunek 11 – rozdział 4) za pomocą trzech wzorców łącznie z tekstów wyekstrahowano 214 rzeczowych informacji z 200 zamówień z BZP zawierających różne formy fleksyjne wyrazów. Następnie rzeczowe informacje zgodnie z rysunkami 11 i 12 poddano analizie fleksyjnej. Na podstawie obliczonych miar skojarzeniowych wyeliminowano rzeczowe informacje zawierające najmniej poprawne powiązania form fleksyjne wyrazów. W tym celu eksperymentalnie został dobrany próg, czyli

wartość graniczną miary skojarzeniowej, która umożliwiła uwzględnienie lub odrzucenie rzeczowych informacji w dalszej części procedury eksploracji. Przykładowy fragment listy skojarzeniowej dla pierwszego wzorca informacyjnego z rysunku 34 przedstawiono w tabeli 52.

Tabela 52. Lista skojarzeniowa dla trójki: podmiot – *zbyć*, właściwość – *co*, obiekt – *kontrakt*.

Lp.	Podmiot	Obiekt	cw	lw	Sk [%]
1	*realizacja*	*procesów*	18	32	0,563
2		*procesu*	9		0,281
3		*procesem*	3		0,094

Źródło: opracowanie własne

Po przeprowadzeniu eliminacji najmniej poprawnych powiązań form fleksyjnych wyrazów pozostało 185 rzeczowych informacji będących elementami γ-gramowej reprezentacji dokumentów tekstowych w modelu przestrzeni wektorowej, stanowiącej zbiór $Z'_{T.}$

W etapie 3 procedury z rysunku 11 dokonano transformacji danych ze zbioru Z'_T do Z''_T za pomocą właściwej eksploracji danych tekstowych w modelu przestrzeni wektorowej VSM. W celu realizacji tego etapu przez eksperta dziedzinowego dobrane zostały odpowiednie techniki eksploracji danych tekstowych. Są to techniki opisane w rozdziale 4.2. W efekcie działania tego etapu został utworzony zbiór danych Z''_T, który w kolejnym etapie procedury integracji został poddany dyskretyzacji.

W ramach etapu czwartego procedury integracji (Rysunek 11 – *Etap 4*) dane ze zbiorów Z_N oraz Z''_T poddano dyskretyzacji według równej ilości obiektów w przedziałach oraz doborowi wartości nominalnych, co zaprezentowano w tabelach od 53 do 56.

Tabela 53. Zamiana wartości ciągłych w dyskretne wartości lingwistyczne oraz formę kodową dla atrybutu warunkowego – *wielkość sprzedaży.*

Liczba mieszkańców		
Wartość ciągła [osoby]	**Wartość lingwistyczna**	**Forma zakodowana**
Poniżej 461531	mała	1
461531-1111704	średnia	2
powyżej 1111704	duża	3

Źródło: opracowanie własne

Tabela 54. Zamiana wartości lingwistycznych na formę kodową dla atrybutu warunkowego – *stanowisko pracy.*

Stanowisko pracy	
Wartość lingwistyczna	**Forma zakodowana**
Stażysta	1
Specjalista	2
Kierownik	3

Źródło: opracowanie własne

Tabela 55. Zamiana wartości lingwistycznych w formę kodową dla atrybutu decyzyjnego warunkowego – *rentowność zamówienia*.

Atrakcyjność	
Wartość lingwistyczna	**Forma zakodowana**
TAK	1
NIE	2

Źródło: opracowanie własne

W przypadku wyniku eksploracji danych tekstowych przedziały zostały zdefiniowane przez eksperta (Tabela 65).

Tabela 56. Zamiana wartości lingwistycznych w formę kodową dla atrybutu warunkowego – *wynik eksploracji danych tekstowych.*

Wynik eksploracji danych tekstowych		
Wartość ciągła [podobieństwo]	**Wartość lingwistyczna**	**Forma zakodowana**
Poniżej 0,5	Klasa I	1
0,5 i powyżej	Klasa II	2

Źródło: opracowanie własne

Ostatecznie w etapie 5 procedury z rysunku 11 opracowano reprezentację danych numerycznych (zbiór Z_E) w postaci systemu informacyjnego – tablicy decyzyjnej, której fragment zawiera się w tabeli 57.

W etapie 6 procedury integracji z rysunku 11 przeprowadzono analizę istotności danych ze zbioru Z_E przy użyciu współczynników wynikających z metody Teorii Zbiorów Przybliżonych. W systemie informacyjnym *SI* uwzględniono 22-elementowy zbiór przypadków (obiektów). Bazując na wartościach atrybutów poszczególnych przypadków otrzymano tabele informacyjną w formie zakodowanej (tabela 58), w której wyodrębniono elementarne zbiory warunkowe E_i oraz zbiory decyzyjne (koncepty) X_i.

Tabela 57. Fragmentaryczna reprezentacja danych numerycznych w postaci tablicy decyzyjnej

Lp.	Atrybuty warunkowe			Atrybut decyzyjny
	Stanowisko pracy (q_1)	Liczba ludności (q_2)	Wynik eksploracji danych tekstowych (q_3)	Atrakcyjność
1	3	3	1	1
2	3	2	1	1
3	3	2	1	1
4	3	1	1	1
5	2	3	1	1
6	3	1	1	1
…	…	…	…	…

Źródło: opracowanie własne

Tabela 58. Tabela informacyjna z przekształconymi wartościami atrybutów do formy zakodowanej.

Lp.	Atrybuty warunkowe			Zbiory elementarne	Atrybut decyzyjny	Koncept
	Stanowisko pracy (q_1)	Liczba ludności (q_2)	Wynik eksploracji danych tekstowych (q_3)	*Eie*	Atrakcyjność	*Xix*
1	3	3	1	E_1	1	X_1
2	3	2	1	E_2	1	X_1
3	3	2	1	E_2	1	X_1
4	3	1	1	E_3	1	X_1
5	2	3	1	E_4	1	X_1
6	3	1	1	E_3	1	X_1
7	3	3	1	E_1	1	X_1
8	2	3	1	E_4	1	X_1
9	3	3	1	E_1	1	X_1
10	2	3	1	E_4	1	X_1
11	3	3	1	E_1	1	X_1
12	2	3	1	E_4	2	X_2
13	3	2	2	E_6	2	X_2
14	2	3	2	E_5	2	X_2
15	3	1	2	E_7	2	X_2
16	2	3	2	E_5	2	X_2
17	2	3	2	E_5	2	X_2
18	3	3	2	E_8	2	X_2
19	2	1	2	E_9	2	X_2
20	2	3	2	E_5	2	X_2
21	3	3	2	E_8	2	X_2
22	2	1	2	E_9	2	X_2

Źródło: opracowanie własne

gdzie:

q_1, q_2, q_3 – atrybuty warunkowe,
E_{ie} – elementarne zbiory warunkowe odpowiadające klasom abstrakcji,
ie – indeks zbioru warunkowego,
X – koncept decyzyjny,
ix – indeks konceptu decyzyjnego.

Poszczególne przypadki zostały przyporządkowane do odpowiednich elementarnych zbiorów warunkowych zaprezentowanych w *(132)*.

(132)

E_1={1, 7, 9, 11}, E_2={2, 3}, E_3={4, 6}, E_4={5, 8, 10, 12}, E_5={14, 16, 17, 20}, E_6={13}, E_7={15}, E_8={18, 21}, E_9={19, 22},

Kolejno przeprowadzono obliczenia, w celu wyznaczenia dolnych przybliżeń konceptów decyzyjnych $\underline{B}X_{ix}$ oraz liczby przypadków $card(\underline{B}X_{ix})$ należących do dolnych przybliżeń, wzory *(133)* oraz *(134)*.

(133)

X_1 = {1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11}
$\underline{B}X_1 = E_1 \cup E_2 \cup E_3$ = {1, 2 ,3, 4, 6, 7, 9, 11}
$card(\underline{B}X_1)$ = 8

(134)

X_2 = {12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22}
$\underline{B}X_2$ = E5 ∪ E6 ∪ E7 ∪ E8 ∪ E9 = {13, 14, 15, 16, 17, 18, 19, 20, 21, 22}
$card(\underline{B}X_2)$ = 10

Na podstawie liczby przypadków należących do dolnych przybliżeń wyznaczono pozytywny obszar rodziny konceptów decyzyjnych $POS_B(X)$, zgodny ze wzorem (135).

(135)

$POS_B(X) = E_1 \cup E_2 \cup E_3 \cup E_5 \cup E_6 \cup E_7 \cup E_8 \cup$ E9 = {1, 2 ,3, 4, 6, 7, 9, 11, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22}

Ilość wszystkich przypadków w pozytywnym obszarze card($POS_B(X)$) jest równa 18. Jakość przybliżenia rodziny konceptów decyzyjnych $\gamma_B(X)$ jest zatem równa wartości 0,82 co oznacza, że 82% przypadków wykorzystanych w obliczeniach pozwala na generowanie reguł pewnych. Następnie przeprowadzono obliczenia, które zostały wykorzystane do

określenia górnych przybliżeń konceptów $\overline{B}X_{ix}$ oraz ilości przypadków $card(\overline{B}X_{ix})$ należące do górnych przybliżeń, wzory *(136)* i *(137)*.

(136)

X_1 = {1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11}

$\overline{B}X_1 = E_1 \cup E_2 \cup E_3 \cup E_4$ = {1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12}

$card(\overline{B}X_1)$ = 12

(137)

X_2 = {12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22}

$\overline{B}X_2 = E_4 \cup E_5 \cup E_6 \cup E_7 \cup E_8 \cup E_9$ = {5, 8, 10, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22}

$card(\overline{B}X_2)$ = 14

Na podstawie obliczeń otrzymano dokładności przybliżeń kolejnych konceptów decyzyjnych równe $\mu(X_1)$ = 0,67, $\mu(X_2)$ = 0,71. Dokładność przybliżenia rodziny konceptów decyzyjnych $\beta_B(X)$, czyli przeciętny stopień zrozumienia decyzji wyniósł 0,69.

W dalszej części badań testowych wyliczono istotność kolejnych atrybutów warunkowych q_1, q_2, q_3, zgodnie ze wzorem *(77)*. Wyniki obliczeń przedstawiono w tabeli 59. Na podstawie wyznaczonych istotności poszczególnych atrybutów warunkowych stwierdzono, że każdy z atrybutów ma znaczący wpływ na definicję wiedzy wynikającą z systemu informacyjnego *SI*. Istotność szczególnie ważnego ze względu na integrację atrybutu q_3 *– wynik eksploracji danych tekstowych,* wynosi 0,89. Oznacza to, że znacząco wpływa on na wiedzę w postaci zbioru reguł wygenerowanych na podstawie danych reprezentowanych przez system informacyjny *SI* i potwierdza zasadność uwzględnienie w systemie *SI* atrybutu wynikającego z eksploracji danych tekstowych.

Tabela 59. Jakość i dokładność przybliżenia konceptów decyzyjnych *X* oraz istotności usuwanego atrybutu warunkowego.

Atrybut warunkowy	**Istotność atrybutu**	**Dokładność przybliżenia rodziny konceptów $\beta_B(X)$**	**Jakość przybliżenia rodziny konceptów $\gamma_B(X)$**
q_1	0,18	0,47	0,67
q_2	0,06	0,42	0,77
q_3	0,89	0,05	0,09

Źródło: opracowanie własne

W wyniku realizacji etapu 7 procedury z rysunku 11 został zbudowany algorytm decyzyjny, który opiera się na wyodrębnionych regułach decyzyjnych. Dla reguły pewnych

oraz niepewnych obliczono wsparcie i współczynnik pewności, które przedstawiono tabelach 60 i 61.

Tabela 60. Reguły pewne R*eg* algorytmu decyzyjnego w formie tabeli informacyjnej.

Reguła (*Reg_{ir}*)	**Atrybuty warunkowe**			**Atrybut decyzyjny**	**Pewność**	**Wsparcie**
	Stanowisko pracy (*q_1*)	**Liczba ludności (*q_2*)**	**Wynik eksploracji danych tekstowych (*q_3*)**	**Atrakcyjność**		
Reg_1	3	3	1	1	1	4
Reg_2	3	2	1	1	1	2
Reg_3	3	1	1	1	1	2
Reg_4	2	3	2	2	1	4
Reg_5	3	2	2	2	1	1
Reg_6	3	1	2	2	1	1
Reg_7	3	3	2	2	1	2
Reg_8	2	1	2	2	1	2

Źródło: opracowanie własne

gdzie:

q_1, *q_2*, *q_3* – atrybuty warunkowe,
Reg_{ir} – reguła decyzyjna,
ir – indeks reguły decyzyjnej.

Tabela 61. Reguły niepewne *Reg* algorytmu decyzyjnego w formie tabeli informacyjnej.

Reguła (*Reg_{ir}*)	**Atrybuty warunkowe**			**Atrybut decyzyjny**	**Pewność**	**Wsparcie**
	Stanowisko pracy (*q_1*)	**Liczba ludności (*q_2*)**	**Wynik eksploracji danych tekstowych (*q_3*)**	**Atrakcyjność**		
Reg_1	2	3	1	1	0,75	4

Źródło: opracowanie własne

gdzie:

q_1, *q_2*, *q_3* – atrybuty warunkowe,
Reg_{ir} – reguła decyzyjna,
ir – indeks reguły decyzyjnej.

Formalny zapis zbioru reguł pewnych określono wzorem *(138).*

(138)

Reg_1: IF ((q_1 = 3) AND (q_2 = 3) AND (q_3 = 1)) THEN d = 1

Reg_2: IF ((q_1 = 3) AND (q_2 = 2) AND (q_3 = 1)) THEN d = 1

Reg_3: IF ((q_1 = 3) AND (q_2 = 1) AND (q_3 = 1)) THEN d = 1

Reg_4: IF ((q_1 = 2) AND (q_2 = 3) AND (q_3 = 2)) THEN d = 2

Reg_5: IF ((q_1 = 3) AND (q_2 = 2) AND (q_3 = 2)) THEN d = 2

Reg_6: IF ((q_1 = 3) AND (q_2 = 1) AND (q_3 = 2)) THEN d = 2

Reg_7: IF ((q_1 = 3) AND (q_2 = 3) AND (q_3 = 2)) THEN d = 2

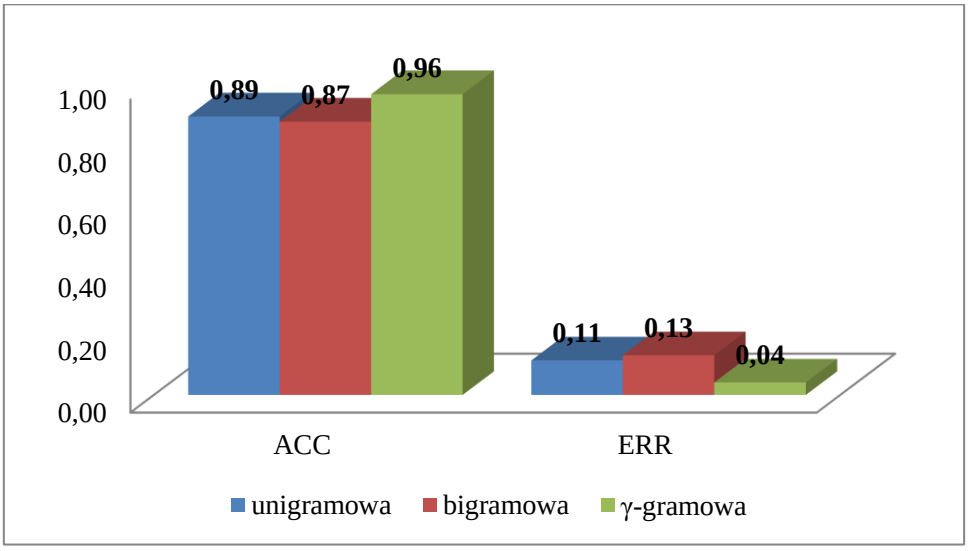
Reg_8: IF ((q_1 = 2) AND (q_2 = 1) AND (q_3 = 2)) THEN d = 2

Formalny zapis reguły niepewnej określono wzorem *(1139)*.

(139)

Reg_1: IF ((q_1 = 2) AND (q_2 = 3) AND (q_3 = 1)) THEN d = 1

Ostatecznie zbiór 200 przypadków zamówień publicznych poddano klasyfikacji w etapie 8 procedury integracji z rysunku 11. Następnie obliczono miary jakości decyzji (ACC oraz ERR) dotyczących sklasyfikowania 200 przypadków do dwóch kategorii tj. do kategorii reprezentującej rentowne zamówienia publiczne oraz do kategorii z pozostałymi zamówieniami. Wyniki obliczeń w postaci średniej arytmetycznej poszczególnych miar jakości (ACC, ERR) dla 11 wartości współczynnika k klasyfikatora kNN, zaprezentowano na rysunku 37.



Rysunek 37. Średnie wartości miar jakości decyzji (ACC oraz ERR) dla trzech reprezentacji danych tekstowych (unigramowej, bigramowej oraz γ-gramowej) w przypadku eksploracji danych tekstowych i numerycznych (wariant A z rysunku 2)
Źródło: opracowanie własne

W celu realizacji wariantu B z rysunku 2 eksploracji z zastosowaniem wyłącznie metod eksploracji danych numerycznych pominięte zostały etapy 1, 2, 3 procedury integracji z rysunku 11. Do eksploracji danych numerycznych wykorzystano reprezentację danych opracowaną, w etapach 4 i 5 procedury z rysunku 2, dla wariantu A badań testowych, które przedstawiono w tabeli 57. Na podstawie tej reprezentacji danych numerycznych, jednak bez wykorzystania atrybutu stanowiącego wynik eksploracji danych tekstowych, opracowano tablicę decyzyjną zawierającą się w tabeli 62.

Tabela 62. Tabela informacyjna z przekształconymi wartościami atrybutów do formy zakodowanej.

Lp.	Atrybuty warunkowe		Elementarne zbiory warunkowe	Atrybut decyzyjny	Koncept
	Stanowisko pracy (q_1)	Liczba ludności (q_2)	*Eie*	Atrakcyjność	*Xix*
1	3	3	E_1	1	X_1
2	3	2	E_2	1	X_1
3	3	2	E_2	1	X_1
4	3	1	E_3	1	X_1
5	2	3	E_4	1	X_1
6	3	1	E_3	1	X_1
7	3	3	E_1	1	X_1
8	2	3	E_4	1	X_1
9	3	3	E_1	1	X_1
10	2	3	E_4	1	X_1
11	3	3	E_1	1	X_2
12	2	3	E_4	2	X_2
13	3	2	E_2	2	X_2
14	2	3	E_4	2	X_2
15	3	1	E_3	2	X_2
16	2	3	E_4	2	X_2
17	2	3	E_4	2	X_2
18	3	3	E_1	2	X_2
19	2	1	E_5	2	X_2
20	2	3	E_4	2	X_2
21	3	3	E_1	2	X_2
22	2	1	E_5	2	X_2

Źródło: opracowanie własne

gdzie:

q_1, q_2, q_3 – atrybuty warunkowe,
E_{ie} – elementarne zbiory warunkowe odpowiadające klasom abstrakcji,
ie – indeks zbioru warunkowego,
X – koncept decyzyjny,
ix – indeks konceptu decyzyjnego.

W etapie 6 procedury integracji z rysunku 11 przeprowadzono analizę istotności danych ze zbioru Z_E przy użyciu współczynników wynikających z metody Teorii Zbiorów Przybliżonych. W tym celu Poszczególne przypadki zostały przyporządkowane do odpowiednich elementarnych zbiorów warunkowych zaprezentowanych w *(140)*.

(140)

E_1={1, 7, 9, 11, 18, 21}, E_2={2, 3, 13}, E_3={4, 6, 15}, E_4={5, 8, 10, 12, 14, 16, 17, 20}, E_5={19, 22},

Kolejno przeprowadzono obliczenia, które zostały wykorzystane do określenia dolnych przybliżeń konceptów decyzyjnych $\underline{B}X_{ix}$ oraz ilości przypadków $card(\underline{B}X_{ix})$ należących do dolnych przybliżeń, wzory *(141)* oraz *(142)*.

(141)

X_1 = {1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11}

$\underline{B}X_1$= {}

$card(\underline{B}X_1)$ = 0

(142)

X_2 = {12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22}

$\underline{B}X_2$ = E_5 = {19, 22}

$card(\underline{B}X_2)$ = 2

Na podstawie ilości przypadków należących do dolnych przybliżeń wyznaczono pozytywny obszar rodziny konceptów decyzyjnych $POS_B(X)$, zgodny ze wzorem *(143)*.

(143)

$POS_B(X)$ = E_5 = {19, 22}

Ilość wszystkich przypadków w pozytywnym obszarze card($POS_B(X)$) jest równa 2. Jakość przybliżenia rodziny konceptów decyzyjnych jest zatem równa wartości 0,09 co oznacza, że 9% przypadków wykorzystanych w obliczeniach pozwala na generowanie reguł pewnych. Następnie przeprowadzono obliczenia, które zostały wykorzystane do określenia górnych przybliżeń konceptów $\overline{B}X_{ix}$ oraz ilości przypadków $card(\overline{B}X_{ix})$ należące do górnych przybliżeń, wzory *(144)* oraz *(145)*.

(144)

X_1 = {1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11}

$\overline{B}X_1$ = $E_1 \cup E_2 \cup E_3 \cup E_4$ = {1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 20, 21}

$card(\overline{B}X_1)$ = 20

(145)

X_2 = {12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22}

$\overline{B}X_2 = E_1 \cup E_2 \cup E_3 \cup E_4 \cup E_5$ = {1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22}

$card(\overline{B}X_2) = 22$

Na podstawie obliczeń otrzymano dokładności przybliżeń kolejnych konceptów decyzyjnych równe $\mu(X_1) = 0$, $\mu(X_2) = 0{,}09$. Dokładność przybliżenia rodziny konceptów decyzyjnych $\beta_B(X)$, czyli przeciętny stopień zrozumienia decyzji wyniósł 0,05.

W dalszej części badań testowych wyliczono istotność kolejnych atrybutów warunkowych q_1,q_2,q_3, zgodnie ze wzorem *(77)*. Wyniki obliczeń przedstawiono w tabeli 63.

Tabela 63. Jakość i dokładność przybliżenia konceptów decyzyjnych *X* oraz istotności usuwanego atrybutu warunkowego.

Atrybut warunkowy	**Istotność atrybutu**	**Dokładność przybliżenia rodziny konceptów $\beta_B(X)$**	**Jakość przybliżenia rodziny konceptów $\gamma_B(X)$**
q_1	1	0	0
q_2	1	0	0

Źródło: opracowanie własne

Na podstawie wyznaczonych istotności poszczególnych atrybutów warunkowych stwierdzono, że każdy z atrybutów ma wpływ na definicję wiedzy na podstawie informacji zawartych w systemie informacyjnych – tablicy decyzyjnej zaprezentowanej w tabeli 62.

W wyniku realizacji etapu 7 z rysunku 11 został zbudowany algorytm decyzyjny, który opiera się na wyodrębnionych regułach decyzyjnych.. Dla reguły pewnych oraz niepewnych obliczono wsparcie i współczynnik pewności, które przedstawiono tabelach 64 oraz 65.

Formalny zapis reguły pewnej określono wzorem *(146)*.

(146)

Reg_1: IF ((q_1 = 2) AND (q_2 = 1)) THEN d = 2

Tabela 64. Reguły pewne R*eg* algorytmu decyzyjnego w formie tabeli informacyjnej.

Reguła (*Reg_{ir}*)	**Atrybuty warunkowe**		**Atrybut decyzyjny**	**Pewność**	**Wsparcie**
	Stanowisko pracy (q_1)	**Liczba ludności (q_2)**	**Atrakcyjność**		
Reg_1	2	1	2	1	2

Źródło: opracowanie własne

gdzie:

q_1, q_2 – atrybuty warunkowe,
Reg_{ir} – reguła decyzyjna,
ir – indeks reguły decyzyjnej.

Tabela 65. Reguły niepewne R*eg* algorytmu decyzyjnego w formie tabeli informacyjnej.

Reguła (*Reg*$_{ir}$)	**Atrybuty warunkowe**		**Atrybut decyzyjny**	**Pewność**	**Wsparcie**
	Stanowisko pracy (q_1)	**Liczba ludności (q_2)**	**Atrakcyjność**		
Reg_1	3	3	1	0,66	6
Reg_2	3	2	1	0,66	3
Reg_3	3	1	1	0,66	3
Reg_4	2	3	2	0,63	8

Źródło: opracowanie własne

gdzie:

q_1, q_2 – atrybuty warunkowe,
Reg_{ir} – reguła decyzyjna,
ir – indeks reguły decyzyjnej.

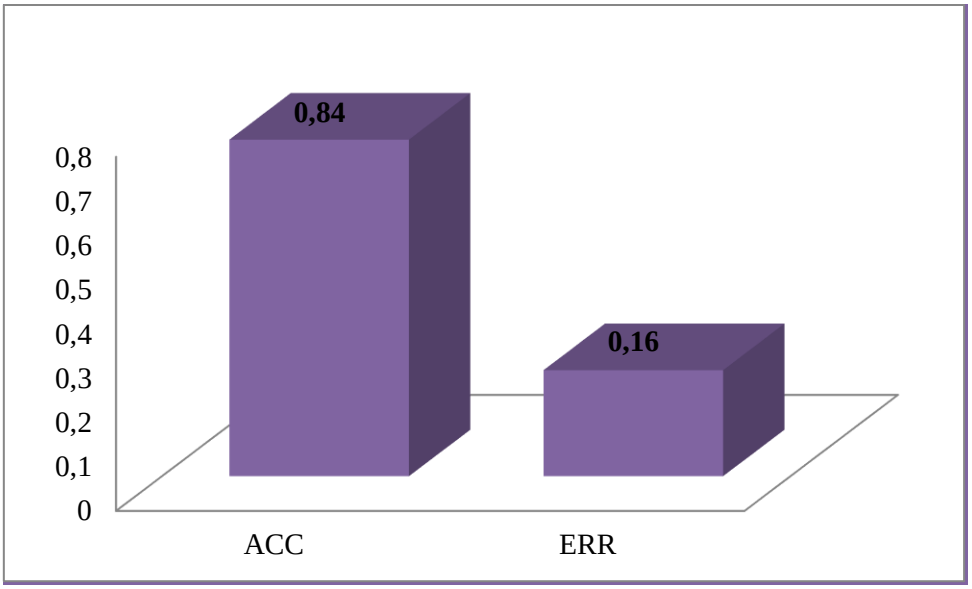
Formalny zapis zbioru reguł niepewnych określono wzorem *(147).*

(147)

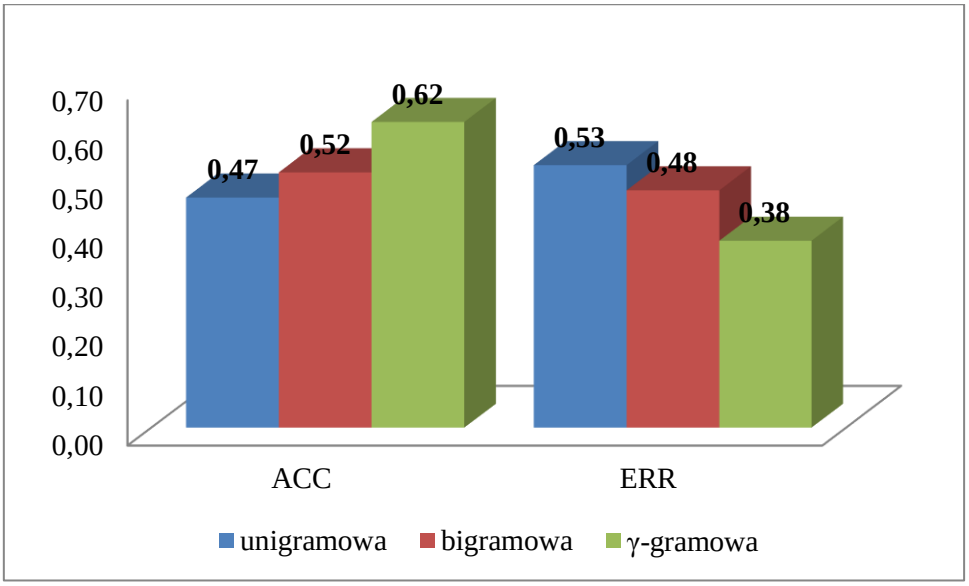
Reg_1: IF ((q_1 = 3) AND (q_2 = 3)) THEN d = 1
Reg_2: IF ((q_1 = 3) AND (q_2 = 2)) THEN d = 1
Reg_3: IF ((q_1 = 3) AND (q_2 = 1)) THEN d = 1
Reg_4: IF ((q_1 = 2) AND (q_2 = 3)) THEN d = 2

Ostatecznie zbiór 200 przypadków zamówień publicznych poddano klasyfikacji w etapie 8 procedury integracji z rysunku 11. Wyniki eksploracji danych numerycznych w postaci miar jakości decyzji (ACC, ERR) przedstawiono na rysunku 38.

W celu realizacji wariantu C z rysunku 2 eksploracji z zastosowaniem wyłącznie metod eksploracji danych tekstowych pominięte zostały etapy 4, 5, 6, 7 i 8 procedury integracji z rysunku 11. Do eksploracji wyłącznie danych tekstowych wykorzystano reprezentacje danych opracowane w pierwszym wariancie eksploracji (wariant A z rysunku 2). Zgodnie z wybranymi technikami eksploracji danych tekstowych w wariancie z wykorzystaniem zintegrowanych metod eksploracji danych tekstowych i numerycznych (wariant A z rysunku 2) dokonano eksploracji danych tekstowych, a wyniki w postaci miar jakości decyzji przedstawiono na rysunkach 39.



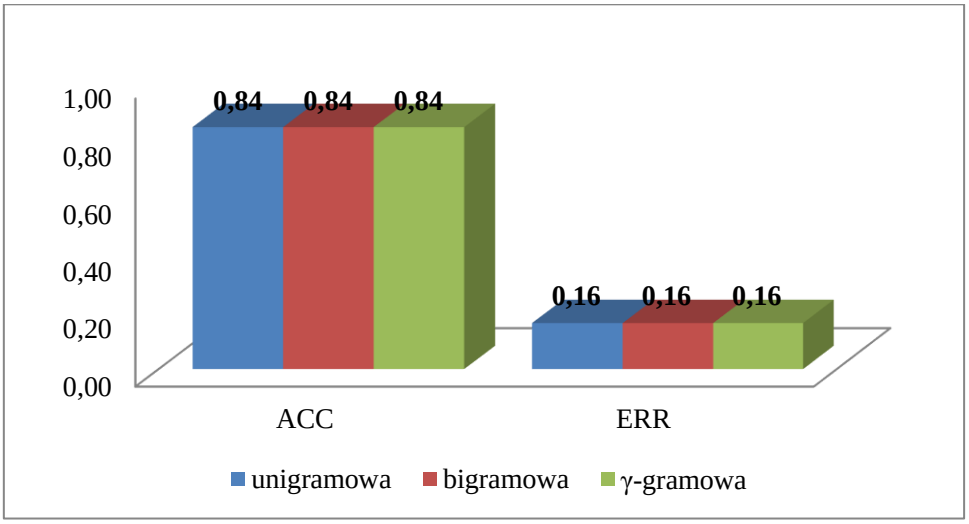
Rysunek 38. Średnie wartości miar jakości decyzji (ACC, ERR) w przypadku eksploracji danych numerycznych (wariant B z rysunku 2)
Źródło: opracowanie własne



Rysunek 39. Wartości miar jakości (ACC oraz ERR) dla trzech reprezentacji danych tekstowych (unigramowej, bigramowej oraz γ-gramowej) dla wariantu eksploracji z wykorzystaniem integracji metod eksploracji danych tekstowych (wariant C z rysunku 2)
Źródło: opracowanie własne

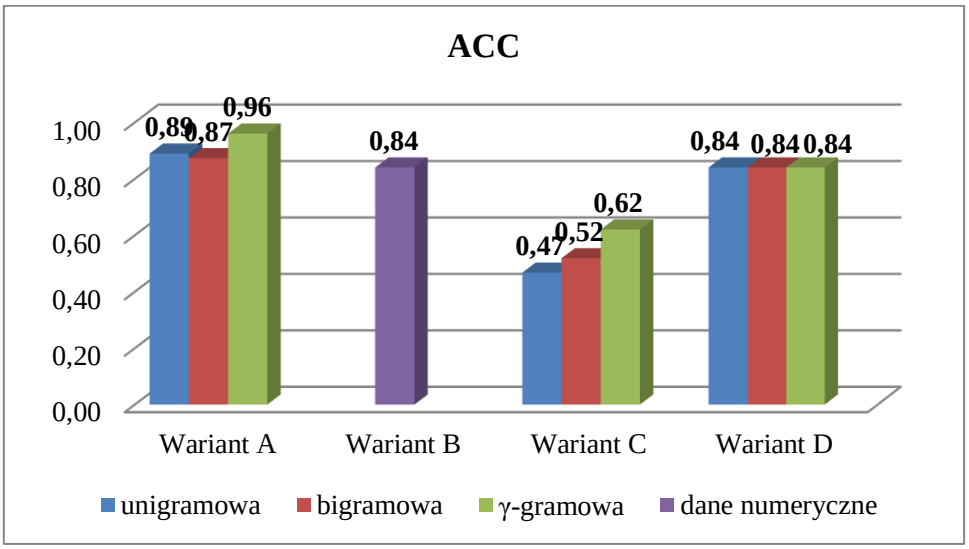
Wariant D z rysunku 2 dotyczy metody integracji wyników oddzielnych eksploracji danych numerycznych (wariant B z rysunku 2) oraz eksploracji danych tekstowych (wariant C z rysunku 2). Integracja wyników eksploracji w tym wariancie polega na wyborze korzystniejszego (z wyższą miarą ACC i niższą miarą ERR) wariantu eksploracji z pośród wariantów B i C dla poszczególnych poziomów zwrotu (od 1 do 11). Następnie miary jakości

decyzji zostały uśrednione dla wszystkich 11 poziomów zwrotu, tak jak to przebiegało w poprzednich wariantach A, B oraz C. Wyniki w postaci miar jakości decyzji przedstawiono na rysunkach 40.

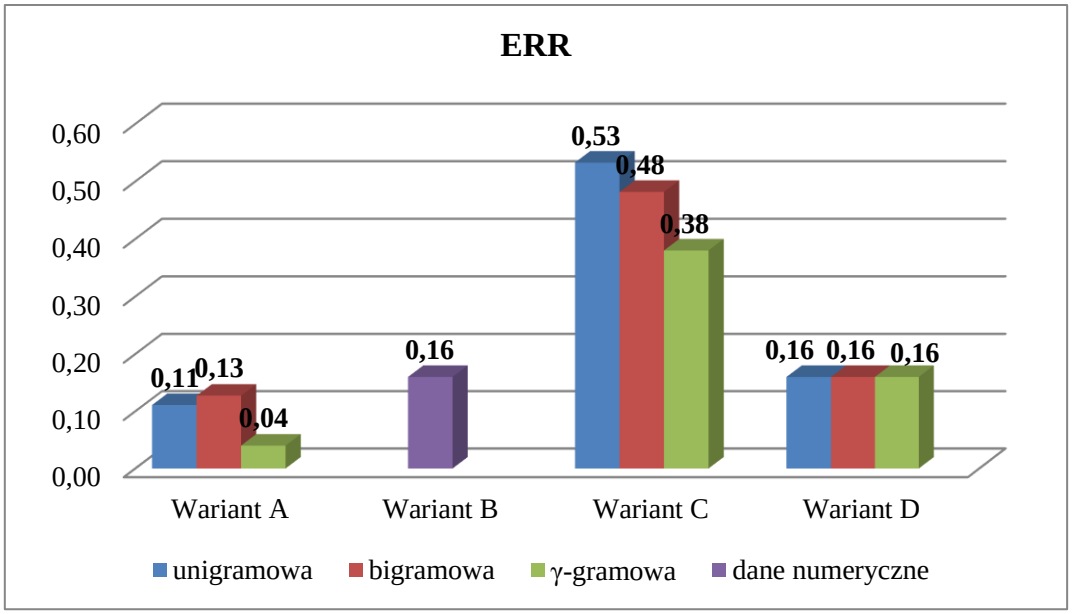


Rysunek 40. Wartości miar jakości (ACC oraz ERR) dla trzech reprezentacji danych tekstowych (unigramowej, bigramowej oraz γ-gramowej) dla metody integracji wyników eksploracji danych uzyskanych w wariancie B i C (wariant D z rysunku 2)
Źródło: opracowanie własne

Reasumując w wyniku obliczeń parametrów jakości kwalifikacji w rozpatrywanym przykładzie uzyskano wyniki zgodne z rysunkami 41 i 42.



Rysunek 41. Średnie wartości miar jakości decyzji ACC osiągnięte dla III przykładu procesu *PD*
Źródło: opracowanie własne



Rysunek 42. Średnie wartości miar jakości decyzji ERR osiągnięte dla III przykładu procesu *PD*
Źródło: opracowanie własne

6. Dyskusja wyników badań i weryfikacja hipotezy

W pierwszym etapie badań dla wszystkich trzech przykładowych procesów *PD*, opisanych w rozdziałach 5.2, 5.3 i 5.4, zgodnie z etapami 1 i 2 procedury integracji metod eksploracji danych tekstowych i numerycznych z rysunku 11, została opracowana reprezentacja danych tekstowych. Do przygotowania γ-gramowej reprezentacji danych tekstowych wykorzystano wzorce informacyjne zdefiniowane przez eksperta dziedzinowego. Przy użyciu wzorców informacyjnych w etapie 2 procedury z rysunku 11 z dokumentów tekstowych zostały wyekstrahowane rzeczowe informacje, które następnie poddano weryfikacji za pomocą analizy fleksyjnej polegającej na sprawdzeniu poprawności form fleksyjnych wyrazów zawartych w poszczególnych rzeczowych informacjach. W efekcie analizy fleksyjne dokonano istotnej redukcji elementów reprezentacji, które generowały tzw. szum informacyjny. Wyeliminowane elementy reprezentacji ze względu na niepoprawne powiązania form fleksyjnych wyrazów stanowiły błędnie wyekstrahowane rzeczowe informacje. Różnice wynikające z liczby elementów reprezentacji przed i po eliminacji niepoprawnych elementów reprezentacji przedstawiono w tabeli 66.

Tabela 66. Weryfikacja za pomocą analizy fleksyjnej elementów reprezentacji γ-gramowej – wyekstrahowanych za pomocą wzorców rzeczowych informacji.

Przykładowe procesy PD	Ilość wyekstrahowanych rzeczowych informacji	Ilość rzeczowych informacji po weryfikacji (eliminacja informacji na podstawie analizy fleksyjnej)
I	168	91
II	32	24
III	214	185

Źródło: opracowanie własne

W etapie 3 procedury z rysunku 11 dokonano transformacji danych ze zbioru Z'_T do zbioru Z''_T za pomocą właściwej eksploracji danych tekstowych w modelu przestrzeni wektorowej.

W etapie 4 procedury integracji (Rysunek 15 – *Etapy 4*) dane ze zbiorów Z_N oraz Z''_T poddano dyskretyzacji oraz doborowi wartości nominalnych, a następnie w etapie 5 procedury z rysunku 11 uwzględniono je w systemie informacyjnym *SI*.

W etapie 6 z rysunku 11 określono istotność danych uwzględnionych w systemie informacyjnym *SI*. W szczególności zbadano istotność atrybutu wynik eksploracji danych tekstowych, co umożliwiło ocenę wpływu integracji metod eksploracji danych tekstowych i numerycznych na definicję wiedzy wykorzystywanej w procesie *PD*. W tym celu za pomocą

współczynnika istotności wynikającego z metody Teorii Zbiorów Przybliżonych określono istotność atrybutów warunkowych w poszczególnych przykładowych procesach *PD* (przykłady: I, II oraz III), która została przedstawiona w tabeli 67.

Tabela 67. Istotność atrybutów warunkowych z badań testowych.

Przykładowy proces *PD*	**Nazwa atrybutu**	**Istotność atrybutu**
I	Obszar do koszenia	0,61
	Odległość od siebie firmy	0,24
	Wynik eksploracji danych tekstowych	0,31
II	Cena sprzedaży	0,6
	Wielkość sprzedaży	0,8
	Wynik eksploracji danych tekstowych	1
III	Stanowisko pracy	0,18
	Liczba mieszkańców	0,06
	Wynik eksploracji danych tekstowych	0,89

Źródło: opracowanie własne

Na podstawie wyznaczonych istotności poszczególnych atrybutów warunkowych stwierdzono, że każdy z atrybutów miał duży wpływ na definicję wiedzy w postaci reguł decyzyjnych. Z obliczeń istotności atrybutów wynika, że w przykładach II i III procesu *PD*, wynik eksploracji danych tekstowych miał największy wpływ z pośród wszystkich atrybutów warunkowych na definicję wiedzy wykorzystywaną w procesie *PD*. Z badań wynika również, że jakość przybliżenia rodziny konceptów decyzyjnych wynosiła, dla kolejnych przykładowych procesów *PD* (przykłady: I, II oraz III): 59%, 45% i 82%, co oznacza, że taki procent przypadków (obiektów), wziętych pod uwagę przy definiowaniu wiedzy, generuje reguły pewne. Z kolei dokładność przybliżenia rodziny konceptów decyzyjnych wyniosła dla kolejnych procesów *PD* (przykłady: I, II oraz III): 42%, 29%, 69%, co świadczy o stopniu zrozumienia decyzji wynikających z ze zbioru analizowanych przypadków (obiektów), na podstawie których wygenerowane zostały reguły decyzyjne.

Następnie w etapie 7 procedury integracji z rysunku 11, na podstawie wartości atrybutów warunkowych, dokonano ekstrakcji reguł decyzyjnych.

W ostatnim etapie 8 procedury z rysunku 11 dokonano klasyfikacji obiektów z wykorzystaniem danych testowych.

Wyniki badań testowych dotyczące przykładowych procesów *PD* (przykłady: I, II oraz III) w postaci poszczególnych średnich wartości miar jakości decyzji (*ACC, ERR*) osiągniętych dla czterech wariantów eksploracji, zgodnych z rysunkiem 2, tj.:

1. zintegrowanej eksploracji danych tekstowych i numerycznych (Wariant A),

2. eksploracji wyłącznie danych numerycznych (Wariant B),
3. eksploracja wyłącznie danych tekstowych (Wariant C),
4. metody integracji wyników eksploracji danych uzyskanych w wariancie B i C (Wariant D),

z uwzględnieniem trzech różnych reprezentacji danych tekstowych (reprezentacji unigramowej, bigramowej i γ-gramowej) zostały porównane i zaprezentowane na rysunkach 23, 24, 32, 33, 41 oraz 42.

Z porównań wynika, że w przypadku każdej reprezentacji (unigramowej, bigramowej i γ-gramowej) osiągnięte wartości miar jakości decyzji (*ACC, ERR*) są najkorzystniejsze w przypadku autorskiej procedury eksploracji z wykorzystaniem integracji metod eksploracji danych tekstowych i numerycznych zaprezentowanej w rozdziale 4 (Wariant A z rysunku 2). Procentową różnicę pomiędzy wartościami miar jakości decyzji *ACC* oraz *ERR* osiągniętymi dla tego wariantu (wariant A z rysunku 2), a wartościami tych miar jakości decyzji uzyskanymi dla pozostałych wariantów eksploracji (warianty B, C oraz D z rysunku 2), przedstawiono w tabelach 68, 69 oraz 70. W porównaniu wykorzystano współczynniki Δ_{ACC} oraz Δ_{ERR} wyrażone procentowo, zgodne ze wzorami *(148)* oraz *(149)*.

$$\Delta_{ACC} = ACC_{WN} - ACC_{WM} = 1 - \Delta_{ERR} \qquad (148)$$

lub

$$\Delta_{ERR} = ERR_{WM} - ERR_{WN} = 1 - \Delta_{ACC} \qquad (149)$$

gdzie:

ACC_{WN} , ACC_{WM} – współczynnik całkowitej dokładności dla wybranego wariantu *WN* oraz *WM* z rysunku 2 (wariant A, B, C, D),

ERR_{WM} , ACC_{WN} – współczynnik całkowitego poziomu błędu dla wybranego wariantu *WM* oraz *WN* z rysunku 2 (wariant A, B, C, D).

Na podstawie zestawień z tabel od 68 do 70 można jednoznacznie stwierdzić, że wariant A z rysunku 2 tj. eksploracji danych z użyciem procedury z rozdziału 4 integracji metod eksploracji danych tekstowych i numerycznych we wszystkich przykładowych procesach *PD* (przykłady I, II oraz III) jest najkorzystniejszy, a wartości miar jakości decyzji (*ACC i ERR*) w przypadku tego wariantu eksploracji są znacząco wyższe dla miary *ACC* oraz niższe dla miary

ERR w stosunku do wartości miar jakości uzyskany w pozostałych wariantach eksploracji (warianty B, C oraz D z rysunku 2). Z kolei przy zastosowaniu wariantu A z rysunku 2, najwyższa nośność informacyjna danych wyrażona miarami jakości decyzji *ACC* oraz *ERR* została osiągnięta przy wykorzystaniu γ-gramowej reprezentacji danych tekstowych. Porównanie z uwzględnieniem współczynnika Δ pomiędzy wartościami miar jakości decyzji dla reprezentacji γ-gramowej w stosunku do pozostałych reprezentacji danych tekstowych w wariancie zintegrowanej eksploracji danych tekstowych i numerycznych (wariant A z rysunku 2) i wariancie eksploracji wyłącznie danych tekstowych (wariant C z rysunku 2), zaprezentowano w tabelach 71, 72 oraz 73.

Tabela 68. Porównanie wartości miar jakości decyzji osiągniętych w 1 przykładzie procesu *PD* w wariancie A (metoda zgodna z procedurą z rozdziału 4) eksploracji z wartościami miar jakości decyzji uzyskanymi w wariantach B (eksploracja jedynie danych numerycznych), C (eksploracja tylko danych tekstowych), D (integracja wyników metody w wariancie B i C)

Reprezentacja (dotyczy wariantów, w których wykorzystywane są dane tekstowe)	Δ_{ACC} (WN=**A**, WM=**B**) [%]	Δ_{ERR} (WM=**B**, WN=**A**) [%]	Δ_{ACC} (WN=**A**, WM=**C**) [%]	Δ_{ERR} (WM=**C**, WN=**A**) [%]	Δ_{ACC} (WN=**A**, WM=**D**) [%]	Δ_{ERR} (WM=**D**, WN=**A**) [%]
unigramowa	15	15	19	19	15	15
n-gramowa	14	14	16	16	14	14
γ-gramowa	20	20	16	16	16	16

Źródło: opracowanie własne

Tabela 69. Porównanie wartości miar jakości decyzji osiągniętych w 2 przykładzie procesu *PD* w wariancie A (procedura z rozdziału 4) eksploracji z wartościami miar jakości decyzji uzyskanymi w wariantach B, C, D

Reprezentacja (dotyczy wariantów, w których wykorzystywane są dane tekstowe)	Δ_{ACC} (WN=**A**, WM=**B**) [%]	Δ_{ERR} (WM=**B**, WN=**A**) [%]	Δ_{ACC} (WN=**A**, WM=**C**) [%]	Δ_{ERR} (WM=**C**, WN=**A**) [%]	Δ_{ACC} (WN=**A**, WM=**D**) [%]	Δ_{ERR} (WM=**D**, WN=**A**) [%]
unigramowa	3	3	23	23	3	3
n-gramowa	12	12	36	36	12	12
γ-gramowa	16	16	25	25	16	16

Źródło: opracowanie własne

Tabela 70. Porównanie wartości miar jakości decyzji osiągniętych w 3 przykładzie procesu *PD* w wariancie A (metoda zgodna z procedurą z rozdziału 4) eksploracji z wartościami miar jakości decyzji uzyskanymi w wariantach B, C, D

Reprezentacja (dotyczy wariantów, w których wykorzystywane są dane tekstowe)	Δ_{ACC} (WN=**A**, WM=**B**) [%]	Δ_{ERR} (WM=**B**, WN=**A**) [%]	Δ_{ACC} (WN=**A**, WM=**C**) [%]	Δ_{ERR} (WM=**C**, WN=**A**) [%]	Δ_{ACC} (WN=**A**, WM=**D**) [%]	Δ_{ERR} (WM=**D**, WN=**A**) [%]
unigramowa	5	5	42	42	5	5
n-gramowa	3	3	35	35	3	3
γ-gramowa	12	12	34	34	12	12

Źródło: opracowanie własne

Tabela 71. Porównanie wartości miar jakości decyzji osiągniętych w 1 przypadku procesu *PD* z wykorzystaniem reprezentacji γ-gramowa w stosunku do wartości miar jakości decyzji uzyskanymi dla reprezentacji unigramowej i n-gramowej w wariantach A (procedura z rozdziału 4) i C (eksploracji tylko danych tekstowych)

Reprezentacja danych tekstowych	Δ_{ACC} (WN=**A**, WM=**A**) [%]	Δ_{ERR} (WM=**A**, WN=**A**) [%]	Δ_{ACC} (WN=**C**, WM=**C**) [%]	Δ_{ERR} (WM=**C**, WN=**C**) [%]
γ-gramowa dla WN i **unigramowa** dla WM	5	5	8	8
γ-gramowa dla WN i **n-gramowa** dla WM	6	6	6	6

Źródło: opracowanie własne

Z porównań umieszczonych w tabelach od 71 do 73 wynika znacząca przewaga jakości wyników uzyskanych dla reprezentacji γ-gramowej w stosunku do pozostałych reprezentacji danych wykorzystanych w wariancie eksploracji danych tekstowych i numerycznych (wariant A z rysunku 2 zgodny z procedurą z rozdziału 4) oraz wariancie eksploracji wyłącznie danych tekstowych (wariant C z rysunku 2).

Ostatecznie dokonano porównania najniższych wartości miar jakości decyzji (osiągniętych w przykładowych procesach *PD*: I, II oraz III) w wariancie eksploracji danych tekstowych i numerycznych (wariant A z rysunku 2 zgodny z procedurą z rozdziału 4)

w stosunku do najwyższych wartości miar jakości osiągniętych w pozostałych wariantach (warianty B, C oraz D z rysunku 2). Procentowe porównanie miar jakości decyzji zaprezentowano w tabeli 74.

Tabela 72. Porównanie wartości miar jakości decyzji osiągniętych w 2 przypadku procesu *PD* z wykorzystaniem reprezentacji γ-gramowa w stosunku do wartości miar jakości decyzji uzyskanymi dla reprezentacji unigramowej i n-gramowej w wariantach A (metoda zgodna z procedurą z rozdziału 4) i C eksploracji danych (tylko eksploracja danych tekstowych)

Reprezentacja danych tekstowych	Δ_{ACC} (WN=**A**, WM=**A**) [%]	Δ_{ERR} (WM=**A**, WN=**A**) [%]	Δ_{ACC} (WN=**C**, WM=**C**) [%]	Δ_{ERR} (WM=**C**, WN=**C**) [%]
γ-gramowa dla WN i **unigramowa** dla WM	13	13	11	11
γ-gramowa dla WN i **n-gramowa** dla WM	4	4	15	15

Źródło: opracowanie własne

Tabela 73. Porównanie wartości miar jakości decyzji osiągniętych w 3 przypadku procesu *PD* z wykorzystaniem reprezentacji γ-gramowa w stosunku do wartości miar jakości decyzji uzyskanymi dla reprezentacji unigramowej i n-gramowej w wariantach A (metoda zgodna z procedurą z rozdziału 4) i C (tylko eksploracja danych tekstowych)

Reprezentacja danych tekstowych	Δ_{ACC} (WN=**A**, WM=**A**) [%]	Δ_{ERR} (WM=**A**, WN=**A**) [%]	Δ_{ACC} (WN=**C**, WM=**C**) [%]	Δ_{ERR} (WM=**C**, WN=**C**) [%]
	ACC/ERR		ACC/ERR	
γ-gramowa dla WN i **unigramowa** dla WM	7	7	15	15
γ-gramowa dla WN i **n-gramowa** dla WM	9	9	10	10

Źródło: opracowanie własne

Na podstawie danych umieszczonych w tabeli 74 można stwierdzić, że nawet w przypadku najmniej korzystnej reprezentacji dla wariantu eksploracji danych tekstowych i numerycznych (wariant A z rysunku 2) we wszystkich przykładach procesów *PD* (przykładowe procesy *PD*: I, II oraz III) wykorzystanych w badaniach testowych, jakość wyniku jest wyższa w stosunku do jakości najkorzystniejszego wyniku wyrażonego miarami jakości decyzji *ACC* oraz *ERR* dla pozostałych wariantów eksploracji (warianty B, C i D z rysunku 2).

Tabela 74. Porównanie wartości miar jakości decyzji osiągniętych w wariancie A (procedura integracyjna z rozdziału 4 oparta na eksploracji danych tekstowych i numerycznych) z wartościami miar jakości decyzji uzyskanymi w wariancie eksploracji B (eksploracja tylko danych tekstowych), C (eksploracja tylko danych numerycznych) i D (integracja wyników uzyskanych w wariancie B i C)

Przypadek procesu *PD*	Δ_{ACC} **(WN=A, WM=B) [%]**	Δ_{ERR} **(WM=B, WN=A) [%]**	Δ_{ACC} **(WN=A, WM=C) [%]**	Δ_{ERR} **(WM=C, WN=A) [%]**	Δ_{ACC} **(WN=A, WM=D) [%]**	Δ_{ERR} **(WM=D, WN=A) [%]**
I	14	14	10	10	10	10
II	3	3	12	12	3	3
III	3	3	25	25	3	3

Źródło: opracowanie własne

Mając na uwadze powyższe wyniki badań testowych można uznać, że hipoteza postawiona w pracy dla rozważanych przypadków użycia (przykładowe procesy *PD*: I, II oraz III) została zweryfikowana, ponieważ:

- miary jakości decyzji (*ACC, ERR*) eksploracji danych dla przykładowych procesów podejmowania decyzji *PD* (przykładowe procesy decyzyjne I, II oraz III zaprezentowanych odpowiednio w rozdziałach 5.2, 5.3 oraz 5.4) w wariancie eksploracji z wykorzystaniem integracja metod eksploracji danych tekstowych i numerycznych (wariant A z rysunku 2 zgodny z procedurą zawartą w rozdziale 4) są wyższe w przypadku miary *ACC* oraz niższe w przypadku miary *ERR* od takich miar w pozostałych wariantach B (eksploracja oparta tylko na danych numerycznych w procesie *PD*) , C (eksploracja oparta tylko na danych tekstowych) oraz wariancie D integrującym wyniki wariantu B, C, co potwierdzają zestawienia zawarte w tabelach od 68 do 70,
- nośność informacyjna mierzona miarą *ACC* oraz *ERR* eksploracji danych w przypadku reprezentacji γ-gramowej, opracowanej za pomocą wzorców informacyjnych oraz analizy fleksyjnej języka polskiego jest wyższa w porównaniu do nośności informacyjnej danych osiągniętej dla reprezentacji unigramowej i bigramowej, co potwierdzają zestawienia danych w tabelach od 71 do 73,

- nawet w przypadku najmniej korzystnego wyniku (wyrażonego miarami *ACC* i *ERR*) dla wariantu eksploracji danych tekstowych i numerycznych autorską metodą integracyjną zgodną z procedurą z rozdziału 4 (wariant A z rysunku 2) we wszystkich przypadkach procesów *PD* z rozdziału 5, jakość wyniku jest wyższa w stosunku do jakości najkorzystniejszych wyników eksploracji danych pozostałych wariantów (warianty C, B i D z rysunku 2), co potwierdza zestawienie zawarte w tabeli 74.

7. Podsumowanie

W niniejszej pracy ujęto obszerny opis aktualnego stanu badań wraz z przeglądem poszczególnych metod i technik dotyczących eksploracji bazującej na klasyfikacji danych tekstowych oraz numerycznych wykorzystywanych w procesach podejmowania decyzji. W opracowaniu skoncentrowano się na klasyfikacji ze względu na jej znaczącą rolę w zbiorze metod eksploracji danych wspierających procesy podejmowania decyzji *PD*. W analizie szczególną uwagę poświęcono procesowi opracowania reprezentacji eksplorowanych danych, w którym również stosowane są metody tzw. wstępnej eksploracji danych. Za pomocą wstępnej eksploracji, w której wykorzystywane jest zarówno wiedza eksperta dziedzinowego jak i metody uczenia maszynowego możliwe jest opracowanie reprezentacji danych tekstowych dostosowanej (w sensie wykorzystania jako elementów reprezentacji jedynie informacji, które mają istotny wpływ na podejmowaną decyzję w procesie *PD*) do rozpatrywanego problemu decyzyjnego. Przy wyborze reprezentacji danych szczególnie istotny jest kontekst decyzyjny. Z tego względu we wstępie pracy podkreślono zależność eksploracji danych od tego zagadnienia. Scharakteryzowano najczęściej wykorzystywane reprezentacje danych tekstowych:

- unigramową,
- n-gramową,
- γ-gramową.

Podkreślono podział metod eksploracji danych tekstowych na dwie główne grupy, z których jedna bazuje na uczeniu maszynowym, a druga na wiedzy eksperta. Następnie szczegółowo opisano eksplorację danych numerycznych bazującą na Teorii Zbiorów Przybliżonych. W pracy skoncentrowano się na metodach opracowania reprezentacji danych numerycznych w systemie informacyjnym, ze szczególnym uwzględnieniem technik eliminujących szum informacyjnych.

W ramach niniejszej pracy opracowano autorską procedurę integracji metod klasyfikacji danych tekstowych i numerycznych (rozdział 4 pracy), która zwiększa nośność informacyjną danych w procesie podejmowania decyzji. Metoda ta została oparta na wiedzy eksperta, analizie fleksyjnej danych tekstowych dostępnych w procesie decyzyjnym *PD* oraz eksploracji danych numerycznych.

W pracy przeprowadzono badania testowe, które potwierdziły, że wariant procedury z rozdziału 4 (wariant A z rysunku 2) jest korzystniejszy (w sensie zwiększenia nośności informacyjnej danych) w stosunku do wyników uzyskanych z niezależnego działania metod

eksploracji danych numerycznych (wariant B z rysunku 2), metod eksplorowania tylko danych tekstowych (wariant C z rysunku 2) oraz metody integracji wyników eksploracji danych uzyskanych w wariancie B i C (wariant D z rysunku 2). Ponadto nośność informacyjna danych mierzona za pomocą miar jakości decyzji (*ACC, ERR*) jest wyższa w przypadku zastosowania γ-gramowej reprezentacji danych tekstowych, która jest opracowana z wykorzystaniem zarówno wiedzy eksperta (zdefiniowany przez eksperta wzorców informacyjnych) jak i metod uczenia maszynowego (ekstrakcja rzeczowych informacji oraz ich weryfikacja przy użyciu analizy fleksyjnej) w stosunku do pozostałych badanych reprezentacji (unigramowej, n-gramowej - bigramowej). Wykazano również, że dzięki użyciu analizy fleksyjnej tekstu możliwe jest wyodrębnienie poprawnych rzeczowych informacji, które mają istotny wpływ na podejmowane decyzji w procesie *PD*.

Badania testowy przeprowadzone w ramach niniejszej pracy potwierdziły zatem przedstawioną w rozdziale 1.2 hipotezę. Integracja metod analizy fleksyjnej tekstu oraz metod eksploracji danych numerycznych zwiększa nośność informacyjną danych w procesie podejmowania decyzji.

W związku z sformułowanymi we wstępie pracy (rozdział 1.2) pytaniami związanymi ze zidentyfikowanymi brakami metod eksploracji danych w procesie podejmowania decyzji *PD,* na podstawie niniejszego opracowania, a w szczególności przeprowadzonych badań testowych, można stwierdzić, że:

1. Opracowane w pracy metody eksploracji danych jednocześnie uwzględnia zarówno dane numeryczne jak i tekstowe.
2. Dzięki opracowanej w pracy metodzie można osiągnąć wyższą nośność informacyjną danych w procesie eksploracyjnym.
3. Integracja metod eksploracji danych tekstowych i numerycznych wpływa na poprawę jakości wyniku procesu *PD*, poprzez wzrost nośności informacyjnej dostępnych danych.
4. Nośność informacyjną danych w eksploracji wspomagającej proces *PD* można zwiększyć poprzez opracowanie odpowiedniej względem rozpatrywanego problemu decyzyjnego reprezentacji danych.
5. Nośność informacyjna danych determinuje wybór najkorzystniejszego modelu reprezentacji danych dla rozpatrywanego problemu decyzyjnego.
6. W opracowaniu reprezentacji γ-gramowej można uwzględnić specyfikę polskiego języka fleksyjnego.

W ramach niniejszej pracy wniesiony został następujący wkład autorski:

- opracowano wieloaspektową i systemową procedurę integracji uwzględniającą metodę eksploracji danych tekstowych i numerycznych w kontekście procesu *PD*, w którym dostępne są dwa typy danych,
- opracowano metodę budowy reprezentacji danych tekstowych w modelu przestrzeni wektorowej VSM z wykorzystaniem zintegrowanych metod bazujących na wiedzy eksperta (definiowanie wzorców informacyjnych) oraz uczeniu maszynowym (ekstrakcja i weryfikacja rzeczowych informacji na podstawie wzorców),
- opracowano metodę analizy fleksyjnej danych tekstowych wyrażonych w języku fleksyjnym polskim wykorzystywaną do weryfikacji poprawności wyekstrahowanych za pomocą wzorców rzeczowych informacji (elementów reprezentacji γ-gramowej).

Wydaje się, że w kontekście rozwiązywanego w pracy problemu badawczego interesujące jest podjęcie prac nad opracowaniem systemu wspomagającego proces decyzyjny *PD*, opartego na opracowanej w pracy (rozdział 4) metodzie (procedurze) eksploracyjnej, uwzględniającej γ-gramową reprezentację danych tekstowych oraz na automatycznym wyszukiwaniu informacji tekstowych a także numerycznych związanych z procesem *PD* i dostępnych w treściach stron internetowych tworzonych w ramach tzw. semantycznego Internetu z użyciem języka OWL.

Referencje

[1] Adhikari A., Adhikari J., *Advances in Knowledge Discovery in Databases*. Springer, 2015

[2] Aggarwal C.C., Zhai C.: *Mining Text Data.* Springer, 2012

[3] Ahonen-Myka H.: *Information Retrieval Methods*. [on line: 16.09.2016]. Dostępny w Internecie: https://www.cs.helsinki.fi/u/hahonen/irm07/lectures/irm07_5.pdf

[4] Aurangzeb K., Baharum B., Lam Hong L., Khairullah K.: *A Review of Machine Learning Algorithms for Text-Documents Classification*. Journal of Advances in Information Technology, tom 1, nr 1, 2010

[5] Azevedo A.: *Integration of Data Mining in Business Intelligence Systems*. IGI Global, Hershey, 2015

[6] Bechhofer S., Harmelen F., Hendler J., Horrocks I., McGuinness L., Patel-Schneider P.F., Stein A. L. *OWL*. [on line: 16.09.2016]. Dostępny w Internecie: www.w3.org/TR/owl-ref

[7] Berendt B., Hotho A., Mladenic D., Someren M., Spiliopoulou M., Stumme G.: *Web Mining: From Web to Semantic Web: First European Web Mining Forum*, EWMF 2003, Springer, Berlin 2004

[8] Berry M., Linoff G.: *Mastering Data Mining: The Art and Science of Customer Relationship Management*. John Wiley & Sons, New York, 2000

[9] Berry M.: *Survey of Text Mining: Clustering, Classification, and Retrieval*. Springer, 2004

[10] Bökemeier J., Koerber J.: *PHP/Java Bridge*. [on line: 16.09.2016]. Dostępny w Internecie: http://php-java-bridge.sourceforge.net/pjb/

[11] Buhmann L., Lehmann J.: *Pattern Based Knowledge Base Enrichment*. Springer, Berlin, 2013

[12] Burstein F., Brézillon P., Zaslavsky A., *Supporting Real Time Decision-Making: The Role of Context in Decision Support on the Move*. Springer, 2010

[13] Chakraborty D. G., Pagolu M., Garla S., *Text Mining and Analysis: Practical Methods, Examples, and Case Studies Using SAS*. SAS Institute, 2014

[14] Chu H.: *Information Representation and Retrieval in the Digital Age*. Information Today, Inc., Medford, 2003

[15] Diks K: *Eksploracja Danych*. [on line: 16.09.2016]. Dostępny w Internecie: http://wazniak.mimuw.edu.pl/images/3/3d/ED-4.2-m01-1.0.pdf

[16] Dominik A.: *Analiza danych z zastosowaniem teorii zbiorów przybliżonych",* Politechnika Warszawska, 2004

[17] Dong X. L., Gabrilovich E., Murphy K., Dang V., Horn W., Lugaresi C., Sun S., Zhang W.: *Knowledge-Based Trust: Estimating the Trustworthiness of Web Sources*. Cornell University Library, 2015

[18] Drawmiński M.: *Algorytm indukcji reguł decyzyjnych w problemach klasyfikacji i wyboru cech w zadaniach wysokowymiarowych*. Polska Akademia Nauk, Warszawa, 2007

[19] Fatyga P., Podraza P.: *Klasyfikacja danych – przegląd wybranych metod*. Zeszyty Naukowe Wydziału ETI Politechniki Gdańskiej, 2010

[20] Fellbaum C., Tengi R.: *WordNet*. [on line: 16.09.2016]. Dostępne w Internecie: http://wordnet.princeton.edu
[21] Gabrys B., Howlett R. J., Jain L. C.: *Analysis of Stock Price Retudn Using Textual Data and Numerical Data Through Text mining*. Springer, Berlin, 2006
[22] Gawrysiak P.: *Automatyczna kategoryzacja dokumentów*. Uniwersytet Warszawki, 2001
[23] Gawrysiak P.: *Eksploracja danych tekstowych w środowisku WWW*. Politechnika Warszawska, 2007
[24] Gawrysiak P.: *Klasyfikacja dokumentów*. Politechnika Warszawska, 2005
[25] Gibert M.: *The influence of data information-carrying capacity on quality of text mining*. Advances in Data Mining/15th Industrial Conference on Data Mining, Hamburg, 2015
[26] Grudzień Ł.: *Koncepcja oceny jakości informacji o procesach w systemach zarządzania*. Zarządzanie Przedsiębiorstwem, Zakopane, 2012
[27] Hand D. J., Mannila H., Padhraic S.: *Principles of Data Mining*. Massachusetts Institute of Technology Press, 2001
[28] Hearst M. A.: *TextTiling: Segmenting Text into Multi-paragraph Subtopic Passages*. Computational Linguistics, tom 23, nr 1, 1997
[29] Hicklin J., Moler C., Webb P., Boisvert R.F., Miller B., Pozo R., Remington K.: *JAMA : A Java Matrix Package*. [on line: 16.09.2016]. Dostępny w Internecie: http://math.nist.gov/javanumerics/jama/
[30] Hofmann T.: *Unsupervised Learning by Probabilistic Latent Semantic Analysis*., Machine Learning, nr 42, Kluwer Academic Publishers, Hingham, 2001
[31] Hu W. i Feng J.: *Data and Information Quality: an Information-theoretic Perspective*. Computing and Information Systems, nr 9(3), 2005
[32] Hyndman R. J. *The problem with Sturges' rule for constructing histograms*. [on line: 16.09.2016]. Dostępny w Internecie: http://robjhyndman.com/papers/sturges.pdf
[33] Jackson P., Moulinier I.: *Natural Language Processing for Online Applications: Text Retrieval, Extraction, and Categorization*. John Benjamins Publishing, Amsterdam, 2007
[34] Janakiraman V. S., Sarukesi K.: *Decision Support Systems*. PHI Learning Pvt. Ltd., New Delphi, 2008
[35] Katu V., Deshpande B.: *Predictive Analytics and Data Mining: Concepts and Practice with RapidMiner*. Elsevier, Waltham, 2015
[36] Kent A., Williams J. G.: *Encyclopedia of Microcomputers: Volume 27: Supplement 6*. Marcel Dekker, Inc., New York, 2001
[37] Klimkiewicz M., Moczulska K.: *Zastosowanie zbiorów przybliżonych do analizy satysfakcji klienta serwisu pojazdów*., Inżynieria Rolnicza, nr 1(99), 2008
[38] Kobos M.: *Data mining - Przegląd metod eksploracji danych*., Politechnika Warszawska, 2005
[39] Kosmulski M.: *Reprezentacja dokumentów tekstowych w modelu przestrzeni wektorowej*. Politechnika Warszawska, 2005
[40] Kotsyba N.: *Korpusy tekstów w lingwistyce. Wyrażenia regularne. Cz. I*. Uniwersytet Warszawski, 2010
[41] Kulig A.: *Ilościowe charakterystyki złożoności języka naturalnego*. Instytut Fizyki Jądrowej im. Henryka Niewodniczańskiego PAN, Kraków, 2014

[42] Landauer T. K., McNamara D. S., Dennis S., Kintsch W., *Handbook of Latent Semantic Analysis*. Routledge, New York, 2011
[43] Langford G. O.: *Engineering Systems Integration: Theory, Metrics, and Methods*. CRC Press, Boca Raton, 2012
[44] Layka V.: *Learn Java for Web Development: Modern Java Web Development*. Apress, 2014
[45] Libal U.: *Algorytmy rozpoznawania obrazów - Praktyczna ocena jakości klasyfikacji*. Politechnika Wrocławska, 2015
[46] Libra M., Waligóra W.: *Zastosowanie Teorii Zbiorów Przybliżonych do oceny wpływu właściwości materiałowych elementów ze stali 100Cr6 na ich powierzchniową trwałość zmęczeniową*. Archiwum Technologii Maszyn i Automatyzacji, tom 30, nr 2, 2010
[47] Lipiński M.: *O technologii i organizacji IT: SCJP - Tokenizacja tekstu*. [on line: 16.09.2016]. Dostępny w Internecie: http://www.mariuszlipinski.pl/2009/04/scjp-tokenizacja- tekstu.html
[48] Lubaszewski W., Gajęcki M.: *Automatyczna ekstrakcja powiązań semantycznych z tekstu polskiego.*, Computer Science, tom 4, 2002
[49] Lubaszewski W.: *Słowniki komputerowe i automatyczna ekstrakcja informacji z tekstu*. Wydawnictwo AGH, Kraków, 2009
[50] Lula P.: *Text mining jako narzędzie pozyskiwania informacji z dokumentów tekstowych*. Akademia Ekonomiczna w Krakowie, 2005
[51] Lupton R., *Statistics in Theory and Practice*. Princeton University Press, Princeton, 1993
[52] Lutfi M., Aris I.: *Inconsistent Decision System: Rough Set Data Mining Strategy to Extract Decision Algorithm of a Numerical Distance Relay – Tutorial*. Advances in Data Mining Knowledge Discovery and Applications, 2012
[53] Łukaszuk T.: *Techniki eksploracji danych oparte na funkcjach kryterialnych typu CPL w informatycznym systemie pracy zdalnej*. [on line: 16.09.2016]. Dostępny w Internecie: http://www.wi.pb.edu.pl/index.php/nauka/seminaria/31-techniki-eksploracji-danych-oparte-na-funkcjach-kryterialnych-typu-cpl-w-informatycznym-systemie-pracy-zdalnej, 2016
[54] Macdonald C., Ounis I., Plachouras V.: *Advances in Information Retrieval: 30th European Conference on IR Research*. ECIR 2008, Springer, Berlin, 2008
[55] Małecka D.: *Implementacja metod eksploracji danych - Oracle Data Mining*. [on line: 16.09.2016]. Dostępny w Internecie: http://docplayer.pl/4825532-Implementacja-metod-eksploracji-danych-oracle-data-mining.html
[56] Markowski K.: *Podmiotowe uwarunkowania decyzji inwestycyjnych*. Zarządzanie finansami firm - teoria i praktyka nr 965, Wrocław, 2002
[57] Merkelis R.: *Philosophy and Linguistics*, [on line: 16.09.2016]. Dostępny w Internecie: http://www.slideshare.net/robertasmerkelis/philosophy-and-linguistics-28940425
[58] Miner G.: *Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications*. Academic Press, Waltham, 2012
[59] Ming Fai Wang F., Liu Z., Chiang M.: *Stock Market Prediction from WSJ: Text Mining via Sparse Matrix Factorization*. Cornell University Library, 2014

[60] Misztal M..: *Wybrane metody oceny jakości klasyfikatorów – przegląd i przykłady zastosowań*. Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu, nr 328, Wrocław, 2014
[61] Morzy T. : *Eksploracja danych: problemy i rozwiązania,* V Konferencja PLOUG, Zakopane, 1999
[62] Morzy T., Morzy M., Leśniewska A.: *Eksploracja tekstu I*. [on line: 16.09.2016]. Dostępny w Internecie: http://wazniak.mimuw.edu.pl/images/6/62/ED-4.2-m12-1.01.pdf
[63] Murphy K. P.: *Machine Learning: A Probabilistic Perspective*. MIT Press, London, 2012
[64] Nettleton D.: *Commercial Data Mining: Processing, Analysis and Modeling for Predictive Analytics Projects.* Elsevier, Waltham , 2014
[65] Neustein A.: *Text Mining of Web-Based Medical Content*. Walter de Gruyter GmbH & Co KG, 2014
[66] Nilsson A. G., Gustas R., Wojtkowski G., Wojtkowski W., Wrycza S., Zupancic J.: *Advances in Information Systems Development: Bridging the Gap between Academia and Industry*. Springer, New York, 2006
[67] Nowak A.: *Teoretyczne podstawy zbiorów przybliżonych*. [on line: 16.09.2016]. http://zsi.tech.us.edu.pl/~nowak/se/konspektTD.pdf
[68] Nowak-Brzezińska A. *Przygotowanie danych w środowisku R*. [on line: 16.09.2016]. Dostępny w Internecie: http://zsi.tech.us.edu.pl/~nowak/ed/ped_PD.pdf
[69] Palmer D. D.: *Chapter 2: Tokenisation and Sentence Segmentation*. [on line: 16.09.2016]. Dostępny w Internecie: https://s3.amazonaws.com/tm-town-nlp-resources/ch2.pdf
[70] Piegat A.: *Materiały z wykładów Teorii Zbiorów Przybliżonych*. Zachodniopomorski Uniwersytet Technologiczny w Szczecinie, 2010
[71] Potopia P.: *Metody i narzędzia automatycznego przetwarzania informacji tekstowej i ich wykorzystanie w procesie zarządzania wiedzą*. Automatyka, tom 15, nr 2, Wydawnictwa AGH, Kraków, 2011
[72] Power D. J., *Decision Support Systems: Concepts and Resources for Managers*. Greenwood Publishing Group, Westport, 2002
[73] Przepiórkowski A., *Korpus IPI PAN*. Instytut Podstaw Informatyki PAN, Warszawa, 2004
[74] Raghavan V., Bollmann P., Jung G. S.: *A critical investigation of recall and precision as measures of retrieval system performance*. ACM Transactions on Information Systems (TOIS), tom 7, nr 3, New York, 1989
[75] Ramasubramanian C., Ramya R. : *Effective Pre-Processing Activities in Text Mining using Improved Porter's Stemming Algorithm*. International Journal of Advanced Research in Computer and Communication Engineering, tom 2, nr 12, Chennai, 2013
[76] Rambaud M.G., Moreno A.I., Hoste V., Mattens S., Coninck P.: *On the architecture of words. Applications of meaning studies*. Editorial UNED, Madrid, 2015
[77] Rejer I.: *Integracja źródeł wiedzy w modelach rozmytych zależności ekonomicznych*. Wydawnictwo Naukowe Uniwersytetu Szczecińskiego, 2008
[78] Rubin A.: *Statistics for Evidence-Based Practice and Evaluation*. Cengage Learning, 2009

[79] Samuelson W. F., Marks S. G.: *Ekonomia menedżerska* II. Polskie Wydawnictwo Ekonomiczne, Warszawa, 2009
[80] Schumaker R. P., Chen H.: *Textual Analysis of Stock Market Prediction Using Breaking Financial News: The AZFin Text System*, ACM Transactions on Information Systems (TOIS), tom 27, nr 2, New York, 2009
[81] Sikora M.: *Wybrane metody oceny i przycinania reguł decyzyjnych*. Studia Informatica, Wydawnictwo Politechniki Śląskiej, 2012
[82] Silva C., Ribeiro B.: *Inductive Inference for Large Scale Text Classification: Kernel Approaches and Techniques*. Springer, 2009
[83] Sołdacki P.: *Zastosowanie metody płytkiej analizy tekstu do przetwarzania dokumentów w języku polskim*. Politechnika Warszawska, 2006
[84] Srivastava A. N., Sahami M.: *Text Mining: Classification, Clustering, and Applications*. CRC Press, Boca Raton, 2009
[85] Stankiewicz M.: *Modelowanie profili klientów w informatycznym systemie wspomagania decyzji*. Zachodniopomorski Uniwersytet Technologiczny w Szczecinie, 2013
[86] Synak P.: Temploralne aspekty eksploracji danych: metody zbiorów przybliżonych. Warszawa, 2003
[87] Śmiałkowska B., Gibert M.: „The classification of text documents by using Latent Semantic Analysis for extracted information", Ekonomiczne Problemy Usług, tom 6, 2013
[88] Śmiałkowska B., Gibert M.: *The classification of text documents in Polish language by using Latent Semantic Analysis for extracted information*. Theoretical Applied Information, tom 25, nr 239–250, 2013
[89] Thomas J. D., Sycara K.: *Integration Genetic Algorithms and Text learning for Financial Prediction*, GECCO-2000 Workshop on Data Mining with Evolutionary Algorithms, 1999
[90] Tittel E., Noble J.: *HTML, XHTML and CSS For Dummies*. Wiley Publishing Inc., Indiana, 2008
[91] Toffler A.: *Szok przyszłości*. Wydawnictwo Kurpisz, 2007
[92] Tomanek T.: *Analiza sentymentu – metoda analizy danych jakościowych. Przykład zastosowania oraz ewaluacja słownika RID i metody klasyfikacji Bayesa w analizie danych jakościowych*. Przegląd Socjologii Jakościowej tom 10 nr 2, Łódź, 2014
[93] Unold J.: *System informacyjny a jakościowe ujęcie informacji.* Teoretyczne Podstawy Tworzenia SWO i strategie Budowy E-Biznesu, Katowice, 2004
[94] Wang J. X.: *What Every Engineer Should Know About Decision Making Under Uncertainty*. Marcel Dekker, Inc., New York, 2002
[95] Wang J.: *Data Mining: Opportunities and Challenges*. IRM Press, London, 2003
[96] Wang R. Y., Strong D. M.: *Beyond Accuracy: What data quality means to data consumers*. Journal of Management Information Systems, tom 12, nr 4, 1996
[97] Weiss S. M., Indurkhya N., Zhang T.: *Fundamentals of Predictive Text Mining*. Springer, London, 2010
[98] Weiss S. M., Indurkhya N., Zhang T., Damerau F.: *Text Mining: Predictive Methods for Analyzing Unstructured Information*. Springer, New York, 2010

[99] Williams H. E., Lane D.: *Web Database Applications with PHP and MySQL*. O'Reilly Media, Inc., 2004
[100] Wuthrich B., Permunetilleke D., Leung S., Cho V., Zhang J., Lam W.: *Daily Prediction of Major Stock Indices from textual WWW Data*, KDD-98 Proceedings, AAAI, New York, 1998
[101] Yang J., Yang X.: *Incomplete Information System and Rough Set Theory: Models and Attribute Reductions*. Springer, 2012
[102] Yin Y., Kaku I., Tang J., Zhu J., *Data Mining: Concepts, Methods and Applications in Management and Engineering Design.* Springer, London, 2011
[103] Yuan J.: *Image and Video Data Mining*, Northwestern University, Illinois, 2009
[104] Zeng A., Huang Y.: *A text classification algorithm based on Rocchio and Hierarchical clustering*. Springer, Berlin, 2011
[105] Zięba M.: *Klasyfikacja*. [on line: 16.09.2016]. Dostępny w Internecie: https://www.ii.pwr.edu.pl/~zieba/Lista5.pdf
[106] *Biuletyn Zamówień Publicznych*. [on line: 16.09.2016]. Dostępny w Internecie: http://bzp1.portal.uzp.gov.pl
[107] *Eksploracja danych, Governica.com*. [on line: 16.09.2016]. Dostępne w Internecie: www.governica.com/Eksploracja_danych
[108] *Kody CPV*. [on line: 16.09.2016]. Dostępny w Internecie: http://kody.uzp.gov.pl
[109] Słownik Języka Polskiego. [on line: 16.09.2016]. Dostępne w Internecie: www.sjp.pl
[110] *Uczenie maszynowe i sztuczne sieci neuronowe/Wykład Ocena jakości klasyfikacji*, [on line: 16.09.2016]. Dostępny w Internecie: http://brain.fuw.edu.pl/edu/index.php/Uczenie_maszynowe_i_sztuczne_sieci_neuronowe/Wyk%C5%82ad_Ocena_jako%C5%9Bci_klasyfikacji

Spis rysunków

Spis tabel

Spis symboli

Lp.	Symbol	Opis
1	Z	Zbiór danych w procesie *PD*
2	Z_N	Zbiór danych numerycznych
3	Z_T	Zbiór danych tekstowych
4	Z_P	Zbiór danych innych niż dane numeryczne i tekstowe
5	Z_E	Zbiór danych tekstowych i numerycznych w procesie *PD*
6	R	Reprezentacja dokumentu tekstowego
7	r	Element reprezentacji dokumentu tekstowego
8	j	Indeks elementu reprezentacji
9	mr	Maksymalna liczba elementów reprezentacji dokumentu tekstowego
10	$Ninf$	Nośność informacyjna danych
11	M	Zbiór wybranych miar jakości decyzji w procesie *PD*
12	W_M	Wynik procesu decyzyjnego w postaci wybranych miar jakości klasyfikacji
13	ζ	Nieznany i niezidentyfikowany zbiór informacji mający wpływ na nośność informacyjną danych *Ninf*
14	g	Nieznana funkcja, której wartość określa nośność informacyjną *Ninf*
15	t	Dokument tekstowy
16	w	Wyraz występujący w dokumencie tekstowym
17	mw	Maksymalna liczba wyrazów wydobytych z dokumentu tekstowego
18	iw	indeks wyrazu, wydobytego z dokumentu tekstowego
19	V	Słownik wyrazów
20	mz	maksymalna liczba znaków wydobyta z dokumentu tekstowego
21	z	Znak rozdzielający zdania np. kropka, wykrzyknik itd.
22	Z	Zbiór wszystkich znaków, które mogą kończyć zdanie
23	iz	Indeks wyrazu z, wydobytego z dokumentu tekstowego
24	sk	Miara skojarzenia wyrazu definiującego z definiowanym
25	cw	Częstość względna wyrazu definiującego
26	lp	Częstość bezwzględna wyrazu definiowanego
27	iv	Indeks wyrazu w słowniku V
28	mv	Maksymalna liczba wyrazów w słowniku
29	$wpw.$	Skrót oznaczający „w przeciwnym wypadku”.
30	lw	Długość sekwencji wyrazów
31	is	Indeks sekwencji wyrazów
32	tf	Częstość termu (ang. Term Frequency)
33	idf	Odwrotna częstość dokumentu (ang. Inverse Document Frequency)
34	$tfidf$	Częstość termu - odwrotna częstość dokumentu
35	Mrt	Macierz termów i dokumentów tekstowych
36	Mr	Macierz termów
37	O	Macierz wartości osobliwych
38	Mt	Macierz dokumentów tekstowych
39	o	Liczba wartości osobliwych
40	T	Transpozycja macierzy
41	tw	Wzorcowy dokument tekstowy
42	$cos()$	Miara kosinusowa
44	rw	Element reprezentacji wzorcowego dokumentu tekstowego

45	mw	Liczba dokumentów wzorcowych
46	itw	Indeks dokumentu wzorcowego
47	kl	Klasa dokumentów tekstowych
48	P	Prawdopodobieństwo wystąpienia
49	WR	Wektor odwzorowujący kategorię *kl*
50	ϕ	Przesłanka w regule decyzyjnej
51	ψ	konkluzja w regule decyzyjnej
52	h	Liczba przedziałów
53	U	Uniwersum – zbiór obiektów
54	u	Obiekt
55	a	Atrybut
56	mu	Liczba obiektów z uniwersum
57	A	Zbiór atrybutów
58	v	Wartość atrybutu
59	V	Zbiór wartości atrybutów
60	f	Funkcja informacyjna
61	SI	System informacyjny
62	$IND_{SI}(B)$	Relacja równoważności określona na zbiorze B
63	B	podzbiór zbioru wszystkich atrybutów A
64	Aw	Zbiór atrybutów warunkowych
65	$U/\ IND_{SI}(B)$	Zbiór klas abstrakcji
67	$I_{SI,B1}()$	Klasa abstrakcji
68	TD	Tablica decyzyjna
69	$\underline{B}X$	Dolne przybliżenie
70	$\overline{B}X$	Górne przybliżenie
71	$POS_B(X)$	Obszar pozytywny
72	$NEG_B(X)$	Obszar negatywny
73	$BN_B(X)$	Obszar brzegowy
74	$\gamma_B(X)$	Współczynnik jakości przybliżenia konceptów decyzyjnych
75	$\beta_B(X)$	Współczynnik dokładności przybliżenia konceptów decyzyjnych
76	$card(POS_B(X))$	Liczba przypadków (obiektów) w pozytywnym obszarze zbioru *X*
77	$card(U)$	Liczba przypadków w uniwersum
78	$card(\underline{B}X)$	Liczba przypadków należących do dolnego przybliżenia
79	$card(\overline{B}X)$	Liczba przypadków należących do górnego przybliżenia
80	Ad	Zbiór atrybutów decyzyjnych
81	Aw	Zbiór atrybutów warunkowych
82	Bw	Podzbiór atrybutów warunkowych
83	$\emptyset$	Zbiór pusty
84	ad	Atrybut decyzyjny
85	aw	Atrybut warunkowy
86	iu, ju	Indeksy obiektów z uniwersum
87	d	Klasa decyzyjna, wartość atrybutu decyzyjnego *ad*
88	$Wo_{TD}(x,d)$	Współczynnik wiarygodności obiektu
89	$Kw_{u,d}$	Współczynnik pomocniczy
90	$Wo_{u,d,a}$	Współczynnik pomocniczy
91	$\sigma_C(a)$	Współczynnik istotności atrybutu a
92	$RED()$	Redukt aproksymacyjny

93	$CORE()$	Rdzeń aproksymacyjny
94	*awmin*	Minimalna ilość atrybutów w jądrze
95	*iaw*	Indeks atrybutów warunkowych
96	*Mo*	Macierz rozróżnialności
97	*a**	Zmienna w funkcji rozróżnialności
98	*ma*	Maksymalny indeks atrybutu
99	$card(\#\phi\ \#)$	Liczba reguł w pszesłankach
100	$sup_{Regi}(\phi,\psi)$	Wsparcie reguły o indeksie Regi
101	$cer(\phi,\psi)$	Pewność reguły
102	$card(\|\ \phi \wedge \psi\ \|)$	Liczba reguł o takich samych przesłankach i konkluzjach
103	*Regi*	Indeks reguły
104	*sk*	Miara skojarzeniowa
105	*cw*	Częstość względna
106	*lw*	Częstość bezwzględna
107	*TP*	Współczynnik prawdziwy pozytywny
108	*FP*	Współczynnik fałszywy pozytywny
109	*FN*	Współczynnik fałszywy negatywny
110	*TN*	Współczynnik prawdziwy negatywny
111	*SE*	Współczynnik czułości/precyzja
112	*PPV*	Pozytywny współczynnik predykcji/precyzja
113	*SP*	Współczynnik specyficzności
114	*NPV*	Negatywny współczynnik predykcji
115	*ACC*	Współczynnik całkowitej dokładności
116	*ERR*	Współczynnik całkowitego poziomu błędu
117	*F*	Współczynnik F
118	χ^2	Statystyka testowa w teście McNeamara
119	H_0, H_1	Hipotezy statystyczne
120	Δ	Współczynnik różnicy wyniku wyrażony procentowo
121	*Mc1*	Współczynnik testu Mcnemara
122	*Mc2*	Współczynnik testu Mcnemara
123	*Mc3*	Współczynnik testu Mcnemara
124	*Mc4*	Współczynnik testu Mcnemara