http://dx.doi.org./10.17651/SOCJOLING.31.6

Socjolingwistyka XXXI, 2017
PL ISSN 0208-6808

AGNIESZKA SZYMAŃSKA

Uniwersytet Kardynała Stefana Wyszyńskiego, Warszawa

WYKORZYSTANIE ALGORYTMÓW *TEXT MINING* DO ANALIZY DANYCH TEKSTOWYCH W PSYCHOLOGII

Słowa kluczowe: algorytmy, dane tekstowe, *text mining*.

STRESZCZENIE

W psychologii analizy danych zapisanych w postaci tekstów stanowią ważny element prac badawczych. Niemniej nadal poszukuje się narzędzi, metod, które mogą umożliwić szybką analizę danych zarejestrowanych w postaci tekstów, gdyż analizy te są najczęściej bardzo czasochłonne. W prezentowanym artykule przybliżono metodę *text mining*, która ma szczególne zastosowanie w analizie informacji zapisanych w postaci danych tekstowych. Wykorzystanie metody *text mining* jest omawiane na przykładzie analizy obieranych przez rodziców celów wychowawczych.. W artykule przedstawiono sposób, w jaki algorytmy *text mining*: a) dokonują analizy tekstu przez zliczenie słów i nadanie im wag, b) przeprowadzają analizę relacji między słowami za pomocą składowych głównych (Principal Component Analysis), c) przekształcają dane słowne w liczbowe, przygotowując zbiór danych do kolejnych obliczeń.

1. WPROWADZENIE

W naukach psychologicznych bardzo duża część analizowanych danych pochodzi z wywiadów, narracji. Zapisane w postaci tekstów dane wymagają czasochłonnych analiz. Stworzono wiele programów, których zadaniem jest pomoc w prowadzeniu analiz tekstów, np. ATLAS, NVIVO i inne (Franzosi 2010). Programy te umożliwiają użytkownikom klasyfikację słów, badanie zależności znajdujących się w danych słownych. Najczęściej programy te pomagają strukturalizować informacje i skracać czas prowadzonych analiz. Nie mogą jednak ułatwić przekształcenia danych tekstowych w liczbowe. Umiejętności takie posiadają niektóre algorytmy, w które wyposażony jest moduł *text mining*. Dzięki temu można skrócić czas prowadzonej analizy tekstów. Poza liczeniem słów, ujawnianiem ich skupisk metoda ta umożliwia przekształcenie materiału tekstowego w liczbowy i przygotowanie zbioru do kolejnych analiz. Ta funkcja oprogramowana jest w module *text mining* pakietu STATISTICA (Elder i in. 2012). *Text mining* jest również oprogramowany w programie R (Bouchet-Valat, Bastin 2013), a także SPSS Clementine i Orange (Nisbet, Elder, Minerv 2009). W naukach psychologicznych aplikacja rozwiązań, jakie dają metody *data mining* i *text mining*, jest jeszcze rzadkością, na co zwracają uwagę naukowcy na świecie. Do 2014 r. w naukach dotyczących psychologii

rozwojowej i wychowawczej odnotowano zaledwie sześć przykładów wykorzystania metod *data mining* do analizy danych (Yim, Boo, Ebbeck 2014). Informacje tu podane nie są dokładne, gdyż badacze, którzy dokonywali tego porównania, nie mogli dotrzeć do wszystkich raportów z badań (np. stosowanych w Polsce).

W Polsce jedną z pierwszych aplikacji metody *data mining* w psychologii przeprowadziła w 2004 r. Ewa Rzechowska, która algorytm indukcyjny Quinlana wykorzystała w strategii Rekonstrukcji Transformacji Procesu (Rzechowska 2004, 2011a, 2011b). Wkrótce pojawiły się kolejne aplikacje algorytmów: w psychologii wychowawczej i psycholingwistyce (Szymańska 2012a; Tarwacka-Odolczyk i in. 2014; Ważyńska i in. 2015).

Celem tego artykułu jest przedstawienie możliwości zastosowań niektórych algorytmów *text mining* w badaniach psychologicznych. Próby aplikacji *data mining* i *text mining* w psychologii pokazują, że nie tylko skracają one znacznie czas pracy, ale co więcej są niezmiernie użyteczne wobec specyficznych danych, wobec których nie można (ze względu na niespełnione założenia formalne) zastosować statystyk (Tarwacka-Odolczyk i in. 2014; Szymańska 2012b; Torebko, Szymańska 2015; Ważyńska i in. 2015; Szymańska 2015). W pierwszej części tekstu przybliżone zostanie, czym są algorytmy sztucznej inteligencji. W kolejnej omówione zostanie działanie trzech algorytmów *text mining*. W ostatniej na przykładzie danych dotyczących obieranych przez rodziców celów wychowawczych zobrazowane zostanie wykorzystanie trzech algorytmów scharakteryzowanych we wcześniejszej części: a) zliczającego frekwencje i nadającego wagi, b) przedstawiającego wyniki w postaci analizy składowych głównych, c) dokonującego transformacji słów na dane liczbowe.

2. ALGORYTMY SZTUCZNEJ INTELIGENCJI

Algorytmy to pewne działania prowadzące do wykonania jakiegoś zadania. Zadaniem algorytmu jest uruchomienie serii działań celem przeprowadzenia systemu z punktu A do punktu B. Słowo algorytm pochodzi od słowa *algorism* (ang.) oznaczającego wykonywanie działań za pomocą liczb arabskich. Istnieje bardzo wiele różnych algorytmów i wiele klas ich podziału. Jedna z klasyfikacji algorytmów wynika z ich zastosowania. Mówi się więc o algorytmach genetycznych[1], algorytmach równoległych[2], algorytmach kwantowych[3] oraz algorytmach sztucznej inteligencji (Rutkowski 2006; Krupa 1995; Luger, Stubblefield 1989).

Funkcją algorytmów sztucznej inteligencji jest rozwiązywanie problemów na wzór istot inteligentnych, np. człowieka. Algorytmy te posiadają możliwości samouczenia się. Rozwiązują najtrudniejsze problemy klasy problemów NP (niedeterministycznie wielomianowych[4]), a więc NP-trudne. Algorytmy sztucznej inteligencji powstały w wyniku

[1] U ich podstaw znajdują się dobór naturalny oraz dziedziczność.

[2] Wykonywanych na wielu maszynach liczących.

[3] Wykonywanych na komputerach kwantowych.

[4] To problem, dla którego poszukuje się rozwiązania przy pewnej ilości niezbędnych zasobów: czasu i pamięci.

rozwijania się dziedziny, jaką jest sztuczna inteligencja, która zajmowała się tworzeniem modeli zachowań inteligentnych. Jej głównym celem było sprawdzenie, czy można nauczyć komputer myśleć i podejmować decyzje na wzór człowieka. Wykorzystywała ona wiedzę z różnych obszarów: cybernetyki, informatyki, robotyki, psychologii itp. (Nisbet, Elder, Miner 2009).

Szybko zorientowano się, że moc obliczeniowa współczesnych komputerów, ich umiejętności samouczenia się, rozpoznawania ludzkiej mowy umożliwiają tworzenie systemów eksperckich i diagnostycznych, których zadaniem jest odnajdywanie rozwiązań i pomoc w podejmowaniu decyzji (Alqarni i in. 2011; Rutkowski 2006; Luger, Stubblefield 1989; Żurada, Barski, Jędruch 1996), szczególnie tych decyzji, które wymagają brania pod uwagę jednocześnie bardzo wielu różnych przesłanek. Algorytmy można podzielić na:

1. **Algorytmy podstawowe**, do których należą na przykład: Automatyczne Sieci Neuronowe (Automated Neural Networks), Zgeneralizowane Modele Addytywne (Generalized Additive Models), Zgeneralizowane EM i k-średnich analiza skupień (Generalized EM k-Means Cluster Analysis).
2. **Algorytmy zaawansowane**, do których należą na przykład: Drzewa Interakcyjne (CART, C & RT, CHAID), Maszyny Wektorów Wspierających (Support Vector Machines).
3. **Algorytmy specjalnego zastosowania**, np. *text mining*, algorytm Quinlana.

Jak już zostało wspomniane, nie należy pracy algorytmów utożsamiać z ludzką inteligencją, choć sposób, w jaki uczą się one i rozwiązują problemy, pochodzi bezpośrednio od tego, jak poznają rzeczywistość i uczą się istoty inteligentne[5]. Dlatego nazywamy je algorytmami sztucznej inteligencji. Opierają się bowiem na sposobie działania istot inteligentnych.

3. ALGORYTMY *TEXT MINING*

Text mining służy do analizowania tekstu w celu wydobycia nieustrukturalizowanych informacji znajdujących się w zbiorze. W procedurze tej można analizować słowa lub całe skupiska słów oraz sprawdzać ich powiązania z innymi zmiennymi w zbiorze danych. Można porównywać ze sobą całe dokumenty, sprawdzając podobieństwa i różnice między nimi. Według doniesień liczba danych zapisana w plikach tekstowych to 85–90% wszystkich danych istniejących na świecie (Hotho, Nürnberger, Paaß 2005). Jest to więc duża baza niestrukturalizowanych informacji, do których potrzebna jest metoda, która umożliwia ich analizę. Algorytmy *text mining* wydobywają z tekstu informacje dotyczące związków między słowami oraz trendy występujące w tekście. Ogólne cele

[5] Istoty inteligentne — wcale nie oznacza, że zalicza się do nich jedynie człowiek (choć dla wielu algorytmów sposób rozumowania człowieka był wzorem). Jeden ze sławniejszych algorytmów, tzw. algorytm mrówkowy, w swoim działaniu nawiązuje do zachowań mrówek, a jego zadaniem jest poszukiwanie optymalnych rozwiązań w grafach.

algorytmów *text mining* obejmują: a) identyfikację powiązanych w dokumencie słów, b) identyfikację powiązanych dokumentów, c) wykrywanie ukrytych wzorów w dokumentach tekstowych, d) identyfikację skupisk słów powiązanych z analizowanymi słowami kluczowymi (*key words*) wskazanymi przez badacza. Działanie algorytmów *text mining* zaczyna się od:

a) **zliczania liczby słów i nadawania słowom wag**. Algorytm zlicza słowa (rysunek 1) i nadaje im wagę według liczby ich występowania w zbiorze. Algorytm może zastosować różne metody nadawania słowom wag, np. *inverse document frequencies* (opracowaną przez Karen Sparck Jones (1972)) polegającą na zliczeniu liczby występujących słów i nadaniu najsilniejszych wag (*importance*) tym słowom, które pojawiały się w jednym dokumencie, ale które zarazem nie były powtarzane we wszystkich dokumentach (chodzi o kontrolowanie powtórzeń) (Elder i in. 2012). Jak podają Christopher Manning i Hinrich Schutze:

> Document frequency is also scaled logarithmically. The formula log $ = log *N* – log dfi gives full weight to words that occur in 1 document (log N – log dfi = 1og N – log 1 = log N). A word that occurred in all documents would get zero weight (log N – log dfi = log N – log N = 0) (Manning, Schutze 2002: 545).

Gdyby więc np. rodzic cały czas przywoływał w kolejnych powtórzeniach cechę, jaką chce rozwinąć u dziecka, ta cecha otrzymałaby wbrew pozorom niższą wagę, niż można by oczekiwać z jej częstości. Transformacja *inverse document frequencies* dokonywana jest według wzoru 1.

$$\text{idf}(i, j) = \begin{cases} 0, & \text{gdy } wfi,j = 0 \\ (1 + \log(wfi,j))\, \log N / dfi, & \text{gdy } wfi,j > 1 \end{cases}$$

gdzie:

N, całkowita liczba analizowanych dokumentów,

wfi,j, częstość występowania i-tego terminu (słowa) w j-tym dokumencie,

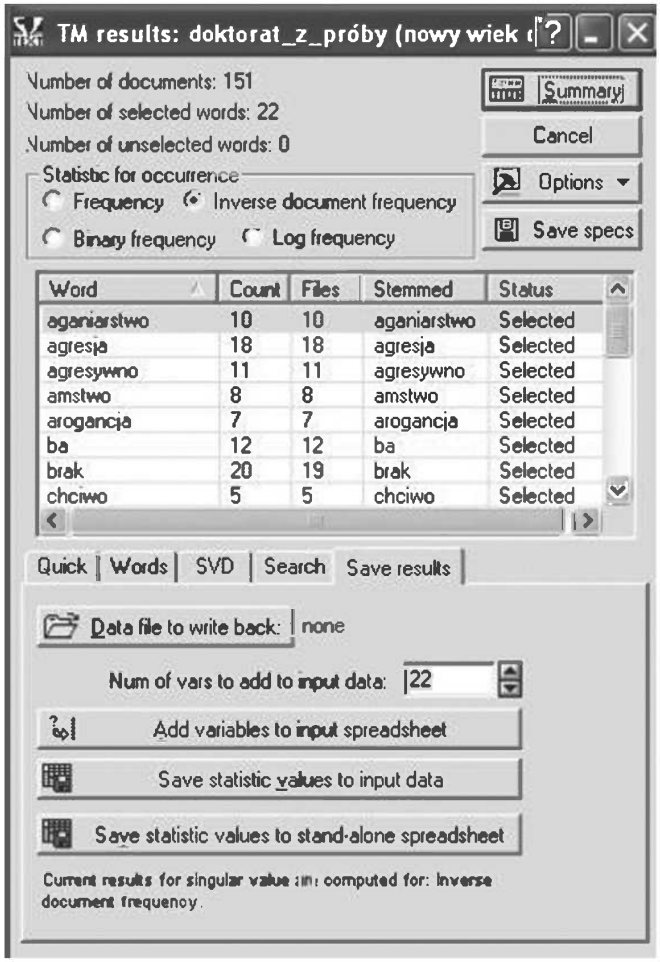
dfi, liczba dokumentów dla i-tego terminu.

W tabelach 8 i 9 przedstawiony został przykład takich wyliczeń, można w niej odczytać frekwencję i wagę nadawane słowom w tekście.

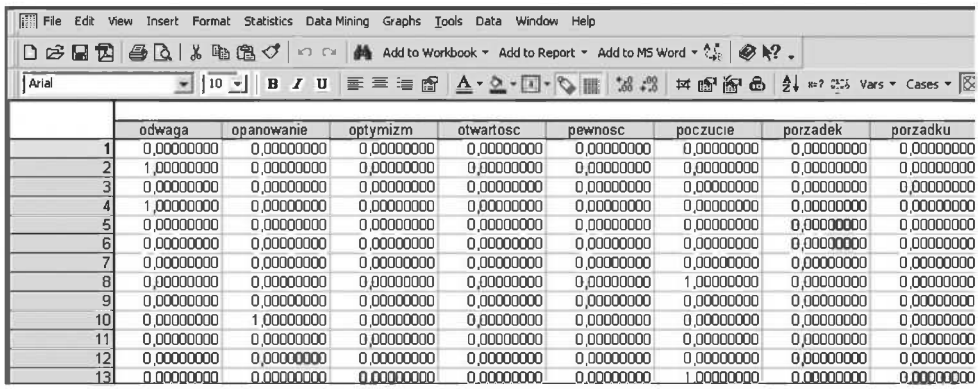
b) **tworzenia skupisk słów** — wykorzystywana przez algorytm analiza składowych głównych (Principal Component Analysis — PCA) służy do zidentyfikowania skupisk słów (Nisbet, Elder, Miner 2009). Wyniki PCA prezentowane są w postaci wykresu rozrzutu. Celami analizy składowych głównych są zredukowanie liczby skorelowanych zmiennych do mniejszej liczby zmiennych zwanych składowymi, a później interpretacja najważniejszych składowych (Aranowska, Ciok 1992; Aranowska 1996, 2005). Analiza może ujawnić istnienie kilku składowych wyjaśniających zmienność całkowitą. Jednakże to pierwsza składowa tłumaczy największy procent zmienności, a każda kolejna składowa mniejszy (Aranowska, Ciok 1992; Bartholomew i in. 2008).

Interpretacji dokonuje się najczęściej między pierwszą i drugą składową, gdyż one dwie wyjaśniają największy procent zmienności wyników w zbiorze. Taka analiza przedstawiona została na rysunkach 3 i 4. Można odczytać, które słowa tworzą dane skupiska, dzięki prezentacji na wykresie.

c) **kodowania słów na zupełnie nowe zmienne**, którym tym razem przypisane są wartości liczbowe. Algorytm może przekształcić dane słowne na liczbowe. Czyni to przez nadanie słowom frekwencji i zapisanie ich w zbiorze danych. Na rysunku 2 przedstawiony został przykład fragmentu bazy z frekwencją przypisaną słowom.



Rysunek 1. *Output* prezentujący frekwencję słów
Źródło: opracowanie własne



File Edit View Insert Format Statistics Data Mining Graphs Tools Data Window Help

Add to Workbook Add to Report Add to MS Word

Arial 10 B I U Vars Cases

	odwaga	opanowanie	optymizm	otwartosc	pewnosc	poczucie	porzadek	porzadku
1	0,00000000	0,00000000	0,00000000	0,00000000	0,00000000	0,00000000	0,00000000	0,00000000
2	1,00000000	0,00000000	0,00000000	0,00000000	0,00000000	0,00000000	0,00000000	0,00000000
3	0,00000000	0,00000000	0,00000000	0,00000000	0,00000000	0,00000000	0,00000000	0,00000000
4	1,00000000	0,00000000	0,00000000	0,00000000	0,00000000	0,00000000	0,00000000	0,00000000
5	0,00000000	0,00000000	0,00000000	0,00000000	0,00000000	0,00000000	0,00000000	0,00000000
6	0,00000000	0,00000000	0,00000000	0,00000000	0,00000000	0,00000000	0,00000000	0,00000000
7	0,00000000	0,00000000	0,00000000	0,00000000	0,00000000	0,00000000	0,00000000	0,00000000
8	0,00000000	0,00000000	0,00000000	0,00000000	0,00000000	1,00000000	0,00000000	0,00000000
9	0,00000000	0,00000000	0,00000000	0,00000000	0,00000000	0,00000000	0,00000000	0,00000000
10	0,00000000	1,00000000	0,00000000	0,00000000	0,00000000	0,00000000	0,00000000	0,00000000
11	0,00000000	0,00000000	0,00000000	0,00000000	0,00000000	0,00000000	0,00000000	0,00000000
12	0,00000000	0,00000000	0,00000000	0,00000000	0,00000000	0,00000000	0,00000000	0,00000000
13	0,00000000	0,00000000	0,00000000	0,00000000	0,00000000	1,00000000	0,00000000	0,00000000

Rysunek 2. *Output* ukazujący przypisanie słowom frekwencji w bazie danych
Źródło: opracowanie własne

Text mining jest więc potężnym narzędziem również do transformacji danych tekstowych na liczbowe. Zapisuje on nowo transformowane dane do pliku z danymi liczbowymi, umożliwiając wykonywanie na nim kolejnych obliczeń. Nowe zmienne zapisane przez algorytmy *text mining* w bazie danych mogą zostać wykorzystane do dalszych analiz. Przy zastosowaniu algorytmów *text mining* można otrzymać w ciągu krótkiej chwili nowy zbiór danych, co zaoszczędza wielomiesięcznej niekiedy pracy. Co więcej, nie ma możliwości, żeby człowiek, badając setki tekstów, dokonał analizy skupisk słów tak, jak robią to algorytmy TM, ze względu na ograniczoną moc obliczeniową umysłu ludzkiego. Algorytmy znajdują w tekstach wzory i szukają słów sąsiadujących tworzących skupiska, dodatkowo analizę wykonują z matematyczną dokładnością[6]. Podsumowując, algorytmy *text mining* transponują dane jakościowe (słowa) na dane liczbowe, zliczają je, nadają im wagi i tworzą interpretowalne skupiska. Program STATISTICA Data Miner może analizować cały dokument tekstowy i posiada stronę internetową, która stanowi bazę dla prowadzonych wyliczeń. Można więc wprowadzać do programu cały plik tekstowy i nie trzeba kodować danych wcześniej. Algorytmy *text mining* same utworzą bazę do wyliczeń. Badacz nie musi więc poświęcać czasu na kodowanie słów, zapisywanie ich wartości w oprogramowaniu. Całą analizę może zrobić algorytm. Jest to ogromna oszczędność czasu w porównaniu z innymi metodami analizy tekstów. Przybliżmy działanie trzech algorytmów *text mining* na przykładzie badań psychologicznych.

[6] Należy dodać, że badacz może wskazać, jakich słów algorytm ma nie brać pod uwagę, takich jak np. spójniki, zaimki, może również zaznaczyć synonimy i parafrazy, jak również wymienić słowa, które go najbardziej interesują.

4. CEL BADANIA PRÓBA I PROCEDURA BADAWCZA

4.1. Plan badań

Badanie miało charakter analizy materiału tekstowego, którego dokonano za pomocą algorytmów *text mining*. Celem prowadzonych analiz była odpowiedź na pytanie badawcze: Jakie cele wychowawcze obierają rodzice dzieci trzyletnich i czteroletnich?

4.2. Zmienne zależne i niezależne uwzględnione w hipotezach

Zmienną niezależną stanowił wiek dziecka (grupa dzieci trzyletnich i czteroletnich). Zmienną zależną były cele wychowawcze obierane przez rodziców. Cele wychowawcze to cechy psychiczne, które w trakcie procesu wychowawczego rodzic chce ukształtować w dziecku (Brzezińska 2002; Glenn 2005; Miller 1966; Muszyński 1972; Sośnicki 1966).

4.3. Procedura badawcza i próba badana

Badanie prowadzone było drogą internetową na terenie Polski. Na stronie internetowej umieszczono opracowany wcześniej kwestionariusz dla rodziców dotyczący celów wychowawczych. Rodzice proszeni byli na początku badania o pomyślenie o swoim dziecku (tym, które uczęszcza do przedszkola) i do końca badań udzielanie odpowiedzi tylko na temat tego dziecka. Ta procedura chroniła przed krzyżowością udzielania odpowiedzi w wypadku, gdy rodzic miał więcej niż jedno dziecko. W badaniu udział wzięło 151 osób zarówno ojców, jak i matek dzieci przedszkolnych będących w wieku 3 i 4 lat. Przedział wieku osób badanych wynosił od 22 do 54 lat, z najliczniejszą reprezentacją osób pomiędzy 29. a 35. rokiem życia (można więc przyjąć, że była to grupa młodych dorosłych). Dominanta wyniosła 33 lata, a mediana 27 lat. Informacje na temat wykształcenia badanych osób zostały zaprezentowane w tabeli 1.

Wykształcenie	**Procent**
Średnie	33,73
Wyższe	61,45
Doktorat	4,82
∑	100

Tabela 1. Poziom wykształcenia osób w próbie badanej
Źródło: opracowanie własne

Na podstawie danych zaprezentowanych w tabeli 1 można stwierdzić, że w grupie osób badanych dominowały osoby dobrze wykształcone. Dominanta wskazuje na to, że najbardziej liczna była grupa osób po studiach. Jak widać (tabela 2), osoby badane

pochodziły głównie z dużych miast, ale w badaniu wzięły również udział osoby ze wsi i z mniejszych miejscowości.

Miejsce zamieszkania	Procent
Wieś	16,87
Miasto do 10 tys. mieszkańców	2,41
Miasto 15 tys.	15,66
Miasto 50–200 tys.	26,51
Miasto 200–500 tys.	16,87
Miasto > 500 tys.	21,68
∑	100

Tabela 2. Miejsce zamieszkania osób badanych
Źródło: opracowanie własne

W próbie badanej znalazła się podobna liczba rodziców chłopców i dziewczynek (tabela 3).

Płeć dziecka	Frekwencja	Procent
Chłopiec	68	45
Dziewczynka	83	55
∑	151	100

Tabela 3. Rozkład frekwencji płci dziecka
Źródło: opracowanie własne

W badaniu udział wzięli rodzice trzylatków i czterolatków uczęszczających do przedszkoli (tabela 4).

Wiek	Frekwencje	Procent
3 lata	66	43,7
4 lata	85	56,3
∑	151	100

Tabela 4. Frekwencja wieku dzieci
Źródło: opracowanie własne

Rozkład płci dzieci w poszczególnych grupach wiekowych prezentuje tabela 5.

Wiek	Płeć	Częstość	Procent
3 lata	Chłopcy	32	48,5
	Dziewczynki	34	51,5
	∑	66	100

Wiek	Płeć	Częstość	Procent
4 lata	Chłopcy	51	60
	Dziewczynki	34	40
	∑	85	100

Tabela 5: Rozkład płci dzieci w poszczególnych grupach wiekowych
Źródło: opracowanie własne

Dzieci rodziców z próby badanej uczęszczały do przedszkoli państwowych, ale również do placówek prywatnych. Dane na ten temat przedstawia tabela 6. Najwięcej dzieci chodziło do przedszkoli państwowych, a mniej do prywatnych.

Przedszkole	Frekwencje	Procent
Publiczne	79	52,3
Prywatne	45	29,8
Inne	27	17,9
∑	151	100

Tabela 6. Frekwencja dzieci uczęszczających do różnych przedszkoli
Źródło: opracowanie własne

4.4. Opis techniki operacjonalizacji zmiennej doboru celów wychowawczych

4.4.1. Skala Rozbieżności i jej własności psychometryczne

Do pomiaru celów wychowawczych rodziców wykorzystano skalę Rozbieżności, która przeznaczona jest do pomiaru celów wychowawczych rodziców oraz różnicy między celami wychowawczymi rodziców (cechami psychicznymi, które rodzic chce ukształtować w dziecku) a obecnym poziomem dziecka w zakresie rozwoju tych cech (Szymańska 2011, 2012b). Skala Rozbieżności mierzy więc cele wychowawcze oraz dystans między celem wychowawczym i aktualnym poziomem rozwoju dziecka w zakresie kształtowanych cech. Skala składa się z 12 pytań, które ułożone są parami. Każda para pytań mierzy odpowiedzi rodzica dotyczące jednego celu wychowawczego. Pierwsze pytanie w każdej parze odnosi się do celu wychowawczego. Rodzice proszeni są o wymienienie cechy, którą pragną ukształtować w dziecku (tabela 7). Jednocześnie na skali od –7 do 7 szacują, jak bardzo pragną, aby dziecko daną cechę w przyszłości posiadało. Drugie pytanie w parze dotyczy stopnia, w jakim dziecko ma obecnie rozwiniętą wspomnianą cechę. Rodzice na skali od –7 do 7 określają stopień posiadania tej cechy przez dziecko.

INSTRUKCJA
Proszę wymienić trzy cechy, które są dla Pani/Pana
jako rodzica szczególnie ważne i podejmuje Pani/Pan wysiłki, aby dziecko te cechy rozwinęło.
Cecha pierwsza (nazwa cechy):
Oceń ważność tej cechy dla Ciebie jako rodzica,
w jakim stopniu chciałbyś, aby Twoje dziecko takie było?
–7, –6, –5, –4, –3, –2, –1, 0, 1, 2, 3, 4, 5, 6, 7
(–7) zdecydowanie nie takie, (7) zdecydowanie takie
Oceń, w jakim stopniu (wpisz imię dziecka) posiada daną cechę?
–7, –6, –5, –4, –3, –2, –1, 0, 1, 2, 3, 4, 5, 6, 7
(–7) zdecydowanie nie posiada (7) zdecydowanie posiada

Tabela 7. Pierwsza para pytań skali Rozbieżności dotycząca celu wychowawczego
Źródło: opracowanie własne

Trzy pary pytań skali dotyczą cech pozytywnych (cech, które rodzice chcieliby, aby dziecko rozwinęło), kolejne trzy pary — cech negatywnych (niepożądanych przez rodzica). Różnicę między celem wychowawczym a obecnym stanem dziecka mierzy się za pomocą kwadratu odległości Euklidesowej. W ten sposób uzyskuje się sześć miar rozbieżności (trzy od pożądanego celu wychowawczego i trzy od celu niepożądanego).

5. WYNIKI

Analizy za pomocą algorytmów *text mining* przeprowadzono w trzech etapach, wykorzystując trzy algorytmy *text mining*:

1) pierwszy algorytm zliczył frekwencje i nadał wagi wymienionym przez rodziców celom wychowawczym,
2) drugi algorytm za pomocą analizy składowych głównych (Principal Component Analysis) zredukował słowa do głównych składowych i ujawnił ich najsilniejszy wkład w składowe,
3) trzeci algorytm dla wyłonionych słów zliczył frekwencje, a słowa zapisał w postaci zmiennych w bazie danych, którym przypisano częstość wymieniania przez osoby badane. W ten sposób algorytm zbudował nową bazę danych liczbowych na podstawie pierwszej bazy danych słownych. Ta baza będzie mogła zostać wykorzystana w przyszłości do kolejnych analiz.

5.1. Wyniki dla celów wychowawczych pożądanych

5.1.1. Frekwencje i ważność wymienionych przez rodziców celów wychowawczych

Dla cech, które rodzice chcą ukształtować w dziecku, wskaźnik ważności oraz frekwencje wymieniania przedstawiono w tabeli 8.

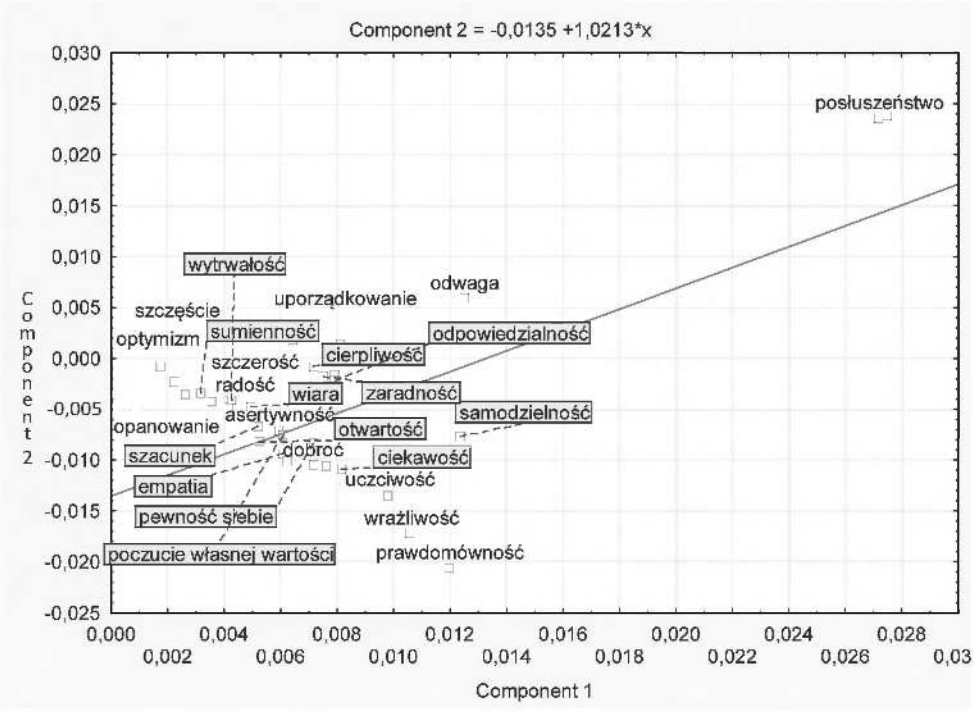
Lp.	Cechy	Frekwencje	Wskaźnik ważności	Klasyfikacja cechy
1	samodzielność	27,00	100,00	kompetencje
2	odwaga	15,00	98,89	cnota
3	posłuszeństwo	13,00	94,55	nieklasyfikowany
4	uczciwość	13,00	94,03	cnota
5	cierpliwość	11,00	92,82	cnota
6	empatia	11,00	91,55	kompetencje
7	prawdomówność	11,00	93,05	cnota
8	wrażliwość	9,00	90,39	temp. osobowy
9	ciekawość	8,00	88,80	temp. osobowy
10	odpowiedzialność	8,00	88,97	kompetencje
11	poczucie własnej wartości	8,00	89,18	temp. osobowy
12	otwartość	7,00	85,80	temp. osobowy
13	pewność siebie	7,00	85,65	temp. osobowy
14	szacunek	7,00	86,48	cnota
15	szczerość	7,00	88,04	cnota
16	wytrwałość	7,00	96,95	cnota
17	zaradność	7,00	86,61	kompetencje
18	asertywność	6,00	85,06	kompetencje
19	dobroć	6,00	83,39	cnota
20	radość	6,00	76,56	temp. osobowy
21	wiara	6,00	83,83	cnota
22	opanowanie	5,00	86,07	temp. osobowy
23	optymizm	5,00	80,86	temp. osobowy
24	uporządkowanie	5,00	87,42	cnota
25	sumienność	5,00	81,13	cnota
26	szczęście	5,00	80,46	temp. osobowy

Tabela 8. Wskaźniki wag i frekwencji cech, które rodzice chcą ukształtować w dziecku
Źródło: opracowanie własne

Cele wychowawcze, które zostały wskazane przez rodziców, można zaklasyfikować do trzech głównych obszarów: a) kompetencji (asertywność, samodzielność, empatia, zaradność, odpowiedzialność), b) cech temperamentalno-osobowych (pewność siebie, wrażliwość, otwartość, ciekawość, optymizm, szczęście, radość, poczucie własnej wartości, opanowanie) oraz c) cnót, które umożliwiają przestrzeganie zasad etycznych (odwaga, cierpliwość, uczciwość, szczerość, prawdomówność, szacunek, sumienność, uporządkowanie, wiara, dobroć, wytrwałość). Jedna cecha (posłuszeństwo) nie została przypisana do żadnego z wymienionych obszarów. Obliczenie frekwencji wymienionych przez rodziców celów z różnych wymiarów pozwala ustawić je w hierarchii. Najczęściej, bo aż 102 razy, rodzice wskazywali na cele związane z rozwojem cnót, a zatem z rozwojem dyspozycji dziecka do kierowania się względami moralnymi umożliwiającymi mu czyny zgodne z wartościami etycznymi. Na drugim miejscu wymieniano cele związane z kompetencjami (51 razy), a na trzecim — z cechami osobowo-temperamentalnymi

(37 razy). Oczywiście samo przypisanie cech do poszczególnych wymiarów może być przedmiotem dyskusji. Na przykład odwaga może zostać uznana za jeden z przejawów osobowości otwartej na doświadczenie. Częściej jednak cechę tę kojarzy się z cnotą kardynalną męstwa niż z poszukiwaniem wrażeń.

5.1.2. Wyniki analizy składowych głównych



Rysunek 3. Cechy, które rodzice chcą ukształtować u dziecka
Źródło: opracowanie własne

Rysunek 3 pokazuje skupiska podawanych przez badaną grupę pozytywnych celów wychowawczych. Można zauważyć, że cechą posiadającą największy ładunek jest posłuszeństwo, potem wskazano odwagę, samodzielność i prawdomówność. Te cechy są najważniejsze i najsilniej łączą się z pierwszym komponentem. W następnej kolejności dopiero wymieniono wrażliwość i uczciwość. Wszystkie cechy są pozytywnie związane z komponentem pierwszym. Możemy więc powiedzieć, że jest to komponent celów wychowawczych. Komponent drugi różnicuje już jednak cele. Najsilniej łączy się z posłuszeństwem i odwagą. Można również dostrzec, że takie cechy, jak prawdomówność, wrażliwość i uczciwość, są ze sobą silnie związane i tworzą skupisko, co oznacza, że

były przez rodziców wybierane razem. Ponieważ jednak są one ujemnie skorelowane z komponentem drugim, wiadomo, że nie były wybierane z cechami dodatnio związanymi z tym komponentem, takimi jak posłuszeństwo, odwaga. Linia przedstawiona na rysunku 3 dzieli słowa według dodatniego i ujemnego związku z komponentem drugim. Słowa występujące nad nią były z komponentem drugim związane dodatnio, a słowa poniżej tej linii — ujemnie. Słowa znajdujące się blisko siebie pojawiały się w wypowiedziach rodziców razem.

5.2. Wyniki dla celów wychowawczych niepożądanych

5.2.1. Frekwencje i ważność wymienionych przez rodziców celów wychowawczych

Podobną jak w wypadku celów pozytywnych analizę przeprowadzono również dla cech, których rodzic nie chce ukształtować w swoim dziecku. Wskaźnik ważności oraz frekwencje wymieniania poszczególnych cech przedstawiono w tabeli 9.

Lp.	Cechy	Frekwencje	Wskaźnik ważności	Klasyfikacja cechy
1	egoizm	26,00	98,45	przeciwieństwo cnoty
2	lenistwo	20,00	95,48	przeciwieństwo cnoty
3	agresja	18,00	95,30	temp. osobowy
4	agresywność	11,00	91,75	temp. osobowy
5	bałaganiarstwo	10,00	90,66	temp. osobowy
6	nieposłuszeństwo	9,00	89,35	nieklasyfikowany
7	kłamstwo	8,00	87,76	przeciwieństwo cnoty
8	arogancja	7,00	85,82	przeciwieństwo cnoty
9	nerwowość	7,00	85,82	temp. osobowy
10	samolubność	7,00	85,82	przeciwieństwo cnoty
11	brak szacunku	7,00	85,82	przeciwieństwo cnoty
12	chciwość	5,00	80,48	przeciwieństwo cnoty
13	zarozumiałość	5,00	80,48	temp. osobowy

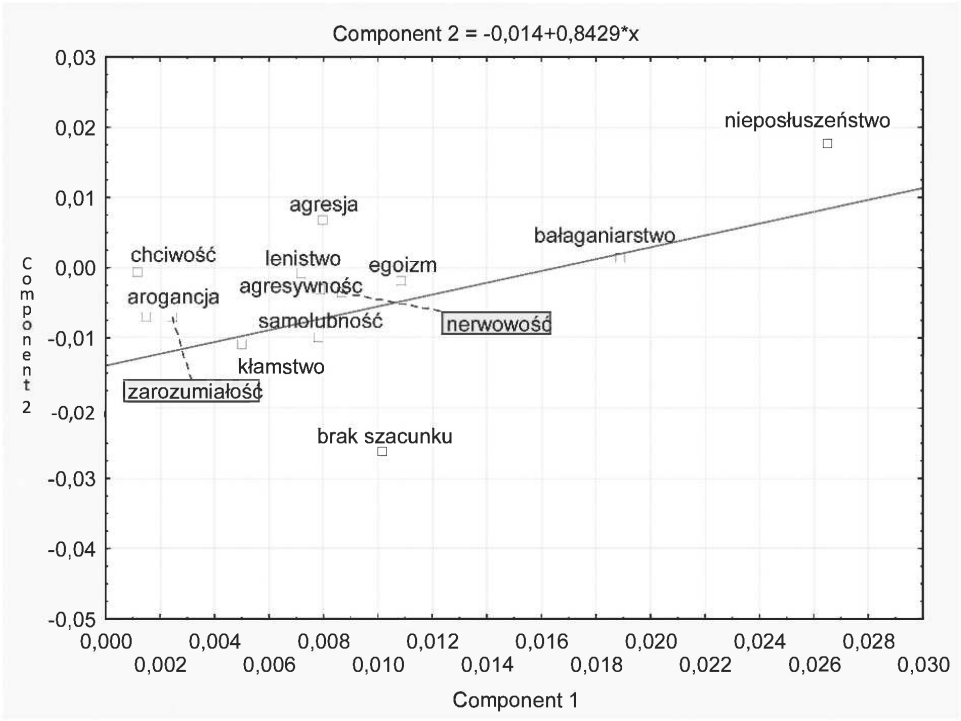
Tabela 9. Wskaźniki wag i frekwencji cech, których rodzice nie chcą ukształtować w dziecku
Źródło: opracowanie własne

Zauważmy, że nazwy cech negatywnych, podobnie jak pozytywnych, również są pojęciami z kręgu wymiarów osobowości i temperamentu oraz cnót lub ich przeciwieństw.

Tym razem rodzice najczęściej wymieniali cechy świadczące o przeciwieństwie cnót: lenistwo (przeciwieństwo pracowitości) i egoizm, kłamstwo (przeciwieństwo prawdomówności, szczerości i uczciwości) oraz brak szacunku (przeciwieństwo szacunku). Pierwsze miejsce zajmują cele związane z brakiem cnót (95). Na drugim pod względem frekwencji miejscu (51) rodzice podawali cechy, które powiązalibyśmy z wymiarami osobowości i temperamentu: agresywność, zarozumiałość, bałaganiarstwo, nerwowość.

5.2.2. Wyniki analizy składowych głównych

Na rysunku 4 widać, że najbardziej niepożądaną cechą jest nieposłuszeństwo. Ma ono najsilniejszy ładunek dla komponentu pierwszego, a kolejno pojawiają się: bałaganiarstwo, egoizm, brak szacunku. Wszystkie te cechy korelują dodatnio z pierwszym komponentem. Komponent drugi natomiast różnicuje cechy. Najsilniej związany jest on z nieposłuszeństwem, a ujemnie z brakiem szacunku. Można również zauważyć, że pozostałe cechy są bardzo blisko siebie i tworzą skupisko.



Rysunek 4. Cechy, których rodzice nie chcą ukształtować u dziecka
Źródło: opracowanie własne

Podsumowując, należy powiedzieć, że rodzice wymieniają cechy należące do obszaru cnót, cech osobowo-temperamentalnych oraz kompetencji dziecka. Daje się zauważyć, że cechy negatywne często są wymieniane jako przeciwieństwo cech pozytywnych, ale nie zweryfikowano, czy dzieje się tak w parach podawanych przez tę samą osobę. Warto też podkreślić, że rodzice w swoim systemie wychowawczym przywiązują dużą wagę do wartości etycznych.

5.3. Wnioski z badania

Wyniki ujawniają, że cele wychowawcze rodziców poza podstawową klasyfikacją na pożądane przez rodzica i niepożądane można podzielić na skupiska ze względu na rodzaje obieranych cech, które rodzice chcą, aby dziecko rozwinęło. Trzy podstawowe skupiska cech pożądanych tworzą cele dotyczące na pierwszym miejscu rozwoju etyczno-moralnego, na drugim kompetencji dzieci, a na trzecim rozwoju cech osobowo-temperamentalnych. Cele niepożądane można podzielić na dwa skupiska: pierwsze dotyczy rozwoju cech, które stanowiły przeciwieństwo cnót, drugie cech osobowo-temperamentalnych. Tabela 10 przedstawia te zestawienia.

Cele pożądane	**Cele niepożądane**
Etyczno-moralne (cnoty)	Przeciwieństwa cnót
Kompetencje	Osobowo-temperamentalne
Osobowo-temperamentalne	

Tabela 10. Zestawienia skupisk celów wychowawczych
Źródło: opracowanie własne

6. PODSUMOWANIE

Wykonanie zaprezentowanej analizy dotyczącej celów wychowawczych stało się możliwe jedynie dzięki aplikacji metody *text mining*. Bez jej wykorzystania niemożliwe byłoby przeprowadzenie analizy składowych głównych i ustalenie, które słowa są ze sobą powiązane. Omówiona już metoda tworzenia nowej bazy na podstawie frekwencji występowania słów może (a nawet powinna) zostać wykorzystana do kolejnych analiz. Z jakimi metodami może być łączona metoda *text mining*? Właściwie z wszystkimi. Baza skonstruowana przez algorytmy text mining może posłużyć do obliczeń ilościowych, może być również wykorzystana w specjalnej klasie modeli SEM, jakimi są modele MIMIC (Szymańska 2016). Szczególnie cenne wydają się połączenia analiz prowadzonych przez algorytm *text mining* z innymi metodami *data mining*. Bardzo ciekawe rozwiązanie można uzyskać z połączenia bazy przygotowanej przez algorytmy *text mining* z algorytmem decyzyjnym, jakim jest Classification & Regression Tree (Nisbet, Elder, Miner 2009). Inne metody analizy, z którymi może być wykorzystany *text mining*, to maszyna wektorów wspierających, sztuczne sieci neuronowe czy metoda Generalized EM & k-Means Cluster Analysis.

Należy zwrócić uwagę, że wszystkie wymienione powyżej metody analizy, z którymi proponuje się łączyć *text mining*, są nieparametryczne, a więc nie posiadają założeń dotyczących normalności rozkładu, homogeniczności czy równoliczności grup. Z wyjątkiem układów równań strukturalnych należą one do metod *data mining*.

Podsumowując, wykorzystanie algorytmów *text mining* w naukach psychologicznych może istotnie i silnie zredukować czas analizy i pozwolić na wykonanie obliczeń, które bez korzystania z tej klasy algorytmów byłyby niemożliwe. Dzięki algorytmom przekształcającym dane słowne w liczbowe można również przeprowadzić wiele innych analiz. Należy zwrócić uwagę, że dane pochodzące ze słów rzadko kiedy spełniają założenia leżące u podstaw statystyk parametrycznych. Algorytmy *data mining* nie posiadają takich założeń, stąd doskonale nadają się do łączenia ich z algorytmami *text mining*.

BIBLIOGRAFIA

Alqarni M., Arabi Y., Kakiashvili T., Khedr M., Koczkodaj W.W., Leszek J., Przelaskowski A., Rutkowski K. 2011: Improving the predictability of ICU illness severity scales, *Computer Science and Information Systems*, 11–17.

Aranowska E. 1996: *Metodologiczne problemy zastosowań modeli statystycznych w psychologii. Teoria i praktyka*, Warszawa: Studio 1.

Aranowska E. 2005: *Pomiar ilościowy w psychologii*, Warszawa: Wydawnictwo Naukowe Scholar.

Aranowska E., Ciok A. 1992: Związki między zmiennymi w interpretacji analizy składowych głównych i analizy korespondencji, [w:] Aranowska E. (red.), *Wybrane problemy metodologii badań*, Warszawa: Wydawnictwa Uniwersytetu Warszawskiego, 133–181.

Bartholomew D.J., Steele F., Moustaki I., Galbraith J.I. 2008: *Analysis of multivariate social science data*, Boca Raton, FL: Chapman & Hall/CRC Press.

Bouchet-Valat M., Bastin G. 2013: RcmdrPlugin.temis, a Graphical Integrated Text Mining Solution in R, *The R Journal* 5(1), 188–196.

Brzezińska A. 2002: *Społeczna psychologia rozwoju* [Social psychology of development], Warszawa: Wydawnictwo Naukowe Scholar.

Elder J., Hill T., Miner G., Nisbet B., Delen D., Fast A. 2012: *Practical Text Mining and Statistical Analysis for Nono-structured Text Data Application*, Oxford: Elsevier.

Franzosi R. 2010: *Quantitative narrative analysis*, Thousand Oaks, CA: SAGE Publications.

Glenn E. 2005: Incorporating parental goals in parenting programs through collaborative relationships with parents, *Journal of Extension* 43(1).

Hotho A., Nurnberger A., Paaß G. 2005: *A Brief Survey of Text Mining. A Brief Survey of Text Mining* (Ldv. Forum), <http://www.jlcl.org/2005_Heft1/19-62_HothoNuernbergerPaass.pdf>.

Krupa T. 1995: Sztuczna inteligencja, *Prace Naukowe Instytutu Technologii Maszyn i Automatyzacji Politechniki Wrocławskiej* 57, 146–165.

Luger G.F. Stubblefield W.A. 1989: *Artificial Intelligence and the Design of Expert Systems*, Redwood City, California: The Benjamin/Cummings Publishing Company, Inc.

Manning Ch.D., Schutze H. 2002: Foundations of statistical natural language processing, *ACM.SIGMOD Record* 31(3), 37, <http://doi.org/10.1145/601858.601867>.

Miller R. 1966: *Proces wychowania i jego wyniki* [The upbringing process and its results], Warszawa: Biblioteka Nauczyciela PZWS.

Muszyński H. 1972: *Ideał i cele wychowania* [The ideal and goals of upbringing], Warszawa: Biblioteka Nauczyciela PZWS.

Nisbet R., Elder J., Miner G. 2009: *Handbook of Statistical Analysis and Data Mining Applications*, Burlington, MA: Academic Press (Elsevier).

Rutkowski L. 2006: *Metody i techniki sztucznej inteligencji*, Warszawa: Wydawnictwo Naukowe PWN.

Rzechowska E. 2004: *Potencjalność w procesie rozwoju. Mikroanaliza konstruowania wiedzy w dziecięcych interakcjach rówieśniczych*, Lublin: Wydawnictwo KUL.

Rzechowska E. 2011a: *Dojrzały pracownik na rynku pracy. Jak zabezpieczyć przed wykluczeniem społecznym osoby 50+?*, Lublin: Wydawnictwo Lubelskiej Szkoły Biznesu.

Rzechowska E. 2011b: Podejście procesualne. Warianty badań nad procesami w mikro- i makroskali, *Roczniki Psychologiczne* 14(1), 127–157.

Sośnicki K. 1966: *Istota i cele wychowania* [The essence and goals of upbringing], Warszawa: Nasza Księgarnia.

Sparck Jones K. 1972: A Statistical Interpretation of Term Specificity and its Retrieval, *Journal of Documentation* 28(1), 11–21, <http://doi.org/10.1108/eb026526>.

Szymańska A. 2011: Parental Stress in an Upbringing Situation and Giving Children Help: A Model of the Phenomenon, *International Journal of Interdisciplinary Social Sciences* 6(3), 141–153.

Szymańska A. 2012a: Parental Directiveness as a Predictor of Children's Behavior at Kindergarten, *Psychology of Language and Communication* 16(3), 1–24.

Szymańska A. 2012b: Doświadczanie trudności w sytuacji wychowawczej a reprezentacja dziecka w umyśle rodzica: model zjawiska, *Psychologia Rozwojowa* 17(4), 79–91.

Szymańska A. 2015: Wykorzystanie algorytmu Text Mining do analizy przyjmowanych przez rodziców celów wychowawczych, Warszawa: XXIV Konferencja Psychologii Rozwojowej.

Szymańska A. 2016: Przejście od danych jakościowych do ilościowych — aplikacja algorytmu Text Mining do budowy modeli SEM, *XXV Jubileuszowa Ogólnopolska Konferencja Psychologii Rozwojowej*, Kraków: XXV Jubileuszowa Ogólnopolska Konferencja Psychologii Rozwojowej.

Tarwacka-Odolczyk A., Tomaszewski P., Szymańska A., Bokus B. 2014: Deaf children building narrative texts. Effect of adult-shared vs. non-shared perception of a picture story, *Psychology of Language and Communication* 18(2), 149–177.

Torebko K., Szymańska A. 2015: Zastosowanie algorytmu Classification & Regression Tree (C & RT) do wyjaśnienia relacji pomiędzy błędami wychowawczymi popełnianymi przez rodziców a rozwojem kompetencji emocjonalnych dzieci w wieku wczesnoszkolnym, Warszawa: XXIV Konferencja Psychologii Rozwojowej.

Ważyńska A., Szymańska A., Bartczak M., Bokus B. 2015: Przy okrągłym stole dialogowego Ja. Gdzie siedzi sceptyk?, [w:] Bokus B., Kosowska E. (red.), *O wątpieniu*, Piaseczno: Studio Lexem, 63–82.

Yim H.Y.B., Boo Y.L., Ebbeck M. 2014: A Study of Children's Musical Preference: A Data Mining Approach, *Australian Journal of Teacher Education* 39(2), 21–34.

Żurada J., Barski M., Jędruch W. 1996: *Sztuczne sieci neuronowe*. Warszawa: Państwowe Wydawnictwo Naukowe.

ABSTRACT

Usage of text mining algorithms to analyze textual data in psychology

Keywords: algorithms, text data, text mining.

In the psychology the analysis of data written in the form of texts are an important element of research work. Nevertheless, tools are still sought, methods that can enable rapid analysis of data recorded in the form of texts, because these analyzes are usually very time consuming. This article approximates the text mining method, which is particularly applicable in the analysis of information recorded in the form of text data. Analysing textual data using text mining algorithms is shown on the example of parents' choice of educational goals. The paper presents the way in which text mining algorithms: a) perform text analysis by counting words and weighting them, b) analyze relationships between words by means of Principal Component Analysis, c) convert verbal data into numerals by preparing a set data for subsequent calculations.