Krzysztof Tomanek
Uniwersytet Jagielloński
https://doi.org/10.18778/1733-8069.13.2.07

Metodyka dla analizy treści w projektach stosujących techniki text mining i rozwiązania CAQDAS piątej generacji

Abstrakt Projekty, w których przychodzi nam pracować z dużymi wolumenami danych tekstowych, pochodzących z rożnych źródeł i zapisanych w różnorodnych formatach, rodzą wiele dylematów natury metodologicznej, wymagają często niestandardowych decyzji i rozwiązań. W szczególności zadanie polegające na opracowaniu danych o różnorodnej jakości, nieustrukturyzowanych typu *quan* i *qual* wymagać może pracy, w której dynamicznie zmieniają się strategie analizy danych, sposoby przekształcania danych tekstowych. Artykuł opisuje przykład takiej właśnie „dynamicznej" metodyki. Wykazała ona swoją wartość w zadaniu polegającym na klasyfikacji wypowiedzi pisanych. W tak zarysowanym kontekście autor artykułu mierzy się z następującymi celami: (a) czy można zastosować oprogramowanie klasy CAQDAS do pracy półautomatycznej lub automatycznej zastępującej część manualnej pracy nad klasyfikacją wypowiedzi? (b) jak skonstruować metodykę klasyfikacji dla danych o różnorodnej jakości? (c) kiedy klasyfikacja automatyczna jest przydatna, a kiedy nie ma szans powodzenia?

W artykule zaznaczone zostaną momenty, w których analityk sięga po wiedzę typową dla analiz danych jakościowych oraz te, kiedy wiedza z tego obszaru nie jest już wystarczająca do realizacji wskazanych celów (*natural language processing*, uczenie maszynowe). Przykład projektu będący tłem artykułu wymusił zastosowanie kilku narzędzi i języków wspierających pracę na danych. Praca nad transformacją, klasyfikacją oraz wizualizacją wyników wymagała zastosowania bazy MySQL oraz programów: R, QDA Miner, Wordstat, QlikSense. Roli i ograniczeniom narzędzi klasy CAQDAS poświęconych zostało także kilka uwag.

Słowa kluczowe analiza treści, *Mixed Methods*, *Big Data*, techniki *text mining*, CAQDAS

Krzysztof Tomanek, napisał doktorat z zakresu nauk społecznych na Uniwersytecie Jagiellońskim. Główne obszary zainteresowania i rozwoju naukowego: metodologia nauk społecznych, *big data*, *data mining*, *text mining*, badania z zakresu zaufania i lojalności, wizualizacja danych oraz interaktywne wizualizacje danych. Autor artykułów naukowych i popularnonaukowych z zakresu praktycznych zastosowań analiz tekstu, *sentyment analysis*, CAQDAS.

Adres kontaktowy:
Uniwersytet Jagielloński, Instytut Socjologii
ul. Grodzka 52, 31-044 Kraków
e-mail: k_tomanek@wp.pl

Analiza treści stosowana jest wobec zróżnicowanych typów danych tekstowych. Innych strategii analitycznych wymaga praca z tekstami opublikowanymi na blogach czy wypowiedziami zapisanymi na forach dyskusyjnych, a innych praca z tekstami pamiętników czy transkrypcjami pochodzącymi z wywiadów grupowych. Dlatego też analiza treści tozbiór technik, które stosowane są do rozwiązywania różnorodnych problemów badawczych. Przykłady takich obszarów badawczych, które dynamizują rozwój technik analizy treści, to: *culturomics*, analiza opinii, analiza sentymentu.

Jedną z najczęściej cytowanych w ostatnich latach w obszarze *culturomics* jest publikacja pokazująca, jak zmieniała się literatura XX wieku pod względem zawartego w niej „ładunku emocjonalnego" – czyli, innymi słowy, pod względem stosowania słów nacechowanych emocjonalnie (Acerbi i in. 2013). Praca zespołu Alberto Acerbi'ego mieści się w obszarze analiz sentymentu. Sformułowanie „analiza sentymentu" (SA) bywa używane zamiennie (Liu 2012) z bardziej ogólnym – „analiza opinii" (AO). W obu przypadkach pierwszy człon frazy dotyczy automatycznych i półautomatycznych metod analizy treści. Celem tych metod jest identyfikowanie i klasyfikowanie wypowiedzi ze względu na pojawiające się w nich słowa i frazy kluczowe. W przypadku SA są to słowa nacechowane emocjonalnie, a w AO zakres poszukiwań obejmuje nie tylko opinie identyfikowane ze względu na temat czy formę wypowiedzi, ale także ze względu na obiekt, którego opinia dotyczy, oraz na przykład styl, w jakim wypowiedź została sformułowana. Analiza opinii ma zatem szerszy zakres przedmiotowy, a analiza sentymentu jest jednym z elementów tego obszaru (Liu 2012).

Jeszcze inny, stosunkowo nowy obszar, w którym analiza treści jest stosowana w sposób półautomatyczny i automatyczny, to stylometria (Rybicki, Heydel 2013; Eder 2014). Ten rodzaj analiz treści za cel stawia sobie charakterystykę sposobu posługiwania się językiem przez autora wypowiedzi. Celem takiej analizy może być zatem zbudowanie profilu autora tekstów, rozpoznawanie autorstwa tekstów (Zaśko-Zielińska 2014; Rosa 2015). Systemy dokonujące automatycznych stylometrycznych budowane są także dla języka polskiego (Eder, Piasecki 2015).

Cechą wspólną wspomnianych powyżej obszarów, w których stosowane są techniki analizy treści, jest fakt, że rozwijają się wraz z rozwojem technologii komputerowych. Analizując doniesienia z badań nad tekstami, z niewielką dozą ryzyka można stwierdzić, że analizy tekstu wspierane komputerowo coraz częściej służą naukowcom z różnorodnych dziedzin (Lieberman i in. 2007; Baptiste i in. 2011; Taboada i in. 2011). Powszechnie zautomatyzowane lub półautomatyczne analizy tekstu stosują naukowcy zajmujący się analizami kulturowymi (Baptiste i in. 2011), lingwiści (Lieberman i in. 2007: 713–716), historycy (Pagel, Atkinson, Meade 2007: 717–720), czy zajmujący się antropologią (DeWall i in. 2011: 200–207). Coraz częściej też analiza treści wspierana komputerowo stosowana jestprzez socjologów (Niedbalski 2014).

W każdym wspomnianym powyżej badaniu analizy treści realizowane były przy wsparciu narzędzi z obszaru CAQDAS (*Computer-Assisted-Qualitative-Data-Analysis-Software*). Wśród dostępnych istnieją takie rozwiązania CAQDAS, które dysponują

algorytmami kodującymi materiał tekstowy automatycznie. Niektóre z takich technik kodowania działają niczym czarne skrzynki. Analityk nie zna ich budowy ani sposobu, w jaki przetwarzają one dane. Co prawda oprogramowanie takie pozwala na przykład zdefiniować jednostkę analizy; określić słowa kluczowe; wskazać fragment tekstu, który wykorzystywany będzie jako wzorzec do kodowania (np. QDA Miner, R), ale nie daje pełnej wiedzy o sposobie działania techniki analitycznej. Istnieją też takie rozwiązania CAQDAS, które oferują transparentne metody i techniki wspierające pracę z kodowaniem tekstów. Taka sytuacja oznacza dla analityka możliwość nie tylko zapoznania się z definicją algorytmu, ale także jego modyfikację. Mamy więc do czynienia z sytuacją, w której oprogramowanie niesie ze sobą możliwość samodzielnego budowania systemów uczących się kodowania tekstów (R, Qualrus). Taka sytuacja sprzyja rozwojowi metod analiz danych jakościowych. CAQDAS, które pozwala projektować metody i algorytmy, to niemal „nieograniczone" środowisko dla wyobraźni analityka.

Omawiana w niniejszym artykule analiza treści sięga po narzędzia CAQDAS, które pozwalają na projektowanie technik, algorytmów w sposób, dla którego ograniczeniem jest jedynie wyobraźnia analityka. Fakt ten pozwala rozszerzyć propozycję Jakuba Niedbalskiego dotyczącą klasyfikacji CAQDAS (Niedbalski 2013: 153–166). Zastosowanie narzędzi otwartych programistycznie otwiera rozdział piątej generacji CAQDAS. Ten nowy etap ewolucji oprogramowania wspierającego analizy danych jakościowych poza możliwościami analitycznymi, jakie daje, posiada jeszcze jedną istotną cechę. Spełnia mianowicie istotne – z punktów widzenia poznania naukowego – kryterium transparentności metodologicznej w prowadzonych analizach.

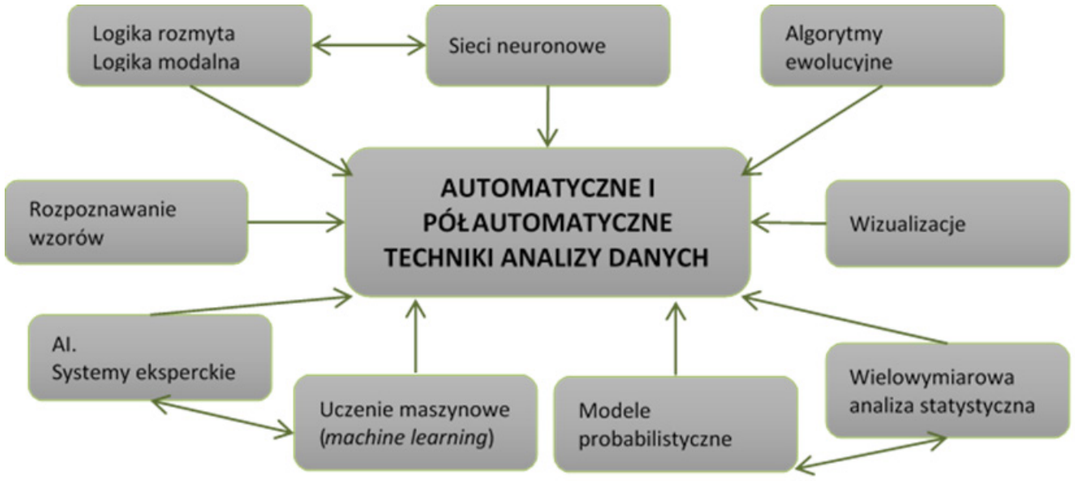
Słowo o zastosowaniu technik półautomatycznych i automatycznych

Wzrost mocy obliczeniowej komputerów osobistych oraz pojemności ich pamięci stworzył w ostatnich latach możliwości zarówno gromadzenia olbrzymich ilości informacji, jak i ich szybkiego przetwarzania. W konsekwencji obserwujemy szybki rozwój różnorodnych automatycznych i półautomatycznych metod analizy danych, technik odkrywania wiedzy również w danychtekstowych. Dla uproszczenia na ilustracji 1 pokazuję obszary tych rozwiązań.

Automatyczne i półautomatyczne metody analiz stosowane są w bardzo wielu dziedzinach: od badań medycznych do przewidywania kursów giełdowych, od przemysłu do gier komputerowych. Stosowane są również w analizie treści. Oto kilka przykładowych zastosowań dla różnorodnych problemów badawczych:

a. rozpoznawanie metod argumentacji (Tomanek 2016),
b. wykrywanie wypowiedzi prowadzących do napięć i konfliktów podczas FGI (Jordan i in. 2007),
c. identyfikacja wypowiedzi tekstowych nacechowanych emocjonalnie i wywołujących emocje (Duggleby 2005),
d. identyfikacja form perswazji w tekstach prasowych (Appel, Richter 2010),

Diagram 1. Obszary, z których korzystają i w których rozwijane są automatyczne i półautomatyczne metody analizy danych



Źródło: opracowanie własne na podstawie Duch (1997), Bolc, Cytowski (1989–1991), Cichosz (2000).

e. rozpoznawanie typowych wypowiedzi w wywiadach IDI,
f. analiza schematów poznawczych w komentarzach oceniających (Kidd, Parshall 2000),
g. identyfikacja problemów drażliwych w wywiadach FGI (Kaplowitz 2000),
h. klasyfikacja dużych zbiorów artykułów prasowych w oparciu o schematy kodowania (Schanie, Pino-Foltz, Logsdon 2008),
i. diagnostyka specyfiki wypowiedzi zwiększających popularność postów na forach dyskusyjnych (Tomanek 2015),
j. identyfikacja liderów opinii, grup opiniotwórczych na forach internetowych (Smith i in. 2009),
k. rozwój tematycznych słowników analitycznych w oparciu o reguły leksykalne (Tomanek, Bryda 2015).

Charakterystyka projektu

Techniki półautomatycznej i automatycznej analizy treści opisane w tym artykule zastosowane zostały wobec dużego zbioru wypowiedzi i opinii pozyskanych w trakcie badań ankietowych. Badania, o których mowa, realizowane były w tradycyjnej formule (wywiady *face to face*) oraz z zastosowaniem podejścia CAWI. Zadania respondentów biorących udział w badaniu polegały na zapisaniu: skojarzeń związanych z markami producentów i produktami konsumenckimi; uzasadnień oceny marki i produktu[1].

[1] Ze względu na prośbę zleceniodawcy projektu nazwa badań zostanie pominięta, a sam opis badania z konieczności pozostaje skrócony do minimum.

Ankieta zrealizowanazostała wśród względnie licznych grup respondentów (średnio rocznie uzyskano 10 000 odpowiedzi). Zbiór danych składał się z informacji liczbowych, jak i tekstowych. Ten drugi typ danych to wypowiedzi pisane będące odpowiedziami na pytania otwarte. Projekt, który realizowany był w cyklu rocznym (w latach 2012–2015), zwiększał liczbę respondentówi gromadził coraz większą liczbęwypowiedzi tekstowych. Pierwsza edycja badań przyniosła stosunkowo dużą liczbę 34 453 opinii[2]. W kolejnych latach liczba danych tekstowych przyrasta o ponad 10, 5 i 27 tysięcy. Między 2012 i 2015 rokiem korpus liczy 207 907 wypowiedzi (przyrost pomiędzy pierwszym i ostatnim rokiem badań to 44%[3]). W każdej edycji projektu analizy tekstów mają na celu wychwycenie wypowiedzi „cennych" z punktu widzenia prezentacji danych w raporcie końcowym (liczba wybieranych do raportu wypowiedzi oznaczona jest w tabeli jako „wybrane"). Szczegółowe dane prezentuje tabela 1.

Przyrost danych w opisanym powyżej tempie ma wpływ na kilka aspektów realizacji projektu. Istotnej zmianie ulegają: czas opracowania danych (Cole i in. 1996), etap analizy oraz wizualizacji, a także koszt realizacji projektu. W takiej sytuacji naturalnymi wydają się pytania:

a. Czy część najbardziej czasochłonnej pracy wykonywanej przy analizie tekstów można w jakimś stopniu zautomatyzować?

b. Czy w wyniku automatyzacji uzyskamy dane o takiej samej jakości jak w przypadku analiz wykonywanych manualnie? Czy automatyzacja opracowania danych da w efekcie trafne rezultaty?

c. Czy metoda i narzędzie wykonujące pracę automatycznie mogą korzystać z wiedzy badaczy, koderów?

Tabela 1. Liczba zebranych wypowiedzi tekstowych w badaniach realizowanych w latach 2012–2015

		Skojarzenia		Uzasadnienia		Łączna liczba wypowiedzi		
		Łącznie w zbiorze	Wybranych	Łącznie w zbiorze	Wybranych	Łącznie w zbiorze	Wybranych	Wybranych %
II etap analiz	I edycja projektu rok 2012	22 431	450	12 022	450	34 453	900	2,6%
	I edycja projektu rok 2013	27 682	630	17 145	630	44 827	1 260	2,8%
	I edycja projektu rok 2014	31 820	1 200	18 685	1 600	50 505	2 800	5,5%
I etap	I edycja projektu rok 2015	45 506	1 452	32 616	1 473	78 122	2 925	3,7%
	Łącznie	127 439	3 732	80 468	4 153	207 907	7 885	3,8%

Źródło: opracowanie własne.

[2] W tabeli 1 jest to kategoria oznaczona jako „Łączenie w zbiorze" – ta kategoria odnosi się osobno do skojarzeń, jak i uzasadnień, jest też zsumowana w części tabeli opisanej jako „Łączna liczba wypowiedzi".

[3] Każda edycja badania oparta jest na realizacji wywiadów ankietowych: metodą tradycyjną (spotkanie ankietera z respondentem i rozmowa twarzą w twarz) oraz poprzez kwestionariusz dostępny *online*). Proporcja wyników w badaniach tradycyjnych i *online* zmieniała się z roku na rok z 25% do 10% na rzecz badań *online*.

Dalsza część rozważań poświęcona będzie poszukiwaniu odpowiedzi na sformułowane powyżej pytania. Kontekstem i przykładem dla tych dociekań będą dane zgromadzone podczas realizacji opisanego projektu. Najogólniej cel analityczny dotyczący pracy na wypowiedziach tekstowych brzmiał następująco: w jaki sposób możliwy jest wybór wypowiedzi „wartościowych"/„cennych"? Cel ten pierwotnie realizowany był manualnie przez zespół koderów. Ich praca przyniosła wiele wzorców wypowiedzi, które uznane były za wartościowe i prezentowane były w raporcie końcowym z przeprowadzonych badań. Wyniki pracy koderów stały się podstawą do zaprojektowania zbioru algorytmów, które posłużyły do klasyfikacji opinii respondentów. Zadanie to realizowane było z zastosowaniem metod półautomatycznych, jak i automatycznych. W dalszej części rozważań opisane zostaną dwie wykorzystane w praktyce strategie analiz (w tabeli nr 1 – I strategia to „I etap analiz", a II strategia to „II etap analiz"). Obie zakończyły się budową dwóch odrębnych klasyfikatorów treści.

Metodologia

Podczas czterech lat realizacji projektu zastosowanych zostało kilka strategii analiz oraz przetestowanych zostało kilka narzędzi analitycznych. Pierwszą i podstawową w początkowych edycjach projektu była strategia rozumiejącego czytania tekstów, manualnego kodowania i klasyfikacji wypowiedzi (tradycyjna strategia analizy treści). Wraz z przyrostem danych tekstowych stosowane były metody półautomatycznego i automatycznego kodowania tekstów. Etap klasyfikacji półautomatycznej oznaczał, iż wyniki pracy algorytmu weryfikowane były pod kątem ich trafności przez koderów. Etap weryfikacji za każdym razem przynosił uwagi krytyczne, które następnie formułowane były jako dodatkowe warunki logiczne, zgodnie z którymi działał automatyczny klasyfikator treści. Zastosowanie takiej logiki działania (automatyczna klasyfikacja → weryfikacja → poprawa klasyfikacji automatycznej) było niezbędne do prowadzenia eksperymentów metodologicznych, które miały na celu odpowiedź na pytanie sformułowane w tej pracy (por. wcześniejsza strona, pytania: a, b, c).

Warto dodać jednak, iż w trakcie realizacji projektu strategie metodologiczne ewoluowały w każdym niemal obszarze związanym z opracowaniem danych tekstowych:

1. Od manualnej pracy nad wyborem tekstów o wysokiej jakości do automatycznej eliminacji wypowiedzi małowartościowych (kryteria eliminacji wypowiedzi omówione zostaną w dalszej części tekstu).

2. Od strategii kodowania otwartego jednoetapowego do kodowania dwustopniowego (wstępnej klasyfikacji, a następnie kodowania zogniskowanego).

3. Od pracy manualnej przy kodowaniu i klasyfikacji tekstów do klasyfikacji automatycznej i do metod półautomatycznych weryfikowanych przez koderów (I etap analizy oparty na danych z lat 2014–2015; I etap analizy oparty na danych z lat 2012–2015).

4. Od pracy zorganizowanej liniowo, gdzie analizy manualne poprzedzają automatycznedo pracy przebiegającej symultanicznie na kilku zadaniach analitycznych równocześnie (podejście specyficzne dla tak zwanych metody zwinnych).

5. Od stosowania zamkniętych programistycznie narzędzi CAQDAS do stosowania narzędzi

pozwalających na samodzielne projektowanie: reguł klasyfikacji, funkcji ważenia, miar określających jakość uzyskanych rezultatów – tak zwana 5 generacja narzędzi CAQDAS (Tomanek 2014b).

Opisane powyżej kierunki zmian strategii były wynikiem dwóch obserwacji. Po pierwsze, już po zakończeniu pierwszej edycji projektu wiadome było, że dwa analizowane typy wypowiedzi różnią się w sposób znaczący. Uzasadnienia osiągają w najlepszych przypadkach kilkanaście słów (przeciętna długość zdania złożonego). Skojarzenia są natomiast wypowiedziami krótkimi w postaci jednego słowa, frazy, kilku słów (niezmiernie rzadko zdarzały się wypowiedzi zapisane w formie pełnego zdania). Te dwa typy wypowiedzi wymagają zmian zarówno w doborze algorytmów przeszukujących treści, jak i w sposobie ich wykorzystania. Te dwie lekcje wyciągnięte z procesu diagnostyki wypowiedzi zmieniają również proces prowadzenia analiz. Zmiany zastosowane w analizie opisane są poniżej.

Ad 1.

Praca nad wyborem wypowiedzi skupiała się pierwotnie na podejmowaniu decyzji opartych na czytaniu wypowiedzi. Szybko jednak można było się przekonać, że zarówno wśród skojarzeń, jak i uzasadnień pojawiają się wypowiedzi będące wyrazem emocji niezwiązanych z treścią pytania zadanego w badaniu; zapisem przypadkowych ruchów palców po klawiaturze (np. ciągi typu „dsdsdssds", „eqweweqw"). W związku z możliwością wychwycenia niektórych z tych nietypowych wypowiedzi skonstruowany został algorytm, który je eliminował automatycznie. Podstawowe warunki pracy takiego algorytmu zakładały:

1. Wykluczenie wszystkich obserwacji zawierających mniej niż 5 znaków (najkrótsze spośród zidentyfikowanych słów, które okazywały się atrakcyjne, składały się z więcej niż pięciu znaków, na przykład – skojarzenie: piękna. Decyzja ta przeszła przez kilka testów, w których koder oceniał eliminowane i klasyfikowane przez algorytm wyniki.

2. Eliminacja wszystkich wypowiedzi, w których pojawił się przynajmniej jeden wulgaryzm (w tym celu zastosowane zostało tak zwane podejście słownikowe, które w tekstach wyszukuje słowa uwzględnione w słowniku) (Bolasco, Ratta-Rinaldi 2004; Tomanek 2014a).

3. Wykluczenie wszystkich obserwacji, w których nie pojawiło się przynajmniej jedno słowo możliwe do lematyzacji.

4. Eliminacja wypowiedzi, które bez spacji zawierają więcej niż 13 znaków (przykładem wypowiedzi, która bliska jest tej granicy i jest klasyfikowana może być słowo: „fantastycznie". Przykłademciągu znaków eliminowanych może być: sadadasdadassad.

Ad 2.

Kodowanie otwarte ma ten niewątpliwy walor, że dostarcza do dalszych analiz większej dawki tekstu wraz z kontekstem, w którym pojawia się kluczowy, najistotniejszy dla wypowiedzi fragment. O ile jednak strategia ta jest cenna w pracy nad tekstami

literackimi, zapisami indywidualnych wywiadów pogłębionych lub wywiadów grupowych, o tyle staje się kłopotliwa dla zadań skupionych na automatyzacji analiz, w szczególności dla tych, które prowadzone są na krótkich wypowiedziach w formie równoważników zdań, fraz. Kodowanie dwustopniowe pozwala na precyzyjniejszy wybór tekstów przeznaczonych do prezentacji w raporcie.

Ad 3.

Manualna praca koderów w trakcie kolejnych edycji projektu trwała coraz dłużej. Od jednego tygodnia w pierwszym projekcie do trzech w ostatniej, czwartej edycji. Zastosowanie automatycznych metody czyszczenia tekstów oraz ich wstępnej klasyfikacji (ad 1) pozwoliło na skrócenie czasu pracy bez straty jakości wyników. Bardziej ambitne zadanie polegające na identyfikacji i klasyfikacji „cennych" wypowiedzi prowadzić miało do w pełni automatycznej klasyfikacji. Zamysł ten (I etap oparty na danych z lat 2014–2015) realizowany był w następujących etapach:

1. Analiza danych z lat 2014–2015 i wyodrębnienie wypowiedzi eksperckich jako wzorcowych wypowiedzi stosowanych w procesie uczenia klasyfikatora automatycznego.

2. Budowa zbioru uczącego (do szkolenia algorytmu klasyfikacyjnego) i testowego (do testowania trafności klasyfikacji), na których prowadzone były analizy.

Wstępne prace nad algorytmem, który automatycznie „czyścił" dane tekstowe z wypowiedzi niepożądanych, przynosiły wyniki podobne do wyborów losowych. Na 100 wybranych wypowiedzi (w kilku niezależnych losowaniach) wyniki uzyskiwały trafność na poziomie 48% do 51% (4 lub 5 wypowiedzi na 10 nadawało się do prezentacji w raporcie), a pozostałe wymagały usunięcia z analiz. Tak skonstruowana metoda ani nie dawała wartościowych wyników (lepszych niż wybór losowy), ani nie skracała czasu pracy. Stąd też konieczna była poprawka do strategii opartej na pełnej automatyzacji. Ta zmiana oparta została na rozwinięciu podejścia półautomatycznego. Polegało ono na weryfikacji nietrafnie wybranych przez algorytm wypowiedzi oraz wypowiedzi odrzuconych przez algorytm. Wyniki tej pracy pozwoliły na wprowadzenie praktycznych zmian w budowie metody klasyfikującej wypowiedzi (zmiany te opisane są w dalszej części tekstu).

Warto podkreślić, że sukces wskazanej tu strategii opiera się na konstrukcji algorytmu, który czerpie z wniosków dostarczanych przez koderów. Analiza oparta na tych wnioskach rozwijana była na próbce tekstów (zbiór uczący), a następnie wyniki uczenia algorytmu weryfikowane były na zbiorze wypowiedziwcześniej nieanalizowanych (zbiór testowy). Weryfikacja obu tych analiz (uczenie, testowanie) dała lepszą kontrolę nad budową metody klasyfikacji, a co za tym idzie – nad jakością uzyskanych wyników.

Ad 4.

W pracy z algorytmami uczącymi się i metodami półautomatycznymi kluczowa jest możliwość realizacji badań nad tekstem w sposób symultaniczny

(z zastosowaniem różnych metod analiz na tym samym zbiorze tekstów). W związku z tym przyjęto założenie, iż prace nad automatycznym klasyfikatorem prowadzone będą z zastosowaniem równolegle kilku metod. Wśród nich znalazły się: regresja logistyczna, regresja logistyczna karana – typu Lasso, drzewa decyzyjne (CART), Support Vector Machines (C-SVM oraz One-Class SVM), Naiwny Klasyfikator Bayesowski. Praca kilku osób w jednym czasie nad tym samym zagadnieniem wymagała również sprawnej wymiany wiedzy i wzajemnego informowania się o postępach (sukcesach, jak i porażkach) w testowaniu różnych algorytmów. Tak zrodziła się potrzeba stosowania metodyk zwinnych (*agile methods*, na przykład metodyki *scrum*) (Schwaber 2013). To zwinne podejście w prowadzeniu prowadziło analityków od analiz jakościowych do analiz ilościowych, i ponownie od podejścia *quan* do *qual*.

Ad 5.

Zastosowanie automatycznej klasyfikacji wypowiedzi na te poddawane dalszej analizie i te eliminowane z analiz pozwoliło na zastosowanie podstawowych technik stosowanych w ramach *text mining*. Były to:

1. Parsowanie – unifikacja struktury tekstu: dekompozycja danych tekstowych, ilościowa reprezentacja zbioru dokumentów.

2. Transformacja i redukcja wymiarów: transformacja reprezentacji tekstu do formy ilościowej; redukcja wymiarów do zwartego formatu informacyjnego.

3. Analiza: zastosowanie algorytmów analizujących tekst i budujących reguły klasyfikacji.

Aby możliwe było zrealizowanie wskazanych prac na tekście, konieczne było stosowanie otwartych narzędzi analiz tekstu. Potrzebne było zatem oprogramowanie piątej generacji CAQDAS. W przypadku tego projektu wybór padło na program R[4]. To rozwiązanie programistyczne daje przede wszystkim możliwości rozwoju i implementacji różnorodnych algorytmów z zakresu analiz jakościowych.

Opis badań i analiz

Kiedy analiza dotyczy kilkudziesięciu tysięcy wypowiedzi, pojawia się pokusa, by zastosować metody, które wykonają zadanie analityczne „częściowo" za nas – badaczy. Wśród metod tych istnieją dwa charakterystyczne podejścia (Sołdacki 2006). Są to:

a. Głęboka analiza tekstu (ang. *Deep Text Processing*, DTP): to podejście opiera się na komputerowej analizie lingwistycznej wielu możliwych interpretacji, powiązań między słowami, frazami, relacji gramatycznych występujących w tekście. Z powodu tych „wielu możliwych interpretacji" taka analiza nie zawsze jest potrzebna, a także możliwa do osiągnięcia w czasie, jaki mamy na nią przeznaczony, czy też w oparciu o narzędzia, jakimi dysponujemy. Z tych też powodów coraz częściej realizowana jest częściowa czy płytka analiza tekstu.

b. Płytka analiza tekstu (ang. *Shallow Text Processing*, STP): efekt tej analizy jest niepełny w stosunku do rezultatów DTP (Piskorski 2001). Analiza tego typu rozpoznaje: słowa i ich odmiany; zapis w liczbie mnogiej i pojedynczej; przymiotniki; nazwy własne; podmiot, do którego wypowiedź się odnosi oraz identyfikuje jego rodzaj. Pomijane są tu bardziej złożone problemy, takie jak: rozpoznawanie ironii, identyfikacja emocji, metafory.

[4] Zob. https://www.rstudio.com/.

Zadanie, które jest tu opisywane (identyfikacja „cennych" skojarzeń i uzasadnień), nie jest takim, które wymaga analizy DTP. Pożądanym wynikiem pracy nad wypowiedziami jest wyodrębnienie tych, które wnoszą „istotną" informację o podmiocie wypowiedzi. Istotnym, dla przykładu, nie jest określenie: „wydarzenie było super" (jest to wypowiedź raczej trywialna ze względu na wartość informacyjną), ale już „takie wydarzenia zawsze gromadzą liczną publiczność" wnosi treść istotną z punktu widzenia oceny wydarzenia. Do identyfikacji krótkich wypowiedzi wystarczająca jest często analiza STP, która z kolei realizowana była na dwa odmienne sposoby (Tomanek 2014c):

a. Analiza oparta na metodach słownikowych: w tej strategii skupiamy się na identyfikacji słów (Key-Word-in-Context), fraz (Key-Phrase-in-Context), słów z określonych obszarów tematycznych (Bag-of-Words). Identyfikacji wypowiedzi istotnych służą reguły logiczne, syntaktyczne i uproszczona analiza kontekstu, w jakim słowa i frazy występują.

b. Metody statystyczne: w tym przypadku skupiamy się na automatycznym przetwarzaniu treści w oparciu o przyjęte uprzednio wzory wypowiedzi pożądanych (czyli takich, jakich szukamy) lub też posługujemy się automatyczną klasyfikacją bezwzorcową.

Oba zarysowane podejścia znacząco ograniczają zakres analizy tekstu. Dla zadania, z jakim się tu mierzymy, jest to ich silna strona. Oba jednak mogą dostarczać znacząco odmiennych wyników. Oto kilka najważniejszych powodów pozwalających uzasadnić ten wniosek:

1. O ile analiza słownikowa może czerpać z analizy tematycznej – to znaczy bierze pod uwagę znaczenie słów (posługuje się słowami sklasyfikowanymi w ramach określonych tematów), o tyle analiza statystyczna może w ogóle nie sięgać do znaczeń. Podstawą analizy statystycznej może stać się tylko i wyłącznie liczbowa reprezentacja słowa, na przykład długość słowa albo jego unikalność (rzadkie występowanie) w różnych wypowiedziach.

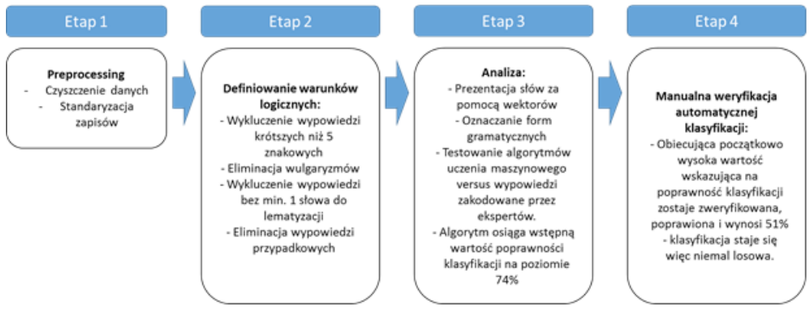
2. Analiza słownikowa: sięgapo wiedzę ekspercką (czyli po wypowiedzi zidentyfikowane przez koderów), ale też rozszerza ją o dodatkowe możliwe wypowiedzi. Dzieje się to przez rozbudowanie algorytmu przeszukującego tekst o frazy, które są równoważne znaczeniowo słowom kluczowym stosowanym we wcześniejszych analizach (regułą podstawową dla rozszerzania zbioru słów kluczowych jest synonimia); innym rozszerzeniem jest identyfikacja wypowiedzi biegunowo różnych (regułą pozwalającą na identyfikację słów znaczeniowo-biegunowo różnych jest antonimia).

3. Analiza słownikowa: posługuje się regułami logicznymi pozwalającymi na analizę kontekstu,

w jakim słowa i frazy występują. Dzięki prostym zapisom uwzględniającym operatory logiki możemy zidentyfikować: wypowiedzi sprzeczne znaczeniowo; wypowiedzi o konkretnych miastach, osobach, wydarzeniach (Gonzalez, Dankel 1993). Analiza statystyczna, posługując się takimi miarami jak korelacja, unikalność słów, pomaga w znalezieniu pewnych wzorów mówiących o występowaniu słów w tekście, ale nie podpowiada niczego w kontekście znaczeń analizowanych treści. W szczególności analiza statystyczna podpowiadać może zależności gramatycznepomiędzy elementami wypowiedzi.

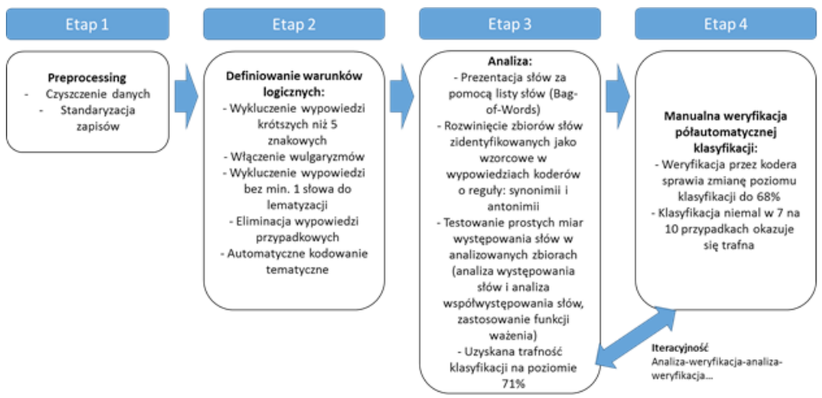
Wskazane powyżej powody stały u podstaw przeprowadzenia dwóch niezależnych analiz. Jednej – analizy statystycznej opartej na metodach wskazanych wcześniej w tekście (por. ad 4) oraz na metodzie słownikowej. Obrazowo przebieg obu analiz można zaprezentować tak, jak na diagramach 2 i 3.

Diagram 2. Przebieg procesu analiz z zastosowaniem analiz statystycznych



Źródło: opracowanie własne.

Diagram 3. Przebieg procesu analiz z zastosowaniem analizy słownikowej



Źródło: opracowanie własne.

Etapy 1, 2 i 4 opisane zostały we wcześniejszych fragmentach tekstu. Etap 3 to wielowymiarowa analiza tekstu. Na tym etapie każdej analizowanej wypowiedzi przypisane zostają wartości liczbowe, które oznaczają: liczbę słów w wypowiedzi, unikalność słowa (brak słowa we wzorcowych wypowiedziach wskazanych przez koderów), fakt wystąpienia słowa we wzorcowych wypowiedziach wskazanych przez koderów, wystąpienie w wypowiedzi wyrazu identyfikującego nazwę usługodawcy lub produktu.

Etap 3 w tej strategii – poza wskazanymi w analizie statystycznej – uwzględniał dodatkowo następujące zabiegi analityczne: zastosowanie stop listy (listy słów nieistotnych), identyfikacja wypowiedzi zawierających słowa oceniające (plus ich synonimy), identyfikacja wypowiedzi zawierających minimum dwa słowa o przeciwnym wydźwięku (wypowiedzi o zabarwieniu ambiwalentnym), identyfikacja wypowiedzi oceniających wraz z przeczeniem.

Dwie zarysowane powyżej strategie dały znacząco odmienne wyniki w kontekście poprawności klasyfikacji. W przypadku automatycznych analiz statystycznych poprawność wyboru wypowiedzi wyniosła 74% i została zweryfikowana przez koderów do 51%. Klasyfikacja stosująca metodę słownikową osiągnęła poziom poprawności wyboru równy 71%. Po weryfikacji wynik ten utrzymał się na poziomie 68%. Kluczowym w tej strategii okazała się iteracyjność przebiegu weryfikacji: praca kodera kończyła się zaleceniami wskazującymi, jak należy zmienić klasyfikator automatyczny – po zmianie dokonanej w algorytmie następowała kolejna ocena dokonywana przez koderów oraz kolejna automatyczna klasyfikacja. Strategia ta wykonana została trzykrotnie. Dzięki temu zabiegowi wzrosła trafność identyfikacji wypowiedzi bliskich znaczeniowo, a także biegunowo odmiennych. Ta podstawowa przewaga jednak nie daje 100% trafności klasyfikacji. Idealna klasyfikacja na poziomie 100% możliwa jest do osiągnięcia dla: języków nienaturalnych (na przykład dla języków skryptowych, języków programowania); wypowiedzi w dobrze opisanym języku (być może na przykład w łacinie); wypowiedzi silnie sformalizowanych (być może np. instrukcje obsługi urządzeń mechanicznych). Mimo iż nie udało się osiągnąć tego idealnego poziomu trafności klasyfikacji, wydaje się, że metoda półautomatyczna wykorzystująca strategię analizy słownikowej oraz weryfikację wykonaną przez koderów pozwala na redukcję prac manualnych w pracy nad tekstami – co można poczytywać za sukces tego podejścia.

Wnioski

Zastosowanie metod półautomatycznej i automatycznej w analizie treści niesie ze sobą nie tylko sukcesy poznawcze, ale także szereg porażek i problemów. Wśród korzyści zastosowania metody półautomatycznej wymienić można następujące: za podstawę przyjmuje ona opracowanie tekstu oparte o tradycyjne podejście powszechne w analizie treści (kodowanie tekstu przez człowieka) i jakościową analizę danych; daje możliwość wykonania szybkiej analizy na dużych zbiorach wypowiedzi; jest procesem iteracyjnym, co pozwala na zmniejszenie liczby błędów w regułach klasyfikacji oraz zwiększenie trafności klasyfikacji; daje możliwość stosowania istniejących narzędzi klasyfikacyjnych jako elementów inspirujących analizę lub wzbogacających klasyfikację (na przykład listy słów specyficzne dla danych

tematyk – inspiracją jest tu więc analiza tematyczna); jest punktem wyjścia do wielowymiarowych analiz treści (to w związku z możliwością liczbowej reprezentacji wypowiedzi); jest podejściem rozwijającym się w ramach programu badawczego (wtedy, gdy jeden słownik klasyfikacyjny rozwijany jest w różnorodnych projektach i testowany na różnych zbiorach tekstów) – w efekcie pozwala na ewolucję słowników klasyfikujących, które stosowane mogą być dla nieeksplorowanych jeszcze tekstów (zbiorów testowych).

Wybrane w toku analiz podejście wiąże się jednak z pewnymi problemami: wybór kategorii analitycznych trafnie odzwierciedlających analizowaną treść nie może opierać się tylko na analizie frekwencji słów i fraz, ale wymaga również zastosowania odmian słów, zapisów słów w różnych rodzajach oraz czasach – zabieg taki zwiększa czasochłonność analiz oraz wymaga wydajnych komputerów z nowoczesnymi procesorami; opracowanie tekstów zgodnie z zasadami *preprocessingu* danych tekstowych jest procesem czasochłonnym; konieczność wielokrotnego wykonywania analiz (iteracyjność) zwiększa czasochłonność procesu analitycznego.

Analiza tekstów opierająca się o metodę, jaką stanowi słownik klasyfikacyjny, daje dobre wyniki wtedy, gdy: po pierwsze, realizowana jest jako proces iteracyjny (czasochłonny, ale zwiększający trafność klasyfikacji); po drugie, wykorzystuje strategie mieszane (*bottom up*, *top down*); po trzecie, jest metodą nadzorowaną i opiera się na wielokrotnej weryfikacji.

Zaznaczyć należy, że przeprowadzony eksperyment i analiza obejmowały teksty stosunkowo proste językowo, składniowo i stylistycznie. Analizowany język zapisany był w naturalnej formie (spontanicznych, nieustrukturyzowanych) wypowiedzi. Łatwo poddawał się analizie (był wyczyszczony z błędów zapisu, rozpoznawane były wszystkie formy wyrazów – bez względu na odmianę, rodzaj, liczbę). Mimo osiągniętego „sukcesu poznawczego" (klasyfikacja na poziomie 68%), trudno ogłosić tu sukces metody. Nie jest bowiem możliwe ekstrapolowanie osiągniętego rezultatu na zadania, w których mielibyśmy dokonać klasyfikacji wypowiedzi dla języków niszowych, slangowych, subkulturowych, emocji czy choćby ustrukturyzowanych dłuższych lub językowo trudniejszych wypowiedzi. Takimi wypowiedziami są te zawierające żart, parafrazę, metaforę, aforyzm (Weizenbaum 2008). Dla takich wypowiedzi niezbędne byłoby opracowanie dodatkowych rozwiązań (Forslid, Wiken 2015).

Podsumowując powyższe rozważania, odpowiedzi na postanowione w artykule pytania można sformułować w sposób następujący:

a. Czy część najbardziej czasochłonnej pracy wykonywanej przy analizie tekstów można w jakimś stopniu zautomatyzować?

Wykonane eksperymenty na analizowanych wypowiedziach pokazują, że stosunkowo prostą jest eliminacja wypowiedzi niewnoszących wartości do raportów badawczych. Aby cel ten osiągnąć, niezbędnym jest wykonaniediagnostyki tekstów skupionej na: formie zapisu stosowanego języka, występowaniu nieznaczących ciągów znaków odtwarzanych przez wielokrotne przyciskanie klawiszy występujących blisko na klawiaturze. Trudniejszym zadaniem jest natomiast zidentyfikowanie wypowiedzi ambiwalentnych, eliminacja złożonych i długich wypowiedzi, które nie są wartościowe dla dalszych analiz. Istotnym na tym etapie jest przejrzenie korpusu wypowiedzi przez kodera, który może sformułować wstępne warunki eliminacji wypowiedzi z dalszych analiz.

b. Czy w wyniku automatyzacji uzyskamy dane o takiej samej jakości jak w przypadku analiz wykonywanych manualnie? Czy automatyzacja opracowania danych da w efekcie trafne rezultaty?

Automatyczna analiza wypowiedzi tekstowych formułowanych w języku naturalnym prowadzona bez nauczyciela (wsparcia wiedzy kodera) wydaje się być – na tym etapie rozwoju wiedzy – skazana na błąd klasyfikacji. Dlatego też udział koderów w analizach tekstów wydaje się warunkiem *sine qua non* do osiągania wysokiego poziomu trafności klasyfikacji.

c. Czy metoda i narzędzie wykonujące pracę automatycznie mogą korzystać z wiedzy badaczy, koderów?

Im więcej wiedzy ludzkich koderów wykorzystanezostanie na etapie projektowania klasyfikatorów półautomatycznych, tym większa jest trafność klasyfikacji. Koderzy stanowią także bardzo cenne źródło wiedzy na etapie weryfikacji wyników metody półautomatycznej. Pytanie o to, jaki zakres wiedzy możliwy jest do opisu za pomocą algorytmów pozostaje nadal otwartym.

Bibliografia

Acerbi Alberto i in. (2013) *The Expression of Emotions in 20th Century Books*. „PLoS ONE", vol. 8, no. 3, s. 1–6.

Appel Markus, Richter Tobias (2010) *Transportation and Need for Affect in Narrative Persuasion: A Mediated Moderation Model*. „Media Psychology", vol. 13, s. 101–135.

Bolasco Sergio, Ratta-Rinaldi della Francesca (2004) *Experiments on Semantic Categorisation of Texts: Analysis of Positive and Negative Dimension*. „JADT 2004: 7es Journées internationales d'Analyse statistique des Données Textuelles" [dostęp 1 maja 2014 r.]. Dostępny w Internecie‹http://lexicometrica.univ-paris3.fr/jadt/jadt2004/pdf/JADT_018.pdf›.

Bolc Leonard, Jerzy Cytowski (1989–1991) *Metody przeszukiwania heurystycznego, t. 1–2*. Warszawa: PWN.

Cichosz Paweł (2000) *Systemy uczące się*. Warszawa: WNT.

Cole Ron i in. (1996) *Survey of the State of the Art in Human Language Technology*. Cambridge University Press [dostęp 5 maja 2017 r.]. Dostępny w Internecie: ‹http://www.dfki.de/~hansu/HLT-Survey.pdf›.

DeWall Nathan C. i in. (2011) *Tuning in to Psychological Change: Linguistic Markers of Psychological Traits and Emotions over Time in Popular U.S. Song Lyrics*.„Psychology of Aesthetics, Creativity, and the Arts", vol. 5, no. 3, s. 200–207.

Duch Włodzisław (1997) *Fascynujący świat programów komputerowych*. Poznań: Wydawnictwo Nakom.

Duggleby Wendy (2005) *What about Focus Group Interaction Data?*„Qualitative Health Research", vol. 15, no. 6, s. 832–840.

Eder Maciej (2014) *Metody ścisłe w językoznawstwie i pułapki pozornego obiektywizmu. Przykład stylometrii*. „Teksty Drugie", t. 2, s. 90–105.

EderMaciej, Piasecki Maciej (2015) *System do klasyfikacji tekstu i analizy stylometrycznej, referat wygłoszony podczas warsztatów CLARIN* [dostęp 20 marca 2017 r.]. Dostępny w Internecie: ‹http://clarin-pl.eu/pliki/warsztaty/Stylometria%20i%20klasyfikacja%20-%20warsztaty.ppt›.

Forslid Erik, Wiken Niklas (2015) *Automatic Irony and Sarcasm Detection in Social Media, UPPTEC F 15045 Examensarbete 30* [dostęp 30 listopada 2016 r.]. Dostępny w Internecie: ‹http://uu.diva-portal.org/smash/get/diva2:852975/FULLTEXT01.pdf›.

Gonzalez Avelino J., Dankel Douglas D. (1993) *The Engineering of Knowledge-Based Systems: Theory and Practice*. Upper Saddle River, NJ: Prentice-Hall International.

Jordan Joanne i in. (2007) *Using Focus Groups to Research Sensitive Issues: Insights from Group Interviews on Nursingin the Northern Ireland "Troubles"*. „International Journal of Qualitative Methods", vol. 6, no. 4 [dostęp 14 kwietnia 2017 r.]. Dostępny w Internecie: ‹http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.533.61&rep=rep1&type=pdf›.

Kaplowitz Michael D. (2000) *Statistical Analysis of Sensitive Topics inGroup and Individual Interviews*. „Quality & Quantity", vol. 34, s. 419–431

Kidd Pamela S., Parshall Mark B. (2000) *Getting the Focus and the Group: Enhancing Analytical Rigor in Focus Group Research*. „Qualitative Health Research", vol. 10, no. 3, s. 293–308.

Lieberman Erez i in. (2007) *Quantifying the Evolutionary Dynamics of Language*. „Nature", vol. 449, no. 7163, s. 713–716.

Liu Bing (2012) *Sentiment Analysis and Opinion Mining* [dostęp 1 maja 2014 r.]. Dostępny w Internecie: ‹www.cs.uic.edu/~liub/FBS/SentimentAnalysis-and-OpinionMining.pdf›.

Michael Jean Baptiste (2011) *Quantitative Analysis of Culture Using Millions of Digitized Books. Program for Evolutionary Dynamics*. Cambridge: Harvard University.

Niedbalski Jakub (2013) *CAQDAS – oprogramowanie do komputerowego wspomagania analizy danych jakościowych. Historia, ewolucja i przyszłość*. „Przegląd Socjologiczny", t. 62, nr 1, s. 153–166.

Niedbalski Jakub, red. (2014) *Metody i techniki odkrywania wiedzy. Narzędzia CAQDAS w procesie analiz danych jakościowych*. Łódź: Wydawnictwo Uniwersytetu Łódzkiego.

Pagel Mark, Atkinson Quentin D., Meade Andrew (2007) *Frequency of Word-Use Predicts Rates of Lexical Evolution Throughout Indoeuropean History*. „Nature", vol. 449, s. 717–720.

Piskorski Jakub (2001) *Shallow Text Processor for Information Extraction from Free-Text Business Documents*. Poznań: Akademia Ekonomiczna w Poznaniu.

Rosa Krzysztof (2015) *Listy pożegnalne niedoszłych samobójców. Analiza treści*. „Przegląd Socjologiczny", t. 64, nr 4, s. 103–128.

Rybicki Jan, Heydel Magdalena (2013) *The Stylistics and Stylometry of Collaborative Translation: Woolf's "Night and Day" in Polish*. „Digital Humanities 2012: digital diversity: cultures, languages and methods", vol. 28, no. nr 4, s. 708–717.

Schanie Carrie L., Pinto-Foltz Melissa D., Logsdon Cynthia M. (2008) *Analysis of Popular Press Articles Concerning Postpartum Depression: 1998-2006*. „Issues Ment. Health Nurs.", vol. 29, no. 11, s. 1200 –1216.

Schwaber Ken (2013) *Scrum Gide* [dostęp 20 marca 2017 r.]. Dostępny w Internecie: ‹http://www.scrumguides.org/docs/scrumguide/v1/Scrum-Guide-PL.pdf›.

Smith Marc in. (2009) *C&T '09: Proceedings of the Fourth International Conference on Communities and Technologies* [dostęp 5 maja 2017 r.]. Dostępny w Internecie: ‹http://www.connectedaction.net/wp-content/uploads/2009/08/2009-CT-NodeXL-and-Social-Queries-a-social-media-network-analysis-toolkit.pdf›.

Sołdacki Przemysław (2006) *Zastosowanie metod płytkiej analizy tekstu do przetwarzania dokumentów w języku polskim*. Niepublikowana praca doktorska, Politechnika Warszawska, Wydział Elektroniki i Technik Informacyjnych, Instytut Informatyki [dostęp 27 listopada 2017 r.]. Dostępny w Internecie: ‹https://www.google.pl/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&ved=0ahUKEwjli7q5iMvQAhXDFSwKHfNPAPwQFggkMAA&url=https%3A%2F%2Frepo.pw.edu.pl%2Fdocstore%2Fdownload.seam%253Bjsessionid%3DF74241A1317DC5E22F87A22B33BE1F6F%3FfileId%3DWEiTI-b192c072-00cc-41df-9bba-a2b0a211e9bc&usg=AFQjCNH-0laWKScvxkvp6FNAqRhpaHuKr0A&bvm=bv.139782543,d.bGg&cad=rja›.

Taboada Maite i in. (2011) *Lexicon-Based Methods for Sentiment Analysis*. „Journal of Computational Linguistics", vol. 37, no. 2, s. 267–307.

Tomanek Krzysztof (2014a) *Analiza sentymentu: historia i rozwój metody w ramach CAQDAS* [w:] Niedbalski Jakub, red., *Metody i techniki odkrywania wiedzy. Narzędzia CAQDAS w procesie analiz danych jakościowych*. Łódź: Wydawnictwo Uniwersytetu Łódzkiego, s. 155–172.

Tomanek Krzysztof (2014b) *Jak nauczyć metodę samodzielności* [w:] Niedbalski Jakub, red., *Metody i techniki odkrywania wiedzy. Narzędzia CAQDAS w procesie analiz danych jakościowych*. Łódź: Wydawnictwo Uniwersytetu Łódzkiego, s. 174.

Tomanek Krzysztof (2014c) *„Analiza sentymentu" – metoda analizy danych jakościowych. Przykład zastosowania oraz ewaluacja słownika RID i metody klasyfikacji Bayesa w analizie danych jakościowych*. „Przegląd Socjologii Jakościowej", t. 10, nr 2, s. 118–136.

Tomanek Krzysztof (2015) *Społeczność fanów nauki w świecie wirtualnym. Analiza eksploracyjna treści i aktywności społeczności internetowej skupionej wokół fanpage'a „I fucking love science"*. „Edukacja Humanistyczna", nr 1(32), s. 123–138.

Tomanek Krzysztof (2016) *Analiza argumentacji. Praktyczne implikacje zastosowania modelu argumentacji Stephena Toulmina do analiz danych tekstowych* [w:] Wojciech Doliński i in., red., *Rzeczywistość i zapis. Problemy badania tekstów w naukach społecznych i humanistycznych*. Łódź: Wydawnictwo Uniwersytetu Łódzkiego, s. 229–242.

Tomanek Krzysztof, Bryda Grzegorz (2015) *Odkrywanie postaw dydaktyków zawartych w komentarzach studenckich. Analiza treści z zastosowaniem słownika klasyfikacyjnego*. „Przegląd Socjologiczny", t. 64, nr 4, s. 51–81.

Weizenbaum Joseph (2008) *Moglibyśmy mieć raj*. „Forum", nr 28, s. 28–29.

Zaśko-Zielińska Monika (2014) *Ślady idiolektów w listach pożegnalnych zgromadzonych w Polskim Korpusie Listów Pożegnalnych Samobójców* [w:] R. Cieśla, red., *Dokument i jego badania*. Wrocław: Katedra Kryminalistyki Wydziału Prawa, Administracji i Ekonomii Uniwersytetu Wrocławskiego, s. 425–435.

Cytowanie

Tomanek Krzysztof (2017) *Metodyka dla analizy treści w projektach stosujących techniki textmining i rozwiązania CAQDAS piątej generacji*. „Przegląd Socjologii Jakościowej", t. 13, nr 2, s. 128–143 [dostęp dzień, miesiąc, rok]. Dostępny w Internecie: ‹www.przegladsocjologiijakosciowej.org›.

Methodology for Content Analysis in Text Mining Projects and Fifth Generation of CAQDAS

Abstract: Projects which we work with—large volumes of text data that are acquired from various sources and stored in a variety of formats—rise many dilemmas of a methodological nature, often require unstandardized decisions and solutions. In particular, compiling data of various quality, unstructured types, and of quan and qual nature requires dynamic strategies, ideas, and ways of analysis. The article describes an example of this approach. It shows its value in classification of written statements. In such context, the author of the article faces the following objectives: (a) can we use CAQDASso that semiautomatic or automatic work would replace some manual work regarding classification of the expressions; (b) how to construct a classification methodology for data of various quality; (c) when the automatic classification is useful and when there is no chance of success?
The article will be marked with moments in which the analyst reaches for knowledge typical for qualitative data analysis, and when the knowledge of this area is no longer sufficient to classify content (natural language processing, machine learning). An example of a project being the background of this article forced the use of several tools and languages to support work with the data. Work on the transformation, classification, and visualization of results required applications such as: MySQL, R, QDA Miner, WordStat, Qlik Sense. Role and limits of the computer-assisted qualitative data analysis software tools have also been noted.

Keywords: Content Analysis, Mixed Methods Approach, Big Data, Text Mining, CAQDAS