http://dx.doi.org/10.18778/7969-549-2.08

Krzysztof Tomanek
Uniwersytet Jagielloński

Analiza sentymentu: historia i rozwój metody w ramach CAQDAS

Streszczenie. Analiza sentymentu (SA) jest jednym z tych obszarów analiz tekstowych, którego rozwój jest silnie związany z rozwojem CAQDAS. Istnieje kilka różnych metod analiz opinii i analiz emocji zawartych w tekstach. Jedną z nich jest metoda słownikowa (*Dictionary-based Approach DbA*), z której SA korzysta najczęściej. Istnieje wiele różnorodnych słowników wspierających automatyczną analizę tekstów oraz analizę sentymentu. W artykule opisano popularne rozwiązania stosowane przez analityków i badaczy zajmujących się analizami opinii. Wskazane zostały również kierunki rozwoju SA w ramach CAQDAS. Ten drugi wątek artykułu wiąże się z: poszukiwaniem coraz bardziej zaawansowanych reguł wyszukiwania tekstów, budową reguł odkrywania wzorów wypowiedzi oraz pojawianiem się coraz większej liczby słowników klasyfikacyjnych. Analitycy pracujący nad DbA, wykorzystują coraz częściej wiedzę z zakresu językoznawstwa, wielowymiarowych metod analizy danych sprawiając, że słowniki analityczne wyszukują treści w sposób coraz bardziej efektywny. Jednym z obszarów, w którym słowniki rozwijają się dynamicznie, jest identyfikacja emocji czy ocena siły ładunku emocjonalnego zawartego w tekście. Artykuł skupia się na porównaniu dostępnych dla badaczy rozwiązań słownikowych w ramach SA. Opisana została: specyfika słowników, możliwości implementacyjne, zakres tematyczny słowników oraz przykładowe zastosowania.

Słowa kluczowe: analiza sentymentu, analiza opinii, analiza danych jakościowych, analiza treści, Text Mining, kodowanie tekstów, słownik analizy sentymentu, przetwarzanie języka naturalnego, KADJ, CAQDAS.

Wprowadzenie

W ostatnich latach pojawiło się wiele publikacji zarówno z zakresu metod, jak i zastosowań analizy sentymentu (Lieberman i in. 2007; Taboada i in. 2011). Jedną z najczęściej cytowanych i chyba najlepiej znaną jest publikacja z obszaru *culturomics* (Acerbi i in. 2013). Ta popularna analiza pokazuje, jak zmieniała się literatura XX w. pod względem ładunku emocjonalnego zawartego w używanych w niej słowach. Liczba cytowań artykułu Alberto Acerbi w różnych obszarach nauki sprawia, że można z niewielką dozą ryzyka stwierdzić, że analizy tekstu wspierane komputerowo coraz częściej służą naukowcom z różnorodnych dziedzin. Powszechnie

analizy tekstu stosują naukowcy zajmujący się analizami kulturowymi (Jean-Baptiste i in. 2011), lingwistyką (Lieberman i in. 2007: 713–716), historią (Pagel i in. 2007: 717–720), antropologią (DeWall i in. 2011: 200–207). Coraz częściej też po metody KADJ sięgają socjologowie, o czym świadczy chociażby ten tom.

O definicji

Sformułowanie „analiza sentymentu" (SA) bywa używane zamiennie (Liu 2012) z bardziej ogólnym – „analiza opinii"[1] (AO). W obu przypadkach pierwszy człon frazy dotyczy automatycznych i półautomatycznych metod analizy tekstów. Ich celem jest identyfikowanie i klasyfikowanie wypowiedzi ze względu na pojawiające się w nich słowa kluczowe. W przypadku SA są to słowa nacechowane emocjonalnie, w AO zakres poszukiwań obejmuje nie tylko opinie, lecz także obiekt, którego opinia dotyczy oraz profil autora, którego opinię zapisano. Analiza opinii ma zatem szerszy zakres przedmiotowy, a analiza sentymentu jest jednym z elementów tego obszaru. Definicja operacyjna Binga Liu pokazuje precyzyjnie tę zależność (Liu 2012: 19):

[…] opinia składa się z pięciu elementów: e_i, a_{ij}, s_{ijkl}, h_k, t_l, gdzie: e_i jest nazwą obiektu, a_{ij} to aspekt obiektu e_i, s_{ijkl} jest opinią/emocją dotyczącą aspektu a_{ij} danego obiektu e_i, h_k jest autorem opinii, a t_l to czas, w którym opinia została wypowiedziana przez h_k. Opinia s_{ijkl} ma charakter pozytywny, negatywny, neutralny i cechuje się różnym poziomem siły/intensywności.

Klarowną reprezentację obszaru, którego dotyczą SA i AO oraz podejścia, jakie stosuje się w komputerowych analizach danych jakościowych (KADJ), pokazuje piramida wiedzy (Awad, Ghaziri 2004).

Analizy opinii i emocji w tekstach ulokowane są na drugim szczeblu od podstawy piramidy. To miejsce w hierarchii sugeruje, że istnieje możliwość wykonania SA za pomocą metod w pełni zalgorytmizowanych lub półautomatycznych. Podstawą do stosowania algorytmów wyszukujących informacje w tekście jest wiedza o: typie oraz sposobie zapisu i formach, w jakich informacje te występują w tekście. Trzy najpopularniejsze podejścia do tego typu zadania analitycznego to:

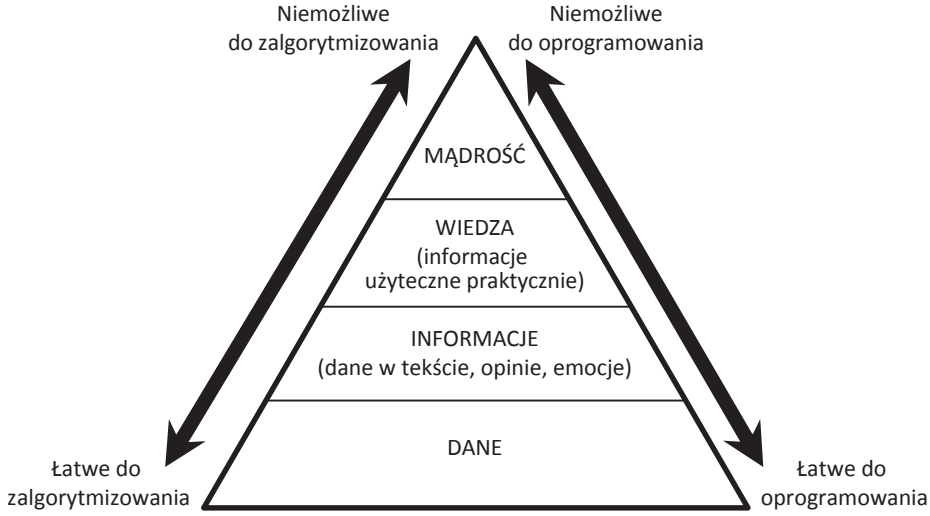
1) analiza słownikowa (*Dictionary-based Approach* (Bolasco, Ratta-Rinaldi 2004)),

2) uczenie maszynowe z nauczycielem (*Supervised machine-learning* (Généreux, Evans 2006)),

[1] Pierwsza część tego artykułu rozwija kilka wątków związanych z analizą sentymentu, o której zastosowaniu pisałem szerzej w tekście pt. *Analiza sentymentu – metoda analizy danych jakościowych. Przykład zastosowania oraz ewaluacja słownika RID i metody klasyfikacji Bayesa w analizie danych jakościowych*, który został przyjęty do druku w „Przeglądzie Socjologii Jakościowej", t. X, nr 2. Oba artykuły (cytowany i aktualny) mają odmienne cele i mogą być czytane niezależnie.

3) uczenie maszynowe bez nadzoru (*Unsupervised machine-learning* (Gamon, Aue 2005)).

W tym artykule przybliżę pierwsze stanowisko. Zarysuję historię rozwoju słowników stosowanych w ramach SA. Scharakteryzuję słowniki, których implementacja w środowiskach CAQDAS[2] jest powszechnie dostępna dla badaczy.



Ilustr. 1. Piramida wiedzy i możliwości analityczne w ramach KADJ

Źródło: opracowanie własne na podstawie: Ohana 2009

Historia i rozwój analizy sentymentu

Analiza sentymentu z jednej strony ma na celu zidentyfikowanie stanów emocjonalnych autora wypowiedzi, z drugiej – służy do określenia emocjonalnego efektu, jaki dana wypowiedź może wywołać u jego odbiorcy. W przywołanym tu rozumieniu termin „analiza sentymentu" wprowadzili w roku 2001 Das i Chen (Das, Chen 2001) oraz Tong (Tong 2001).

[2] Powyższe zastrzeżenie jest istotne z punktu widzenia dostępnych współcześnie rozwiązań analitycznych. Poza CAQDAS mamy do dyspozycji wiele otwartych środowisk programistycznych, które dostarczają o wiele więcej możliwości analitycznych, narzędzi klasyfikacyjnych niż mamy z tym do czynienia w ramach CAQDAS (por. np. darmowe środowisko IDLE Python's Integrated Development Environment do pracy z danymi tekstowymi). Gdyby artykuł ten miał być poświęcony językom, za pomocą których możliwe są analizy SA, AO, treści tu prezentowane miałyby zdecydowanie innych charakter.

Już w pierwszych latach swojego rozwoju analiza sentymentu korzystała z szeregu metod zbudowanych w obszarze przetwarzania języka naturalnego (*natural language processing* – NLP) (Nasukawa, Yi 2003; Pang, Lee 2008). Naturalnym środowiskiem, w którym badacze mogli w prosty i efektywny sposób stosować osiągnięcia NLP, były aplikacje wspierające analizę danych zarówno ilościowych (QUAN), jak i jakościowych (QUAL). Stąd też prekursorzy rozwoju analizy sentymentu od momentu pojawienia się SA skupiali się na budowie i ewaluacji algorytmów dokonujących analiz w ramach środowisk IT (Yi, Nasukawa, Bunescu, Niblack 2003). Wraz z rozwojem automatycznych i półautomatycznych metod klasyfikacyjnych nastąpił rozwój podejścia opartego na analizie wykorzystującej słowniki. Takie narzędzia analityczne służyły początkowo dwóm zadaniom: (1) miały identyfikować w tekście słowa i frazy kluczowe, a następnie (2) rozróżniać je i klasyfikować. Z biegiem czasu słownikom klasyfikacyjnym stawiano coraz bardziej złożone zadania analityczne. Wraz ze słownikami rozwijały się metody i algorytmy wykorzystujące wiedzę z zakresu językoznawstwa i filozofii języka. Kierunek tych zmian można określić na osi: od najprostszych zadań klasyfikacyjnych do złożonych zadań polegających na identyfikacji natężenia emocjonalnego przekazu. Oto cztery klasy problemów, które najczęściej rozwiązywane są w ramach analizy sentymentu:

1. Klasyfikacja biegunowa (*polarity classification*): różnicuje wypowiedzi, teksty, fragmenty tekstów w zależności od typu emocji, które zawarte są w słowach. Wynikiem klasyfikacji biegunowej są dwie klasy słów nacechowanych emocjami (pozytywne, negatywne) oraz słowa neutralne z punktu widzenia ładunku emocjonalnego w nich zawartego.

Przykładowe zastosowania słowników wykorzystywanych w ramach klasyfikacji biegunowej: analiza ocen produktów, identyfikacja silnych i słabych stron usług, prosta analiza wypowiedzi w mediach społecznościowych.

2. Klasyfikacja tematyczna: pozwala na rozróżnienie więcej niż dwóch typów emocji zawartych w wypowiedziach (np. złość, duma, radość, miłość).

Przykładowe zastosowania: rekonstrukcja schematów pisania recenzji filmowych, analiza profilu autora wypowiedzi, identyfikacja oceny atrybutów produktów.

3. Klasyfikacja wypowiedzi ze względu na siłę emocjonalną przekazu. Klasyfikacja pozwala określić poziom natężenia emocjonalnego wypowiedzi.

Przykładowe zastosowania: analiza zgodności ocen w wielu opiniach, identyfikacja siły perswazyjnej przekazu, identyfikacja natężenia emocji w wypowiedziach polityków, zaawansowana analiza mediów społecznościowych: identyfikacja grup opiniotwórczych, identyfikacja podobieństwa preferencji dotyczących produktów czy zdarzeń.

4. Klasyfikacja tematyczna tekstów oraz ocena sentymentu. SA staje się elementem szerszych analiz tematycznych.

Przykładowe zastosowania: automatyczne streszczenia tekstów, analiza sentymentu w takich obszarach tematycznych, jak: marketing, ekonomia, automatyczne kodowanie tekstów[3], zaawansowana analiza mediów społecznościowych: identyfikacja lidera opinii, wyłonienie grup opiniotwórczych, identyfikacja krytyków oraz źródeł opinii kreatywnych.

Wskazane cztery typy problemów wyznaczają jednocześnie etapy rozwoju analizy sentymentu.

Ad 1. Klasyfikacja biegunowa. W pierwszych latach rozwoju SA analitycy skupiali się głównie na poprawie trafnej identyfikacji i klasyfikacji słów nacechowanych emocjonalnie. W pierwszych pracach poświęconych SA uwaga autorów skupiona była na rzetelnym rozróżnieniu dwóch klas słów, które jednoznacznie określają pozytywne i negatywne emocje (Hatzivassiloglou, McKeown 1997). To podejście nie pozwalało na uchwycenie nowych, oryginalnych sposobów wyrażania emocji, nie uwzględniano także kontekstu, w jakim występują słowa. Stąd też kolejny etap w historii SA dotyczy: (1) rozbudowy słowników o nowe wyrazy wyrażające emocje, (2) budowy klas/typów emocji zawartych w tekstach. Ten rozwój słowników odbywał się między innymi dzięki wynikom analizy współwystępowania słów, ale też dzięki uwzględnianiu relacji, jakie zachodzą między słowami. W szczególności relacji opartych na synonimii, antonimii oraz reguł syntaktycznych.

Ad 2. Klasyfikacja tematyczna. Dwie wskazane w poprzednim punkcie strategie analityczne doprowadziły nie tylko do rozszerzenia zakresu słów w słownikach (Popescu, Etzioni 2005), ale pozwoliły też na bardziej precyzyjne określenie rodzaju emocji zawartych w wypowiedziach oraz trafniejsze określenie przedmiotu wypowiedzi, wobec którego emocje są kierowane. Efektem tego etapu rozwoju analiz emocji jest na przykład słownik RID, klasyfikujący wypowiedzi w ramach 7 kategorii emocji.

Ad 3. Klasyfikacja wypowiedzi ze względu na siłę emocjonalną przekazu. Możliwość określenia typu emocji zawartych w wypowiedziach otworzyło wyobraźnię analityków na kolejne wyzwanie. Skoro możliwe jest określenie ładunku emocjonalnego, to czy możliwe jest również określenie siły tego ładunku czy natężenia emocji w tekście? W odpowiedzi na to pytanie przyjęto rozwiązanie pochodzące z badań opinii sędziów niezależnych, którzy mieli za zadanie określić w języku naturalnym słowa wyrażające emocje z różnym natężeniem (Pang, Lee 2005; Turney 2002). Przykładem słownika, który klasyfikuje wypowiedzi nacechowane emocjonalnie, wskazując jednocześnie natężenie emocji zawartych w tekście, jest AFINN. Słownik ten przypisuje słowom wartości na skali od –5 do –1 (negatywne emocje) do 1 do 5 (pozytywne emocje).

[3] Por. wykorzystanie słowników opartych na relacjach synonimii do budowy reguł kodowania półautomatycznego (Tomanek 2014).

Ad 4. Klasyfikacja tematyczna tekstów oraz ocena sentymentu. Współcześnie słowniki określają zarówno typ, jak i poziom natężenia emocji, wskazując dodatkowo obszar tematyczny, którego emocje dotyczą. Coraz trafniej identyfikowane są więc obiekty, których emocje dotyczą. Ten ostatni cel jest osiągany dzięki:

a) rozpoznaniu ontologii analizowanego tekstu (Grassi i in. 2011: 480–489; Cambria i in. 2013: 41–53),

b) analizie tematyki wypowiedzi,

c) identyfikacji kontekstu występowania słów niosących emocje,

d) określeniu relacji semantycznych oraz syntaktycznych zachodzących pomiędzy słowami kluczowymi a słowami bliskimi im znaczeniowo.

Istnieje wiele różnorodnych słowników wykorzystywanych w analizie sentymentu. Na potrzeby tego opracowania wyróżnię ich dwa podstawowe typy. Pierwszy obejmuje słowniki tematyczne, które między innymi pozwalają identyfikować emocje. Analiza sentymentu nie jest jedynym ani też głównym zadaniem, dla którego słowniki te zostały stworzone. Drugi typ słowników to narzędzia służące tylko i wyłącznie zadaniom SA. W tab. 1 podaję nazwy słowników, które zostaną omówione w dalszej części tekstu. Wszystkie wymienione słowniki istnieją w wersjach elektronicznych i mogą – zazwyczaj po niewielkich modyfikacjach – być stosowane przez oprogramowania wspierające KADJ.

Tabela 1. Dwie kategorie słowników wspierających analizę sentymentu

Słowniki tematyczne zawierające kategorie identyfikujące emocje	Słowniki przeznaczone do analizy sentymentu
1. Harvard IV Dictionary 2. LIWC 3. LASWELL Values Dictionary 4. RID 5. General Inquirer	1. WordStat Sentiment 2. SentiWordNet 3. AFINN 4. Loughran & McDonald Financial Sentiment Dictionary 5. Lexicoder Sentiment Dictionary (LSD)

Źródło: opracowanie własne.

Słowniki tematyczne zawierające kategorie identyfikujące emocje

1. Harvard IV Dictionary (H4D)

Słownik H4D zbudowany jest ze 105 kategorii (83 to kategorie unikalne, pozostałe 22 zawierają kombinacje słów zaczerpniętych z kategorii unikalnych), wśród których istnieją dwa generalne zbiory zawierające słowa identyfikujące emocje (emocje pozytywne, emocje negatywne). Idea leżąca u podstaw

klasyfikacji słów w ramach H4D to skala Osgooda (Osgood, Snider 1969). Słownik ten bierze pod uwagę kontekstową zmienność znaczeń słów. Oznacza to, że jedno słowo może mieć więcej niż jedno znaczenie. Dla przykładu słowo „uroczystość" kwalifikowane jest do takich kategorii, jak: pozytywne emocje, przynależność-afiliacja, aktywność, rytuał. H4D uwzględnia również dodatkowe kategorie. Pozwalają one na identyfikację słów kluczowych związanych z takimi emocjami, jak: podniecenie, popęd, uczucie, przyjemność, zmartwienie, gniew.

Tabela 2. Liczba słów w wybranych kategoriach w słowniku H4D

Kategoria	Liczba słów w kategorii
Emocje pozytywne	1045
Emocje negatywne	1160

Źródło: opracowanie własne na podstawie www.wjh.harvard.edu /~inquirer/homecat.htm.

Słownik dostępny jest w dwóch wersjach językowych: angielska, francuska. Format słownika pozwala na zastosowanie go bez modyfikacji w takich rozwiązaniach CAQDAS, jak: General Inquirer, Protan, TextQuest, WordStat.

2. LIWC (LInguistic Word Count)

Słownik Jamesa W. Pennebakera pozwala na analizę częstości, z jaką słowa używane są w tekstach. Narzędzie to klasyfikuje wypowiedzi w ramach 70 różnych kategorii, takich jak emocje i procesy (percepcyjne, społeczne, biologiczne). LIWC porównuje częstotliwości słów w analizowanym materiale do częstotliwości użycia tych słów w danym języku. Inne zadanie, które realizuje LIWC, to hierarchiczna organizacja kategorii zidentyfikowanych w tekście. Przykładowo wszystkie zaimki są zawarte w nadrzędnej kategorii słów funkcyjnych. Z kolei osobiste zainteresowania ludzi to kategoria obejmująca użycia słów kontekstowo zależnych, w takich obszarach życia, jak: praca, czas wolny, dom, religia.

Tabela 3. Liczba słów w wybranych kategoriach w słowniku LIWC

Kategoria	Liczba słów w kategorii
Emocje pozytywne	406
Emocje negatywne	499

Źródło: opracowanie własne na podstawie: www.liwc.net/.

LIWC jest dostępny w następujących językach: angielski, niemiecki, hiszpański, włoski, koreański, polski. Słownik bez większych zmian współpracuje z takimi narzędziami, jak: LIWC, TextQuest, WordStat.

3. Laswell Value Dictionary (LVD)

Metoda klasyfikacyjna Harolda Lasswella została opracowana w formie słownikowej w książce *Dynamic of Culture* (Namenwirth, Weber 1987). Słownik klasyfikuje wypowiedzi w ramach czterech podstawowych kategorii związanych z wartościami: władza, uczciwość, szacunek, przynależność oraz czterech kategorii związanych z ideą dobrobytu/opieki społecznej: bogactwo, pomyślność, oświecenie, umiejętności. W ramach każdej kategorii istnieją subkategorie dodatkowo grupujące słowa dla wyróżnionych w słowniku typów wartości: zyski, straty, uczestnicy, dokonania, miejsca. Autorzy słownika skategoryzowali słowa w sposób jednoznaczny. Oznacza to, że jedno słowo bez względu na znaczenia, jakie może posiadać w różnych kontekstach, przypisane jest do jednej kategorii słownikowej.

LVD zawiera pięć kategorii użytecznych w analizach SA: (1) kategoria o nazwie pozytywne uczucia zawiera listę słów identyfikujących takie stany emocjonalne, jak: akceptacja, uznanie, wsparcie emocjonalne itp.; (2) negatywne emocje obejmują słowa oznaczające: płacz, przerażenie, wstręt, złośliwość; (3) kategoria NIE składa się ze zbioru reguł syntaktycznych, które pozwalają na klasyfikację wypowiedzi poprzez zastosowanie logiki: negacja + słowo identyfikujące emocje; (4) czwarta kategoria identyfikuje wypowiedzi związane z poczuciem pewności (w tej kategorii wyróżniono dwa wymiary: pewność, stałość); (5) ostatnia kategoria oznaczona słowem JEŻELI odnosi się do takich stanów emocjonalnych, jak: niepewność, wątpliwość, niejasność.

Tabela 4. Liczba słów w wybranych kategoriach w słowniku LVD

Kategoria	Liczba słów w kategorii
Emocje pozytywne	126
Emocje negatywne	193
NIE	25
PEWNOŚĆ	175
JEŻELI	132

Źródło: opracowanie własne na podstawie: www.wjh.harvard.edu/~inquirer/homecat.htm.

Słownik jest dostępny w języku angielskim w formacie współpracującym bez większych zmian z następującymi narzędziami: General Inquirer, Protan, WordStat, TextQuest.

4. RID (Regressive Imagery Dictionary)

RID zawiera ponad 3200 słów przypisanych do trzech głównych kategorii. Są to: (1) pierwotne procesy poznawcze (29 subkategorii); (2) wtórne procesy poznawcze (7 subkategorii) i (3) emocje (7 subkategorii). Słownik Martindale'a służy do identyfikacji schematów myślenia, których reprezentacją są rodzaje słów używanych w wypowiedzi. Dwa modele myślenia, które identyfikuje w tekstach RID, to: (1) myślenie koncepcyjne i (2) myślenie pierwotne. Ten pierwszy model charakteryzuje myślenie: abstrakcyjne, logiczne, zorientowane na obiekty świata nas otaczającego i zorientowane na rozwiązywanie problemów. Myślenie pierwotne jest skojarzeniowe, konkretne i niekoniecznie związane z otaczającą nas rzeczywistością. Analiza z zastosowaniem RID stosowana jest na przykład do charakterystyki profilu autora tekstu ze względu na wyróżnione w słowniku kategorie.

Tabela 5. Liczba słów w wybranych kategoriach w słowniku RID

Kategoria	Liczba słów w kategorii
Emocje pozytywne	70
Niepokój	49
Smutek	75
Uczuciowość	65
Agresja	222
Zachowania ekspresyjne	52
Chwała	76

Źródło: opracowanie własne na podstawie: http://provalisresearch.com/.

Słownik współpracuje z takimi narzędziami, jak: Protan, WordStat, TextQuest i jest dostępny w następujących wersjach językowych: angielska, francuska, portugalska, szwedzka, niemiecka, łacińska, węgierska, rosyjska (w trakcie przygotowania), polska (w przygotowaniu)[4].

[4] Nad tłumaczeniem słownika pracują badacze w Instytucie Socjologii Uniwersytetu Jagiellońskiego (dr Annamaria Orla-Bukowska, dr Grzegorz Bryda, dr Krzysztof Tomanek).

5. General Inquirer (GI)

Słownik GI łączy w sobie kategorie i słowa zaczerpnięte z czterech źródeł, a są nimi: (1) Harvard IV-4 Dictionary; (2) Lasswell Value Dictionary; (3) kategorie autorstwa Rogera Hurwitza (autor słownika GI); (4) kategorie jednoznacznie klasyfikujące słowa – tzw. markery. GI rozpoznaje i kwalifikuje słowa, biorąc pod uwagę zmienność ich znaczeń w różnych kontekstach. Dla przykładu słowo „race" klasyfikowane jest inaczej wtedy, gdy oznacza wyścig, inaczej kiedy odnosi się do grupy ludzi o wspólnym pochodzeniu, a jeszcze inaczej, kiedy jest użyte w sformułowaniu idiomatycznym „rat race".

Tabela 6. Liczba słów w wybranych kategoriach w słowniku GI

Kategoria	Liczba słów w kategorii
Emocje pozytywne	1915
Emocje negatywne	2291

Źródło: opracowanie własne na podstawie: www.wjh.harvard.edu/~inquirer/.

GI jest dostępny w języku angielskim. Narzędzia GI (słownik, oprogramowanie) dostępne są on-line oraz, po wcześniejszym kontakcie z autorami, jako oprogramowanie możliwe do zainstalowania na komputerach osobistych i działające off-line[5].

Słowniki przeznaczone do prowadzenia analizy sentymentu

1. WordStat Sentiment Dictionary (WSSD)

WSSD, podobnie jak GI, jest słownikiem łączącym kilka istniejących rozwiązań w ramach SA. Autorzy WSSD wykorzystali słowa identyfikujące emocje w następujących słownikach: (1) Harvard IV Dictionary; (2) RID; (3) LIWC. WordStat Sentiment został rozwinięty poprzez dodanie: synonimów, wyrazów pokrewnych oraz odmian słów kwalifikowanych w ramach (1)–(3). Finalna wersja zawiera trzy kategorie słów: emocjonalnie pozytywne, emocjonalnie negatywne, negacje. Ostatnia z kategorii wspiera kontekstową analizę słów kluczowych poprzez dwa typy reguł językowych zastosowanych w słowniku.

1. Słowa negatywne są identyfikowane jako:
   a) słowa negatywne niepoprzedzone negacją (nie, nigdy) w odległości trzech słów w tym samym zdaniu,

[5] www.wjh.harvard.edu/~inquirer/server_blognote.html.

b) słowa pozytywne poprzedzone negacją w odległości trzech słów w tym samym zdaniu;

2. Słowa pozytywne są identyfikowane jako:
   a) słowa negatywne poprzedzone negacją w odległości trzech słów w tym samym zdaniu,
   b) słowa pozytywne niepoprzedzone negacją w odległości trzech słów w tym samym zdaniu.

Tabela 7. Liczba słów w wybranych kategoriach w słowniku WSSD

Kategoria	Liczba słów w kategorii
Emocje pozytywne	4733
Emocje negatywne	2428

Źródło: opracowanie własne na podstawie: http://provalisresearch.com/.

WSSD jest ogólnodostępny, a jego format pozwala na edycję zawartości słownika, co umożliwia stosowanie go w dowolnym środowisku CAQDAS. Stosowany w ramach programu WordStat umożliwia samodzielne budowanie reguł syntaktycznych. WSSD jest dostępny w języku angielskim.

2. SentiWordNet (SWN)

SWN wyróżnia się wśród słowników służących do analizy sentymentu dwiema cechami. Po pierwsze klasyfikacja oprócz dwóch kategorii opisujących emocje uwzględnia klasę słów niebędących wyrazem opinii (słowa „obiektywne" – Esuli, Sebastiani 2006). Po drugie budowa SWN opierała się na wykorzystaniu systemu uczącego się z nauczycielem. Efekt tego podejścia to: niejednoznaczne przyporządkowanie słów, możliwość analizy kontekstu występowania słowa, ale przede wszystkim ocena i przypisanie poziomu prawdopodobieństwa, z jakim dane słowo niesie pozytywne lub negatywne emocje (Esuli, Sebastiani 2006).

Tabela 8. Liczba słów w wybranych kategoriach w słowniku SWN

Kategoria	Liczba słów w kategorii
Emocje pozytywne	11 067
Emocje negatywne	12 080

Źródło: opracowanie własne na podstawie: http://sentiwordnet.isti.cnr.it/.

SWN jest jednym z największych słowników (117 659 wyrazów) dostępnym w ramach *creative commons*[6]. Wykorzystywany jest nie tylko w ramach SA, ale także do automatycznych streszczeń, w automatycznych tłumaczeniach, automatycznej klasyfikacji tekstów. Słownik dostępny jest w językach: angielskim, polskim (w przygotowaniu)[7].

3. AFINN

AFINN to wyjątkowe dla analizy SA narzędzie. Słownik ten nie tylko klasyfikuje wyrazy do dwóch kategorii (emocjonalnie pozytywne, emocjonalnie negatywne), ale każdemu słowu przypisuje natężenie emocji (*sentiment strength*), lub siłę emocji, jaką potencjalnie może to słowo wywołać. W ramach wskazanych dwóch klas mamy zatem pięć zbiorów, w których znajdują się słowa o różnym natężeniu emocjonalnym (od –1 do –5 i od 1 do 5). AFINN stosowany był w analizach postów na twitterze, a także w analizach mikroblogów (Nielsen 2011).

Tabela 9. Liczba słów w wybranych kategoriach w słowniku AFINN

Kategoria	Liczba słów w kategorii
Emocje pozytywne	878
Emocje negatywne	1 598

Źródło: opracowanie własne na podstawie: www2.imm.dtu.dk/pubdb/views/publication_details.php?id=6010.

Format słownika pozwala na edycję jego treści, co umożliwia stosowanie tego rozwiązania w dowolnym środowisku CAQDAS. Słownik dostępny jest w języku angielskim, a w trakcie przygotowań jest wersja polska słownika[8].

4. Loughran & McDonald Financial Sentiment Dictionary (L&M)

Słownik L&M różni się od dotychczas omawianych narzędzi klasyfikacyjnych zarówno sposobem, w jaki został zbudowany, jak i obszarem, którego dotyczy. Loughran i Macdonald zauważyli, że słowniki budowane w ramach

[6] CC to międzynarodowa organizacja, która wspiera darmowe dzielenie się twórczością, współpracę i innowacje dzięki popularyzacji otwartych rozwiązań prawnych: http://creativecommons.pl/.

[7] Nad tłumaczeniem słownika pracują badacze w Instytucie Socjologii Uniwersytetu Jagiellońskiego (dr Annamaria Orla-Bukowska, dr Grzegorz Bryda, dr Krzysztof Tomanek).

[8] Nad tłumaczeniem słownika pracują badacze w Instytucie Socjologii Uniwersytetu Jagiellońskiego (dr Annamaria Orla-Bukowska, dr Grzegorz Bryda, dr Krzysztof Tomanek).

różnych dyscyplin dokonują błędnych klasyfikacji wypowiedzi i tekstów z obszaru ekonomii i finansów. W szczególności słownik harwardzki (H4D) w 75% przypadków jako negatywne klasyfikuje słowa, które w obszarze ekonomii nie są za takie uznawane (Loughran, McDonald 2010). Ze swoich obserwacji autorzy wyciągnęli wnioski, konstruując słownik tematyczny, który z większą trafnością klasyfikuje słowa niosące różne ładunki emocjonalne w obszarze ekonomii.

Tabela 10. Liczba słów w wybranych kategoriach w słowniku L&M

Kategoria	Liczba słów w kategorii
Emocje pozytywne	353
Emocje negatywne	2337
Słowa wyrażające niepewność	285
Słowa wzmacniające siłę przekazu (mocne)	19
Słowa wzmacniające siłę przekazu (słabe)	27
Słowa kontekstowo zależne	731

Źródło: opracowanie własne na podstawie: www3.nd.edu/~mcdonald/Word_Lists.html.

Format słownika pozwala na stosowanie go w dowolnym środowisku CAQDAS. W jednej z wersji L&M jest przystosowany do użycia w programie WordStat. L&M jest dostępny w języku angielskim.

5. Lexicoder Sentiment Dictionary (LSD)

LSD (Sevenans, Soroka 2013) integruje kilka istniejących w obszarze SA rozwiązań. Składa się mianowicie ze słów zaczerpniętych z następujących słowników: (1) harwardzkiego (H4D); (2) Thesaurus Rogeta (Roget 2011); (3) Martindale'a (RID). LSD powstał na bazie analiz: kampanii wyborczych, wiadomości telewizyjnych, artykułów prasowych dotyczących polityki publicznej. Słownik przeszedł fazę ewaluacji, w której wyniki automatycznego kodowania 900 artykułów porównane były z kodowaniem wykonanym przez badaczy. Autorzy przekonują o wyższej jakości kodowania (kodowanie bliższe nawykom ludzkich koderów), jaką uzyskuje LSD w porównaniu do innych automatycznych metod klasyfikacji dostępnych w obszarze SA (Young, Soroka 2012).

Tabela 11. Liczba słów w wybranych kategoriach w słowniku LSD

Kategoria	Liczba słów w kategorii
Emocje pozytywne	1709
Emocje negatywne	2858

Źródło: opracowanie własne na podstawie: www.lexicoder.com/.

LSD jest przystosowany do pracy z oprogramowaniem, w ramach którego został zbudowany (Lexicoder). Jednocześnie jest to słownik otwarty na prace edycyjne, dlatego też może być stosowany w dowolnym środowisku CAQDAS.

Analiza sentymentu i CAQDAS

Wśród dostępnych narzędzi wspierających SA można wyróżnić trzy typy:

1) platformy internetowe świadczące usługi związane z analizą treści (np. https://sentimentalytics.com, http://sentione.pl/, http://simplymeasured.com/),

2) narzędzia, które umożliwiają wykonanie analizy sentymentu dzięki wbudowanym słownikom,

3) tak zwana piąta generacja CAQDAS[9], czyli środowiska pozwalające na posługiwanie się słownikami klasyfikacyjnymi, ale też pozwalające na samodzielne projektowanie metod i algorytmów wykorzystujących słowniki analityczne.

Charakteryzując słowniki wykorzystywane w analizach SA, podawałem przykłady narzędzi drugiego i trzeciego typu. Więcej informacji o darmowych programach do analiz sentymentu znaleźć można na blogu http://blog.mashape.com[10].

Podsumowanie

Analiza sentymentu traktowana jako metoda automatycznej lub półautomatycznej analizy wypowiedzi jest powszechnie stosowana w wielu obszarach nauk społecznych i nauk o kulturze. Słowniki do analiz SA są często wbudowane zarówno w darmowych, jak i komercyjnych CAQDAS. Różnią się od siebie liczbą słów, sposobem, w jaki są przygotowane dla użytkownika oraz poziomem zaawansowania analiz, które mogą wykonywać. W tab. 12 zamieszczam podsumowanie charakterystyk omawianych w tym artykule słowników.

[9] Por. Tomanek (2014).

[10] http://blog.mashape.com/post/48757031167/list-of-20-sentiment-analysis-apis.

Tabela 12. Porównanie słowników wspierających analizę sentymentu

Kryterium / Słownik	Liczba słów identyfikujących emocje	Słowa w formie rdzeni	Stopniowanie przymiotników	Darmowy dostęp
Harvard IV Dictionary	2 205	–	–	Tak
LIWC	905	–/✓*	–	Nie
LASWELL	651	–	–	Tak
RID	539	✓	✓	Tak
General Inquirer	4 206	–/✓	–/✓	Tak
WordStat Sentiment	7 161	✓	✓	Tak
SentiWordNet	23 147	✓	–	Tak
AFINN	2 476	–	✓	Tak
Loughran & McDonald	3 752	–	✓	Tak
Lexicoder	4 567	–	–	Tak

* Symbole –/✓oznaczają, że dane rozwiązanie jest częściowo wdrożone/dostępne.

Źródło: opracowanie własne.

Wybór słownika do analizy sentymentu jest decyzją wymagającą wzięcia pod uwagę kilku kryteriów. Poza wskazanymi w tab. 12 są to:

1) ogólna ocena trafności klasyfikacji (Thelwall i in. 2010),

2) ocena trafności klasyfikacji w ramach danego obszaru tematycznego (Loughran, McDonald 2010),

3) istnienie ewaluacji procesu adaptacji językowej słownika.

Korzystanie ze słowników klasyfikacyjnych jest gwarantem transparentności wyników i metod stosowanych w analizach treści, a możliwość oceny błędu klasyfikacji uznać należy za silną stronę analiz tekstowych (Hopkins, King 2010) stosujących słowniki klasyfikacyjne. Problematyczne w SA jest konstruowanie słowników i reguł słownikowych trafnie identyfikujących poszukiwanych przez analityka wypowiedzi (Tomanek 2014).

Bibliografia

Acerbi Alberto, Lampos Vasileios, Garnett Philip, Bentley R. Alexander (2013), *The Expression of Emotions in 20th Century Books*, „PLoS ONE", vol. 8, no. 3, s. 1–6; www.plosone.org/article/fetchObject.action?uri=info%3Adoi%2F10.1371%2Fjournal.pone.0059030&representation=PDF [dostęp: 1.05.2014].

Awad Elias M., Ghaziri Hassan M. (2004), *Knowledge Management*, Pearson – Prentice Hall, New Jersey.

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; http://lexicometrica.univ-paris3.fr/jadt/jadt2004/pdf/JADT_018.pdf [dostęp: 1.05.2014].

Cambria Erik, Mazzocco Thomas, Hussain Amir (2013), *Application of Multi-Dimensional Scaling and Artificial Neural Networks for Biologically Inspired Opinion Mining*, "Biologically Inspired Cognitive Architectures", vol. 4, s. 41–53.

Das Sanjiv R., Chen Mike J. (2007), *Yahoo! for Amazon: Sentiment Extraction from Small Talk on the Web*, "Management Science", vol. 53, no. 9, s. 1375–1388.

DeWall C. Nathan, Pond Jr. Richard S., Campbell W. Keith, Twenge Jean M. (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.

Esuli Andrea, Sebastiani Fabrizio (2005), *Determining the Semantic Orientation of Terms Through Gloss Analysis*, [w:] *Proceedings of CIKM-05, 14th ACM International Conference on Information and Knowledge Management*, Bremen; http://ontotext.fbk.eu/Publications/CIKM05--short.pdf [dostęp: 1.05.2014].

Esuli Andrea, Sebastiani Fabrizio (2006), *SENTIWORDNET: A Publicly Available Lexical Resource for Opinion Mining*, [w:] *Proceedings of the 5th Conference on Language Resources and Evaluation*, LREC'06, Genoa; http://gandalf.aksis.uib.no/lrec2006/pdf/384_pdf.pdf [dostęp:1.05.2014].

Gamon Michael, Aue Anthony (2005), *Automatic Identification of Sentiment Vocabulary: Exploiting Low Association with Known Sentiment Terms*, [w:] *Proceedings of the ACL-05 Workshop on Feature Engineering for Machine Learning in Natural Language Processing*, Ann Arbor, MI; http://research.microsoft.com/pubs/65462/sentiment_feats_camera.pdf [dostęp: 1.05.2014].

Généreux Michael, Evans Roger (2006), *Towards a Validated Model for Affective Classification of Texts*, [w:] *Proceedings of the Workshop of Sentiment and Subjectivity in Text*, Association for Computational Linguistics, Sydney; http://aclweb.org/anthology//W/W06/W06-0308.pdf [dostęp: 1.05.2014].

Grassi Marco, Cambria Erik, Hussain Amir, Piazza Francesco (2011), *Sentic Web: A New Paradigm for Managing Social Media Affective Information*, „Cognitive Computation", vol. 3, no. 3, s. 480–489; http://sentic.net/sentic-web.pdf [dostęp: 1.05.2014].

Hatzivassiloglou Vasileios, McKeown Kathleen (1997), *Predicting the Semantic Orientation of Adjectives*, [w:] *Proceeding EACL'97 Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics*, Stroudsburg, PA; http://acl.ldc.upenn.edu/P/P97/P97-1023.pdf [dostęp: 1.05.2014].

Hopkins Daniel, King Gary (2010), *A Method of Automated Nonparametric Content Analysis for Social Science*, "American Journal of Political Science", vol. 54, no. 1, s. 229–247; http://dash.harvard.edu/bitstream/handle/1/5125261/method%20.pdf?sequence=1 [dostęp: 1.05.2014].

Lieberman Erez, Michel Jean-Baptiste, Jackson Joe, Tang Tina, Nowak Martin A. (2007), *Quantifying the Evolutionary Dynamics of Language*, „Nature", vol. 449, no. 7163, s. 713–716.

Liu Bing (2012), *Sentiment Analysis and Opinion Mining*, Morgan & Claypool Publishers, California; www.cs.uic.edu/~liub/FBS/SentimentAnalysis-and-OpinionMining.pdf [dostęp: 1.05.2014].

Loughran Tim, McDonald Bill (2010), *When is a Liability not a Liability? Textual Analysis, Dictionaries, and 10-Ks*, "The Journal of Finance", vol. 66, no. 1, s. 35–65.

Michel Jean-Baptiste et al. (2011), *Quantitative Analysis of Culture Using Millions of Digitized Books*, "Science", vol. 331, s. 176–182.

Namenwirth J. Zvi, Weber Robert Philip (1987), *Dynamics of Culture*, Unwin Hyman, Boston.

Nasukawa Tetsuya, Yi Jeonghee (2003), *Sentiment Analysis: Capturing Favorability Using Natural Language Processing*, [w:] *Proceedings of the Conference on Knowledge Capture (K-CAP)*, Banff, Canada.

Nielsen Finn Å. (2011), *A New ANEW: Evaluation of a Word List for Sentiment Analysis in Microblogs*, [w:] Matthew Rowe, Milan Stankovic, Aba-Sah Dadzie, Mariann Hardey (eds), *Proceedings of the ESWC2011 Workshop on 'Making Sense of Microposts': Big Things Come in Small Packages 718 in CEUR Workshop Proceedings*, Heraklion.

Ohana Bruno (2009), *Opinion Mining with the SentWordNet Lexical Resource*, Institute of Technology, School of Computing – Dissertations, Dublin; http://arrow.dit.ie/cgi/viewcontent.cgi?article=1019&context=scschcomdis [dostęp: 1.05.2014].

Olsher Daniel J. (2012), *Full Spectrum Opinion Mining: Integrating Domain, Syntactic and Lexical Knowledge*, [w:] Jilles Vreeken, Charles Ling, Mohammed J. Zaki, Ar no Siebes, Jeffrey Xu Yu, Bart Goethals, Geoff Webb, Xindong Wu (eds), *ICDMW 2012 The 12th IEEE International Conference on Data Mining Workshops*, Brussels.

Osgood Charles E., Snider James G. (eds), (1969), *Semantic Differential Technique: A Sourcebook*, Aldine, Chicago.

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.

Pang Bo, Lee Lillian (2005), *Seeing Stars: Exploiting Class Relationships for Sentiment Categorization with Respect to Rating Scales*, [w:] *Proc. 43rd Ann. Assoc. for Computational Linguistics.* Assoc. for Computational Linguistics, Cambridge, Massachusetts; www.cs.cornell.edu/home/llee/papers/pang-lee-stars.pdf [dostęp: 1.05.2014].

Pang Bo, Lee Lillian (2008), *Opinion Mining and Sentiment Analysis*, "Foundations and Trends in Information Retrieval", vol. 2, s. 1–135.

Popescu Ana-Maria, Etzioni Oren (2005), *Extracting Product Features and Opinions from Reviews*, [w:] *Proc. Human Language Technology Conf./Conf. Empirical Methods in Natural Language Processing*, Assoc. for Computational Linguistics; http://turing.cs.washington.edu/papers/emnlp05_opine.pdf [dostęp: 1.05.2014].

Roget Peter Mark (2011), *Roget's Thesaurus*, EBook #22, MICRA, Inc; www.gutenberg.org/cache/epub/22/pg22.html [dostęp: 1.05.2014].

Sevenans Julie, Soroka Stuart (2013), *Lexicoder Topic Dictionaries*, McGill University, Montreal; www.lexicoder.com/download.html [dostęp: 1.05.2014].

Taboada Maite, Brooke Julian, Tofiloski Milan, Voll Kimberly, Stede Manfred (2011), *Lexicon-Based Methods for Sentiment Analysis*, "Journal of Computational Linguistics", vol. 37, no. 2, s. 267–307; http://oldsite.aclweb.org/anthology-new/J/J11/J11-2001.pdf [dostęp: 1.05.2014].

Thelwall Mike, Buckley Kevan, Paltoglou Georgios, Cai Di, Kappas Arvid (2010), *Sentiment Strength Detection in Short Informal Text*, "Journal of the American Society for Information Science and Technology", vol. 61, no. 12, s. 2544–2558.

Tomanek Krzysztof (2014), *Jak nauczyć metodę samodzielności? O uczących się metodach analizy treści*, [w:] Jakub Niedbalski (red.), *Metody i techniki odkrywania wiedzy. Narzędzia CAQDAS w procesie analizy danych jakościowych*, Wydawnictwo Uniwersytetu Łódzkiego, Łódź.

Tong Richard M. (2001), *An Operational System for Detecting and Tracking Opinions in On-Line Discussion* [w:] *Working Notes of the SIGIR. Workshop on Operational Text Classification*, New Orleans.

Turney Peter (2002), *Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews*, [w:] *Proc. 40th Ann. Assoc. for Computational Linguistics, Assoc. for Computational Linguistics*; http://acl.ldc.upenn.edu/P/P02/P02-1053.pdf [dostęp: 1.05.2014].

Yi Jeonghee, Nasukawa Tetsuya, Bunescu Razvan, Niblacki Wayne (2003), *Sentiment Analyzer: Extracting Sentiments about a Given Topic Using Natural Language Processing Techniques*, [w:] *Proceedings of the Third IEEE International Conference on Data Mining* (*ICDM'03*), IEEE Computer Society Washington; http://oucsace.cs.ohiou.edu/~razvan/papers/icdm2003.pdf [dostęp: 1.05.2014].

Young Lori, Soroka Stuart (2012), *Affective News: The Automated Coding of Sentiment in Political Texts*, "Political Communication", vol. 29, s. 205–231; www.snsoroka.com/files/2012Young-Soroka%28PolComm%29.pdf [dostęp: 1.05.2014].

Sentiment Analysis: History and Development of the Method within CAQDAS

Summary. Sentiment analysis (SA) is one of those areas of text analysis, the development of which is strongly associated with the development of CAQDAS. There are several different methods of opinion analysis and sentiment analyzes. One of them is the dictionary analysis (Dictionary-based Approach). There are many different dictionaries to support automatic text analysis and sentiment analysis. The article goal is to describe the popular solutions used by analysts and researchers involved in the sentiment analysis. The text describes the development of SA within CAQDAS solutions. The development is associated with: the construction of more and more capacious in a word dictionaries, embedding in dictionaries syntactic rules, taking into account the context of words that identify the occurrence of emotions and the assessment of the strength and the ability to assign emotional charge contained in the text. Text compares available solutions understood as dictionary classifiers.

Keywords: qualitative data analysis, sentiment analysis, opinion mining, content analysis, Text Mining, coding techniques, sentiment analysis dictionary, natural language processing, CAQDAS.