Krzysztof Tomanek

Analiza sentymentu - metoda analizy danych jakościowych : przykład zastosowania oraz ewaluacja słownika RID i metody klasyfikacji Bayesa w analizie danych jakościowych

Przegląd Socjologii Jakościowej 10/2, 118-136

2014

Krzysztof Tomanek
Uniwersytet Jagielloński

Analiza sentymentu – metoda analizy danych jakościowych. Przykład zastosowania oraz ewaluacja słownika RID i metody klasyfikacji Bayesa w analizie danych jakościowych

Abstrakt Celem artykułu jest prezentacja podstawowych metod klasyfikacji jakościowych danych tekstowych. Metody te korzystają z osiągnięć wypracowanych w takich obszarach, jak przetwarzanie języka naturalnego i analiza danych nieustrukturalizowanych. Przedstawiam i porównuję dwie techniki analityczne stosowane wobec danych tekstowych. Pierwsza to analiza z zastosowaniem słownika tematycznego. Druga technika oparta jest na idei klasyfikacji Bayesa i opiera się na rozwiązaniu zwanym naiwnym klasyfikatorem Bayesa. Porównuję efektywność dwóch wspomnianych technik analitycznych w ramach analizy sentymentu. Akcentuję rozwiązania mające na celu zbudowanie trafnego, w kontekście klasyfikacji tekstów, słownika. Porównuję skuteczność tak zwanych analiz nadzorowanych do skuteczności analiz zautomatyzowanych. Wyniki, które prezentuję, wzmacniają wniosek, którego treść brzmi: słownik, który w przeszłości uzyskał dobrą ocenę jako narzędzie klasyfikacyjne, gdy stosowany jest wobec nowego materiału empirycznego, powinien przejść fazę ewaluacji. Jest to, w proponowanym przeze mnie podejściu, podstawowy proces adaptacji słownika analitycznego, traktowanego jako narzędzie klasyfikacji tekstów.

Słowa kluczowe analiza danych jakościowych, analiza sentymentu, analiza treści, text mining, kodowanie tekstów, przetwarzanie języka naturalnego, słownik RID, naiwny klasyfikator Bayesa, CAQDAS

Krzysztof Tomanek, doktorant w Instytucie Socjologii Uniwersytetu Jagiellońskiego. Jego zainteresowania badawcze dotyczą zagadnień lojalności, teorii zaufania, zagadnienia Quality of Life w badaniach społecznych. Najważniejsze zainteresowania metodologiczne obejmują zastosowanie technik text mining do analiz danych jakościowych, analizy danych jakościowych wspierane rozwiązaniami NLP, SVR. Prowadzi grant badawczy MNiSW dotyczący Festiwalu Kultury Żydowskiej w Krakowie (wspólnie z dr Anną Marią Orla-Bukowską). Jest autorem projektów ogólnopolskich badań konsumenckich oraz publikacji dotyczących wykorzystania zaawansowanych technik analizy treści w różnorodnych środowiskach CAQDAS.

Adres kontaktowy:
Instytut Socjologii
Uniwersytet Jagielloński
ul. Grodzka 52, 30-962 Kraków
e-mail: k_tomanek@wp.pl

Wprowadzenie – inspiracje teoretyczne

Badacze sięgający w praktyce po metody analizy tekstów stawiają przed nimi różnorodne cele. Od pozyskiwania prostych informacji tekstowych (*Information Extraction* [IE][1]) po odkrywanie modeli koncepcyjnych i wiedzy zawartej w tekstach (*Knowledge Discovery in Databases* [KDD][2]). Od opracowania i kodowania informacji tekstowych (*Text Encoding* [TE][3]) po klasyfikację (*Text Classification* [TC][4]). W tym artykule poddaję analizie dwie metody klasyfikacji tekstów. Opiszę ich właściwości oraz poddam ocenie wyniki uzyskane dzięki ich zastosowaniu. Zanim przejdę do opisu metod, usytuuję je w szerszej perspektywie metodologicznej, podając typologię metod klasyfikacyjnych stosowanych w analizach *text mining*.

Wyróżnić możemy dwa odmienne podejścia do zagadnienia klasyfikacji tekstów. Pierwsze poddaje analizie zawartości tekstów i wypowiedzi. To podejście wykorzystuje słowa i frazy istotne w danym tekście; posługuje się językiem logiki; wykorzystuje znaczenia analizowanych słów i frazy; bierze pod uwagę reguły leksykalne i składniowe rządzące językiem. W tym podejściu niezbędna jest znajomość: gramatyki analizowanego języka, znaczeń używanych w nim słów, specyfiki wypowiedzi związanej ze stosowanym słownictwem. Drugie podejście sięga po metody statystyczne. W tym przypadku mamy do czynienia z wykorzystaniem technik ilościowych traktujących tekst jako obiekt charakteryzowany za pomocą danych ilościowych, które opisywać mogą zarówno zawartość dokumentu (np. liczbę słów lub fraz, poziom złożoności tekstu, ilość cytowań itp.),jak i sam dokument (np. data publikacji, liczba znaków w dokumencie itp.). Tak traktowany tekst–obiekt reprezentowany jest jako wektor w wielowymiarowej przestrzeni wyznaczonej przez zbiór cech opisujących dokumenty. Poniżej podaję syntetyczny opis metod stosowanych w obrębie dwóch zarysowanych tu podejść.

W dalszej części artykułu skupię się na charakterystyce dwóch podejść. Opiszę metodę klasyfikacji słownikowej dla analizy sentymentu. Scharakteryzuję klasyfikację dokonaną za pomocą naiwnego algorytmu Bayesa.

Metoda słownikowa dla analizy sentymentu

Określenie *analiza sentymentu* odnosi się do automatycznych i półautomatycznych metod analizy tekstów. Podstawowe cele tych metod to: identyfikacja i klasyfikacja fragmentów lub całych wypowiedzi

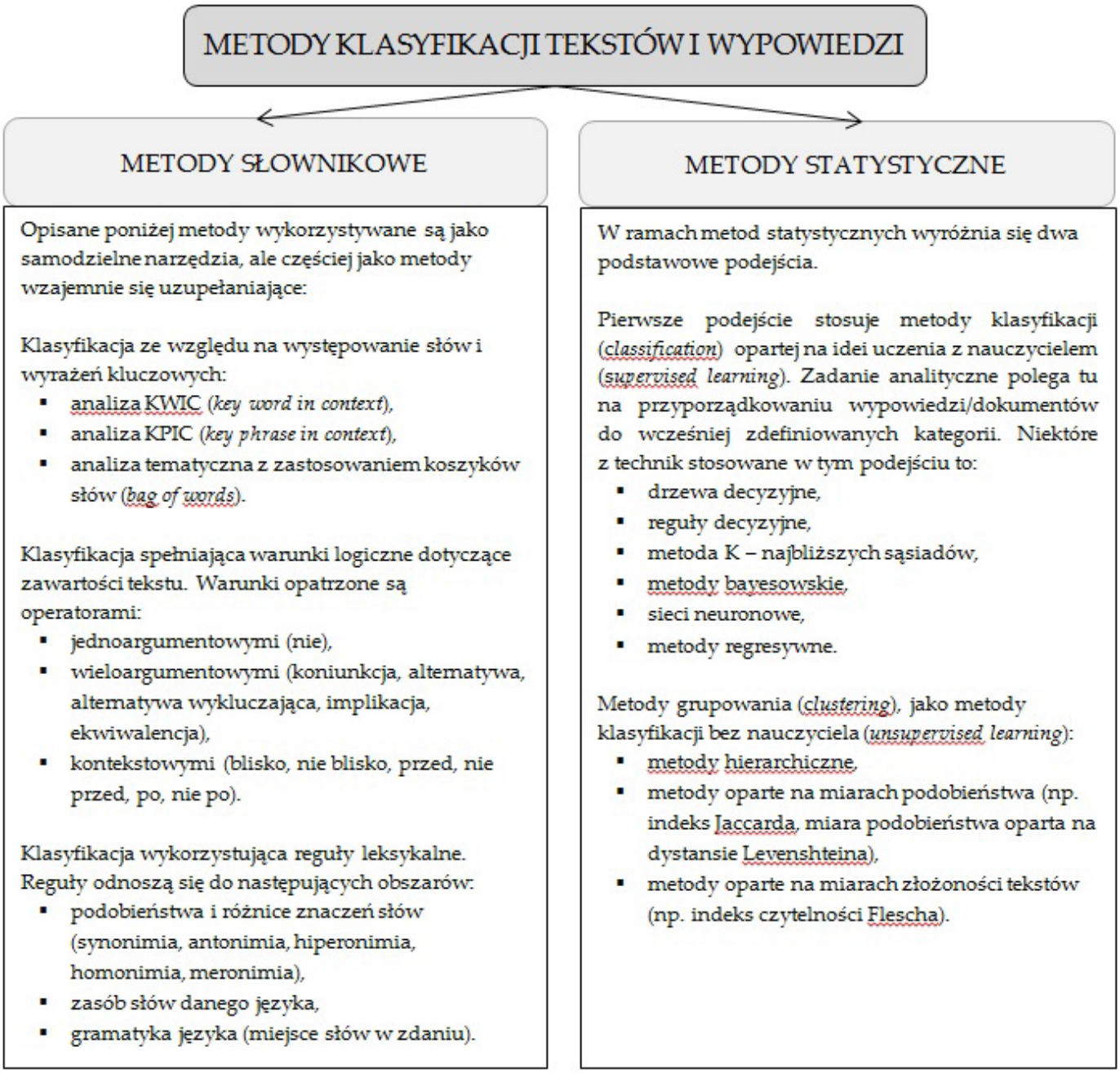
[1] Zbiór technik (leksykalnych lub statystycznych, stosujących język logiki) służących do wydobywania z tekstów informacji, faktów.

[2] Proces wydobywania wiedzy z danych (również tekstowych) oparty za zastosowaniu różnorodnych technik analitycznych, takich jak: selekcja informacji z tekstów, pre-procesowanie danych, transformacje danych, zastosowanie technik *data mining* i *text mining*, interpretacja, ewaluacja.

[3] Techniki TE to zbiór rozwiązań służących do opracowywania zawartości dokumentów. Celem zastosowania TE jest przygotowanie tekstu i struktury dokumentów tak, aby dawały one większe możliwości analityczne niż nieopracowane dokumenty tekstowe. Przykłady technik TE: tokenizacja, lematyzacja, *stemming*, filtrowanie, stop lista, indeksowanie.

[4] Techniki TC to zbiór rozwiązań służących do strukturyzacji dokumentów, wypowiedzi lub części wypowiedzi. Metody te obejmują zarówno automatyczne, półautomatyczne, jak i manualne opracowywanie materiału tekstowego. Popularne i często stosowane techniki to: klasyfikacja oparta na indeksach, naiwny Bayes, *metoda K-najbliższych sąsiadów, drzewa decyzyjne, metody suport vector machines [SVR]. Więcej informacji o metodach klasyfikacji tekstów znaleźć można w: Hotho, Nürnberger, Paaß (2005).*

Diagram 1. Typologia metod klasyfikacji tekstów.



Źródło: opracowanie własne.

ze względu na pojawiające się w nich słowa nacechowane emocjonalnie. Analiza sentymentu opiera się na dwóch założeniach. Po pierwsze, niektóre wypowiedziane słowa wyrażają emocje. Po drugie, istnieją słowa, których wypowiedzenie wywołać może emocje (Pang, Lee 2008). A zatem analiza sentymentu z jednej strony wskazuje na stany emocjonalne autora wypowiedzi, z drugiej –służy również określeniu emocjonalnego efektu, jaki dana wypowiedź może wywołać. W tym rozumieniu termin *analiza sentymentu* wprowadzili Das i Chen (2001) oraz Tong (2001).

Analiza opinii (Pang, Lee 2008), której przykładem jest analiza sentymentu, korzysta z rozwiązań wypracowanych w obszarze przetwarzania języka naturalnego (*natural language processing* [NLP], Nasukawa, Yi [2003]). W ramach CAQDAS pierwsze próby budowy oceny sentymentu wypowiedzi pojawiły się w 2003 roku. Prekursorzy tej metody zbudowali pierwsze algorytmy dokonujące automatycznej klasyfikacji tekstów na te o pozytywnym i te o negatywnym wydźwięku emocjonalnym (Yi i in. 2003). Po tej pierwszej próbie nastąpił szybki rozwój słowników służących do analiz wypowiedzi i dokumentów (Nielsen 2011). Z jednej strony mamy do dyspozycji słowniki tematyczne[5], klasyfikujące wypowiedzi ze względu na tematykę, której dotyczą, z drugiej strony –rozwijają się różnorodne słowniki pozwalające na identyfikację słów i wypowiedzi wyrażających lub wywołujących emocje. Słowniki te pozwalają zarówno na proste klasyfikacje (pozytywne–negatywne), ale też na klasyfikacje bardziej złożone (niepokój–chwała–agresja–smutek–miłość). Przykładem zaawansowanego słownika do analizy sentymentu może być SentiWordNet (Esuli, Sebastiani 2006). Bardziej złożony koncepcyjnie jest słownik RID (Hogenraad, Orianne 1986). Powstały również słowniki mieszane, które łączą dwie przywołane powyżej idee: analizę tematyczną oraz analizę sentymentu. Przykładem takiego narzędzia jest słownik Loughrana i McDonalda (2011), który klasyfikuje wypowiedzi odnoszące się do ekonomii, dziedziny finansów ze względu na zawarty w nich ładunek emocji.

[5] Przykładem takiego słownika może być Visual Thesaurus klasyfikujący wypowiedzi z siedmiu obszarów tematycznych (geografia, ludzie, nauki społeczne, sztuka i literatura, matematyka, nauka, słownictwo). Por.: http://www.visualthesaurus.com/vocabgrabber/.

Liczba publikacji zarówno z zakresu metod, jak i zastosowań analizy sentymentu (Dini, Mazzini 2002; Cardie i in. 2003; Dave, Lawrence, Pennock 2003) jest coraz większa. Warto wspomnieć o jednej pracy– chyba najczęściej cytowanej i najlepiej znanej w tym obszarze – opublikowanej w roku 2013 analizie Acerbiego, Lamposa, Garnetta, Bentleya. Autorzy dokonali opisu literatury XX-wiecznej w duchu analizy sentymentu, a dokładniej –opisali tendencje i zmiany związane z występowaniem słów nacechowanych emocjonalnie w literaturze. Autorzy wykorzystali, co warto dodać, narzędzia z obszaru CAQDAS.

Metody analizy tekstu wspierane komputerowo służą coraz większym gronom naukowców. Powszechnie wykorzystuje się je w analizach zjawisk kultury (Michel i in. 2011), lingwistyce (Lieberman i in. 2007, historii (Pagel, Atkinson, Meade 2007), antropologii (DeWall i in. 2011).

W tym artykule opiszę zastosowanie analizy sentymentu do identyfikacji emocji w tekstach prasowych. Sprawdzę efektywność słownika RID w identyfikacji emocji i porównam ją z efektywnością metody Bayesa.

Klasyfikacja bayesowska

Probabilistyczne metody klasyfikacji tekstów oparte są na założeniu, że słowa w analizowanych tekstach zostały ze sobą zestawione w sposób losowy. W ramach tej metodologii zakłada się, że dana kategoria tekstów k_1 charakteryzuje się określonym słownictwems_1, a inną kategorię tekstów k_2 cechuje słownictwo s_2. Na tej podstawie określamy

prawdopodobieństwo *a priori*, czyli decydujemy o klasyfikacji jakiegoś tekstu, nie dysponując świadectwami przemawiającymi na rzecz ani przeciw poprawności tej klasyfikacji. Zakłada się także, że tekst, który posiada słownictwo określone jako s_1 w większej liczbie niż słownictwo określone przez s_2, powinien być zaklasyfikowany do kategorii k_1, a nie na przykład do k_2. Tym sposobem określamy prawdopodobieństwo *a posteriori*, czyli prawdopodobieństwo klasyfikacji w sytuacji, w której posiadamy pewne świadectwa potwierdzające poprawność klasyfikacji.

Naiwny klasyfikator Bayesa opiera się na założeniu o wzajemnej niezależności predyktorów. W przypadku klasyfikacji tekstów oznacza to, że słowa, które identyfikują określoną kategorię wypowiedzik_1,mogą występować niezależnie w różnych lub w tym samym tekście. Taki naiwny klasyfikator może więc identyfikować i klasyfikować słowa, nie biorąc pod uwagę kontekstu, w jakim one występują. Takie postępowanie często nie odzwierciedla specyfiki wypowiedzi tekstowej. Niemniej jednak podejście bayesowskie często okazuje się trafnym rozwiązaniem ze względu na swoją prostotę. Wzór Bayesa określa bowiem prawdopodobieństwo tego, że szanse realizacji jakiegoś zdarzenia w kolejnej próbie zależą od tego, jak często dochodzi do owego zdarzenia i jak często do niego nie dochodzi. Innymi słowy, jeśli w tekstach naukowych naiwny klasyfikator Bayesa znajdzie takie sformułowania, jak: prawdopodobieństwo, trafność przewidywań, i słowa te zdecydowanie rzadziej występowały będą w jakichkolwiek innych tekstach (np. nienaukowych), to, napotykając nowy tekst, w którym wskazane słowa wystąpią, naiwny Bayes zaklasyfikuje go do tekstów naukowych, a nie na przykład tych z zakresu astrologii.

Najogólniej mówiąc, naiwny klasyfikator Bayesa w analizie tekstów, w wersji, w której ją prezentuję, stosuje zasadę: im więcej w analizowanym tekście zidentyfikuję słów, które zdefiniowane są w danej kategorii tekstów k_1 i jednocześnie zidentyfikuję mniej słów zdefiniowanych w kategorii k_2, tym większe prawdopodobieństwo, że analizowany tekst należy do kategorii k_1, z którą tekst dzieli największą liczbę słów.

W dalszej części artykułu podam przykład zastosowania algorytmu naiwnego Bayesa. Wyniki, jakie uzyskam za pomocą tej metody, porównam z efektywnością klasyfikacji opartej na metodzie słownikowej. Zanim to uczynię, podam jeszcze opis materiału, jaki zastosowałem w analizie oraz opis metod zastosowanych na etapie opracowywania tego materiału.

Źródła danych. Metody opracowywania danych

Dane wykorzystane w prezentowanej tu analizie pochodzą z grantu badawczego przyznanego przez Ministerstwo Edukacji i Szkolnictwa Wyższego[6]. Analizowany tu zbiór danych to:

- 454 artykułów prasowych opublikowanych w prasie polskiej,

[6] Grant realizowany był w latach: 2009–2013 i poświęcony jest badaniu postaw publiczności uczestniczącej w Festiwalu Kultury Żydowskiej w Krakowie. Kierownikiem grantu jest: dr Annamaria Orla-Bukowska, grant realizują: Annamaria Orla-Bukowska, Krzysztof Tomanek.

- artykuły publikowane były w latach: 1992–2002,
- artykuły dotyczą festiwali kultury żydowskiej o numerach: 3–6, 8–9, 11–12,
- analizą objęty jest materiał pochodzący z 61 tytułów prasowych (z czego 370 [81,5%] artykułów opublikowanych zostało w dziennikach).

Dane przekazane przez firmę monitorującą media pierwotnie składowane były w formie elektronicznej w formatach: PDF, JPEG. Pliki sprowadzone zostały do postaci tekstowej, edytowalnej. Proces ten przebiegał w następujący sposób:

- dokonana została konwersja do formatu DOCX (wykorzystano oprogramowanie Abby Fine Reader w wersji 10),
- zweryfikowano poprawność konwersji (sprawdzone zostały m.in. sposób zapisu słów, interpunkcja, kompletność informacji),
- zapis każdego tekstu jako osobnego pliku opierał się na pracy niezautomatyzowanej i był nadzorowany,
- pliki opatrzone zostały nazwami zgodnie ze zbudowaną metodą kodowania (nazwa pliku wykorzystana została jako źródło podstawowych informacji o artykule prasowym i zawierała następujące informacje: data wydania [dzień, miesiąc, rok], tytuł czasopisma, numer festiwalu kultury żydowskiej [FKŻ], którego dotyczył artykuł),
- wszystkie artykuły zostały zarchiwizowane (pliki skategoryzowane są zgodnie z numeracją FKŻ, każdy plik posiada kopię zapasową, składowaną na innym komputerze).

Dla dalszej analizy istotne są zastosowane wobec materiału prasowego techniki kodowania. Wśród zastosowanych najważniejsze to:

- kodowanie tematyczne (identyfikacja i klasyfikacja treści do zdefiniowanych uprzednio obszarów tematycznych, np. wypowiedzi dotyczące: sztuki, religii, FKŻ, relacji polsko-żydowskich itp.),
- kodowanie zogniskowane(technika polegająca na pogłębieniu analizy wykonanej za pomocą kodowania tematycznego, tak aby w kolejnej iteracji kod obejmował fragment danych tekstowych precyzyjnie odpowiadający pytaniu badawczemu, np. w ramach danych tekstowych oznaczonych kodem „sztuka" identyfikowane są fragmenty dotyczące filmu, fotografii, malarstwa, muzyki, wystaw zdjęć itp.),
- kodowanie zerojedynkowe (kodowanie nadające dokumentowi jedną z dwóch wartości, np. wypowiedź wyrażająca pozytywne emocje lub negatywne emocje – technika zwana również *dummy coding*),
- opis wielozmiennowy – specyficzne rozumienie kodowania sprowadzające się do nadania tekstom prasowym kodów identyfikujących ich przynależność do danej klasy obiektów (np. typ artykułu: wywiad, program, relacja,

zapowiedź, recenzja, relacja po zakończeniu FKŻ, inne), taki opis poddawany może być procesowi binaryzacji (por. kolejny punkt),

- kodowanie entropijne lub kontekstowe kodowanie binarne (technika binaryzacji złożonych zmiennych opisujących dane tekstowe, tak zwane *context-adaptive binary arithmetic coding* [CABAC], na przykład redukcja wektora charakteryzującego tekst za pomocą kategorii emocji w koncepcji RID do zmiennej porządkowej w postaci: wypowiedź z przewagą słów negatywnych emocjonalnie – wypowiedź neutralna – wypowiedź z przewagą słów pozytywnych emocjonalnie).

Perspektywa metodologiczna

Zastosowane przeze mnie podejście metodologiczne czerpie z kilku tradycji. Opiszę ich podstawy w sposób syntetyczny.

Najczęściej stosuję perspektywę, jaką w metodologii proponuje pragmatyzm. Charles Sanders Peirce miał nadzieję przenieść nauki laboratoryjne do filozofii, z kolei William James objawiał się jako trzeźwy empirysta, a John Dewey czerpał bezpośrednio z przyrodoznawstwa, próbując unaukowić dociekania filozoficzne. Ci trzej klasycy dali podstawy do rozwiązań, jakie w filozofii i metodologii nauki proponował Richard Rorty. Metodologię naukową traktował on jak skrzynkę z narzędziami. Sięgając do niej, sięgamy po wiele rozwiązań, z których nie tylko jedno będzie przydatne w potrzebie (Rorty 1996; 1999). Stosując ideę Rortyego, staram się stosować różnorodne rozwiązania analityczne z obszaru analiz jakościowych i ilościowych.

Druga wykorzystywana przeze mnie strategia czerpnie z obszaru *Mixed Methods Research*. To podejście przydatne jest mi wtedy, gdy:

- weryfikuję trafność wyników analiz jakościowych (QUAL), odwołując się do metod analiz ilościowych (QUAN),

- wyjaśniam wyniki QUAN, sięgając do danych QUAL.

Trzecia tradycja, do której sięgam, to obszar metod związanych z przetwarzaniem języka naturalnego (NLP). Stąd właśnie zapożyczam ideę analizy sentymentu. W obszarze NLP idea ta jest szczególnym rodzajem analizy opinii (*opinion mining* [Pang, Lee 2008]).

Narzędzia analityczne

W analizie sentymentu posłużę się słownikiem klasyfikującym słowa w dwojaki sposób. Po pierwsze dokonuję analizy sentymentu, klasyfikując słowa do wielu kategorii. Na drugim etapie dokonuję uproszczenia i klasyfikuję słowa do dwóch kategorii: pozytywne (emocje pozytywne, uczuciowość, chwała) i negatywne(niepokój, smutek, agresja). Słownik będący podstawą klasyfikacji to RID (*Regressive Imagery Dictionary*). Autorem słownika jest profesor psychologii Colin Martindale (1976; 1977; 1990).Słownik odnosi się do podstawowych i pierwotnych procesów poznawczych, którym przypisuje specyficzne dla nich słowa. Typy kategorii, które wyróżnił Martindale, prezentuję w tabeli 1.

Tabela 1. Kategorie emocji w słowniku Colina Martindale'a.

EMOCJE	
KATEGORIE	PRZYKŁADOWE SŁOWA
Pozytywne	wzruszenie, wesołość, radość, zabawa
Niepokój	obawa, strach, fobia
Smutek	depresja, niezadowolenie, samotność
Uczuciowość	czułość, małżeństwo, miłość
Agresja	wściekłość, sarkazm, przykrość
Ekspresja	zachowania, sztuka, taniec, śpiew
Chluba	podziw, bohaterskość, duma, król

Źródło: opracowanie własne na podstawie słownika Martindale'a dostępnego na stronie http://provalisresearch.com.

Stosując słownik RID, wykorzystuję narzędzia CAQDAS i inne narzędzia IT. Są to:

- pakiet QDA Miner/Wordstat/Simstat do zestawień i analiz statystycznych,
- darmowe środowisko IDLE Python's Integrated Development Environment do czyszczenia danych tekstowych oraz stosowania procedur deduplikacji.

Dodatkowo stosuję stop listę. Jest to lista słów, które są wykluczone z analiz. Należy wspomnieć, że wybór stop listy do analiz nie powinien nigdy odbywać się na zasadzie polecenia czy dobrej renomy listy. Każda z analiz niesie ze sobą konkretne pytania badawcze, co oznacza konieczność dostosowania do nich wykorzystywanych narzędzi badawczych, a zatem też stop listy.

Techniki i procesy analityczne

Identyfikacja i klasyfikacja słów do dwóch wyodrębnionych kategorii (pozytywne, negatywne) wykonane zostały z zastosowaniem słownika RID. Analiza ta przebiegała w pięciu etapach:

- pierwszy etap to n-gramowa klasyfikacja bez nadzorowania. N-gramowa to w tym przypadku analiza 1-gramowa, czyli polegająca na identyfikacji pojedynczych słów. Jest to najprostsze z rozwiązań metodologicznych polegające na rozpoznaniu słów w tekście oraz porównaniu ich ze słowami znajdującymi się w słowniku RID oraz, w konsekwencji, klasyfikacji słów do jednej z kategorii słownikowych. Słowa znajdujące się słowniku to: pełne wersje słów kluczowych związanych z emocjami (np. złość identyfikuje dokładnie jedno słowo, którego

jest odwzorowaniem) oraz rdzenie słów, wobec których zastosowana została procedura *stemmingu* (proces polegający na wydobyciu z wybranego wyrazu tzw. rdzenia, a więc tej jego części, która jest odporna na odmiany przez przypadki, rodzaje itp., np. uprzejm*[7] identyfikuje i klasyfikuje takie słowa, jak: uprzejmość, uprzejmie, uprzejmy, uprzejma),

- weryfikacja sklasyfikowanych słów odbywała się poprzez analizę kontekstową i analizę nadzorowaną (nieautomatyczną). Na tym etapie poprzez analizę kontekstu, w jakim wystąpiły klasyfikowane słowa, możliwa była ocena trafności klasyfikacji,

- trzeci etap analizy klasyfikacyjnej to rozbudowa słownika RID o reguły analityczne poddające diagnozie analizę kontekstu, w jakim występuje słowo kluczowe. Dodatkowo zakres słownika zostaje rozszerzony o analizę fraz. W ten sposób analiza n-gramowa zostaje rozszerzona o analizę fraz (pod uwagę wzięte zostały frazy dwu i trzy wyrazowe występujące więcej niż 5 razy),

- na kolejnym etapie wykonana zostaje analiza klasyfikacyjna, która bierze pod uwagę rozbudowany słownik RID i wykorzystuje bayesowski model klasyfikacji. Wyniki tej analizy zostają poddane ocenie w kolejnym kroku,

- weryfikacja użycia klasyfikatora bayesowskiego pozwala na ocenę efektywności zastosowanych metody i narzędzi analitycznych.

Wymienione tu techniki analityczne pozwoliły mi na weryfikację poprawności klasyfikacji, poprawę rzetelności analiz oraz optymalizację algorytmów klasyfikacyjnych zastosowanych w pierwotnym podejściu bez nadzorowania. Wnioski z zastosowania opisanych narzędzi i procedur omówię w ostatnim punkcie artykułu.

Wyniki analiz

Po wdrożeniu procedur czyszczenia i normalizacji danych, pierwszy etap w analizie to eksploracyjna analiza danych. W zestawieniach pojawiają się artykuły, których rok wydania nie jest znany (5 artykułów[8]). Pliki te wyłączyłem z dalszej analizy. Zastosowanie pierwszego modelu klasyfikacyjnego (bez nadzorowania) dało mi następujące wynik:

- słowa emocjonalnie negatywne: 521,

- słowa emocjonalnie pozytywne: 400.

Uzyskany wynik poddaję weryfikacji. W pierwszej kolejności kontroluję słowa zidentyfikowane przez słownik. Weryfikuję więc trafność klasyfikacji poprzez analizę kontekstową sklasyfikowanych słów. Poniżej podaję kilka przykładów analizowanych kontekstów.

[7] Zapis słów z gwiazdką w roli sufiksu oznacza, że wyszukiwane są słowa, która zawierają ciąg liter umieszczony przed gwiazdką. Na przykład: dom* wyszukuje takie słowa jak: domowy, domator, domatorka, domownik. Zapis, w którym gwiazdka występuje w roli prefiksu i sufiksu, identyfikuje słowa, które zawierają ciąg litera zapisany pomiędzy gwiazdkami. Na przykład: *dom* wyszukuje takie słowa jak: domownik, zadomowiony, udomowiony, Radom.

8 Podjąłem próby ustalenia pochodzenia wskazanych artykułów. Korespondencja z dostawcą oraz googlowanie fragmentów treści artykułów nie dały pozytywnej odpowiedzi.

Ramka 1. Przykłady analizowanych wypowiedzi.

Źródło: opracowanie własne.

Analiza kontekstu występowania słów pozytywnych pokazała, że w paragrafach, w których widnieją słowa emocjonalnie pozytywne, istnieją też inne słowa o podobnym zabarwieniu. Ta analiza pozwala włączyć zidentyfikowane słowa do kategorii słów pozytywnych emocjonalnie. Koszyk słów pozytywnych został poszerzony między innymi o takie rdzenie słów, jak: atrakc*, ambicj*, piękn*, uroczyst*.

Klasyfikacja słów do kategorii negatywnych emocjonalnie wykazała błędne użycie niektórych ze zdefiniowanych rdzeni słów. Oto przykłady pokazujące potrzebę nowych definicji:

- eliminacja rdzenia na rzecz listy słów: groz*– groza, grozić, groźne, groźba,
- redefinicja rdzenia: cierpi*– cierpie*, zadus* – zadusi*,
- wyłączenie słów o zabarwieniu w dużej mierze zależnym od kontekstu: zabiera* (zabierać komuś czyjeś dobro, zabierać głos w dyskusji), humor.

Dodatkowe rozszerzenia słownika RID uzyskuję w efekcie dodania negacji. Tak więc nie_atrakc*, nie_lojaln* identyfikują wypowiedzi nacechowane negatywnie.

W wyniku wskazanych powyżej procedur poprawność klasyfikacji poprawiła się w sposób istotny. Różnicę pomiędzy liczbą słów zidentyfikowanych

Ramka 2. Przykłady analizowanych wypowiedzi.

Przykłady poprawnie sklasyfikowanych wypowiedzi negatywnych:

Naczytałam się w Polsce, że nie tylko Żydzi **CIERPIELI** podczas wojny, że był to kraj rozdzierany przez **ZABORCÓW**.

Tak się **STRASZNIE** zdarzyło, że naziści umieścili obozy dla Żydów na ziemi polskiej.

Inna pani, która przedstawiła się, jako Żydówka pochodząca z Krakowa, a mieszkająca w Nowym Jorku, oświadczyła, że jest jej zwyczajnie **PRZYKRO**, bowiem tamtejsi źle poinformowani Żydzi twierdzą, że cala Polska jest antysemicka.

Przykłady niepoprawnie sklasyfikowanych wypowiedzi negatywnych:

Festiwal Kultury Żydowskiej, który po raz 11 odbył się tego roku na krakowskim Kazimierzu stanowi nie lada wyzwanie dla wszelkich imprez jazzowych, jakie się dzieją w tym mieście, od pojedynczych koncertów poczynając na szacownych **ZADUSZKACH** skończywszy.

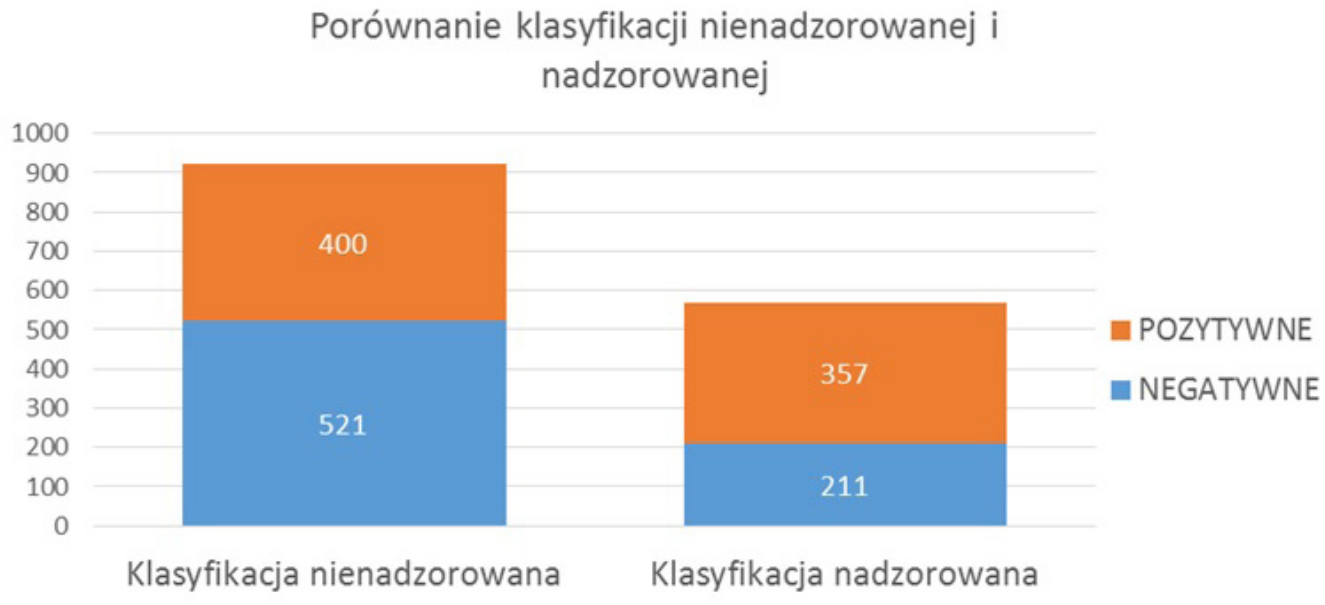
W ciągu pięciu lat namalowałem ze sto takich obrazków. Nie wystawiam ich i nie sprzedaję. Chwała Bogu, że nie muszę **ZABIEGAĆ** o pieniądze.

Wielu dyskutantów **ZABIERAŁO** głos w duchu bardzo osobistym.

Kiedyś mi zrobili piękne lakierowane trzewiki. Ubrałem je i poszedłem na futbol mecz. Kiedy futbol do mnie doszedł i ja go **KOPNĄŁEM**, otworzyła mi się podeszwa od przodu do tyłu. Mama dała mi pasem. To była wspaniała kobieta.

Źródło: opracowanie własne.

Wykres 1. Porównanie rezultatów metod klasyfikacji.



Źródło: opracowanie własne.

i sklasyfikowanych za pomocą dwóch metod widać na poniższym wykresie (wykres 1).

Na tym etapie analiz dysponuję przetestowanym i zweryfikowanym podstawowym słownikiem sentymentów opartym na koncepcji Martindale-'a. Skuteczność klasyfikacji z zastosowaniem tego słownika porównam teraz do trafności klasyfikacji, jaką daje klasyfikator naiwny Bayesa (Pang, Lee 2002). Jego zastosowanie do analizy tekstów zasadne jest w szczególności w sytuacji, w której mamy podstawy, by zakładać, że pomiędzy zmiennymi opisującymi analizowane obiekty istnieje istotny związek (Domingos, Pazzani 1997). W tym przypadku jest to intuicyjne założenie mówiące, że słowa o zabarwieniu emocjonalnym występują częściej w konkretnych typach artykułów. Publikacje prasowe wykorzystane w analizie scharakteryzowane zostały w oparciu o treść, jaką zawierają. Ten zabieg doprowadził do pogrupowania tekstów w następujące kategorie (spis kategorii zawiera tabela 2):

Tabela 2. Częstość występowania zidentyfikowanych typów artykułów.

TYP ARTYKUŁU	CZĘSTOŚĆ WYSTĄPIENIA	PROCENT
zapowiedź	149	32,80%
recenzja	125	27,50%
relacja	73	16,10%
wywiad	59	13,00%
program	35	7,70%
relacja post	13	2,90%
SUMA	**454**	**100%**

Źródło: opracowanie własne.

Intuicja podpowiada, że tekst, który jest programem imprezy, zdecydowanie rzadziej zawierał będzie słowa nacechowane emocjonalnie. Dodatkowo można założyć z niewielkim błędem, że zapowiedzi mogą zawierać takich słów mniej niż recenzje i relacje post. Te wskazówki sprawiają, że klasyfikator bayesowski sprawdza się w analizie lepiej niż słownik RID w jego uproszczonej wersji (uwzględniającej dwie kategorie: słowa pozytywnie i negatywnie nacechowane emocjonalnie). Podpowiedź, jaką otrzymał algorytm bayesowski, zwiększa trafność klasyfikacji paragrafów zawierających w sobie słowa o zabarwieniu emocjonalnym. Podsumowanie tej analizy podaję w tabeli 3.

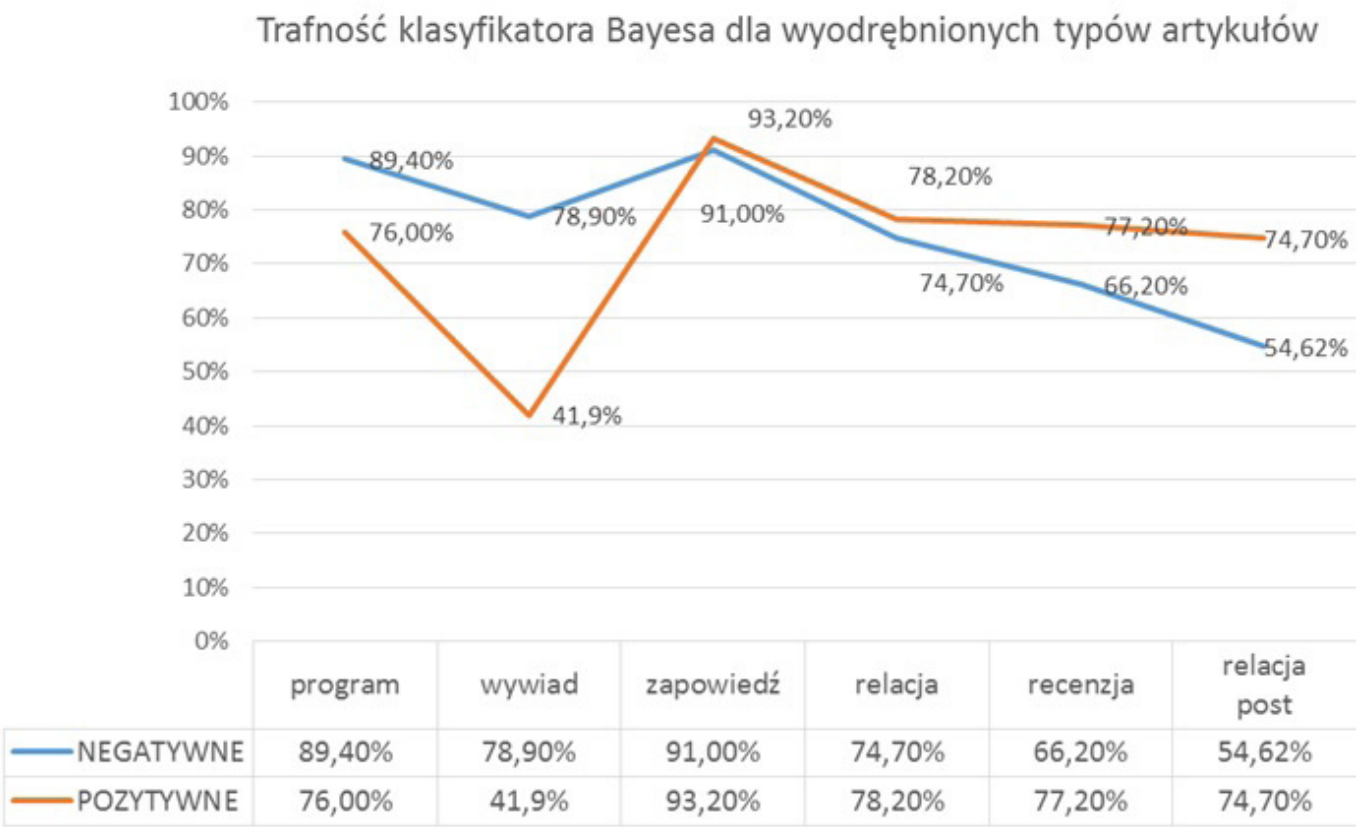
Tabela 3. Porównanie analiz wykonanych za pomocą wykorzystanych metod klasyfikacji.

ZAKRES ANALIZY	LICZBA ZIDENTYFIKOWANYCH	NAIWNY BAYES %	NIENADZOROWANY RID %
1-gram	721	88.12	79.20
1-gram + frazy	1274	74.67	64.87

Źródło: opracowanie własne.

W szczególności większość poprawnych klasyfikacji naiwny Bayes przeprowadził w tekstach, których dotyczyło założenie o zależności pomiędzy typem artykułu a występowaniem słów nacechowanych emocjonalnie. Wyniki klasyfikacji pokazuje wykres 2.

Wykres 2. Trafność klasyfikacji bayesowskiej w zidentyfikowanych typach artykułów (1-gram + frazy).



Źródło: opracowanie własne.

Szczególną uwagę zwraca niska trafność klasyfikacji wypowiedzi w przypadku wywiadów (pozytywne) oraz relacji post (negatywne). W tym pierwszym przypadku użycie przez rozmówców języka potocznego i spontanicznego napotyka na wiele niesklasyfikowanych w słownikach słów oraz problemy interpretacyjne, co w konsekwencji powoduje błędne klasyfikacje. Relacje post (po zakończeniu festiwalu kultury żydowskiej) bogate były w sformułowania wieloznaczne. Dodatkowo zawarte w nich refleksje sprzyjały pojawianiu się obok siebie ocen biegunowych w jednym zdaniu oraz takich, które formułowane były z wielu punktów widzenia (np. radość publiczności i zła organizacja). Przykłady takich wypowiedzi podaję poniżej:

Ramka 3. Przykłady analizowanych wypowiedzi.

Wypowiedzi niejednoznaczne i problematyczne w klasyfikacji

WYWIAD:

Co prawda nie dzwoniły telefony, co chwilę, ale zdarzył się jeden czy drugi cham czy chamka, który oczywiście jej nie wyłączył. Także **WYBITNIE** zakłócały odbiór tego koncertu raz po raz, często w najmniej oczekiwanych i potrzebnych momentach, rozbłyskujące flesze.

Na koncercie **KLASKAŁAM**, ale podczas modlitwy - **PŁAKAŁAM.**

Odkrywam nie tylko kulturę **CIERPIENIA**- poznaję także kulturę **OSIĄGNIĘĆ**.

RELACJE POST:

PRZECIWNIK kary śmierci.

KOCHANI wróćmy do normalności - postuluję, przynajmniej na niektórych koncertach, przynajmniej takich jak ten - zostawcie aparaty fotograficzne w domu

Taki to był Kazimierz. **ZACNY** i podły. Smutny i wesoły. Na jednym jego końcu tłum żegnał zmarłego rabina Kornitzera, a na drugim stary chasyd spieszył się na ślub córki cadyka z Bobowej.

Źródło: opracowanie własne.

Wskazane powyżej, zidentyfikowane w trakcie analiz, błędy klasyfikacyjne oraz modyfikacje słowników wykorzystałem w celu przebudowania słownika RID. Aby pokazać i zaakcentować wagę tego zabiegu, podam klasyfikację wypowiedzi prasowych dokonaną za pomocą słownika RID w jego pierwotnym kształcie oraz klasyfikację dokonaną po jego modyfikacjach.

W pierwotnej postaci słownik RID stanowi strukturę siedmiu kategorii[9]. Klasyfikacja tekstów prasowych za pomocą niezmienionego RID pokazuje, że analizowane teksty zawierają w przeważającej ilości wyrażenia i słowa związane

[9] Nazwy kategorii zmodyfikowałem tak, aby lepiej oddawały ich zawartość. Akceptowalnym kosztem tej modyfikacji jest wydłużenie nazw kategorii.

z zachowaniami ekspresyjnymi oraz te odwołujące się do chwały i chluby. Te dwie kategorie określają 69,9% zidentyfikowanych wypowiedzi prasowych.

Tabela 4. Klasyfikacja tekstów prasowych przez niezweryfikowany słownik RID.

	CZĘSTOŚĆ WYSTĘPOWANIA	%	LICZBA TEKSTÓW	%	TRAFNOŚĆ KLASYFIKACJI %
EKSPRESJA	1779	39,6	376	82,3	64,5
CHLUBA/CHWAŁA	1317	29,3	356	77,9	45,2
UCZUCIOWOŚĆ	474	10,6	219	47,9	41,1
POZYTYWE ODCZUCIA	310	6,9	172	37,6	29,8
AGRESJA/PRZEMOC/ NIECHĘĆ	233	5,2	132	28,9	62,2
NIEPOKÓJ	207	4,6	134	29,3	61,8
SMUTEK	128	2,9	82	17,9	54,2

Źródło: opracowanie własne.

Modyfikacja słownika RID w zakresie trafności klasyfikacji, poprawności zapisów rdzeni słów, trafności skategoryzowanych fraz oraz analizy kontekstu występowania zidentyfikowanych słów pozwala na osiągnięcie nowego wyniku. Po pierwsze, zmodyfikowany słownik identyfikuje mniejszą liczbę wypowiedzi. Po drugie, zwiększa się istotnie procent poprawności klasyfikacji. Dane te pokazuję w tabeli 5.

Tabela 5. Klasyfikacja tekstów prasowych po weryfikacji słownika RID.

	CZĘSTOŚĆ WYSTĘPOWANIA	%	LICZBA TEKSTÓW	%	TRAFNOŚĆ KLASYFIKACJI %
EKSPRESJA	987	0,20	365	27%	73,8
CHLUBA/CHWAŁA	911	0,18	328	24%	77,2
UCZUCIOWOŚĆ	427	0,09	211	15%	69,5
POZYTYWE ODCZUCIA	301	0,06	153	11%	54,3
NIEPOKÓJ	198	0,04	121	9%	78,8
AGRESJA/PRZEMOC/ NIECHĘĆ	193	0,04	124	9%	69,9
SMUTEK	72	0,01	71	5%	72,1

Źródło: opracowanie własne.

Wnioski i rekomendacje

Zagadnienia związane z automatyczną analizą treści wykraczają szeroko poza metody dyskutowane w tym artykule[10]. W tym obszarze problemowym mierzę się zaledwie z jednym zagadnieniem. Jest nim porównanie trafności klasyfikacji tekstów prasowych za pomocą dwóch metod (metoda słownikowa z zastosowaniem RID, naiwny klasyfikator Bayesa). W trakcie analiz obie metody ujawniły swoje słabe i silne strony, które teraz opiszę.

RID wykazuje względną skuteczność klasyfikacyjną, kiedy wzbogacony zostaje o automatyczne metody klasyfikacji. Naiwny Bayes wydaje się być dobrym punktem wyjścia dla analizy sentymentu. To podejście wymaga dodatkowego etapu, którym jest uczenie nadzorowane.

Na przykładzie przeprowadzonej analizy można sformułować hipotezę brzmiącą: RID niewzbogacony o reguły analiz kontekstowych wykazuje względnie słabą trafność klasyfikacyjną w języku polskim w przypadku wypowiedzi spontanicznych i w analizie języka niesformalizowanego (np. wywiady). A zatem efektywność klasyfikacji za pomocą RID może być zależna od typu tekstu i typu języka, wobec których słownik ten jest stosowany.

Dwa przywołane powyżej wnioski wzmacniają twierdzenie o potrzebie ewaluacyjnego podejścia do klasyfikacyjnych analiz tekstów. Przed dokonaniem analizy, procedurom ewaluacji poddane powinny być takie narzędzia, jak: słownik klasyfikacyjny, reguły leksykalne słownika, stop lista, algorytm lematyzacji, rdzenie słów. Dodatkowo proces budowy i doskonalenia słowników klasyfikacyjnych powinien być poddany procedurze wielokrotnej weryfikacji reguł słownikowych. Ten zabieg stosowany w trakcie analiz pozwala na znaczące zwiększenie poprawności klasyfikacji.

W automatycznych analizach tekstów niezbędną staje się miara nietrafności klasyfikacji. Zagadnienie to dotyczy niepewności i błędu pomiaru. Miara błędnej klasyfikacji jest domeną skwantyfikowanych analizy tekstowych (Hopkins, King 2010). Niemniej jednak ocena błędu klasyfikacji może być stosowana również przez metody nieprobabilistyczne, a wśród nich przez metody słownikowe.

Wynik klasyfikacji osiągnięty za pomocą metody Bayesa może być zweryfikowany przez reguły słownikowe. Te dwie metody w zdecydowanie krótszym czasie mogą dać poprawniejszy wynik klasyfikacji niż każda z nich stosowana z osobna.

Poza testowanym słownikiem RID istnieje kilka innych słowników do analizy sentymentu. Wartym sprawdzenia jest podejście, które testowałoby trafność zastosowania różnych słowników do analizy sentymentu dla danego rodzaju wypowiedzi tekstowych (artykułów prasy codziennej, artykułów prasy branżowej, języka subkultury itp.).

Na koniec chciałbym dodać bardziej ogólną refleksję dotyczącą analizy sentymentu. W zdecydowanej większości przypadków opiera się ona na identyfikacji słów i fraz kluczowych. Podejście takie bez wiedzy o zwyczajach językowych autorów

[10] Szersze omówienie tych zagadnień znaleźć można w Jurafsky, Martin (2009).

wypowiedzi, uwzględnienia specyfiki użycia słów, kontekstowej zmienności znaczeń, ale też bez informacji o sposobie wypowiedzi (np. tonie głosu autora wypowiedzi, czym zajmuje się obszar zwany *voice mining*) nastręcza wiele trudnych do rozwiązania problemów. W szczególności pojawiają się one w analizie wypowiedzi spontanicznych oraz wypowiedzi formułowanych w języku niesformalizowanym. W najlepszym wypadku opisane w artykule metody dają poprawność klasyfikacji wypowiedzi na poziomie 80%. Trzeba jednak dodać, że prezentowane tu podejście biorące pod uwagę frekwencję występowania słów nie jest ani jedynym, ani najbardziej trafnym z istniejących. Czym innym jest bowiem zliczenie słów wyrażających emocje, a czym innym rozumienie wypowiedzi przez pryzmat intencji autora wypowiedzi. Interesującym byłoby zatem zaprojektowanie i wykonanie analizy, która pozwoliłaby na porównanie wyników uzyskanych za pomocą dwóch wspomnianych tu podejść.

Bibliografia

Acerbi Alberto i in. (2013) *The Expression of Emotions in 20th Century Books*. „PLoS ONE", vol. 8, no. 3, s. 1–6 [dostęp 1 maja 2014 r.]. Dostępny w Internecie: ‹http://www.plosone.org/article/fetchObject.action?uri=info%3Adoi%2F10.1371%2Fjournal.pone.0059030&representation=PDF›.

Cardie Claire i in. (2003) *Combining low-level and summary representations of opinions for multi-perspective question answering* [w:] *Proceedings of the AAAI Spring Symposium on New Directions in Question Answering*, s. 20–27 [dostęp 1 maja 2014 r.]. Dostępny w Internecie: ‹http://www.aaai.org/Papers/Symposia/Spring/2003/SS-03-07/SS03-07-004.pdf›.

Das Sanjiv R., Chen Mike J. (2001) *Yahoo! for Amazon: Sentiment Extraction fromSmall Talk on the Web*,„Management Science", Vol. 53, No. 9, s. 1375–1388 [dostęp 1 maja 2014 r.]. Dostępny w Internecie: ‹http://algo.scu.edu/~sanjivdas/chat_FINAL.pdf›.

Dave Kushal, Lawrence Steve, Pennock David M. (2003) *Mining the peanut gallery: Opinion extraction and semantic classification of product reviews* [w:] *Proceedings of WWW*, s. 519–528, [dostęp 1 maja 2014 r.]. Dostępny w Internecie: ‹http://www.kushaldave.com/p451-dave.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.

Dini Luca, Mazzini Giampaolo (2002) *Opinion classification through information extraction* [w:] *Proceedings of the Conference on Data Mining Methods and Databases for Engineering, Finance and Other Fields (Data Mining)*, s. 299–310 [dostęp 1 maja 2014 r.]. Dostępny w Internecie: ‹http://www.google.pl/url?sa=t&rct=j-&q=&esrc=s&source=web&cd=1&ved=0CC8QFjAA&url=http%3A%2F%2Fia2010primercuat.googlecode.com%2Fsvn-history%2Fr45%2Ftrunk%2FSEI-GO%2Fdocs%2F10.1.1.109.1736.pdf&ei=D6diU9ahG8ep7AbGu4GYDQ&usg=AFQjCNGlzrqDMZ3aj-M_a-Yv4ITbwdU0KQ&bvm=bv.65788261,d.ZGU&cad=rja›.

Domingos Pedro, Pazzani Michael (1997) *On the optimality of the simple Bayesian classifier under zero-one loss*. Machine Learning", vol. 29, no. 2/3,s.103–130.

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, s. 417–422 [dostęp 1 maja 2014]. Dostępny w Internecie: ‹http://gandalf.aksis.uib.no/lrec2006/pdf/384_pdf.pdf›.

Hogenraad Robert, Orianne Emilie (1986) *Imagery, regressive thinking, and verbal performance in internal monologue*. „Imagination, Cognition, and Personality", vol. 5, no. 2, s. 127–145.

Hopkins Daniel, King Gary (2010) *Extracting systematic social science meaning from text*. „American Journal of Political Science", vol. 54, no. 1, s. 229–247.

Hotho Andreas, Nürnberger Andreas, Paaß Gerhard (2005) *ABrief Survey of Text Mining*. „German Journal for Computer Linguistics and Speech Technology", vol. 20, no. 1, s. 19–62.

Jurafsky Dan, Martin James H. (2009) *Speech and natural language processing: An introduction to natural language processing, computational linguistics, and speech recognition*. Upper Saddle River, NJ: Prentice Hall.

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

Loughran Tim, McDonald Bill (2011) *When is a Liability not a Liability? Textual Analysis, Dictionaries, and 10-Ks*. „The Journal of Finance", vol. 66, no. 1, s. 35–65.

Martindale Colin (1976) *Primitive mentality and the relationship between art and society*. „Scientific Aesthetics", vol. 1, s. 5–18.

------ (1977) *Syntactic and semantic correlates of verbal tics in Gilles de la Tourette's syndrome: A quantitative case study*. „Brain and Language", vol. 4, s. 231–247.

------(1990) *The clockwork muse: The predictability of artistic change*. New York: Basic Books.

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

Nasukawa Tetsuya, Yi Jeonghee (2003) *Sentiment analysis: Capturing favorability using natural language processing* [w:] *Proceedings of the Conference on Knowledge Capture (K-CAP)* s. 70–77 [dostęp 1 maja 2014 r.]. Dostępny w Internecie: ‹http://tredocs.com/tw_files2/urls_41/40/d-39217/7z-docs/7.pdf›.

Nielsen Finn Å. (2011) *A new ANEW: Evaluation of a word list for sentiment analysis in microblogs* [w:] RoweMatthew i in., eds., *Proceedings of the ESWC2011 Workshop on 'Making Sense of Microposts': Big things come in small packages 718 in CEUR Workshop Proceedings*, Heraklion, s. 93–98 [dostęp 1 maja 2014 r.]. Dostępny w Internecie: ‹http://ceur-ws.org/Vol-718/msm2011_proceedings.pdf›

Pagel Mark, Atkinson Quentin D., Meade Andrew (2007) *Frequency of word-use predicts rates of lexical evolution throughout Indo-European history*. „Nature", vol. 449, s. 717–720.

Pang Bo, Lee Lillian (2002) *Thumbs up? Sentiment Classification using Machine Learning Techniques*.„EMNLP '02 Proceedings of the ACL-02 conference on Empirical methods in natural language processing", vol. 10, s. 79–86.

------ (2008) Opinion Mining and Sentiment Analysis. „Foundations and Trends in Information Retrieval", vol. 2, s. 1–135.

Rorty Richard (1996) *Przygodność, ironia i solidarność*. Przełożył Wacław J. Popowski. Warszawa: Spacja.

------ (1999) *Obiektywność, relatywizm i prawda*. Przełożył Janusz Margański. Warszawa: Aletheia.

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 York: ACM, s. 1–6.

Yi Jeonghee i in. (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).* Washington: IEEE Computer Society, s. 427–434.

Cytowanie

Tomanek Krzysztof (2014) *„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 [dostęp dzień, miesiąc, rok]. Dostępny w Internecie: ‹www.przegladsocjologiijakosciowej.org›.

Sentiment Analysis. An Example of Application and Evaluation of RID Dictionary and Bayesian Classification Methods in Qualitative Data Analysis Approach

Abstract: The purpose of this article is to present the basic methods for classifying text data. These methods make use of achievements earned in areas such as: natural language processing, the analysis of unstructured data. I introduce and compare two analytical techniques applied to text data. The first analysis makes use of thematic vocabulary tool (sentiment analysis). The second technique uses the idea of Bayesian classification and applies, so-called, naive Bayes algorithm. My comparison goes towards grading the efficiency of use of these two analytical techniques. I emphasize solutions that are to be used to build dictionary accurate for the task of text classification. Then, I compare supervised classification to automated unsupervised analysis' effectiveness. These results reinforce the conclusion that a dictionary which has received good evaluation as a tool for classification should be subjected to review and modification procedures if is to be applied to new empirical material. Adaptation procedures used for analytical dictionary become, in my proposed approach, the basic step in the methodology of textual data analysis.

Keywords: qualitative data analysis**,** sentiment analysis, content analysis, text mining, coding techniques, natural language processing, RID dictionary, naive Bayes, CAQDAS