MODERN MANAGEMENT REVIEW **2016**
MMR, vol. XXI, 23 (2/2016), pp. 31-42 **April-June**

Anna GŁADYSZ[1]

WYKORZYSTANIE METOD AUTOMATYZACJI TEKSTU W ANALIZIE OPINII KONSUMENCKICH

Analiza opinii konsumenckich jest obszarem badań, który może mieć znaczący wpływ na rozwój działalności biznesowej przedsiębiorstwa. Jest też narzędziem, które może dostarczyć istotnych informacji mających wpływ na wizerunek firmy, co ma duże znaczenie dla firm działających na bardzo konkurencyjnym rynku. Wielu konsumentów przed dokonaniem wyboru towaru lub usługi przeszukuje Internet w poszukiwaniu opinii innych użytkowników sieci. Znalezione rekomendacje często odgrywają decydującą rolę podczas podejmowania decyzji. Aby nadążać za zmieniającymi się oczekiwaniami klientów, warto postawić na badania ich opinii. Narastająca liczba opinii dostępnych w sieci wytworzyła potrzebę ich automatycznej analizy i przetwarzania. Zagadnienie to zyskuje na popularności zarówno wśród badaczy, jak i wśród przedsiębiorców, dla których opinie konsumentów stanowią źródło informacji biznesowej. Dzięki stale rosnącej potrzebie dostępu do opinii klientów, a co za tym idzie – wiedzy i informacji, które można z nich czerpać, narzędzia umożliwiające automatyzację procesu pozyskiwania z nich kluczowych i strategicznych informacji zyskują na znaczeniu. Problem ten wymaga nieco innego spojrzenia na dane i doboru określonego sposobu ich analizowania za pomocą technik eksploracji danych, zwłaszcza tekstowych. Głównym celem pracy jest przeprowadzenie analizy automatycznej klasyfikacji opinii z wykorzystaniem eksploracyjnych metod analizy tekstu oraz metody opartej na wzorcach. Wykorzystane podejścia zostaną porównane z tymi dotychczas stosowanymi w badaniach. Wykorzystanie informacji pozyskanych z opinii klientów przyczynia się do zwiększenia wiedzy pracowników na wszystkich szczeblach organizacji, zapewnia dostęp do odpowiednich informacji we właściwym czasie, dzięki czemu wpływa na trafność podejmowanych decyzji biznesowych.

Słowa kluczowe: opinie konsumenckie, automatyczna analiza opinii, eksploracyjna analiza tekstu, klasyfikacja dokumentów

1. WPROWADZENIE

Żadna firma nie może zatrzymać się na danej fazie rozwoju, uznając, że to wystarczy i osiągnęła wszystko, co mogła. Współcześnie wciąż zmieniają się otoczenie biznesowe i oczekiwania konsumentów, stąd przedsiębiorcy muszą nieustannie rozwijać swoje produkty i usługi, pamiętając o tym, że powinny się one dostosować przede wszystkim do potrzeb ich użytkowników. Aby nadążać za zmieniającymi się oczekiwaniami klientów, warto postawić na badania ich opinii.

Analizaopiniikonsumenckichjestobszarembadań,którymożemiećznaczący wpływ na współczesne zarządzanie, a co za tym idzie – na rozwój działalności biznesowej[2]. Znaczna liczba konsumentów przed dokonaniem wyboru o zakupie towaru lub skorzystaniu

[1]Dr inż. Anna Gładysz, Wydział Zarządzania, Politechnika Rzeszowska, al. Powstańców Warszawy 10, 35-959 Rzeszów, e-mail: anna.gladysz@prz.edu.pl

[2] Zhu F., Zhang, X., *Impact of online consumer reviews on sales: The moderating role of product and consumer characteristics*, „Journal of Marketing" 74/2 (2010), s. 133–148.

z usługi przeszukuje Internet w poszukiwaniu opinii innych użytkowników sieci. Znalezione rekomendacje często odgrywają decydującą rolę podczas podejmowania decyzji. Z tego powodu dla przedsiębiorstwa istotna wydaje się wiedza o tym, w jaki sposób jest ono (a także jego produkty) postrzegane przez konsumentów czy nawet konkurencję. Pozwala to na podejmowanie właściwych działań marketingowych zmierzających do wykreowania jak najlepszej opinii wśród wybranej grupy docelowej.

Narastająca liczba opinii dostępnych w Internecie wytworzyła potrzebę ich automatycznej analizy i przetwarzania. Problem ten wymaga nieco innego spojrzenia na dane i doboru określonego sposobu ich analizowania za pomocą technik eksploracji danych[3]. Znalezienie takich zestawów słów wewnątrz opinii konsumenckich, przy uwzględnieniu ich nacechowania semantycznego, pozwala na stworzenie bazy wiedzy, w oparciu na której możliwe jest także dokonywanie klasyfikacji nowych opinii. Zagadnienie to zyskuje na popularności zarówno wśród badaczy, jak i wśród przedsiębiorców, dla których opinie konsumentów stanowią źródło informacji biznesowej.

Warto zwrócić uwagę także na zjawisko wynikające z globalizacji i powszechnego dostępu do Internetu – napotykamy dostępność opinii w różnych językach. Ich analiza jest dodatkowo utrudniona ze względu na specyfikę każdego języka naturalnego. Zrealizowane do tej pory prace teoretyczne i wypracowane na ich podstawie narzędzia ukierunkowane są głównie na automatyczną analizę opinii przygotowanych w języku angielskim. Dostępność rozwiązań dla innych języków – w tym również dla języka polskiego – jest znacznie bardziej ograniczona[4].

Głównym celem pracy jest przeprowadzenie analizy własności automatycznej klasyfikacji opinii napisanych w języku polskim z wykorzystaniem metod algebraicznych eksploracyjnej analizy tekstu.

2. AUTOMATYCZNA ANALIZA OPINII KONSUMENCKICH

Źródła literaturowe definiują automatyczną analizę opinii konsumenckich jako ogół działań mających na celu zautomatyzowanie procesu wyszukiwania, ekstrakcji i analizy danych pochodzących ze specyficznych tekstów, jakimi są opinie użytkowników. Obszar badań zajmujący się poruszaną problematyką nazywany jest drążeniem opinii (*opinionmining*) lub analizą wydźwięku (*sentimentanalysis*) i jest dobrze znanym problemem z zakresu przetwarzania języka naturalnego (NLP, *naturallanguageprocessing*), lingwistyki komputerowej (*computationallinguistics*) oraz eksploracyjnej analizy tekstu (*textmining*)[5]. Zadaniem stawianym przed automatyczną analizą opinii konsumenckich jest określenie nastawienia autora wypowiedzi do jej przedmiotu.

Opinie konsumenckie obejmują swoim zasięgiem opinie na temat dóbr, opublikowane w pewnym źródle internetowym, wyrażone przez podmioty niebędące ekspertami w danej dziedzinie[6].Opinie przedstawiają specyficzny rodzaj danych tekstowych, które mają subiektywny charakter – wyrażają stosunek autora wypowiedzi do przedmiotu opinii. Opi-

[3] D. Larose, *Odkrywanie wiedzy z danych*, Wydawnictwo Naukowe PWN, Warszawa 2006.

[4] P. Lula, *Automatyczna analiza opinii konsumenckich*,[w:]*Taksonomia 18, Klasyfikacja i analiza danych – teoria i zastosowania*, red. K. Jajuga, M. Walesiak, Wydawnictwo UE we Wrocławiu, Wrocław 2011.

[5] B. Liu, *Opinion Mining and Sentiment Analysis*, [w:]idem, *Web Data Mining, Data-Centric Systems and Applications*, Springer Berlin Heidelberg, Berlin 2011, s. 459–526.

[6] S. Mudambi, D. Schuff, *What makes a helpful online review? A study of customer reviews on Amazon.com*, „MIS Quarterly" 34/1 (2010), s. 185–200.

nie, oprócz subiektywnych odczuć autora, mogą zawierać obiektywne stwierdzenia, które mogąmiećwpływnawynikanalizy.Dlategoczęśćbadaczyrozpoczynaanalizęod określenia, czy badany tekst jest obiektywny, czy subiektywny[7].W niektórych serwisach opinie słowne są wspierane oceną punktową lub gwiazdkami. Punkty mogą być przypisywane do całego produktu lub jego aspektów, gdzie lista takich aspektów jest zdefiniowana dla danego produktu. Popularną reprezentacją graficzną przyznawania ocen są gwiazdki, dzięki którym użytkownik może nadać ocenę w skali 1–5, klikając na odpowiednią gwiazdkę[8].Ważnymi zaletami takiego sposobu oceniania są jego szybkość i intuicyjność, a także łatwość przeprowadzania późniejszych analiz na tak wystawionych ocenach. Przykładowo określenie średniej oceny danego produktu wymaga jedynie policzenia średniej arytmetycznej przyznanych punktów. Podejście takie często traktowane jest jako ogólna ocena danego produktu lub jego aspektu. Niestety, istniejące wyniki badań wskazują, że średnia punktów może być mylnym wskaźnikiem sposobu postrzegania dóbr przez konsumentów (często dla jednego produktu recenzenci przyznali wiele bardzo pozytywnych, jak i jednoznacznie negatywnych ocen)[9].

Opinie można podzielić na grupy według ich formatu[10]:

- zalety i wady – opinie mogą być wyrażane w postaci list zalet i wad. Wartości w każdej z tych list mogą być dowolne lub wybierane z predefiniowanej dla danej kategorii produktów listy. Podsumowanie opinii wyrażonych w tym formacie może być łatwo wygenerowane, choćby poprzez określenie liczby konsumentów, którzy wybrali daną zaletę lub daną wadę w opinii produktu. Dzięki temu można szybko poznać najczęściej wymieniane wady i zalety danego produktu;
- zalety i wady oraz podsumowanie – występuje także postać wypowiedzi tekstowej. Podsumowywanie opinii staje się trudniejszym zadaniem i wymaga wykorzystania technik przetwarzania języka naturalnego;
- dowolny – wykorzystanie innych formatów.

Podsumowanie najważniejszych z punktu widzenia Autora cech poszczególnych formatów zaprezentowano w tabeli 1. Poszczególne formaty omówiono, uwzględniając podział:

- zakresu ocenianych aspektów – w jakim stopniu możliwe jest wyrażanie opinii o aspektach danego dobra, uwzględniając różny poziom szczegółowości;
- stopnia trudności podsumowywania opinii – na ile skomplikowane jest automatyczne przetwarzanie opinii wyrażanych w tym formacie.

[7] B.Pang,L. Lee, *Opinion Mining and Sentiment Analysis*, „Foundations and Trends in Information Retrieval" 2/1–2 (2008), s. 1–135.

[8] N.Hu, J.Zhang, P.A. Pavlou, *Overcoming the J-shaped distribution of product reviews*, „Commun. ACM" 52/10 (2009), s. 144–147.

[9] N.Hu, P. Pavlou, J. Zhang, *Can online reviews reveal a product's true quality?: empirical findings and analytical modeling of Online word-of-mouth communication*, Proceedings of the 7th ACM conference on Electronic commerce, ACM, 2006, s. 324–330.

[10] B.Liu, *Web DataMining. Exploring Hyperlinks, Contents, and Usage Data*, Springer-Verlag Berlin, Heidelberg 2007.

Tabela 1. Porównanie różnych formatów wyrażania opinii

Format	Zakres ocenianych aspektów	Stopień trudności podsumowania opinii
Skala punktowa / gwiazdki	Ściśle ustalony, zazwyczaj wąski; zwykle pojedyncza ocena całego przedmiotu lub kilka ocen dla wybranych aspektów przedmiotu	Niski; proste wyliczenie średnich
Zalety i wady	Listy możliwych zalet i wad mogą mieć wiele pozycji; listy te mogą też być rozszerzane przez konsumentów	Niski lub wysoki, zależy, czy lista możliwych wad i zalet jest predefiniowana, czy też wprowadzana przez każdego użytkownika z osobna; narzucona struktura ułatwia przetwarzanie
Zalety i wady oraz podsumowanie	Nieograniczony; konsument może odnosić się do dowolnych aspektów recenzowanego dobra za pomocą dowolnych sformułowań	Bardzo wysoki; brak narzuconej struktury wypowiedzi oraz wieloznaczność języka naturalnego wymagają wykorzystania zaawansowanych technik przetwarzania języka naturalnego
Dowolny	Nieograniczony; konsument może odnosić się do dowolnych aspektów recenzowanego dobra za pomocą dowolnych użytych technik	Bardzo wysoki

Źródło: opracowanie własne.

2.1. Przegląd podejść do automatycznej analizy opinii konsumenckich

W literaturze przedmiotu przedstawiono możliwe rodzaje działań[11]:

- **klasyfikacja opinii** – podział opinii na grupy według ich nacechowania (np. pozytywne, negatywne, neutralne) lub przypisanie pojedynczej opinii jej polaryzacji (przydzielenie jej do jednej z uprzednio wymienionych grup). Brana jest tu pod uwagę opinia jako całość;
- **analiza ukierunkowana na cechy produktu** – wyszukanie w opinii poszczególnych aspektów (cech) przedmiotu opinii, a następnie zbadanie stosunku autora wypowiedzi do tego właśnie aspektu. Badana jest nie cała opinia, ale poszczególne jej części odnoszące się do kolejnych cech opisywanego produktu czy usługi;
- **analiza porównawcza produktów** – badanie opinii na temat jednego produktu, określonej przez analizę zdania porównującego go do innego produktu. Konieczne jest zidentyfikowanie w opinii zdań porównujących, a następnie ich analiza ukierunkowana na przedmiot porównania.

[11] B. Liu, *op. cit*.; B. Pang,L. Lee, *Opinion Mining...*, s. 1–135.

Najczęściej wykorzystywanym rodzajem automatycznej analizy opinii konsumentów jest klasyfikacja opinii. Każdej opinii przypisywane jest nacechowanie – określenie, w jakim stopniu opinia jest pozytywna, czy też negatywna. Wykorzystywanychjesttuwielepodejść.Podstawoweznichopierasięnasłowach. Każdemu słowu w opinii przypisywane jest nacechowanie, a następnie na tej podstawie dokonywana jest ocena nacechowania całej opinii. Podejście to ma jednak wiele wad i jest dużym uproszczeniem. Najczęściej opinie klasyfikowane są do jednej z dwóch grup: pozytywne lub negatywne. Spotkać można klasyfikację zawierającą dodatkową grupę – neutralne oraz klasyfikacje wykorzystujące wielostopniowe skale (np.3-lub4-stopniowaskalapunktów). Jednak zarówno na podstawie studiów literaturowych[12], jak i własnych badań można stwierdzić, że obecnie wykorzystywane narzędzia nie dają dobrych rezultatów przy klasyfikacji na więcej niż dwie grupy.

Koncentrując się na klasyfikacji opinii, można wyróżnić cztery textmininigowe podejścia do niej[13]:

- podejście oparte na słowach (*word-basedapproach*) –znaczenie wypowiedzi (również jej nacechowanie) jest zakodowane w pojedynczych słowach stanowiących dany tekst;
- podejście bazujące na wzorcach (*pattern-basedapproach*) –nacechowanie opinii wyznaczają nie pojedyncze słowa, ale zbudowane z nich związki frazeologiczne. Tak więc konieczne jest wyszukanie wśród słów związków wyrazowych;
- podejście bazujące na ontologiach (*ontology-basedapproach*) – pojedyncza opinia może zostać przedstawiona jako instancja ontologii. Następnie instancje te mogą zostać porównane, opinie zaś zaklasyfikowane do jednej z grup;
- podejście, u którego podstaw stoi uczenie maszynowe (*machine learning approach*) – dzięki zastosowaniu uczenia maszynowego można zbudować system,którynietylkonapodstawieodpowiedniodobranegouczącegozbioruopinii będzie je klasyfikował do odpowiednich grup, ale również będzie się rozwijał wraz z pojawieniem się nowych, specyficznych opinii.

W pracy Cambria, Schullera, Yunqinga i Havasi[14]znaleźć można podobną klasyfikację podejść do automatycznej analizy opinii konsumentów.

2.2. Podejście oparte na słowach

Traktując każdą opinię konsumencką jako dokument tekstowy niemający określonej struktury, nie można dokonać prostej klasyfikacji i pozyskać z niego określonych informacji. Na tym etapie niezbędne jest wstępne przetworzenie opinii, czego efektem będzie odpowiednia postać tekstu składająca się ze zmniejszonej jego reprezentacji. Możliwości algorytmów eksploracji tekstu są mocno ograniczone, jeżeli chodzi o pracę na dużej ilości danych (duża złożoność obliczeniowa i długi czas pracy),dlatego etap ten obejmuje prze-

[12] B. Pang, L. Lee, *Seeing Stars: Exploiting Class Relationships for Sentiment Categorization with Respect to Rating Scales*, Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, Stroudsburg, PA, 2005,s.115–124.

[13] P. Lula, K. Wójcik, *Sentiment analysis of consumer opinions writen in Polish*, „Economics and Management" 2011, s. 1286–1291.

[14] E. Cambria, B. Schuller, X. Yunqing, C. Havasi, *New avenues in opinion mining and sentiment analysis*, „Intelligent Systems, IEEE" 2013/28, s. 15–21.

kształcenie tekstu do zmniejszonej i uproszczonej postaci. Postać taka umożliwia o wiele szybszą i bardziej efektywną analizę danych. W podejściu tym automatyczne przetwarzanie opinii konsumentów obejmuje następujące fazy:

- podział tekstu wejściowego opinii na zdania, słowa oraz usunięcie wszystkich znaków interpunkcyjnych;
- odrzucenie słów nieistotnych (wykorzystanie stop-listy);
- tematyzacja – wybór słów istotnych i sprowadzenie ich do postaci podstawowej (stemming) przy wykorzystaniu metody reguł gramatycznych w algorytmie lub metody słownikowej;
- zliczanie wystąpień słów;
- obliczanie wag dla wszystkich słów;
- przypisanie każdemu dokumentowi przynależnych słów, które mogą odgrywać rolę słów kluczowych.

2.3. Wstępne przetworzenie opinii – zmniejszenie reprezentacji tekstu

2.3.1. Prawo Zipfa

Zgodnie z teorią informacji w każdym języku naturalnym istnieje zależność, mówiąca o tym, że rozkład częstości słów występujących w danym języku nie jest losowy. Co więcej rozkład ten jest bardzo charakterystyczny – stosunkowo niewiele słów bardzo często pojawia się w treści dokumentu oraz dużo słów występuje bardzo rzadko. Ten nierównomierny rozkład słów w językach naturalnych został potwierdzony przez badanie amerykańskiego lingwisty George'a Zipfa[15]. Prawo to umożliwia odnalezienie zależności w ogromnych ilościach danych tekstowych, które na pierwszy rzut oka mogą wydawać się jednolite. Prawo to można także wykorzystać do określenia ważności słów. Gdy każdemu słowu z rozkładu Zipfa przypisze się wartość oznaczającą pozycję w rankingu ważności takiego słowa na podstawieczęstościjegowystąpienia,toczęstośćwystępowaniasłówbędzie odwrotnie proporcjonalna do pozycji tego słowa w rankingu ważności słów[16].

2.3.2. Stop-lista

Każdy język naturalny charakteryzuje specyficzna konstrukcja o odpowiednich kryteriach składniowych i fleksyjnych. Do budowy zdań używane są różne części mowy i są to (w zależności od języka): zaimki, przyimki, rodzajniki, spójniki, wykrzykniki. Słowa należące do wymienionych kategorii mają bardzo wysoką częstość wystąpień, ale nie niosą żadnej użytecznej wiedzy. Metoda stop-listy[17] polega na pominięciu tych słów na początkowym etapie przygotowania danych w celu usprawnienia pracy algorytmu.

2.3.3. Przycinanie (*pruning*)

Poza ograniczaniem liczności zbioru słów poprzez tworzenie stop-listy można także zmniejszać reprezentację tekstu za pomocą miar statystycznych – jest to tzw. przycinanie (*pruning*). Rozwiązanie to polega na usuwaniu słów najczęściej lub zbyt często występujących w danym dokumencie tekstowym oraz słów, których częstość występowania jest

[15] G. Zipf, *Human Behaviour and the Principle of Least Effort*, Cambridge, 1949

[16] M. Ward, *50 najważniejszych problemów zarządzania*, Wydawnictwo Profesjonalnej Szkoły Biznesu, Kraków 1997.

[17] A. Rajaraman, J.D. Ullman, *Data Mining. Mining of Massive Datasets*, Cambridge University Press, New York 2012.

bardzo mała. Określenie progów oddzielających słowa nieistotne z powodu zbyt dużej lub zbyt małej częstości użycia znacznie zmniejsza rozmiar reprezentacji, poprawiając efektywność przetwarzania danych, redukując szum informacyjny, nie zmieniając przy tym znacząco wyników działania algorytmu eksploracji tekstu.

2.3.4. Funkcje ważące

Podejście bazujące na pojęciu modelu przestrzeni wektorowej może być wykorzystywane jako punkt wyjścia dla zadań związanych z automatycznym przetwarzaniem opinii konsumentów[18]. Zastosowanie modelu reprezentacji wektorowej dla dokumentów tekstowych sprowadza się do wyznaczenia macierzy częstości występowania poszczególnych słów w danej opinii[19].

Po uzyskaniu macierzy częstości wykorzystywane są odpowiednie funkcje ważące (*weightingfunctions*), które mają za zadanie ją ulepszyć. Ważenie jest procesem, który każdemu słowu w dokumencie przypisuje wagę wynikającą z częstości jego wystąpień w dokumencie[20]. Najprostszym sposobem ważenia macierzy jest przypisanie każdej współrzędnej wektora dokumentu częstości występowania słowa w dokumencie. Schemat ten jest określany mianem *termfrequency* i oznacza się go jako $tf_{t,d}$. Opisana operacja prowadzi do definicji wskaźnika istotności słowa w postaci:

$$WIS_{t,d}^{A} = tf_{t,d} \tag{1}$$

gdzie:

$WIS_{t,d}^{A}$ – wskaźnik istotności t-słowa w d-tym dokumencie oparty na częstości wystąpienia.

Ta prosta metoda ma poważną wadę – każde słowo w dokumencie jest uznawane za jednakowo ważne. Należy również zauważyć, że wartość wskaźnika jest uzależniona od długości dokumentu.

Chcąc wyeliminować wpływ długości dokumentu, można dokonać przekształcenia równania (1), zastępując wszystkie dodatnie wartościprzez 1, wartości zerowe zaś pozostawiając niezmienione. Prowadzi to do wskaźnika istotności słowa w postaci:

$$WIS_{t,d}^{B} = \begin{cases} tf_{t,d} = 1 \\ tf_{t,d} = 0 \end{cases} \tag{2}$$

gdzie:

$WIS_{t,d}^{B}$ – wskaźnik istotności t-słowa w d-tym dokumencie oparty na jego wystąpieniu równy jedności, jeśli t-słowo występuje w d-tym dokumencie (jeden bądź więcej razy), oraz równy zero jedności,jeśli t-słowo nie występuje w d-tym dokumencie.

---

[18] T. Kohonen, S. Kaski, K. Lagus, J. Salojrvi, J. Honkela, V. Paatero, A. Saarela, *Self-organization of a massive document collection*, „IEEE Transactions on Neutral Networks" 2000/11, s. 574–585.

[19] C.D. Manning, H. Schütze, *Foundations of Statistical Natural Language Processing*, The MIT Press, Cambridge, Mass., 2001.

[20] C.D. Manning, P. Raghavan, H. Schütze, *Introduction to Information Retrieval*, Cambridge University Press, Cambridge 2008.

Próbą realizacji potrzeby zróżnicowania znaczenia poszczególnych słów w dokumencie może być przeskalowanie wartości macierzy $tf_{t,d}$ przez częstotliwość kolekcji (*collectionfrequency*) cf [21]. Jednakże praktyka badawcza pokazuje, że lepszym rozwiązaniem jest uwzględnienie liczby dokumentów, w których dane słowo występuje – częstotliwość dokumentu df_t. Wartości df_t są tym większe, im słowo t występuje w większej liczbie dokumentów. W formule obliczeniowej stosuje się odwrotną częstotliwość dokumentu idf_t zdefiniowaną jako $1/idf_t$, która jest wysoka dla rzadko występującychsłów, niska zaś dla słów często występujących. W wyniku połączenia opisanych dwóch wag otrzymuje się definicję jednego z najbardziej popularnych schematów ważenia dokumentów w dziedzinie wydobywania informacji TF-IDF[22]. Odpowiednie równanie przyjmuje więc postać:

$$WIS_{t,d}^{C} = tf_{t,d} \cdot idf_t = tf_{t,d} \cdot \log_2(N/df_t) \tag{3}$$

gdzie:

N – łączna liczba dokumentów,

$WIS_{t,d}^{C}$ – wskaźnik istotności t-słowa w d-tym dokumencie oparty na reprezentacjiTF-IDF.

Zastosowanie równania (3) prowadzi do uzyskania wskaźników istotności słowa, które przyjmują:

- wartości maksymalne dla słów występujących często w małej liczbie dokumentów;
- wartości niskie dla słów występujących rzadko w małej liczbie dokumentów, lub występujących w dużej liczbie dokumentów, przez co słowa te mają małą siłę rozróżniającą dokumenty;
- wartości minimalne dla słów pojawiających się w (prawie) wszystkich dokumentach.

3. PRZEPROWADZONE BADANIA

W badaniach empirycznych wykorzystano 759 opinii w formach wyrażających wady, zalety i podsumowanie. Opinie dotyczyły bazy hotelowej w Rzeszowie. Pochodziły z serwisu Booking.com i dotyczyły dwóch hoteli: Grand Hotel Boutique oraz Hotel SchanelResidence. Do każdej opinii dołączona była ocena punktowa w skali 0–10, co w serwisie zostało przełożone na wyliczenie średniego wyniku danego hotelu.Analiza istotnych słów jest możliwa po uwzględnieniu dostępnych możliwości wyboru plusów i minusów dostępnych przy wystawianiu opinii przez klienta.

[21] R. Cummins, C. O'Riordan, *Evolving general term weighting schemes for information retrieval: Tests on larger collections*, "Artif.Intell. Rev." 24/3–4(2005), s. 277–299;C.D.Manning,P. Raghavan, H. Schütze, *Introduction to Information...*

[22] G. Salton, A. Wong,C.S. Yang, *A vector space model for automatic indexing*, „Communications of the ACM"1975/18, s. 613–620.

3.1. Przekształcenie słów do formy podstawowej

Przekształcenie słów do formy podstawowej opiera się na bibliotece – słowniku Morfologik. Jest to słownik do znakowania morfosyntaktycznego i syntezy gramatycznej. Został on opracowany przez Marcina Miłkowskiego przy wykorzystaniu zasobów słownika alternatywnego SJP.pl i udostępniony na takich samych warunkach. W Morfologiku każdej parze słów: forma pochodna – forma bazowa, towarzyszą znaczniki morfosyntaktyczne, które określają relację miedzy słowami. To właśnie obecność tych informacji zdecydowała o wyborze słownika Morfologik.

Przekształcanie wyrazów do ich formy podstawowej zostało wykonane dla całościowej kolekcji opinii konsumenckich. Program został napisany w języku Java. Wykorzystano skrypty Control.java oraz StemPL.javakorzystający z biblioteki morfologik-stemming.

W tabeli 2 zestawiono przykładowe działanie redukcji do rdzenia wybranych słów występujących w zbiorze opinii konsumenckich. Redukcja słów do ich formy podstawowej nie uwzględnia kontekstu użycia danego słowa. Potwierdzeniem tego może być słowo mnie, które zostało sprowadzone do formy podstawowej miąć. Jednak należałoby przeanalizować fragment – zestawienie słów sąsiadujących, aby ocenić, czy nie zostało użyte w odniesieniu na przykładdla mnie, przy którym wynik redukcji do rdzenia jest nieprawidłowy. Należy jednak stwierdzić, że w analizowanym zbiorze opinii konsumenckich uzyskane wyniki redukcji słów do form podstawowych nie wpływają na znaczną utratę ich wartości informacyjnej.

Tabela 2. Przykładowa redukcja do rdzenia słów zaczerpniętych z opinii konsumenckich

Słowo wejściowe --> słowo po redukcji do rdzenia
mnie-->miąć
jestem-->być
zadowolony-->zadowolony
zamówiłem-->zamówić
spełnione-->spełnić
powala-->powalać
polecam-->polecać
przydałaby-->przydać

Źródło: opracowanie własne.

3.2. Metody bazujące na macierzy częstości.

W trakcie badań wyznaczono trzy wartości wskaźników istotności słów: $WIS_{t,d}^{A}$ – równanie (1), $WIS_{t,d}^{B}$ – równanie (2), $WIS_{t,d}^{C}$ – równanie (3). Obliczenia zrealizowano w dwóch wersjach – w pierwszej nie uwzględniono stop-listy. Przyjęto bowiem, że prowadzona analiza ma zweryfikować możliwość automatycznego utworzenia stop-listy. W drugiej wersji obliczeń zastosowano stop-listę.

3.2.1. Badanie bez zastosowania stop-listy

W celu określenia istotności słowa w całym korpusie wyznaczono dla poszczególnych słów sumę wskaźników cząstkowych obliczonych dla poszczególnych dokumentów. Przyjęto, że wyższa wartość wskaźnika świadczy o większym znaczeniu danego wyrazu. W trakcie obliczeń uwzględniono jedynie te wyrazy, które występują przynajmniej

w dwóch dokumentach korpusu. Nie zastosowano stop-listy, gdyż przyjęto założenie, że w kolejnych etapach badań zostaną na niej umieszczone wyrazy wskazane przez omawiany tu algorytm jako nieistotne oraz wyrazy występujące tylko w jednym dokumencie. Obliczenia zrealizowano w pakiecie R.

3.2.2. Badanie z zastosowaniem stop-listy

W celu określenia istotności słowa w całym korpusie wyznaczono dla poszczególnych słów sumę wskaźników cząstkowych obliczonych dla poszczególnych dokumentów. Przyjęto, że wyższa wartość wskaźnika świadczy o większym znaczeniu danego wyrazu. W trakcie obliczeń uwzględniono jedynie te wyrazy, które występują przynajmniej w dwóch dokumentach korpusu. Zastosowano także stop-listę utworzoną dla badanego zbioru danych w języku polskim. Obliczenia zrealizowano w pakiecie R.

W tabeli 3 przedstawiono obliczone współczynniki korelacji liniowej pomiędzy trzema wskaźnikami istotności bez uwzględnienia stop-listy i z jej zastosowaniem.

Tabela 3. Macierz korelacji opinii konsumenckich

Badanie bez zastosowania stop-listy			
	$WIS_{t,d}^{A}$	$WIS_{t,d}^{B}$	$WIS_{t,d}^{C}$
$WIS_{t,d}^{A}$	1,0000000	0,9632928	0,8309164
$WIS_{t,d}^{A}$	0,9632928	1,0000000	0,8827261
$WIS_{t,d}^{C}$	0,8309164	0,8827261	1,0000000
Badanie z zastosowaniem stop-listy			
	$WIS_{t,d}^{A}$	$WIS_{t,d}^{B}$	$WIS_{t,d}^{C}$
$WIS_{t,d}^{A}$	1,000000	0,961723	0,766333
$WIS_{t,d}^{A}$	0,961723	1,000000	0,830387
$WIS_{t,d}^{C}$	0,766333	0,830387	1,000000

Źródło: opracowanie własne.

3.3. Ocena metod bazujących na macierzy częstości

Modelem bazowym dla analizowanej grupy wskaźników istotności słów jest model przestrzeni wektorowej konstruowany na podstawie macierzy częstości. W badaniach wykorzystano dwie wersje macierzy częstości – pierwsza tworzona była bez uwzględnienia stop-listy, w drugiej zastosowanostop-listę.

Wskaźnik $WIS_{t,d}^{A}$ i $WIS_{t,d}^{B}$ może zostać wyznaczony niezależnie dla poszczególnych dokumentów. Wskaźnik $WIS_{t,d}^{C}$ można wyznaczyć jedynie na podstawie całego korpusu (jego obliczenie dla pojedynczego dokumentu wymaga znajomości odwrotnej częstości dokumentowej, która szacowana jest na podstawie korpusu).

Badania pokazały, że podejście bez stosowania stop-listy nie pozwoliło na uzyskanie poprawnych rozwiązań (wiele wyrazów zidentyfikowanych przez metodę jako istotne nie ma dużej wartości informacyjnej).

W wypadku analizy opinii konsumenckich bez uwzględnienia stop-listy słowa o najwyższych wskaźnikach istotności zarówno $WIS_{t,d}^{A}$, $WIS_{t,d}^{B}$,jak i $WIS_{t,d}^{C}$ powinny być uznane za nadmiarowe, ponieważ wprowadzają szum informacyjny. Należą do nich między innymi: być, nie, siebie, ten, ale, jak itd. Zdecydowanie lepsze wyniki uzyskane zostały po analizie opinii konsumenckich z uwzględnieniem stop-listy. Pierwsze pięć słów uzyskanych po wyliczeniu wartości wskaźników istotności dla każdej z analizowanych metod pokrywają się niemal w całości. Należą do nich: śniadanie, pokój, polecać, dobry, parking oraz występujące w metodzie wyliczającej wskaźnik $WIS_{t,d}^{C}$ słowo super (nie występuje zaś słowo polecać). Z powodzeniem można potraktować je jako słowa kluczowe mogące pojawić się w opiniach konsumenckich na temat bazy hotelowej.

4. PODSUMOWANIE

W artykule pokrótce przedstawionobadania dotyczące oceny przydatności metod bazujących na macierzy częstości dla opinii konsumenckich. Uogólniając wyniki badań, można sformułować następujące wnioski w zakresie skuteczności omówionych metod dla zbioru opinii konsumenckich dotyczących bazy hotelowej:

- zdecydowanie lepsze wyniki skuteczności zastosowanych metod osiągnięte zostały po zastosowaniu stop-listy dla opinii konsumenckich;
- wwypadku badania metod bazujących na podstawowej macierzy częstości, jej reprezentacji binarnej i modyfikacji macierzy uwzględniającej TF-IDF uzyskuje się bardzo zbliżone wyniki.

Reasumując dotychczasowe rozważania, należy zauważyć, że w celu określenia skuteczności analizowanych metod algebraicznych opartych na modelu przestrzeni wektorowej należy rozszerzyć badanie na szerszy wachlarz istniejących metod wykorzystywanych do automatycznej analizy opinii konsumenckich. Należy oczekiwać zdecydowanie lepszych wyników, proponując rozwiązania pozwalające na identyfikację słów kluczowych przy wykorzystaniu wiedzy dziedzinowej opisanej w postaci sieci semantycznej lub innej metody reprezentacji wiedzy.

LITERATURA

[1] Cambria E., Schuller B., Yunqing X., Havasi C., *New avenues in opinion mining and sentiment analysis,* „Intelligent Systems, IEEE”2013/28, s. 15–21.

[2] Cummins R., O’Riordan C., *Evolving general term weighting schemes for information retrieval: Tests on larger collections*, „Artif.Intell.Rev.” 24/3–4(2005), s. 277–299.

[3] Hu N., Pavlou P., Zhang J., *Can online reviews reveal a product’s true quality?: empirical findings and analytical modeling of Online word-of-mouth communication*, Proceedings of the 7th ACM conference on Electronic commerce, ACM, 2006, s. 324–330.

[4] Hu N., Zhang J.,Pavlou P.A., *Overcoming the J-shaped distribution of product reviews*, „Commun. ACM” 52/10 (2009), s. 144–147.

[5] Kohonen T., Kaski S., Lagus K., Salojrvi J., Honkela J., Paatero V., Saarela A., *Self-organization of a massive document collection, IEEE Transactions on Neutral Networks*, 2000/11, s. 574–585.

[6] Larose D., *Odkrywanie wiedzy z danych*, Wydawnictwo Naukowe PWN, Warszawa 2006.

[7] Liu B., *Web Data Mining. Exploring Hyperlinks, Contents, and Usage Data*, Springer-Verlag Berlin, Heidelberg 2007.

[8] Liu B., *Opinion Mining and Sentiment Analysis*, [w:] idem,*Web Data Mining, Data-Centric Systems and Applications*, Springer, Berlin–Heidelberg 2011, s. 459–526.

[9] Lula P., *Automatyczna analiza opinii konsumenckich*, [w:]*Taksonomia 18, Klasyfikacja i analiza danych – teoria i zastosowania*, red. K. Jajuga, M. Walesiak, Wydawnictwo UE we Wrocławiu, Wrocław2011.
[10] Lula P., Wójcik K., *Sentiment analysis of consumer opinions written in Polish*, „Economics and Management" 2011, s. 1286–1291.
[11] Manning C.D., Raghavan P., Schütze H., *Introduction to Information Retrieval*, Cambridge University Press, Cambridge 2008.
[12] Manning C.D., Schütze H., *Foundations of Statistical Natural Language Processing*, The MIT Press, Cambridge, Mass., 2001.
[13] Mudambi S., Schu– D., *What makes a helpful online review? A study of customer reviews on Amazon. com*, „MIS Quarterly" 34/1 (2010), s. 185–200.
[14] Pang B., Lee L., *Seeing Stars: Exploiting Class Relationships for Sentiment Categorization with Respect to Rating Scales*, Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, Stroudsburg, PA, 2005,s.115–124.
[15] Pang B., Lee L., *Opinion Mining and Sentiment Analysis*,„Foundations and Trends in Information Retrieval" 2/1–2(2008), s. 1–135.
[16] Rajaraman A., Ullman J.D., *Data Mining. Mining of Massive Datasets*, Cambridge University Press, New York 2012.
[17] Salton G., Wong A., Yang C.S., *A vector space model for automatic indexing*, „Communications of the ACM" 1975/18, s. 613–620.
[18] Zhu F., Zhang X., *Impact of online consumer reviews on sales: The moderating role of product and consumer characteristics*, „Journal of Marketing" 74/2 (2010), s. 133–148.
[19] Zipf G., *Human Behaviour and the Principle of Least Effort*, Cambridge 1949.

METHODS OF USE OF AUTOMATION TEXT ANALYSIS CONSUMER OPINION

The analysis of consumer opinion is an area of research that may mean months impact on the development of business enterprises. It is also a tool that can provide relevant information affecting the company's image, which is important for companies operating in a highly competitive market. Increasing the number of reviews available on the network has created the need for their automatic analysis and processing. This issue is gaining popularity among researchers and among entrepreneurs, for whom consumer reviews are a source of business information. With the ever-growing need for access to customer feedback, and thus the knowledge and information that can derive from them, tools to automate the process of acquiring the key and strategic information they are gaining in importance. This problem requires a slightly different view of the data and the selection of a particular method of analysis using data mining techniques, especially text. The main aim of this work is to analyse automatic classification opinion using exploratory methods of text meaning and methods based on patterns. Used approach will be compared with previously used in the research. Use of information obtained from customer feedback helps to raise awareness of employees at all levels of the organization, provides access to the right information at the right time, which affects the accuracy of business decisions.

Keywords:consumer opinions, automatic analysis of consumer opinion, text mining, document classification, automation of text

DOI: 10.7862/rz.2016.mmr.14

Tekst złożono w redakcji: sierpień 2016
Przyjęto do druku: wrzesień 2016