PRACE NAUKOWE
Uniwersytetu Ekonomicznego we Wrocławiu

RESEARCH PAPERS
of Wrocław University of Economics

Nr 385

Taksonomia 25
Klasyfikacja i analiza danych – teoria i zastosowania

Redaktorzy naukowi
Krzysztof Jajuga
Marek Walesiak

Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu
Wrocław 2015

Redaktor Wydawnictwa: Aleksandra Śliwka
Redaktor techniczny: Barbara Łopusiewicz
Korektor: Barbara Cibis
Łamanie: Beata Mazur
Projekt okładki: Beata Dębska

Tytuł dofinansowany ze środków Narodowego Banku Polskiego
oraz ze środków Sekcji Klasyfikacji i Analizy Danych PTS

Informacje o naborze artykułów i zasadach recenzowania
znajdują się na stronie internetowej Wydawnictwa
www.pracenaukowe.ue.wroc.pl
www.wydawnictwo.ue.wroc.pl

ISSN 1899-3192 (Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu)
e-ISSN 2392-0041 (Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu)
ISSN 1505-9332 (Taksonomia)

Wersja pierwotna: publikacja drukowana

Zamówienia na opublikowane prace należy składać na adres:
Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu
tel./fax 71 36 80 602; e-mail:econbook@ue.wroc.pl
www.ksiegarnia.ue.wroc.pl

Druk i oprawa: TOTEM

Spis treści

Summaries

PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU
RESEARCH PAPERS OF WROCŁAW UNIVERSITY OF ECONOMICS nr 385 • 2015
Taksonomia 25 ISSN 1899-3192
Klasyfikacja i analiza danych – teoria i zastosowania e-ISSN 2392-0041

Katarzyna Wójcik, Janusz Tuchowski
Uniwersytet Ekonomiczny w Krakowie
e-mails: wojcikk@uek.krakow.pl; tuchowsj@uek.krakow.pl

WYKORZYSTANIE METODY OPARTEJ NA WZORCACH W AUTOMATYCZNEJ ANALIZIE OPINII KONSUMENCKICH

Streszczenie: Analiza opinii konsumenckich jest obszarem badań, który może mieć znaczący wpływ na rozwój działalności biznesowej. 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. Głównym celem pracy jest przeprowadzenie analizy automatycznej klasyfikacji opinii z wykorzystaniem metody opartej na wzorcach. Podstawową zaletą tego podejścia jest możliwość identyfikowania całych zwrotów w opiniach. Takim związkom frazeologicznym przypisywane jest nacechowanie agregowane następnie do sentymentu całej opinii. Pozwala to na identyfikację w opiniach charakterystycznych struktur, których konstrukcja determinuje interpretację ich nacechowania. Dotyczy to zarówno polaryzacji, jak i siły nacechowania. Wykorzystane podejście zostanie porównane z podejściami dotychczas wykorzystywanymi w badaniach.

Słowa kluczowe: *text-mining*, Web-mining, taksonomia, klasyfikacja dokumentów tekstowych, opinion mining, sentiment analysis, wzorce, Spejd.

DOI: 10.15611/pn.2015.385.34

1. Wstęp

Analiza opinii konsumenckich jest obszarem badań, który może mieć znaczący wpływ na rozwój działalności biznesowej. 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.

Najczęściej wykorzystywanym rodzajem automatycznej analizy opinii konsumentów jest klasyfikacja opinii. Każdej opinii przypisywane jest nacechowanie[1].

[1] Polaryzacja, sentyment; określenie, czy opinia jest pozytywna czy negatywna i ewentualnie w jakim stopniu (bardzo pozytywna/negatywna czy słabo pozytywna/negatywna).

Wykorzystywanych jest tu wiele podejść. Podstawowe z nich opiera się na słowach. 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- lub 4-stopniowa skala punktów). Jednak zarówno studia literaturowe [Pang, Lee 2005], jak i własne badania wykazały, że obecnie wykorzystywane narzędzia nie dają dobrych rezultatów przy klasyfikacji na więcej niż dwie grupy.

Głównym celem pracy jest przeprowadzenie analizy własności automatycznej klasyfikacji opinii napisanych w języku polskim z wykorzystaniem metody opartej na wzorcach. Podstawową różnicą, a zarazem zaletą tego podejścia jest możliwość identyfikowania całych zwrotów w opiniach. Dopiero takim związkom frazeologicznym przypisywane jest nacechowanie agregowane następnie do sentymentu całej opinii. Pozwala to na identyfikację w opiniach charakterystycznych struktur, których konstrukcja determinuje interpretację ich nacechowania. Dotyczy to zarówno polaryzacji, jak i siły nacechowania. Wykorzystane podejście zostanie porównane z podejściami dotychczas wykorzystywanymi w badaniach.

2. Automatyczna analiza opinii konsumenckich

Automatyczna analiza opinii konsumenckich (*sentiment analysis, opinion mining*) to ogół działań mających na celu zautomatyzowanie procesu wyszukiwania, ekstrakcji i analizy danych pochodzących ze specyficznych tekstów, jakimi są opinie użytkowników. Są to działania z pogranicza przetwarzania języka naturalnego (*Natural Language Processing* – NLP), lingwistyki komputerowej (*computational linguistics*) oraz eksploracyjnej analizy tekstu (*text mining*). Jej celem jest określenie nastawienia autora wypowiedzi do jej przedmiotu .

2.1. Opinie

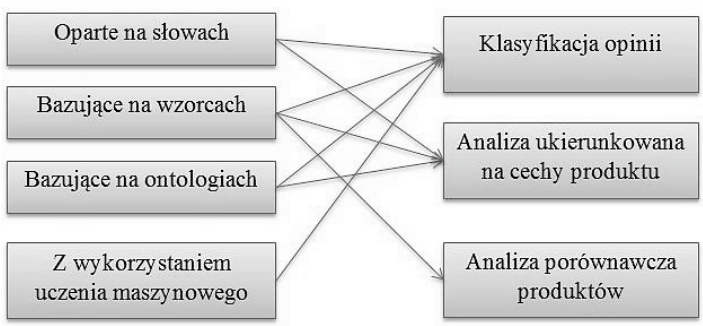
Opinie to specyficzny rodzaj danych tekstowych, które mają subiektywny charakter – wyrażają stosunek autora wypowiedzi do przedmiotu opinii. Opinie, oprócz subiektywnych odczuć autora, mogą zawierać obiektywne stwierdzenia, które mogą mieć wpływ na wynik analizy. Dlatego część badaczy rozpoczyna analizę od określenia, czy badany tekst jest obiektywny czy subiektywny [Pang, Lee 2008]. W niektórych serwisach opinie słowne są wspierane oceną punktową lub gwiazdkami. Opinie można podzielić na grupy według ich formatu [Liu 2007]:

1) zalety i wady oraz podsumowanie,
2) zalety i wady,
3) dowolny.

2.2. Podejścia do automatycznej analizy opinii konsumentów

W ramach automatycznej analizy opinii konsumenckich wyróżnić można trzy rodzaje działań, takie jak [Liu 2007]: klasyfikacja opinii, analiza ukierunkowana na cechy produktu oraz analiza porównawcza produktów. W niektórych pracach analiza ukierunkowana na cechy produktu opisywana jest jako głębszy poziom analizy nacechowania opinii [Liu 2010] – dla produktu czy usługi identyfikowane są cechy/atrybuty i polaryzacja opinii wyznaczana jest na podstawie sentymentu przypisywanego poszczególnym częściom/właściwościom przedmiotu opinii. Szerszy opis każdego z rodzajów analizy można znaleźć w [Wójcik, Tuchowski 2013] oraz [Wójcik, Tuchowski 2014].

Koncentrując się na klasyfikacji opinii, można zauważyć, że wszystkie cztery text miningowe podejścia do automatycznej analizy opinii konsumentów znajdują w niej zastosowanie [Lula, Wójcik 2011]: podejście oparte na słowach (*word-based approach*), podejście bazujące na wzorcach (*pattern-based approach*), podejście bazujące na ontologiach (*ontology-based approach*) oraz podejście, u podstaw którego stoi uczenie maszynowe (*machine learning approach*). W pracy [Cambria i in., 2013] znaleźć można podobną klastfikację podejść do automatycznej analizy opinii konsumentów.



Rys. 1. Wykorzystanie różnych podejść text miningowych do poszczególnych działań w ramach automatycznej analizy opinii konsumenckich

Źródło: opracowanie własne.

Rysunek 1 przedstawia różne text miningowe podejścia do automatycznej analizy opinii konsumenckich przyporządkowane do rodzajów automatycznej analizy opinii konsumenckich, w których mogą zostać wykorzystane.

2.3. Podejście oparte na wzorcach

W podejściu opartym na wzorcach wykorzystywane są reguły bazujące na wyrażeniach regularnych. Podejście to pozwala na identyfikację fraz modyfikujących sentyment, takich jak [Buczyński, Wawer 2008] negacja (*negation*), neutralizacja (*nullification*) czy zastrzeżenie/ograniczenie (*limitation*). Ponadto często w opiniach występują wzmocnienia oraz inne charakterystyczne zwroty.

Podejście oparte na wzorcach jest wstępem do analizy ukierunkowanej na cechy produktu oraz może być wykorzystane do analizy porównawczej produktów.

3. Materiały i metody badań

Celem badania była analiza własności podejścia opartego na wzorcach w kontekście automatycznej analizy opinii konsumentów. Niniejszy rozdział opisuje przykład wykorzystania metody opartej na wzorcach do automatycznej klasyfikacji opinii konsumentów, pozwalający na zilustrowanie wyników badań empirycznych. Wyniki uzyskane dla omawianej metody zostaną porównane z wynikami uzyskanymi przy wykorzystaniu innych metod.

Automatyczna analiza opinii konsumentów przy wykorzystaniu metody opartej na wzorcach polega na zidentyfikowaniu w opiniach charakterystycznych fraz modyfikujących nacechowanie słów je budujących. Na podstawie słowników wyrazów pozytywnych i negatywnych słowom w opinii przypisywane jest nacechowanie. Następnie dzięki regułom (wzorcom) jest ono modyfikowane. Na końcu sentymenty te są agregowane (przy użyciu funkcji agregującej np. średnie) do jednej wartości reprezentującej nacechowanie całej opinii.

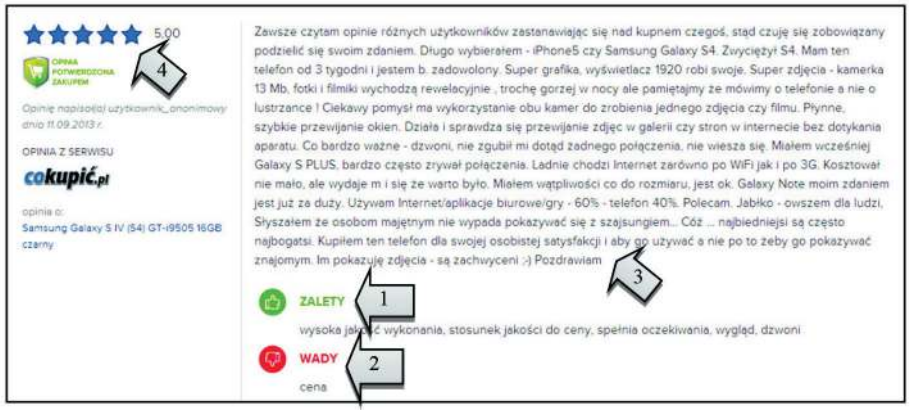
Badania empiryczne podzielone zostały na sześć etapów:

1. Pozyskanie opinii.
2. Analiza podobieństwa opinii.
3. Konstrukcja słowników.
4. Konstrukcja reguł.
5. Analiza nacechowania opinii.
6. Porównanie wyników różnych eksperymentów.

3.1. Zbiór opinii

W badaniach empirycznych wykorzystano 737 opinii w formie pierwszej (wady, zalety, posumowanie). Opinie pochodziły z serwisu Ceneo.pl[2] i dotyczyły smartfonów Samsung Galaxy S II, S III, S4 oraz S5. Do każdej opinii dołączona była ocena punktowa w postaci gwiazdek w przedziale [0,5;5] z krokiem 0,5.

[2] Dostęp dnia 03.09.2014 r.



Rys. 2. Przykładowa opinia wykorzystana w badaniach empirycznych

Źródło: opracowanie własne.

Opinie z serwisu internetowego zostały pobrane do bazy. Następnie z bazy danych zostały one wyeksportowane do plików tekstowych. Każda opinia została zapisana w osobnym pliku tekstowym.

3.2. Spejd

Spejd[3] (*Shallow Parsing and Eminently Judicious Disambiguation*, pl. *Składniowy Parser (Ewidentnie Jednocześnie Dezambiguator)*) to parser powierzchniowy służący do płytkiej analizy składniowej (identyfikacja wewnętrznych struktur w zdaniu bez analizy struktury całego zdania[4]) dostępny na warunkach GNU GPL (powszechnej licencji publicznej). Pozwala on na identyfikowanie konstrukcji składniowych i ujednoznacznianie wyników analizy morfologicznej[5], co oznacza, że nie wymaga tekstu wstępnie przetworzonego (po redukcji do rdzenia).

Spejd został opracowany przez Instytut Podstaw Informatyki Polskiej Akademii Nauk (IPI PAN) i korzysta z Narodowego Korpusu Języka Polskiego (NKJP). Podstawą formalizmu Spejda jest kaskada gramatyk regularnych w postaci reguł. Wbudowane reguły pozwalają na identyfikację zdań, tokenów, analizę morfologiczną oraz tagowanie. Dodatkowy zestaw reguł odpowiada za identyfikację skrótów, dat oraz liczb pisanych w różny sposób.

[3] Początkowo aplikacja miała mieć nazwę Spade (*Shallow Parsing and Dezambiguation Engine*), ale istniał już wówczas parser o tej nazwie.

[4] Określenie, jaką częścią mowy są poszczególne wyrazy występujące w zdaniu, bez identyfikacji tego, jaką rolę w zdaniu odgrywają.

[5] Jednoznaczne określenie formy podstawowej wyrazu często spośród wielu możliwych.

Każda reguła ma następującą strukturę:
Rule „Nazwa reguły”
Left: lewy kontekst
Match: dopasowanie
Right: prawy kontekst
Eval: operacje do wykonania

Tabela 1 przedstawia listę leksemów, operatorów i operacji użytych w badaniach przy definiowaniu reguł.

Tabela 1. Oznaczenia wykorzystywane w Narodowym Korpusie Języka Polskiego

leksem	**notacja**
rzeczownik	subst
przymiotnik	adj
przysłówek	adv
czasownik forma nieprzeszła	fin
partykuło-przysłówek	qub
operatory	
operator	**znaczenie**
[…]	pojedynczy token
atrybut~wartość	istnieje atrybut o podanej wartości
operacje	
operacja	**działanie**
unify	pozostawia tokeny o zgodnych wartościach określonych atrybutów
alter	modyfikuje część tokenu lub wartość jego atrybuty
group	grupuje tokeny

Źródło: opracowanie własne.

3.3. Słowniki

W podejściu opartym na wzorcach do określania sentymentu pojedynczych słów wykorzystywane są słowniki nacechowania. W celu przygotowania słowników wykorzystano język R. Najpierw w aplikacji Spejd sprowadzono wszystkie wyrazy z opinii do rdzenia, a następnie na podstawie tak przygotowanych plików utworzono w języku R macierz częstości. Na etapie wstępnego przetwarzania dokumentów usunięte zostały znaki interpunkcyjne, białe znaki oraz wyrazy znajdujące się na stopliście. Ponadto zamieniono wszystkie litery na małe.

Słowa z macierzy częstości posłużyły do konstrukcji słowników. Wykorzystano jedynie słowa w wersji podstawowej. Utworzono osobno słowniki wyrazów pozytywnych i negatywnych. Każdy z nich liczy około 200 słów. Słowniki przygotowano w dwóch wersjach:

- sentyment o wartości 1 dla słów pozytywnych i –1 dla negatywnych,
- sentyment dodatni dla słów pozytywnych i ujemny dla negatywnych, wartość zależy od siły nacechowania, wartości całkowite od –10 do 10 bez 0.

W słownikach pominięto problematyczne słowa, takie jak: wysoki/niski, szybko/wolno, długo/krótko. Słowa te w zależności od kontekstu będą miały przeciwne nacechowanie.

W sieci znaleźć można przykładowe słowniki wyrazów pozytywnych i negatywnych dla języka angielskiego. Część z nich pozwala jedynie na określenie, czy słowo jest pozytywne czy negatywne (ewentualnie neutralne, jeśli nie występuje w żadnym ze słowników). Są jednak takie, które pozwalają na określenie stopnia nacechowania, a nawet stopnia obiektywizmu danego słowa, jak np. SentiWordNet [Esuli i Sebastiani, 2006]. W pracy [Thelwall i in. 2010] znaleźć można wyniki badań potwierdzające poprawę jakości klasyfikacji przy użyciu słowników pozwalających na określenie siły nacechowania poszczególnych słów.

3.4. Reguły

W badaniach zastosowano dwa rodzaje reguł:
- modyfikujące sentyment pojedynczych słów,
- grupujące słowa w nacechowane frazy.

Do przechowywania wartości nacechowania pobranych ze słowników wykorzystano dodatkowy atrybut sen zdefiniowany w programie Spejd. Pierwsza grupa reguł modyfikowała wartość tego atrybutu.

Rysunki 3 i 4 przedstawiają przykładowe reguły zdefiniowane w programie Spejd. Pierwsza z nich służy do zmiany polaryzacji sentymentu z pozytywnego na negatywny. Wartość nacechowania się nie zmienia. Druga powoduje wzmocnienie sentymentu przez pomnożenie jego wartości przez współczynnik 1,2. Przy regule wzmacniającej konieczne było zdefiniowanie słów, które wzmacniają sentyment innych słów (zarówno pozytywny, jak i negatywny).

```
Rule    "Negative -"
Match:  [orth~nie/i] [sen>0];
Eval:   alter(2, sen=sen*-1,);
```

Rys. 3. Przykładowa reguła w programie Spejd służąca do negacji sentymentu pozytywnego

Źródło: opracowanie własne.

```
Define wzmocnienie =
[base~bardzo|szczerze|gorąco|naprawdę|zdecydowanie];
Rule    "Wzmocnienie +"
Match:  $wzmocnienie [sen>0];
Eval:   alter(2, sen=sen*1.2,);
```

Rys. 4. Przykładowa reguła w programie Spejd służąca do wzmocnienia sentymentu negatywnego

Źródło: opracowanie własne.

Druga grupa reguł identyfikowała w opiniach związki wyrazowe i obliczała nacechowanie całej grupy wyrazów na podstawie nacechowania słów ją stanowiących. Wykorzystywano sentyment zmodyfikowany przez pierwszą grupę reguł. Teksty analizowane były zdanie po zdaniu przez każdą regułę. Rysunek 5 przedstawia jedną z reguł grupujących. Pozwala ona na połączenie czasownika z opisującymi go przymiotnikiem i przysłówkiem.

```
Rule "ADV ADJ SUBST"
Match:  [pos~adv][pos~adj][pos~subst];
Eval:   unify(gender,2,3);
        unify(degree,1);
        group(PAS,1,3, 0.orth);
```

Rys. 5. Przykładowa reguła w programie Spejd grupująca wskazane części mowy w związek wyrazowy

Źródło: opracowanie własne.

Aplikacja Spejd na wejściu przyjmuje między innymi pliki tekstowe. Takie zostały wykorzystane do badań. Wyniki analizy zapisywane są w plikach XML. Dla każdego pliku tekstowego na wejściu powstaje jeden plik XML na wyjściu. Ich nazwy korespondują ze sobą.

W dalszej części badań wartości sentymentu przypisane do słów bądź fraz z plików XML zostały pobrane do autorskiej aplikacji w języku Java i na ich podstawie policzone zostało nacechowanie każdej z opinii.

Rysunek 6 przedstawia przykładową opinię wykorzystaną w badaniach. Z kolei na rys. 7 przedstawiony został fragment pliku XML będącego wynikiem analizy składniowej przykładowej opinii. Widać na nim zbitek słów *Bardzo dobry*. Sentyment słowa *dobry* został zmodyfikowany przez słowo *bardzo* z 5 na 6. Zauważyć można również, że zidentyfikowane zostały słowa pasujące do reguły przedstawionej na rys. 5 – *Bardzo dobry ekran*. Połączone zostały one w grupę, której nacechowanie określone jest przez słowo *dobry* – 6.

```
Super sprzęt. Bateria trzyma naprawdę długo. System ładuje
się szybko. Mam wystarczającą ilość miejsca na wszystkie moje
filmy i zdjęcia. Jedynym mankamentem jest słaba jakość
dźwięku. Głośniki są nie najlepszej jakości. Bardzo dobry
ekran. Ponadto sprzęt nie przegrzewa się co miało miejsce w
przypadku poprzedniego modelu który miałam. Obudowa też jest
zdecydowanie lepsza. Sprawia wrażenie zdecydowanie trwalszej.
Nie mam też żadnych zastrzeżeń co do klawiatury, ale marna
jakość touch pada daje się momentami we znaki. Szczerze
polecam.
```

Rys. 6. Przykładowa opinia wykorzystana w badaniach

Źródło: opracowanie własne.

```
<group base="Bardzo dobry ekran" id="a39" rule="ADV ADJ SUBST"
type="PAS" synh="a34" semh="a36">
<tok id="a34" string-range="string-range(p-1,221,6)">
<orth>Bardzo</orth>
<lex><base>bardzo</base><ctag>adv:pos</ctag></lex>
</tok>
<tok id="a35" string-range="string-range(p-1,228,5)">
<orth>dobry</orth>
<lex><base>dobry</base><ctag>subst:sg:nom:m3:sen=6</ctag></lex>
<lex><base>dobry</base><ctag>subst:sg:acc:m3:sen=6</ctag></lex>
<lex><base>dobry</base><ctag>subst:sg:voc:m3:sen=6</ctag></lex>
<lex><base>dobry</base><ctag>adj:sg:nom:m3:pos:sen=6</ctag></lex>
<lex><base>dobry</base><ctag>adj:sg:acc:m3:pos:sen=6</ctag></lex>
<lex><base>dobry</base><ctag>adj:sg:voc:m3:pos:sen=6</ctag></lex>
</tok>
<tok id="a36" string-range="string-range(p-1,234,5)">
<orth>ekran</orth>
<lex><base>ekran</base><ctag>subst:sg:nom:m3</ctag></lex>
<lex><base>ekran</base><ctag>subst:sg:acc:m3</ctag></lex>
</tok>
</group>
```

Rys. 7. Fragment pliku XML opisującego przykładową opinię

Źródło: opracowanie własne.

4. Wyniki badań empirycznych

Tabela 2. Opisy i wyniki eksperymentów przeprowadzonych w ramach badań empirycznych

Lp.	Oznaczenie	Stemming	Unifikacja sentymentu	Reguły	Słownik	Korelacja	Istotność (*p-value*)	Dokładność (2 grupy)	Dokładność (3 grupy)
1	**S10**	✓			$N_{[-10;10]\setminus 0}$	0,358405	0,0000	80,86%	76,34%
2	**SU10**	✓	✓		$N_{[-10;10]\setminus 0}$	0,388819	0,0000	90,91%	76,93%
3	**SU1**	✓	✓		{-1;1}	0,386515	0,0000	90,77%	77,48%
4	**SUR10**	✓	✓	✓	$N_{[-10;10]\setminus 0}$	0,423316	0,0000	91,59%	78,02%

Źródło: opracowanie własne.

Tabela 3. Dokładność klasyfikacji w podziale na dwie grupy w eksperymencie SUR10

		Użytkownik	
		Negatywna	Pozytywna
Obliczenia	Negatywna	37,74%	4,24%
	Pozytywna	62,26%	95,76%

Źródło: opracowanie własne.

W ramach badań przeprowadzono cztery eksperymenty. Tabela 2 przedstawia opisy symboliczne poszczególnych eksperymentów oraz wyniki uzyskane w każdym z nich. Symbole użyte w oznaczeniach to kolejno S – *stemming* (redukcja do rdzenia), U – unifikacja sentymentu, R – wykorzystanie reguł (wzorców), a 1 lub 10 symbolizują słownik, którego użyto w danym eksperymencie. Z kolei tabela 3 przedstawia, jaki procent opinii został poprawnie, a jaki błędnie zaklasyfikowany z podziałem na opinie pozytywne i negatywne. Wyniki w tab. 3 dotyczą eksperymentu SUR10, w którym jako jedynym użyte zostały wzorce i który dał najlepsze wyniki.

5. Zakończenie

Na podstawie przeprowadzonych badań można stwierdzić, że zastosowanie wzorców zwiększa korelację pomiędzy wynikami oczekiwanymi a uzyskanymi, zwiększając jednocześnie nieznacznie dokładność klasyfikacji. Zastosowana metoda lepiej klasyfikuje opinie pozytywne niż negatywne czy też neutralne. Uwzględnienie siły nacechowanie poprawia jakość klasyfikacji.

Jednakże metoda oparta na wzorcach wymaga dużego nakładu pracy związanego z konstrukcją reguł. Ponadto duży problem stanowią słowa mające różne nacechowanie w zależności od kontekstu.

W kolejnych badaniach planowana jest rozbudowa zbioru reguł, poszerzenie słowników sentymentu, wykorzystanie wzorców w analizie ukierunkowanej na cechy produktów oraz próba połączenia podejścia opartego na wzorcach z podejściem wykorzystującym wiedzę dziedzinową w postaci ontologii.

Literatura

Buczyński A., Przepiórkowski A. (2008), *Demo: An Open Source Tool for Partial Parsing and Morphosyntactic Disambiguation*, *Proceedings of LREC 2008.*

Buczyński A., Wawer A. (2008), *Automated classification of product review sentiments in Polish*, Intelligent Information Systems, s. 213-217.

Cambria E., Schuller B., Yunqing X., Havasi C. (2013, marzec-kwiecień), *New avenues in opinion mining and sentiment analysis*, Intelligent Systems, IEEE, *28*, s. 15-21.

Esuli A., Sebastiani F. (2006), *SENTIWORDNET: A Publicly Available Lexical Resource*, *In Proceedings of the 5th Conference on Language Resources and Evaluation (LREC'06)*, (s. 417-422).

Liu B. (2010), *Sentiment Analysis and Subjectivity*, [w:] N. Indurkhya i F. Damerau (red.), *Handbook of Natural Language Processing, Chapman & Hall/CRC Machine Learning & Pattern Recognition Series* (wyd. drugie, tom 2, s. 627-666). Chapman & Hall/CRC.

Liu B. (2007), *Web DataMining. Exploring Hyperlinks, Contents, and Usage Data.* Heidelberg: Springer-Verlag Berlin.

Lula P., Wójcik K. (2011), *Sentiment analysis of consumer opinions writen in Polish*, Economics and Management (16), s. 1286-1291.

Pang B., Lee L. (2005), *Seeing Stars: Exploiting Class Relationships for Sentiment Categorization with Respect to Rating Scales*, [w:] *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics* (s.115-124), Stroudsburg, PA, USA: Association for Computational Linguistics. doi:10.3115/1219840.1219855.

Pang B., Lee L. (2008), *Opinion Mining and Sentiment Analysis. Foundations and Trends in Information Retrieval*, s. 1-135.

Thelwall M., Buckley K., Paltoglou G., Cai D., Kappas A. (2010, grudzień), *Sentiment in short strength detection informal text*, Journal of the American Society for Information Science.

Wójcik K., Tuchowski J. (2013), *Wpływ automatycznego tłumaczenia na wyniki automatycznej identyfikacji charakteru opinii konsumenckich*, [w:] Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu nr 279*, Taksonomia 21, Klasyfikacja i anaiza danych – teoria i zastosowania*, K. Jajuga i M. Walesiak (red.), s. 124-134.

Wójcik K., Tuchowski J. (2014), *Dobór optymalnego zestawu słów istotnych w opiniach konsumentów na potrzeby ich automatycznej analizy*, Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu nr 328, Taksonomia 23*, Klasyfikacja i analiza danych – teoria i zastosowania*, K. Jajuga i M. Walesiak (red.), s. 106-115.

USING PATTERN-BASED OPINION MINING

Summary: Sentiment analysis or opinion mining is a field of research that can have a significant impact on today's business. The increasing number of consumers' reviews created the need of its automatic analysis. This issue is gaining popularity for both – researchers and entrepreneurs, for whom consumers' reviews are an important source of business information. The main aim of this paper is to examine pattern-based classification of opinions. Pattern-based approach allows identifying certain phrases in opinions to which sentiments can be assigned. An advantage of this approach is a possibility to detect phrases that modify sentiment like negation, nullification, strengthening and others. The approach used in the research is compared with other approaches to opinions classification.

Keywords: text-mining, Web-mining, taxonomy, classification of text documents, opinion mining, sentiment analysis, patterns, Spejd.