

PRACE NAUKOWE

Uniwersytetu Ekonomicznego we Wrocławiu

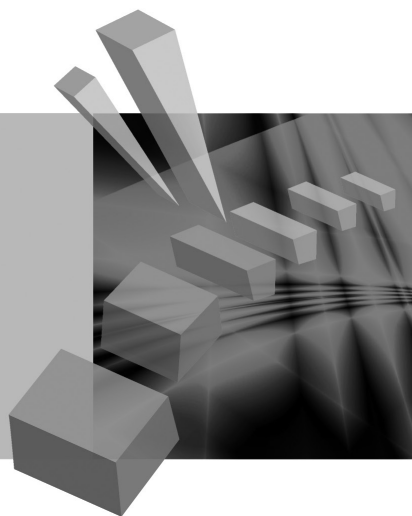
RESEARCH PAPERS

of Wrocław University of Economics

279

Taksonomia 21

Klasyfikacja i analiza danych – teoria i zastosowania



Redaktorzy naukowi

Krzysztof Jajuga

Marek Walesiak



Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu
Wrocław 2013

Redaktor Wydawnictwa: Aleksandra Śliwka

Redaktor techniczny: Barbara Łopusiewicz

Korektor: Barbara Cibis

Łamanie: Małgorzata Czupryńska

Projekt okładki: Beata Dębska

Publikacja jest dostępna w Internecie na stronach:

www.ibuk.pl, www.ebscohost.com,

The Central and Eastern European Online Library www.ceeol.com,

a także w adnotowanej bibliografii zagadnień ekonomicznych BazEkon

http://kangur.uek.krakow.pl/bazy_ae/bazekon/nowy/index.php

Informacje o naborze artykułów i zasadach recenzowania znajdują się

na stronie internetowej Wydawnictwa

www.wydawnictwo.ue.wroc.pl

Tytuł dofinansowany ze środków Narodowego Banku Polskiego

oraz ze środków Sekcji Klasyfikacji i Analizy danych PTS

Kopiowanie i powielanie w jakiegokolwiek formie

wymaga pisemnej zgody Wydawcy

© Copyright by Uniwersytet Ekonomiczny we Wrocławiu

Wrocław 2013

ISSN 1899-3192 (Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu)

ISSN 1505-9332 (Taksonomia)

Wersja pierwotna: publikacja drukowana

Druk: Drukarnia TOTEM

Spis treści

Wstęp	9
Sabina Denkowska, Kamil Fijorek, Marcin Salamaga, Andrzej Sokolowski: Sejm VI kadencji – maszynka do głosowania	11
Barbara Pawelek, Adam Sagan: Zmienne ukryte w modelach ekonomicznych – respecyfikacja modelu Kleina I	19
Jan Paradysz: Nowe możliwości badania koniunktury na rynku pracy	29
Krzysztof Najman: Samouczące się sieci GNG w grupowaniu dynamicznym zbiorów o wysokim wymiarze	41
Kamila Migdał-Najman: Zastosowanie jednowymiarowej sieci SOM do wyboru cech zmiennych w grupowaniu dynamicznym	48
Aleksandra Matuszewska-Janica, Dorota Witkowska: Zróżnicowanie płac ze względu na płeć: zastosowanie drzew klasyfikacyjnych	58
Iwona Foryś, Ewa Putek-Szeląg: Przestrzenna klasyfikacja gmin ze względu na sprzedaż użytków gruntowych zbywanych przez ANR w województwie zachodniopomorskim	67
Joanna Banaś, Małgorzata Machowska-Szewczyk: Klasyfikacja internetowych rachunków bankowych z uwzględnieniem zmiennych symbolicznych.....	77
Marta Jaročka: Wpływ metody doboru cech diagnostycznych na wynik porządkowania liniowego na przykładzie rankingu polskich uczelni	85
Anna Zamojska: Badanie zgodności rankingów wyznaczonych według różnych wskaźników efektywności zarządzania portfelem na przykładzie funduszy inwestycyjnych.....	95
Dorota Rozmus: Porównanie dokładności taksonomicznej metody propagacji podobieństwa oraz zagregowanych algorytmów taksonomicznych opartych na idei metody <i>bagging</i>	106
Ewa Wędrowska: Wrażliwość miar dywergencji jako mierników niepodobieństwa struktur.....	115
Katarzyna Wójcik, Janusz Tuchowski: Wpływ automatycznego tłumaczenia na wyniki automatycznej identyfikacji charakteru opinii konsumenckich ...	124
Małgorzata Misztal: Ocena wpływu wybranych metod imputacji na wyniki klasyfikacji obiektów w modelach drzew klasyfikacyjnych.....	135
Anna Czapkiewicz, Beata Basiura: Badanie wpływu wyboru współczynnika zależności na grupowanie szeregów czasowych	146
Tomasz Szubert: Czynniki różnicujące poziom zadowolenia z życia oraz wartości życiowe osób sprawnych i niepełnosprawnych w świetle badań „Diagnozy społecznej”	154

Marcin Szymkowiak: Konstrukcja estymatorów kalibracyjnych wartości globalnej dla różnych funkcji odległości	164
Wojciech Roszka: Szacowanie łącznych charakterystyk cech nieobserwowanych łącznie	174
Justyna Brzezińska: Metody wizualizacji danych jakościowych w programie R	182
Agata Sielska: Regionalne zróżnicowanie potencjału konkurencyjnego polskich gospodarstw rolnych w województwach po akcesji do Unii Europejskiej	191
Mariusz Kubus: Liniowy model prawdopodobieństwa z regularyzacją jako metoda doboru zmiennych	201
Beata Basiura: Metoda Warda w zastosowaniu klasyfikacji województw Polski z różnymi miarami odległości	209
Katarzyna Wardzińska: Wykorzystanie metody obwiedni danych w procesie klasyfikacji przedsiębiorstw	217
Katarzyna Dębowska: Modelowanie upadłości przedsiębiorstw oparte na próbach niezbilansowanych	226
Danuta Tarka: Wpływ metody doboru cech diagnostycznych na wyniki klasyfikacji obiektów na przykładzie danych dotyczących ochrony środowiska ..	235
Artur Czech: Zastosowanie wybranych metod doboru zmiennych diagnostycznych w badaniach konsumpcji w ujęciu pośrednim	246
Beata Bal-Domańska: Ocena relacji zachodzących między inteligentnym rozwojem a spójnością ekonomiczną w wymiarze regionalnym z wykorzystaniem modeli panelowych	255
Mariola Chrzanowska: <i>Ordinary kriging</i> i <i>inverse distance weighting</i> jako metody szacowania cen nieruchomości na przykładzie warszawskiego rynku	264
Adam Depta: Zastosowanie analizy wariancji w badaniu jakości życia na podstawie kwestionariusza SF-36v2	272
Maciej Beręsewicz, Tomasz Klimanek: Wykorzystanie estymacji pośredniej uwzględniającej korelację przestrzenną w badaniach cen mieszkań	281
Karolina Paradysz: Benchmarkowa analiza estymacji dla małych obszarów na lokalnych rynkach pracy	291
Anna Gryko-Nikitin: Dobór parametrów w równoległych algorytmach genetycznych dla problemu plecakowego	301
Tomasz Ząbkowski, Piotr Jałowicki: Zastosowanie reguł asocjacyjnych do analizy danych ankietowych w wybranych obszarach logistyki przedsiębiorstw przetwórstwa rolno-spożywczego	311
Agnieszka Przedborska, Małgorzata Misztal: Zastosowanie metod statystyki wielowymiarowej do oceny wydolności stawów kolanowych u pacjentów z chorobą zwyrodnieniową leczonych operacyjnie	321
Dorota Perło: Rozwój zrównoważony w wymiarze gospodarczym, społecznym i środowiskowym – analiza przestrzenna	331

Ewa Putek-Szeląg, Urszula Gieraltowska, Analiza i diagnoza wielkości produkcji energii odnawialnej w Polsce na tle krajów Unii Europejskiej..	342
--	-----

Summaries

Sabina Denkowska, Kamil Fijorek, Marcin Salamaga, Andrzej Sokolowski: VIth-term Sejm – a voting machine	18
Barbara Pawelek, Adam Sagan: Latent variables in econometric models – respecification of Klein I model	28
Jan Paradysz: New possibilities for studying the situation on the labour market	40
Krzysztof Najman: Self-learning neural network of GNG type in the dynamic clustering of high-dimensional data.....	47
Kamila Migdał-Najman: Applying the one-dimensional SOM network to select variables in dynamic clustering	57
Aleksandra Matuszewska-Janica, Dorota Witkowska: Gender wage gap: application of classification trees.....	66
Iwona Foryś, Ewa Putek-Szeląg: Spatial classification of communes by usable land traded by the APA in the Zachodniopomorskie voivodeship...	76
Joanna Banaś, Małgorzata Machowska-Szewczyk: Classification of Internet banking accounts including symbolic variables	84
Marta Jarocka: The impact of the method of the selection of diagnostic variables on the result of linear ordering on the example of ranking of universities in Poland.....	94
Anna Zamojska: Empirical analysis of the consistency of mutual fund ranking for different portfolio performance measures.....	105
Dorota Rozmus: Comparison of accuracy of affinity propagation clustering and cluster ensembles based on bagging idea.....	114
Ewa Wędrowska: Sensitivity of divergence measures as structure dissimilarity measurements	123
Katarzyna Wójcik, Janusz Tuchowski: Machine translation impact on the results of the sentiment analysis	134
Małgorzata Misztal: Assessment of the influence of selected imputation methods on the results of object classification using classification trees ...	145
Anna Czapkiewicz, Beata Basiura: Simulation study of the selection of coefficient depending on the clustering time series.....	153
Tomasz Szubert: Factors differentiating the level of satisfaction with life and the life's values of people with and without disabilities in the light of the "Social Diagnosis" survey	162
Marcin Szymkowiak: Construction of calibration estimators of totals for different distance measures	173

Wojciech Roszka: Joint characteristics' estimation of variables not jointly observed.....	181
Justyna Brzezińska: Visualizing categorical data in \mathbf{R}	190
Agata Sielska: Regional diversity of competitiveness potential of Polish farms after the accession to the European Union	200
Mariusz Kubus: Regularized linear probability model as a filter	208
Beata Basiura: The Ward method in the application for classification of Polish voivodeships with different distances.....	216
Katarzyna Wardzińska: Application of Data Envelopment Analysis in company classification process.....	225
Katarzyna Dębowska: Modeling corporate bankruptcy based on unbalanced samples	234
Danuta Tarka: Influence of the features selection method on the results of objects classification using environmental data.....	245
Artur Czech: Application of chosen methods for the selection of diagnostic variables in indirect consumption research.....	254
Beata Bal-Domańska: Assessment of relations occurring between smart growth and economic cohesion in regional dimension using panel models	263
Mariola Chrzanowska: Ordinary kriging and inverse distance weighting as methods of estimating prices based on Warsaw real estate market	271
Adam Depta: Application of analysis of variance in the study of the quality of life based on questionnaire SF-36v2	280
Maciej Beręsewicz, Tomasz Klimanek: Using indirect estimation with spatial autocorrelation in dwelling price surveys.....	290
Karolina Paradysz: Benchmark analysis of small area estimation on local labor markets	300
Anna Gryko-Nikitin: Selection of various parameters of parallel evolutionary algorithm for knapsack problems	310
Tomasz Ząbkowski, Piotr Jałowiecki: Application of association rules for the survey of data analysis in the selected areas of logistics in food processing companies	320
Agnieszka Przedborska, Małgorzata Misztal: Using multivariate statistical methods to assess the capacity of the knee joint among the patients treated surgically for osteoarthritis	330
Dorota Perło: Sustainable development in the economic, social and environmental dimensions – spatial analysis.....	341
Ewa Putek-Szeląg, Urszula Gieraltowska: Analysis and diagnosis of the volume of renewable energy production in Poland compared to EU countries	352

Katarzyna Wójcik, Janusz Tuchowski

Uniwersytet Ekonomiczny w Krakowie

WPLYW AUTOMATYCZNEGO TŁUMACZENIA NA WYNIKI AUTOMATYCZNEJ IDENTYFIKACJI CHARAKTERU OPINII KONSUMENCKICH

Streszczenie: Głównym celem pracy jest ocena wpływu tłumaczenia maszynowego na automatyczną analizę opinii konsumenckich. W kolejnych rozdziałach pracy najpierw scharakteryzowana została automatyczna analiza opinii konsumenckich, a następnie krótko opisane zostało tłumaczenie maszynowe. W kolejnych krokach omówione zostały opinie o produktach i usługach. W dalszej części zaprezentowano wyniki analizy symulacyjnej będącej próbą oceny wpływu tłumaczenia maszynowego na dokładność automatycznej analizy opinii konsumenckich. Pracę kończą wnioski oraz dalsze plany badawcze. W badaniach do automatycznego tłumaczenia wykorzystana została aplikacja Google Translate. Obliczenia dokonywane były z wykorzystaniem aplikacji RapidMiner.

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

1. Wstęp

Analiza opinii konsumenckich jest obszarem badań, który może mieć znaczący wpływ na współczesne zarządzanie. Znaczna liczba konsumentów przed dokonaniem wyboru o zakupie towaru lub skorzystaniu 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 i 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.

Innym zagadnieniem, które wynika z globalizacji i powszechnego dostępu do Internetu, jest dostępność opinii w różnych językach. Sprawia to, że ich analiza jest utrudniona ze względu chociażby na specyfikę każdego języka.

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 [Lula 2011].

Głównym celem pracy jest ocena rozwiązania polegającego na automatycznym przetłumaczeniu opinii z języka źródłowego na język angielski i przeprowadzenie analizy tak uzyskanego tekstu. Pozwoli to na sprawdzenie, czy automatyczne tłumaczenie tekstu nie utrudnia identyfikacji nacechowania opinii.

2. Automatyczna analiza opinii konsumentkich i tłumaczenie maszynowe

Automatyczna analiza opinii konsumentkich (*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 (Wikipedia).

W ramach automatycznej analizy opinii konsumentkich wyróżnić można trzy rodzaje działań, takie jak [Liu 2007]:

- **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 poprzez analizę zdania porównującego go do innego produktu. Konieczne jest zidentyfikowanie w opinii zdań porównujących, a następnie ich analiza ukierunkowana na przedmiot porównania.

Stosowanych jest kilka podejść do klasyfikacji opinii. Koncentrując się na klasyfikacji opinii, można wyróżnić cztery text miningowe podejścia do niej [Lula i Wójcik 2011]:

- **Podejście oparte na słowach (*word-based approach*)** – podstawą tego podejścia jest przekonanie, że znaczenie wypowiedzi (również jej nacechowanie) jest zakodowane w pojedynczych słowach stanowiących dany tekst.
- **Podejście bazujące na wzorcach (*pattern-based approach*)** – w tym podejściu istotne jest przekonanie, że nacechowanie opinii wyznaczają nie pojedyncze słowa, ale zbudowane z nich frazy/związki frazeologiczne. Tak więc konieczne jest wyszukanie wśród słów związków wyrazowych.
- **Podejście bazujące na ontologiach (*ontology-based approach*)** – pojedyncza opinia dotycząca produktu lub usługi może zostać przedstawiona jako instancja ontologii. Następnie instancje te mogą zostać porównane, a na tej podstawie

reprezentowane przez nie opinie mogą zostać zaklasyfikowane do jednej z utworzonych grup.

- **Podejście, u podstaw którego stoi uczenie maszynowe (*machine learning approach*)** – dzięki zastosowaniu uczenia maszynowego można zbudować system, który nie tylko na podstawie odpowiednio dobranego uczącego zbioru opinii będzie je klasyfikował do odpowiednich grup, ale również będzie się rozwijał wraz z pojawieniem się nowych, specyficznych opinii.

Tłumaczenie maszynowe (automatyczne) to dziedzina lingwistyki komputerowej, która zajmuje się opracowywaniem i stosowaniem algorytmów tłumaczenia tekstów z jednego języka naturalnego na drugi (Wikipedia).

Wśród wielu algorytmów tłumaczenia maszynowego najpopularniejsze jest tłumaczenie statystyczne. Jest to tłumaczenie bazujące na wykorzystaniu tzw. korpusów równoległych. Korpus równoległy to zbiór tekstów równoległych. Z kolei tekst równoległy to tekst składający się z zestawionych obok siebie tekstów, w co najmniej dwóch językach. Najczęściej jeden z tekstów jest oryginałem, a pozostałe jego tłumaczeniami, choć zdarza się, że wszystkie teksty są tworzone równoległe.

W tłumaczeniu statystycznym dla każdego zdania szukane jest jego najbardziej prawdopodobne tłumaczenie. Na prawdopodobieństwo to wpływa współwystępowanie słów w zdaniu.

3. Materiały i metody badań

Badania empiryczne można podzielić na następujące etapy:

1. Przygotowanie zbioru opinii w języku źródłowym (polskim).
2. Automatyczne przetłumaczenie opinii na język angielski.
3. Budowa modelu służącego do analizy opinii angielskojęzycznych.
4. Analiza przetłumaczonych opinii.
5. Ocena poprawności proponowanego rozwiązania.

3.1. Opinie o produktach i usługach

Opinie to specyficzny rodzaj danych tekstowych, które mają subiektywny charakter – wyrażają stosunek autora wypowiedzi do przedmiotu opinii. W niektórych serwisach opinie słowne są wspierane oceną punktową lub gwiazdkami.

Opinie można podzielić na grupy według ich formatu [Liu 2007]:

- Format 1: zalety i wady, oraz podsumowanie,
- Format 2: zalety i wady,
- Format 3: dowolny.

W badaniu wykorzystano trzy grupy opinii:

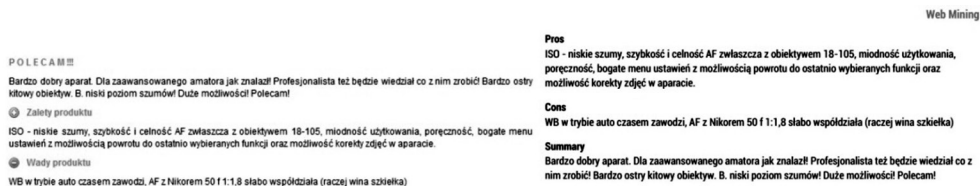
1. 301 opinii w języku angielskim, w tym 176 negatywnych i 125 pozytywnych.
2. 20 opinii w języku polskim, w tym 8 negatywnych i 12 pozytywnych.
3. 30 dodatkowych opinii w języku polskim, w tym 14 negatywnych i 16 pozytywnych.

Wszystkie opinie wykorzystane w badaniu dotyczyły aparatów fotograficznych. Różne były zarówno marki, jak i klasy sprzętu. Opinie w języku angielskim pobrano z serwisu <http://www.cnet.com/> (rys. 1). Natomiast opinie w języku polskim pochodziły z serwisów, takich jak: <http://www.ceneo.pl/> (rys. 2), <http://www.opineo.pl/> i <http://www.skapiec.pl/>. Opinie w obydwu językach reprezentowały format 1, czyli składały się z wad i zalet produktu oraz krótkiego podsumowania. Opinie wybierane były w sposób losowy. Starano się jedynie zachować równowagę pomiędzy opiniami pozytywnymi i negatywnymi.



Rys. 1. Przykładowa opinia w języku angielskim wykorzystana w badaniu empirycznym

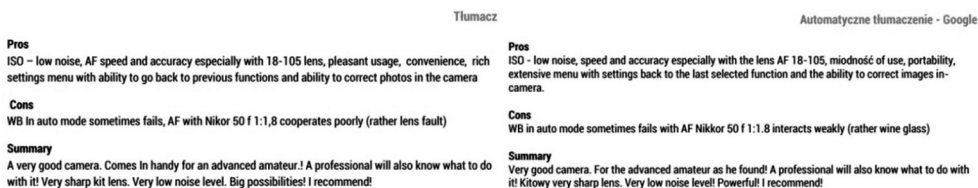
Źródło: <http://reviews.cnet.com>.



Rys. 2. Przykładowa opinia w języku polskim w oryginale oraz po ekstrakcji

Źródło: <http://www.ceneo.pl> oraz opracowanie własne.

Opinie w języku polskim zostały przetłumaczone na język angielski na dwa sposoby: przez profesjonalnego tłumacza oraz maszynowo.



Rys. 3. Przykładowa opinia przetłumaczona na język angielski przez tłumacza oraz za pomocą Google Translate

Źródło: opracowanie własne.

Do tłumaczenia maszynowego wykorzystano Google Translate¹. Narzędzie to realizuje algorytm tłumaczenia statystycznego.

3.2. Modele automatycznej analizy opinii konsumenckich

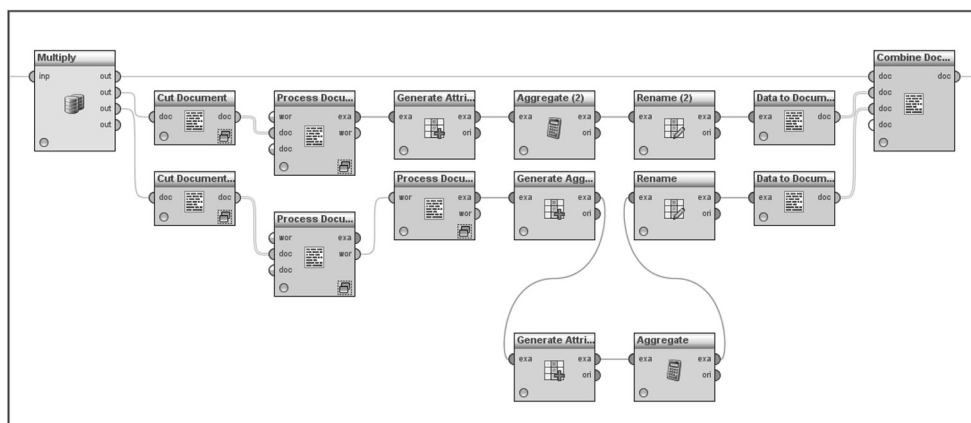
W badaniu skonstruowano dwa modele:

1. Model oparty na podejściu bazującym na słowach

Opinie podczas ekstrakcji zostały podzielone na dwie części. Pierwsza część to zalety i wady produktu, a druga część to podsumowanie. Każda z tych części jest analizowana osobno. W analizie wad i zalet można wyróżnić następujące etapy: 1) podział tekstów na pojedyncze słowa, 2) POS (*Part Of Speech tagging*) – oznaczenie części mowy, 3) przefiltrowanie słów według części mowy, 4) zliczenie rzeczowników i przymiotników po stronie zalet oraz po stronie wad, 5) przemnożenie wyników przez odpowiednie wagi (1 dla zalet i -1 dla wad), 6) agregacja punktów do jednej wartości reprezentującej nacechowanie zalet i wad.

Z kolei analiza podsumowania przebiega według następującego schematu: 1) wstępne przetwarzanie tekstów, 2) klasyfikacja słów według ich nacechowania za pomocą słowników [Ohana, Tierney 2009], 3) przypisanie słowom punktów według ich nacechowania, 4) agregacja punktów do jednej wartości reprezentującej nacechowanie podsumowania.

Ostatni krok w modelu realizującym podejście bazujące na słowach to agregacja wartości uzyskanych dla podsumowania oraz dla zalet i wad. Pozwala to na określenie nacechowania całej pojedynczej opinii (rys. 4).



Rys. 4. Fragment modelu opartego na podejściu bazującym na słowach

Źródło: opracowanie własne w programie RapidMiner.

¹ <http://translate.google.pl/>.

4. Model oparty na uczeniu maszynowym

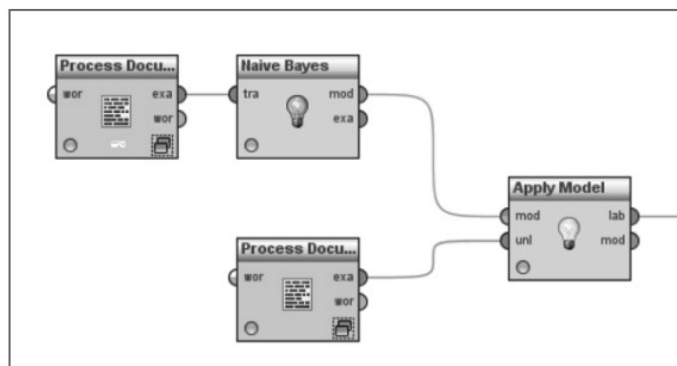
Model oparty na uczeniu maszynowym skonstruowany jest z bloków realizujących zadania według następującego schematu: 1) wstępne przetwarzanie tekstów, 2) uczenie modelu, 3) testowanie modelu, 4) klasyfikacja opinii (rys. 5).

W uczeniu maszynowym wykorzystać można różne algorytmy klasyfikacyjne. W badaniu użyte zostały następujące klasyfikatory [StatSoft, Inc 2010]:

a) naiwny klasyfikator Bayesowski (1) liczone jest prawdopodobieństwo wystąpienia opinii pozytywnej oraz negatywnej, 2) wyznaczana jest odległość pomiędzy obiektami (opiniami), 3) wyznaczana jest szansa na to, że opinia będzie pozytywna bądź negatywna, biorąc pod uwagę określoną liczbę najbliższych opinii, 4) wyznaczane jest prawdopodobieństwo tego, że opinia będzie pozytywna oraz że będzie negatywna, a następnie te dwie wartości są porównywane);

b) metodę k najbliższych sąsiadów (1) wyznaczana jest odległość pomiędzy obiektami (opiniami), 2) wybieranych jest k najbliższych obiektów, 3) na podstawie przynależności k najbliższych sąsiadów określana jest przynależność badanej opinii do grupy pozytywnych lub negatywnych).

Dla każdego z tych klasyfikatorów otrzymano w badaniu różne wyniki.



Rys. 5. Fragment modelu opartego na uczeniu maszynowym

Źródło: opracowanie własne w programie RapidMiner.

Do konstrukcji obydwu modeli wykorzystano aplikację RapidMiner².

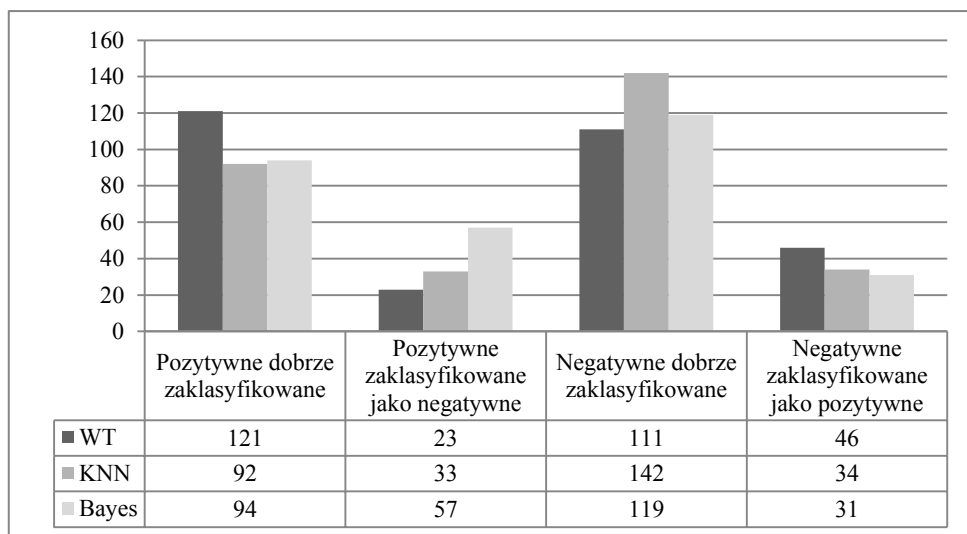
5. Wyniki badań empirycznych

Celem badania jest próba oceny przydatności tłumaczenia maszynowego tekstów w automatycznej analizie opinii konsumenckich. Aby zrealizować ten cel, opinie

² <http://www.rapidminer.com>.

polskojęzyczne zostały przetłumaczone na język angielski na dwa sposoby, a następnie została dokonana ich automatyczna klasyfikacja. Pozwoliło to na porównanie wyników uzyskanych dla tekstów tłumaczonych maszynowo i tradycyjnie.

Do sprawdzenia efektywności każdego z modeli wykorzystano zbiór 301 opinii w języku angielskim (rys. 6).



Rys. 6. Jakość klasyfikacji 301 opinii w języku angielskim przez poszczególne modele

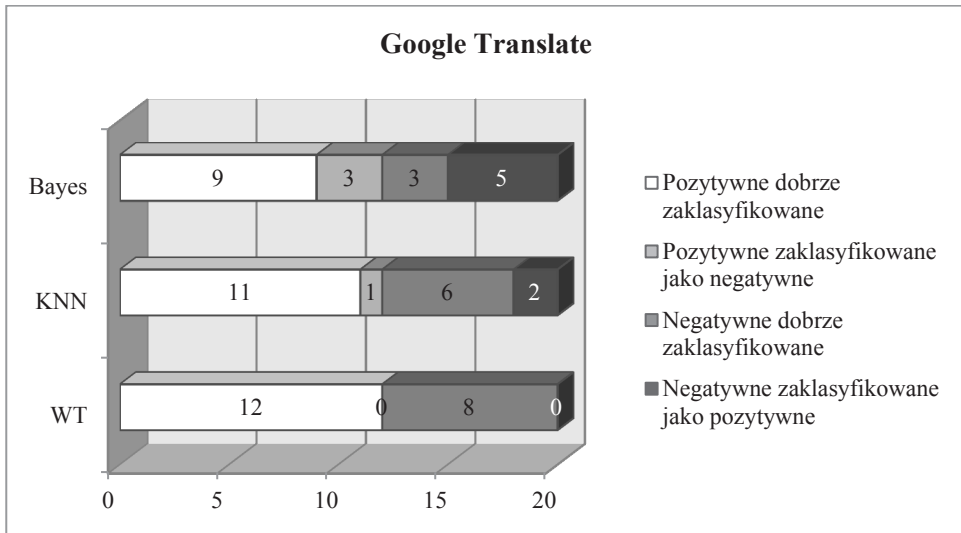
Źródło: opracowanie własne.

Zastosowane na wszystkich wykresach skróty oznaczają odpowiednio: WT – model oparty na podejściu bazującym na słowach, KNN – model oparty na uczeniu maszynowym z zastosowaniem klasyfikatora k najbliższych sąsiadów oraz Bayes – model oparty na uczeniu maszynowym z zastosowaniem naiwnego klasyfikatora Bayesowskiego.

Aby ocenić przydatność automatycznego tłumaczenia, w klasyfikacji opinii zestawiono wyniki uzyskane dla tekstów przetłumaczonych tymi dwoma sposobami (rys. 7 i 8).

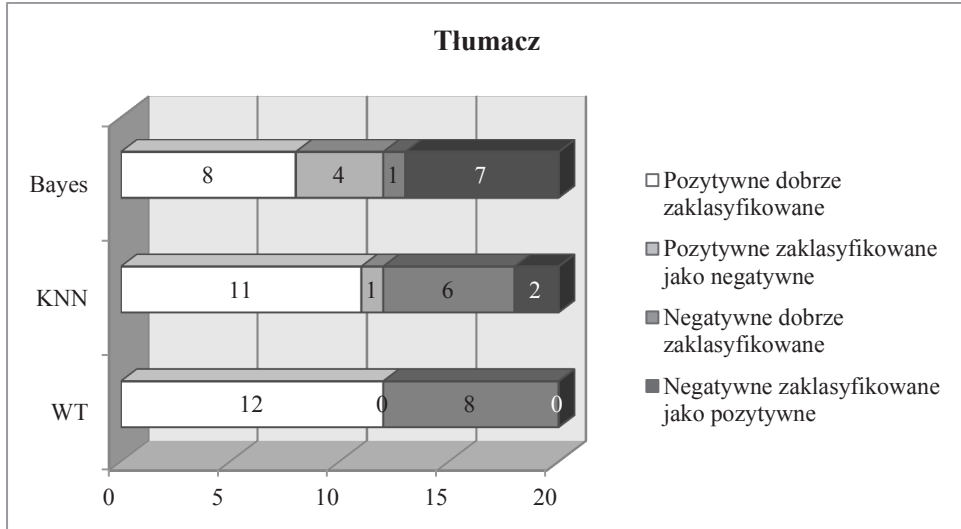
Jak wynika z rys. 7 i 8, nie ma znaczących różnic pomiędzy wynikami uzyskanymi dla tekstów tłumaczonych przez profesjonalnego tłumacza a tymi uzyskanymi dla tekstów tłumaczonych maszynowo. Może to oznaczać, że na automatyczną ocenę nacechowania opinii konsumentów nie ma wpływu rodzaj tłumaczenia.

W celu lepszego przetestowania modelu dodano 30 opinii przetłumaczonych automatycznie z języka polskiego na angielski. Wyniki klasyfikacji tych opinii zaprezentowano na rys. 9.



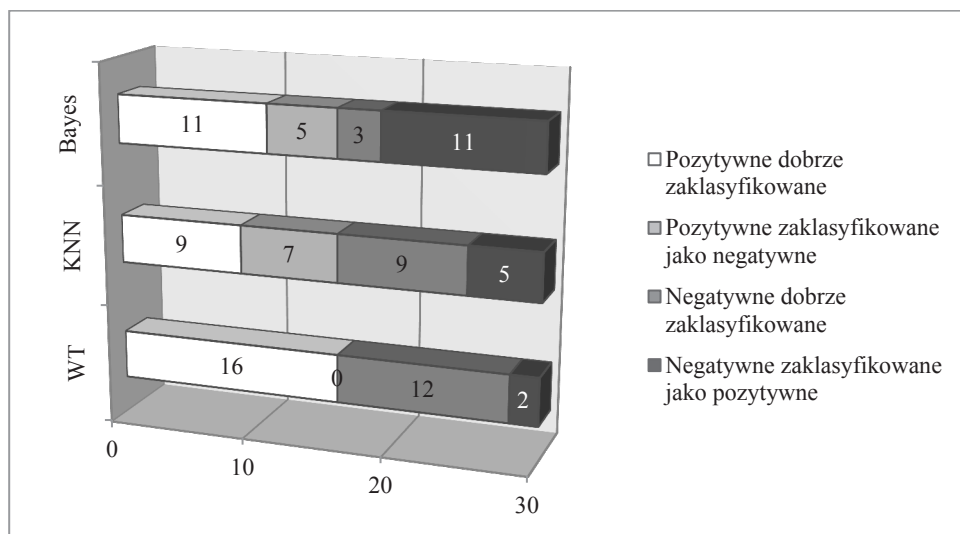
Rys. 7. Jakość klasyfikacji 20 opinii tłumaczonych automatycznie z języka polskiego na angielski

Źródło: opracowanie własne.



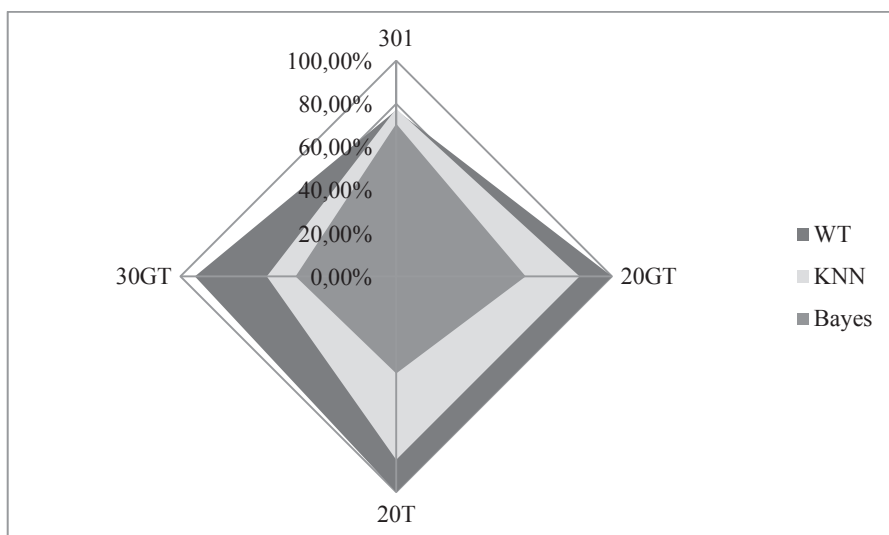
Rys. 8. Jakość klasyfikacji 20 opinii tłumaczonych z języka polskiego na angielski przez tłumacza

Źródło: opracowanie własne.



Rys. 9. Jakość klasyfikacji 30 opinii tłumaczonych z języka polskiego na angielski w Google Translate

Źródło: opracowanie własne.



Rys. 10. Porównanie skuteczności poszczególnych modeli na różnej liczbie opinii testowych

Źródło: opracowanie własne.

Rysunek 10 wraz z tab. 1 obrazują wyniki porównania efektywności opisanych w pracy modeli na różnych zbiorach danych. W przypadku oryginalnych opinii

w języku angielskim wszystkie stosowane modele osiągają podobną skuteczność. W pozostałych przypadkach najlepiej sprawdza się model oparty na podejściu bazującym na słowach. Najgorsze efekty uzyskano dla modelu opartego na uczeniu maszynowym z zastosowanym naiwnym klasyfikatorem Bayesowskim. Może to oznaczać, że sam fakt tłumaczenia tekstu ma wpływ na jakość klasyfikacji, szczególnie w przypadku modeli bazujących na uczeniu maszynowym, gdzie model uczony był na oryginalnych testach angielskich.

Tabela 1. Porównanie skuteczności poszczególnych modeli na różnej liczbie opinii testowych

Zbiór opinii	Oznaczenie	WT	KNN	Bayes
301 opinii w języku angielskim	301	77,08%	77,74%	70,76%
20 polskich opinii przetłumaczonych na język angielski przez tłumacza	20T	100,00%	85,00%	60,00%
20 polskich opinii przetłumaczonych na język angielski przez Google Translate	20GT	100,00%	85,00%	45,00%
30 polskich opinii przetłumaczonych na język angielski przez Google Translate	30GT	93,33%	60,00%	46,67%

Źródło: opracowanie własne.

6. Wnioski i plany badawcze

Tłumaczenie opinii na język angielski umożliwia skorzystanie z gotowych rozwiązań przeznaczonych dla tekstów w tym języku. Pozwala to tym samym zaoszczędzić czas i pieniądze. W zależności od stosowanego modelu tłumaczenie zarówno tradycyjne, jak i maszynowe w różnym stopniu się sprawdza. Przy modelu opartym na słowach tłumaczenie maszynowe daje bardzo dobre wyniki.

W przyszłości planowane jest zwiększenie liczebności zbioru uczącego oraz zwiększenie liczebności zbioru opinii w języku polskim tłumaczonych maszynowo. Ponadto planowane jest przetestowanie pozostałych modeli klasyfikacji opinii.

7. Podsumowanie

W artykule przedstawione zostały pokrótce wyniki badań symulacyjnych dotyczących oceny wpływu tłumaczenia maszynowego na poprawność automatycznej klasyfikacji opinii konsumenckich.

Skonstruowane w badaniu modele osiągają zbliżone wyniki. Również pomiędzy profesjonalnym tłumaczem a aplikacją realizującą tłumaczenie statystyczne nie ma znaczących różnic w wynikach. Zaskakujący dla autorów okazał się fakt, iż wyniki klasyfikacji opinii osiągnięte przy tłumaczeniu maszynowym były lepsze niż te uzyskane dla tekstów tłumaczonych przez profesjonalnego tłumacza. Jednakże może

to być kwestia zbioru danych lub samego tłumacza. Zweryfikować to stwierdzenie mogą planowane badania uwzględniające zwiększenie liczebności badanego zbioru oraz badania opinii dotyczących produktów i usług z innego obszaru.

Model oparty na uczeniu maszynowym okazał się prostszy w konstrukcji. Wymaga jednak dobrze dobranego zbioru uczącego. Zbiór uczący powinien przedmiotem oceny korespondować z opiniami, które mają być badane. Metoda k najbliższych sąsiadów okazała się lepsza do klasyfikacji opinii.

Literatura

- Liu B., *Web DataMining. Exploring Hyperlinks, Contents, and Usage Data*, Heidelberg, Springer-Verlag, Berlin 2007.
- Lula P., *Automatyczna analiza opinii konsumenckich*, [w:] Taksonomia 18, *Klasyfikacja i analiza danych – teoria i zastosowania*, 2011.
- Lula P., Wójcik K., *Sentiment analysis of consumer opinions written in Polish*, “Economics and Management” 2011, 1286-1291.
- Ohana B., Tierney B., *Sentiment Classification of Reviews Using SentiWordNet*, *IT&T Conference*, Dublin Institute of Technology, Dublin 2009.
- StatSoft, Inc., *Electronic Statistic Textbook*, Pobrano 07 31, 2011 z lokalizacji http://www.statsoft.pl/textbook/stathome_stat.html?http://www.statsoft.pl/textbook/stmulzca.html, 2010.
- Wikipedia*. <http://pl.wikipedia.org/wiki> (accessed Październik 12, 2012).

MACHINE TRANSLATION IMPACT ON THE RESULTS OF THE SENTIMENT ANALYSIS

Summary: The main objective of paper is to determine the machine translation impact on the results of the sentiment analysis. In particular parts of the work firstly sentiment analysis is characterized. Secondly machine translation is shortly described. Next part of the work is devoted to opinions about products and services. The results of simulation analysis testing the influence of machine translation on the accuracy of sentiment analysis are the main content of next part of the paper. Finally there are conclusions and further research plans. Google Translate was used in the research as machine translation application. Computations are conducted in RapidMiner application.

Keywords: text-mining, Web-mining, taxonomy, text document classification, opinion mining, sentiment analysis.