Łukasz Lipiński[1], Michał Bernardelli[2]

Anonimowość w Internecie – identyfikacja płci użytkowników na podstawie historii odwiedzanych stron internetowych

Streszczenie

W artykule przedstawiono metodę identyfikacji płci użytkowników Internetu. Proponowana metoda wykorzystuje dane z profili użytkowników zawierających adresy stron internetowych i częstotliwość odwiedzin. Podejście to łączy w sobie analizę leksykalną słów z domen internetowych, sztuczne sieci neuronowe, wyrafinowaną matematycznie wektorową reprezentację profili użytkowników oraz regresję logistyczną jako główny klasyfikator. Analizę empiryczną przeprowadzono na podstawie 10 mln profili polskich użytkowników, osiągając skuteczność klasyfikacji na poziomie 82%. Dodatkowe korzyści z badania to stworzenie listy najczęściej odwiedzanych stron internetowych według płci w Polsce w 2017 r. oraz określenie sposobu wyszukiwania podobnych portali internetowych, który może być wykorzystany w spersonalizowanym marketingu jako źródło oszczędności w postaci zmniejszenia niepotrzebnych wydatków na źle ukierunkowaną reklamę.

Słowa kluczowe: Internet, zagadnienie klasyfikacji, preferencje użytkowników, word2vec, Big Data

JEL: C01, C53, C55

1. Wstęp

Wielkość ruchu internetowego, zaawansowane technologicznie narzędzia wykorzystywane na co dzień przez ludzi oraz zwykła niewiedza przeciętnych użytkowników Internetu często sprawiają złudne wrażenie anonimowości. Tymczasem możliwości śledzenia użytkowników, ich zachowania, lokalizacji czy preferencji są – nawet w granicach obowiązującego prawa – bardzo rozwinięte. Mając odpowiednie narzędzia lub uprawnienia, można precyzyjnie

[1] Cloud Technologies.

[2] Szkoła Główna Handlowa w Warszawie, Kolegium Analiz Ekonomicznych.

zidentyfikować użytkownika lub stworzyć jego odpowiedni profil. Z punktu widzenia biznesowego istotne jest uzyskanie jak najbardziej szczegółowych informacji o użytkowniku. Pozwala to lepiej spersonalizować komunikację, co z kolei zwiększa szansę np. na skuteczniejsze wykorzystanie skłonności zakupowych użytkownika. Do podstawowych cech klienta należą z pewnością płeć oraz wiek, ale również rozszerzone dane demograficzne (narodowość, wyznanie), geolokalizacyjne (miejsce zamieszkania, miejsce pracy), teleinformatyczne (jakość połączenia internetowego, model telefonu komórkowego, laptopa, oprogramowanie) czy ogólnie rozumiane preferencje (polityka, hobby).

Istnieją firmy profesjonalnie zajmujące się badaniem ruchu internetowego. W Polsce badania takie są wykonywane m.in. przez firmę Gemius. Jej miesięczne raporty[3] zawierają m.in. zestawienia najpopularniejszych wydawców internetowych z uwzględnieniem podziału na komputery osobiste, laptopy oraz urządzenia mobilne. Raporty istotne dla reklamodawców są przygotowywane w Związku Pracodawców Branży Internetowej IAB Polska. Część z nich jest opracowywana cyklicznie, np. „Perspektywy rozwojowe reklamy online w Polsce"[4]. Na ogół powstają one jednak na podstawie badań ankietowych przeprowadzanych wśród internautów oraz statystyk ruchu internetowego dostarczanych przez uczestniczące w badaniach portale internetowe. Tym samym pomija się w nich istotny aspekt zachowań internautów, jakim są preferencje konkretnej osoby. Twórcy tych raportów nie pozwalają w łatwy sposób połączyć odwiedzin różnych stron internetowych tej samej osoby, koncentrują się bowiem na aspekcie ilościowym odwiedzin danej strony. Oczywiście takie dane statyczne pomagają w lepszym zrozumieniu rynku internetowego, a przez to trafniejszej alokacji środków finansowych przeznaczonych na reklamę, nie pozwalają jednak na automatyzację działań marketingowych. Świadomość, że dany portal internetowy jest popularniejszy wśród danej grupy ludzi, nie daje jednak wiedzy o konkretnej osobie odwiedzającej taką stronę oraz jej preferencjach. Stąd w wielu działaniach marketingowych wykorzystuje się podejście oparte na dedykowanej reklamie internetowej wyświetlanej w czasie rzeczywistym, w tzw. systemie Real Time Bidding[5]. Reklama ta jest

[3] Najnowszy raport Gemius/PBI dotyczący kwietnia 2018 r. jest dostępny pod adresem https://www.gemius.pl/wszystkie-artykuly-aktualnosci/wyniki-badania-gemiuspbi-za-kwiecien-2018.html (odczyt: 25.05.2018).

[4] Najnowszy, czwarty raport to „Perspektywy rozwojowe reklamy online w Polsce 2017–2018" z 14.12.2017, https://iab.org.pl/badania-i-publikacje/perspektywy-rozwojowe-reklamy-online-w-polsce-2017-2018/ (odczyt: 25.05.2018).

[5] M. Bernardelli, *Cheater detection in Real Time Bidding system – panel approach*, „Roczniki" Kolegium Analiz Ekonomicznych 2015, nr 39, s. 11–23.

lepiej dopasowana, jeżeli ekonometryk posiada więcej informacji o użytkowniku oraz zastosował model efektywniejszy jakościowo pod względem prognostycznym.

Celem badania było przedstawienie algorytmu identyfikacji płci polskich użytkowników Internetu na podstawie historii odwiedzanych przez nich witryn internetowych. W zaproponowanej metodzie wykorzystuje się dane z profili użytkowników zawierających adresy stron internetowych oraz częstości ich odwiedzin. Do opracowania algorytmu zostały wykorzystane dane z ponad 10 mln profili polskich użytkowników oraz prawie 0,5 mld odwiedzin stron internetowych z 2017 r. Skuteczność poprawnej identyfikacji płci z użyciem proponowanej metody oscyluje wokół 80%, bez uwzględniania dodatkowych danych o użytkowniku, np. używanego sprzętu, systemu operacyjnego, przeglądarki, godzin korzystania z Internetu.

Opisana w niniejszym artykule metoda klasyfikacji stanowi połączenie wyrafinowanego algorytmu wykorzystującego sztuczne sieci neuronowe, analizę kontekstową słów oraz regresję logistyczną. Ma to pomóc w ścisłym odzwierciedleniu podobieństwa witryn internetowych na podstawie ich nazw (adresów internetowych), a następnie wykorzystaniu tej wiedzy do oszacowania prawdopodobieństwa zaklasyfikowania danego użytkownika do konkretnej płci. Informacja dotycząca płci może zostać wykorzystana do wyświetlania dedykowanej reklamy na portalach internetowych. Dotarcie do docelowej grupy klientów jest bowiem jednym z podstawowych problemów spersonalizowanego marketingu oraz źródłem oszczędności w postaci ograniczenia zbędnych wydatków na nieskuteczną reklamę, jak również czerpania dodatkowych korzyści z reklamy, która trafiła do odpowiedniego odbiorcy. Należy jednak podkreślić, że przedstawiona w artykule metoda może być z powodzeniem rozszerzona na inne cechy klienta, jak wiek, zapatrywania polityczne, religijne czy rodzaj hobby.

Skuteczność działania tej prostej metody, po pierwsze, skłania do refleksji na temat anonimowości w Internecie, a po drugie, otwiera duże możliwości personalizacji, która może być wykorzystana np. w marketingu internetowym czy rynku e-commerce. Dodatkową wartością badania, oprócz wykorzystania metody klasyfikacji, jest analiza empiryczna na realnych danych pochodzących z ruchu internetowego w Polsce, której wyniki dają pogląd na temat kategorii stron najchętniej odwiedzanych przez Polaków.

Artykuł składa się z pięciu punktów. Po niniejszym wprowadzeniu przedstawiono w punkcie drugim charakterystykę danych oraz podstawową nomenklaturę wykorzystywaną w usługach internetowych. Punkt trzeci zawiera opis metody ze szczególnym uwzględnieniem wchodzących w jej skład algorytmów lingwistycznych, sieci neuronowych oraz regresji logistycznej. Następnie zebrano wyniki

działania prezentowanej metody na danych obejmujących historię odwiedzin witryn internetowych przez polskich internautów. W ostatnim punkcie przedstawiono krótkie podsumowanie oraz możliwości dalszych badań nad prezentowanym zagadnieniem.

2. Charakterystyka danych

Każda ze stron internetowych ma swój unikatowy adres, który pozwala ją odszukać z dowolnego miejsca w sieci Internet. Adres taki, zwany domeną, jest tożsamy z tzw. numerem IP, ale aby ułatwić ludziom zapamiętywanie, w miejsce adresów IP stosuje się bardziej przyjazne z punktu widzenia użytkownika słowne nazwy. Przykłady domen internetowych to np. wikipedia.pl, google.com czy sgh.waw.pl. W ramach jednej domeny może istnieć wiele subdomen, które służą m.in. do rozgraniczenia różnych części danej witryny internetowej. Przykładem subdomen dla domeny sgh.waw.pl są np. poczta.sgh.waw.pl, dziekanat.sgh.waw.pl, firma.sgh.waw.pl, ale też www.sgh.waw.pl[6]. Analogicznie w przypadku domeny google.com subdomenami są mail.google.com, www.google.com czy maps.google.com. Oczywiście adresy stron internetowych są na ogół znacznie dłuższe, ale każda strona internetowa należy do jakiejś subdomeny, a tym samym do konkretnej domeny.

Prezentowana metoda wykorzystuje subdomeny internetowe. Oparcie wyników na domenach wpłynęłoby ujemnie na dokładność klasyfikacji, ponieważ w ramach jednej domeny często występują subdomeny o znacząco różnej tematyce. Przykładowo w ramach domeny onet.pl istnieją subdomeny: kobieta.onet.pl, moto.onet.pl, eurosport.onet.pl czy dziecko.onet.pl (każda skierowana do nieco innej grupy odbiorców). Jednocześnie równie dyskusyjne wydaje się wykorzystanie pełnych adresów internetowych. Przeciw przemawia fakt, że w ramach subdomen dużych portali internetowych istnieją dziesiątki tysięcy stron z artykułami, których oglądalność ogranicza się na ogół do krótkiego okresu. Co więcej, słowa występujące w adresach takich stron są często po prostu ich skróconymi nagłówkami, co sprawia, że nie odzwierciedlają zawartości strony – w przeciwieństwie do nazwy subdomeny, która w większości przypadków jest jednoznaczna, jeżeli chodzi o poruszaną tematykę. Nie bez znaczenia jest również istotny wzrost ilości danych, jeżeli chodzi o użycie pełnych adresów stron

[6] Formalnie adresy www.sgh.waw.pl oraz sgh.waw.pl nie są zatem tożsame.

internetowych, oraz związane z tym większe wymagania obliczeniowe zadania. Stąd wykorzystanie do klasyfikacji subdomen wydaje się naturalnym, a przede wszystkim intuicyjnie najlepszym pomysłem.

Użytkownik odwiedzający wybraną stronę internetową pozostawia, nawet bez swojej wiedzy, wiele informacji związanych ze swoją osobą. Przede wszystkim są to dane związane z używanym sprzętem oraz oprogramowaniem (urządzenie, system operacyjny, przeglądarka). Dostępne są jednak również dane na temat lokalizacji użytkownika[7] (adres IP komputera, kraj, język). W przypadku współczesnych stron internetowych są również przechowywane informacje na temat ustawień strony wprowadzonych przez użytkownika w postaci tzw. ciasteczek (ang. *cookie*), dzięki czemu nie trzeba ponownie wprowadzać tych samych danych przy kolejnych odwiedzinach tej samej strony. Oznacza to jednak, że tworzony jest swego rodzaju identyfikator użytkownika, z którym są powiązane wszystkie zbierane dane. W ten sposób można przechowywać historię odwiedzin użytkowników, przy czym do wymienionych wyżej danych należy dodać informacje związane z zachowaniem użytkownika na stronie, w tym m.in. dokładny czas wejścia na stronę, czas wyświetlania strony, to, czy nastąpiły jakieś akcje typu kliknięcie (ang. *click*) bądź przewinięcie strony (ang. *scroll*), a nawet śledzenie kursora myszki.

Należy zwrócić jednak uwagę na fakt, że stosowany powszechnie sposób przechowywania informacji o użytkowniku w postaci *cookie* nie jest dokładny. Po pierwsze, ciasteczka mogą być usuwane zarówno ręcznie, jak i automatycznie (kwestia ustawień bezpieczeństwa przyjęta przez danego użytkownika), a po drugie, często z jednego urządzenia (komputera czy tabletu) i jednego konta na tym urządzeniu korzysta więcej niż jedna osoba. W takim przypadku współużytkownicy będą nierozróżnialni z punktu widzenia zapisywanych o nich informacji. Biorąc pod uwagę te utrudnienia, nie należy się spodziewać stuprocentowej skuteczności algorytmów klasyfikacyjnych opartych na tych danych.

Dane wykorzystane w badaniu opisanym w tym artykule to profile użytkowników, na które składa się lista odwiedzonych przez nich stron internetowych wraz z licznikiem ich odwiedzin. Przykładowy profil *i*-tego użytkownika może mieć postać:

$$profil_i = \left\{"sport.pl":10,\ "polityka.onet.pl":5,"otomoto.pl":31\right\}.$$

[7] Istnieją sposoby na maskowanie prawdziwej geolokalizacji czy wykorzystywanego oprogramowania.

W tak zdefiniowanych danych panelowych nie zostały uwzględnione czasy odwiedzin stron (podstawowa wersja algorytmu klasyfikacyjnego opisanego w artykule), co implikuje pominięcie informacji na temat kolejności odwiedzania stron przez użytkownika. Czasy odwiedzin mają natomiast znaczenie przy dodatkowym ograniczeniu w postaci limitu do co najwyżej 100 stron ostatnio odwiedzanych przez użytkownika. Ograniczenie to nie jest konieczne, ale w tego rodzaju badaniu istotniejsze wydają się informacje z jak najbliższej przeszłości, co m.in. redukuje możliwość pomyłki ze względu na korzystanie z komputera więcej niż jednej osoby. Co więcej, nie bez znaczenia jest mniejsza złożoność obliczeniowa problemu.

Profile 10 385 003 użytkowników zostały wykorzystane do określenia stopnia podobieństwa pomiędzy subdomenami. Na tej podstawie przeprowadzono analizę, wybierając m.in. najpopularniejsze wśród użytkowników danej płci subdomeny. Opis wyników tej analizy został zaprezentowany w punkcie czwartym. Jednocześnie tak przetworzony zbiór danych posłużył do konstrukcji modelu określającego płeć danego użytkownika. Do 27 907 profili została przyporządkowana płeć użytkownika określona na podstawie formularzy wypełnianych przez użytkowników. Dane pochodziły z 2017 r., a użytkowników i adresy stron internetowych ograniczono do terytorium Polski. Tak przygotowane dane zostały podzielone na dwa zbiory w proporcji 80% do 20%. Pierwszy z nich, o liczności 22 325 użytkowników, to zbiór treningowy, na którym została dokonana estymacja parametrów modelu. Drugi ze zbiorów zawierał 5582 obserwacje i został wykorzystany do walidacji wyników klasyfikacji płci. Algorytm klasyfikacji został opisany dokładniej w kolejnym punkcie.

3. Opis algorytmu

Algorytm klasyfikujący użytkowników ze względu na ich płeć zaproponowany w tym artykule składa się z dwóch autonomicznych elementów. Pierwszy z nich to połączenie sztucznej sieci neuronowej i analizy leksykograficznej, znane pod nazwą algorytmu word2vec[8]. Drugi to model regresji logistycznej, w którym zmienne objaśniające to odpowiednio przekształcone dane wynikowe z word2vec. W kolejnych akapitach tego punktu zostanie przedstawiony opis dwóch części składowych proponowanego algorytmu wraz z transformacjami zmiennych.

[8] https://radimrehurek.com/gensim/models/word²vec.html (odczyt: 11.04.2018).

Podstawą wielu algorytmów uczenia maszynowego jest reprezentacja wektorowa danych wejściowych. Aby przekształcić profil użytkownika do postaci wektorowej, wystarczy przyjąć, że każda z subdomen w jego profilu odpowiada jednej współrzędnej, przy czym wartością tej współrzędnej jest liczba odwiedzin danej subdomeny przez użytkownika. Wartość zero będzie przypisana do tych współrzędnych wektorów, które odpowiadają subdomenom nieznajdującym się w danym profilu użytkownika. Taki uproszczony model reprezentacji danych często wykorzystuje się w przetwarzaniu języka naturalnego i jest on znany pod nazwą *bag-of-words model*[9].

Reprezentacja profilu użytkownika w postaci wektorowej jest problematyczna z co najmniej dwóch powodów. Po pierwsze, wymiar wektora jest duży – w przypadku rozpatrywanego zbioru danych to ponad 4 mln subdomen. Rozwiązaniem tego problemu, stosowanym w praktyce, jest zmniejszenie rozmiaru wektora, np. poprzez określenie minimalnej liczby niezerowych współrzędnych lub zastosowanie bardziej wyrafinowanych technik redukcji wymiaru[10]. Po drugie, brakuje kontekstu dla wybranego słowa, w tym przypadku dla subdomeny. Analiza leksykalna bez uwzględnienia kontekstu stanowi duże uproszczenie, ale tym samym wpływa na mniejszą dokładność aproksymacji rozwiązania badanego problemu.

Istnieje wydajna obliczeniowo oraz efektywna technika reprezentacji słów języka naturalnego w postaci liczbowych wektorów z wielowymiarowej przestrzeni. Jest to metoda opracowana na początku wieku przez Y. Bengio[11], a rozwinięta przez zespół pracowników Google pod przewodnictwem T. Mikolova[12] i w pracach kolejnych naukowców[13]. Pozwala na zastąpienie dużych zbiorów tekstowych przez przestrzeń wektorową liczb o zazwyczaj kilkuset wymiarach, przy czym pod uwagę są brane nie same słowa, ale ich znaczenie oraz kontekst, w którym zostały użyte. Metoda ta, na którą składa się zestaw dwóch modeli używanych w przetwarzaniu języka naturalnego, znana jest pod nazwą word2vec.

---

[9] M. McTear, Z. Callejas, D. Griol Barres, *The Conversational Interface. Talking to Smart Devices*, Springer, 2016.

[10] R.A. Fisher, *The use of multiple measurements in taxonomic problems*, „Annals of Eugenics" 1936, vol. 7(2), s. 179–188; J. Han, M. Kamber, J. Pei, *Data Mining: Concepts and Techniques*, Morgan Kaufman, 2000.

[11] Y. Bengio, R. Ducharme, P. Vincent, Ch. Jauvin, *A neural probabilistic language model*, „Journal of Machine Learning Research" 2003, vol. 3, s. 1137–1155.

[12] T. Mikolov, I. Sutskever, K. Chen, G.S. Corrado, J. Dean, *Distributed representations of words and phrases and their compositionality*. *Advances in Neural Information Processing Systems*, 2013, arXiv:1310.4546.

[13] F. Hill, K. Cho, S. Jean, C. Devin, Y. Bengio, *Embedding Word Similarity with Neural Machine Translation*, 2014, arXiv:1412.6448.

Wynik działania tej techniki, jako reprezentacja wektorowa, pozwala na wykonywanie operacji arytmetycznych na słowach[14]:

ojciec – mężczyzna + kobieta = matka.

Dla języka polskiego dopiero powstają implementacje wykorzystujące technikę word2vec, ale w przypadku subdomen, w których znaki narodowe są pomijane, a kontekst znacząco ograniczony, z powodzeniem można wykorzystać algorytmy przeznaczone do języka angielskiego.

Na metodę word2vec składają się *de facto* dwa algorytmy, zwane CBOW (ang. *continuous bag of words*) oraz *continuous skip-gram*[15]. Pierwszy z algorytmów wyznacza wektor związany z danym słowem na podstawie słów otaczających rozpatrywane słowo bez uwzględnienia ich kolejności (*bag of words*). Drugi z algorytmów natomiast wykorzystuje każde słowo do wyznaczania wektorów dla słów otaczających je, przy czym słowom bliższym danemu przyporządkowana jest wyższa waga. Różnica dotycząca idei stosowania obu algorytmów została zaprezentowana na rysunku 1. Według autorów metody word2vec[16], algorytm *skip-gram* jest wolniejszy, ale bardziej efektywny dla rzadko pojawiających się słów.

Oba algorytmy – CBOW i *skip-gram* – dają wielowymiarową przestrzeń wektorową, w której za metrykę[17] na ogół przyjmuje się odległość kosinusową. Miara ta to kosinus kąta pomiędzy wektorami reprezentującymi w przypadku tego badania subdomeny, a w ogólności pomiędzy słowami, frazami czy całymi dokumentami[18], i dana jest wzorem

$$\mathrm{sim}\left(\mathrm{v}_1, \mathrm{v}_2\right) = \frac{v_1 \cdot v_2}{\left|v_1\right|\left|v_2\right|}, \tag{1}$$

gdzie $\cdot$ oznacza produkt skalarny.

---

14 Np. por. O. Levy, Y. Goldberg, *Linguistic Regularities in Sparse and Explicit Word Representations*, Proceedings of the Eighteenth Conference on Computational Natural Language Learning, Association for Computational Linguistics 2014, s. 171–180, https://aclanthology.coli.uni-saarland.de/papers/W14-1618/w14-1618.

15 T. Mikolov, I. Sutskever, K. Chen, G.S. Corrado, J. Dean, op.cit.

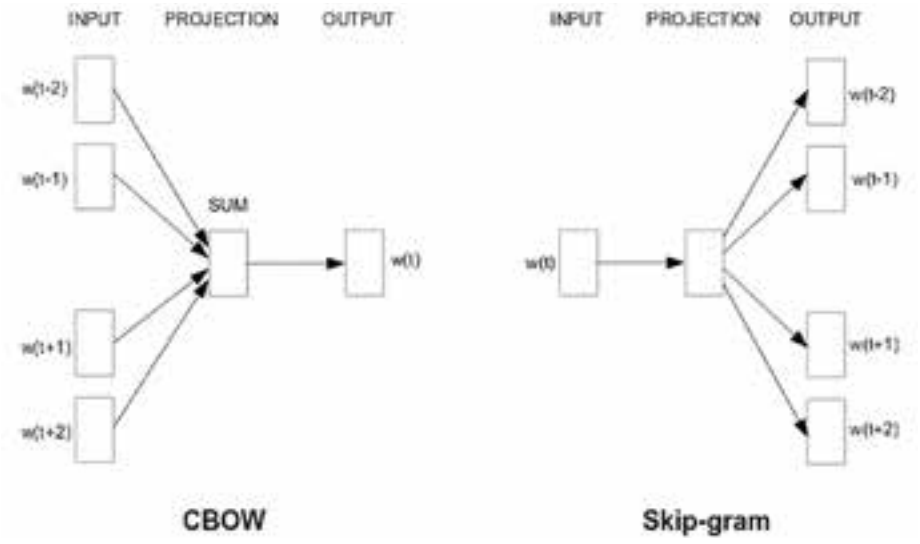
16 https://code.google.com/archive/p/word2vec/ (odczyt: 11.04.2018).

17 Odległość pomiędzy wektorami powinna być w tym przypadku rozumiana jako podobieństwo porównywanych tekstów: dwa dokumenty bliskie sobie w sensie tematyki leżą niedaleko siebie w przestrzeni wektorowej.

18 A. Singhal, *Modern Information Retrieval: A Brief Overview*, „Bulletin of the IEEE Computer Society Technical Committee on Data Engineering" 2011, vol. 24(4), s. 35–43.

Problem zamiany subdomen z profilu użytkownika na reprezentację wektorową można interpretować w kategoriach określenia kontekstu danego słowa w nazwie subdomeny lub też podobieństwa danej subdomeny do innych stron odwiedzanych przez użytkownika. Zastosowanie metodologii word2vec wydaje się zatem zasadne. Ze względu na przyjęte założenie pomijania kolejności, w jakiej były odwiedzane przez użytkownika strony z jego profilu (ale brania pod uwagę tylko ostatnio odwiedzanych stron internetowych), lepszy wydaje się algorytm CBOW i właśnie to podejście zostało wykorzystane w badaniu opisywanym w tym artykule.

Do rozwiązania tej części procedury klasyfikacyjnej została wykorzystana napisana w języku Python implementacja z biblioteki *gensim*[19], wykorzystująca dwuwarstwową sieć neuronową. Zazwyczaj jako rozmiar okna, które obejmuje słowa badane kontekstowo, wybiera się od 5 do 10 słów, natomiast wymiar wynikowej przestrzeni wektorowej pozostaje w zakresie 100–1000. Na potrzeby opisywanego badania przyjęto rozmiar okna równy 10, długość wektora zaś równą 200.



Rysunek 1. Porównanie algorytmów *continuous bag-of-words* oraz *skip-gram*

Źródło: T. Mikolov, K. Chen, G. Corrado, J. Dean, *Efficient Estimation of Word Representations in Vector Space*, 2013, arxiv.org/abs/1301.3781.

W wyniku działania algorytmu CBOW metody word2vec każdy z profili ponad 10 mln użytkowników jest reprezentowany przez macierz, której wierszami

[19] https://radimrehurek.com/gensim/models/word2vec.html (odczyt: 11.04.2018).

są odpowiedniki subdomen w wielowymiarowej przestrzeni wektorowej. Kolejnym etapem proponowanej klasyfikacji użytkowników ze względu na płeć jest przekształcenie takich macierzy na wektory. Podejście to ma co najmniej dwie zalety. Po pierwsze, pozwala na zmniejszenie rozmiaru danych do raptem 200 (przyjęta długość wektora) współrzędnych na użytkownika. Po drugie zaś, wektory reprezentujące profile użytkowników są jednocześnie standardowym formatem wejściowym dla ostatniego z etapów, tj. estymacji parametrów modelu regresji logistycznej. Zastosowanym w badaniu przekształceniem macierzy na wektor jest średnia ważona, przy czym za wagi przyjęto liczby odwiedzin każdej z subdomen z profilu.

Ostatnim etapem procedury klasyfikacji użytkowników ze względu na płeć jest estymacja parametrów modelu. Jako zmienne w modelu regresji logistycznej przyjęto współrzędne wektora odpowiadającego profilom użytkownika (po uśrednieniu dla danego użytkownika). W celu uzyskania bardziej stabilnych i wiarygodnych wyników zastosowano podział obserwacji na zbiór treningowy (80% danych) i testowy (20% danych), a następnie metodę dziesięciokrotnej walidacji[20] na treningowym zbiorze danych. Ten fragment procedury, tj. estymacja wraz z walidacją krzyżową, został implementowany z użyciem pakietu do uczenia maszynowego *scikit-learn*[21], dostępnego w języku Python.

Do oceny jakości predykcyjnej procedury klasyfikacyjnej zostały wybrane następujące metryki[22]: dokładność (ang. *accuracy*, ACC), pole pod krzywą ROC (ang. *Area Under the Receiver Operating Characteristic Curve*, ROC AUC) oraz F1 (F-measure).

4. Analiza empiryczna

Na podstawie danych zebranych o polskich użytkownikach Internetu w 2017 r. zastosowano procedurę kwalifikacji ze względu na płeć opisaną w poprzednim punkcie. Metryki jakości dopasowania modelu zostały przedstawione w tabeli 1.

---

[20] Patrz. np. R. Kohavi, *A study of cross-validation and bootstrap for accuracy estimation and model selection*, „Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence" 1995, vol. 2(12), s. 1137–1143.

[21] http://scikit-learn.org/stable/ (odczyt: 11.04.2018).

[22] D.M.W. Powers, *Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation*, „Journal of Machine Learning Technologies" 2011, vol. 2(1), s. 37–63.

Wszystkie metryki osiągają zbliżone wartości dla zbiorów treningowego i testowego, przekraczając 80%, co należy uznać za bardzo dobry wynik.

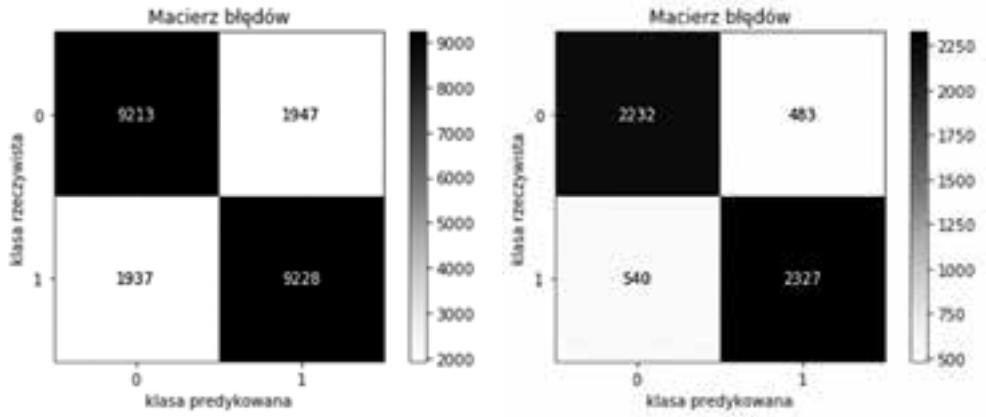
Tabela 1. Wartości metryk dopasowania modelu regresji logistycznej

Zbiór	Dokładność	ROC AUC	F1
Treningowy	0,8260	0,8961	0,8260
Testowy	0,8167	0,8807	0,8168

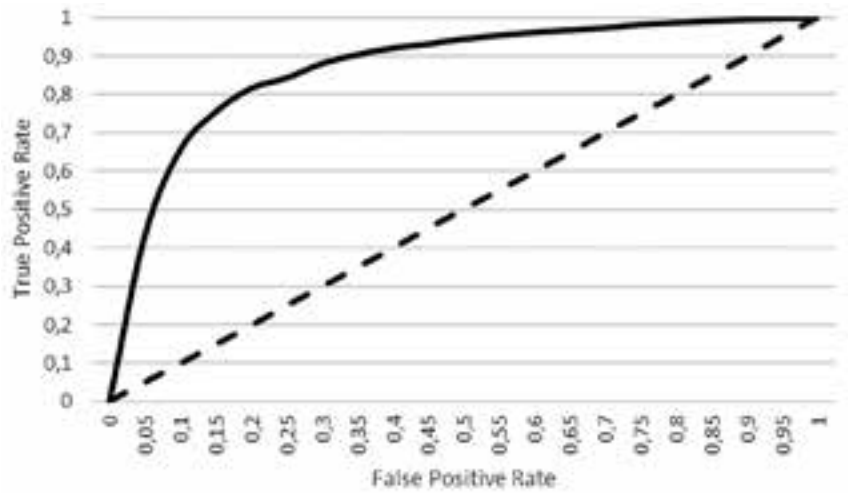
Źródło: opracowanie własne.

Tabele trafności dla obu zbiorów zostały podane w tabeli 2, natomiast wykres krzywej ROC został przedstawiony na rysunku 2. W każdym przypadku porównanie wskazuje na zdecydowaną przewagę proponowanego klasyfikatora (linia kropkowana) nad klasyfikatorem czysto losowym (linia przerywana). Każdy punkt krzywej ROC, odpowiadający tabeli trafności dla innego tzw. punktu odcięcia, wskazuje na dobrą skuteczność prognostyczną podejścia. Należy jednak wziąć pod uwagę, że m.in. ze względu na możliwość korzystania z jednego urządzenia przez więcej niż jedną osobę raczej trudno przypuszczać, iż możliwe jest osiągnięcie blisko stuprocentowej skuteczności klasyfikacji.

Tabela 2. Tabele trafności dla zbioru treningowego (po lewej) oraz testowego (po prawej)



Źródło: opracowanie własne.



Rysunek 2. Porównanie krzywej ROC dla modelu regresji logistycznej (linia ciągła) z modelem czysto losowym (linia przerywana)

Źródło: opracowanie własne.

Dzięki klasyfikatorowi oraz zastosowaniu metody word2vec można również przeprowadzić analizę preferencji oraz różnic i podobieństw w zachowaniu użytkowników Internetu. Ciekawe wydaje się m.in. sprawdzenie, z jakich stron internetowych najczęściej korzystają przedstawiciele obu płci. Listę 30 najczęściej odwiedzanych przez Polaków stron internetowych (subdomen) w 2017 r. z podziałem na płeć podano w tabeli 3. Okazuje się, że – zgodnie z przewidywaniami – mężczyźni znacznie częściej niż kobiety korzystają ze stron internetowych związanych z motoryzacją, sportem i elektroniką. Tymczasem kobiety przejawiają większe zainteresowanie domem, zdrowiem, gotowaniem i urodą.

Tabela 3. Wyniki działania algorytmu klasyfikacyjnego – 30 najczęściej odwiedzanych stron internetowych w 2017 r. przez Polaków z podziałem na płeć użytkownika

Płeć	Subdomeny
Mężczyźni	otomoto.pl, autokult.pl, moto.wp.pl, przegladsportowy.pl, auto-swiat.pl, sport.pl, moto.onet.pl, sportowefakty.wp.pl, autocentrum.pl, elektroda.pl, moto.pl, tech.wp.pl, komputerswiat.pl, gadzetomania.pl, sport.onet.pl, eurosport.onet.pl, opinie.wp.pl, moto.gratka.pl, wykop.pl, poszukaj. elektroda.pl, motoryzacja.interia.pl, technowinki.onet.pl, sprzedajemy.pl, o2.pl, wp.tv, komorkomania.pl, finanse.wp.pl, sport.interia.pl, m.sportowefakty.wp.pl, pilot.wp.pl

Płeć	Subdomeny
Kobiety	allani.pl, polki.pl, pl.nametests.com, domodi.pl, party.pl, forum.gazeta.pl, kobieta.onet.pl, plotek.pl, poradnikzdrowie.pl, fakt.pl, durszlak.pl, zdrowie.gazeta.pl, smaker.pl, wizaz.pl, ofeminin.pl, avanti24.pl, portal.abczdrowie.pl, przyslijprzepis.pl, infozdrowie24.pl, tipy.interia.pl, homebook.pl, nametests.com, mojegotowanie.pl, mojewypieki.com, niezwykle.pl, allrecipes.pl, kwestiasmaku.com, medonet.pl, polubione.pl, se.pl

Źródło: opracowanie własne.

Dodatkową zaletą prezentowanego podejścia i użycia algorytmu word2vec jest możliwość znalezienia stron internetowych o podobnej tematyce (dokładniej: subdomen, które leżą w bliskiej odległości od zadanej subdomeny, przy czym za miarę przyjęto odległość kosinusową w przestrzeni wektorowej). Mając do dyspozycji wektorowe reprezentacje subdomen ze wszystkich profili użytkowników (ponad 4 mln subdomen), można odszukać te najbliższe konkretnej subdomenie. Stanowi to duży krok w kierunku automatycznej segmentacji i kategoryzacji stron internetowych. Przykład zastosowania takiego podejścia został zaprezentowany w tabeli 4.

Tabela 4. Lista podobnych subdomen do danej subdomeny na podstawie ruchu internetowego w Polsce w 2017 r.

Subdomena	Podobne subdomeny
otomoto.pl	autoscout24.pl, forumsamochodowe.pl, chceauto.pl, motodiesel.pl, testy.mojeauto.pl, motonews.pl, citroen.auto.com.pl, renault.auto.com.pl, volvo.auto.com.pl, tuning1.com, gumtree.pl, samochody.mojeauto.pl, mercedes.auto.com.pl, gazeo.pl, forum.fordclubpolska.org, peugeot.auto.com.pl, mazdaspeed.pl
morizon.pl	domy.pl, oferty.net, nportal.pl, dom.money.pl, bezposrednie.com, noweinwestycje.pl, adresowo.pl, komercyjne.pl, mieszkania.trovit.pl, mieszkanie.mitula.com.pl, dom.gratka.pl
demotywatory.pl	kwejk.pl, joemonster.org, sadistic.pl, mistrzowie.org, komixxy.pl, faktopedia.pl, fabrykamemow.pl, piekielni.pl, chamsko.pl, retro.pewex.pl, mklr.pl, wiocha.pl, styl.fm, milanos.pl, fangol.pl, wgrane.pl, memy.pl, besty.pl, funny.pl, mh24.pl, elohell.net, kicze.pl, anonimowe.pl, wykop.pl, jbzdy.pl

Źródło: opracowanie własne.

5. Podsumowanie

Badanie zachowania użytkowników Internetu, ich preferencji oraz determinantów zaobserwowanych prawidłowości jest utrudnione z wielu powodów. Jednym z nich są olbrzymie wolumeny danych do przeanalizowania, innym z pewnością duża różnorodność użytkowników, jak również duży stopień anonimowości. Niemniej technologiczne możliwości śledzenie użytkowników są współcześnie bardzo rozwinięte. Celem badania opisanego w niniejszym artykule było przedstawienie algorytmu identyfikacji płci polskich użytkowników Internetu na podstawie profili złożonych z odwiedzanych przez nich stron internetowych. Wykorzystana metoda pozwala jednak na szersze zastosowania niż tylko klasyfikacja ze względu na płeć. Można wymienić m.in. następujące korzyści z prezentowanego podejścia:

- wysoka skuteczność (82%) klasyfikacji płci użytkownika Internetu na podstawie historii odwiedzanych przez niego stron,
- możliwość prześledzenia preferencji użytkowników danej płci, a tym samym opracowanie lepiej sprofilowanej reklamy lub akcji marketingowych skierowanych do osób danej płci,
- sposób na wyszukiwanie stron o podobnej tematyce, co może posłużyć do ograniczenia kosztów wyświetlania reklam na stronach poprzez dalej idącą personalizację,
- uniwersalność zastosowanego podejścia oraz jego pełna skalowalność (na większą liczbę profili, inne kraje itp.).

Zaprezentowane podejście dopuszcza wiele uogólnień oraz rozszerzeń. Można m.in.:

- wykorzystać do analizy niższy poziom hierarchii stron internetowych niż subdomena,
- użyć rozszerzonych danych o użytkownikach (nie tylko historię odwiedzanych stron), np. geolokalizację, informację o urządzeniu czy oprogramowaniu,
- na podstawie reprezentacji wektorowej subdomen z profili użytkowników dokonać klasyfikacji ze względu na wiek użytkownika, jego status materialny, zainteresowania,
- skonstruować kategorie podobnych stron, jeżeli chodzi o podobieństwo tematyczne adresów internetowych,
- przeprowadzić rozszerzone analizy zachowania użytkowników, biorąc pod uwagę tygodniowy czy dzienny tryb życia (znając dokładne czasy odwiedzin).

Bibliografia

Bengio Y., Ducharme R., Vincent P., Jauvin Ch., *A neural probabilistic language model*, „Journal of Machine Learning Research" 2003, vol. 3, s. 1137–1155.

Bernardelli M., *Cheater detection in Real Time Bidding system – panel approach*, „Roczniki" Kolegium Analiz Ekonomicznych 2015, nr 39, s. 11–23.

Fisher R.A., *The use of multiple measurements in taxonomic problems*, „Annals of Eugenics" 1936, vol. 7(2), s. 179–188.

Han J., Kamber M., Pei J., *Data Mining: Concepts and Techniques*, Morgan Kaufman, 2000.

Kohavi R., *A study of cross-validation and bootstrap for accuracy estimation and model selection*, „Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence" 1995, vol. 2(12), s. 1137–1143.

McTear M., Callejas Z., Griol Barres D., *The Conversational Interface. Talking to Smart Devices*, Springer, 2016.

Powers D.M.W., *Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation*, „Journal of Machine Learning Technologies" 2011, vol. 2(1), s. 37–63.

Singhal A., *Modern Information Retrieval: A Brief Overview*, „Bulletin of the IEEE Computer Society Technical Committee on Data Engineering" 2011, vol. 24(4), s. 35–43.

Źródła sieciowe

Hill F., Cho K., Jean S., Devin C., Bengio Y., *Embedding Word Similarity with Neural Machine Translation*, 2014, arXiv:1412.6448.

http://scikit-learn.org/stable/ (odczyt: 11.04.2018).

https://code.google.com/archive/p/word2vec/ (odczyt: 11.04.2018).

https://iab.org.pl/badania-i-publikacje/perspektywy-rozwojowe-reklamy-online-w-polsce-2017-2018/ (odczyt: 25.05.2018).

https://radimrehurek.com/gensim/models/word2vec.html (odczyt: 11.04.2018).

https://www.gemius.pl/wszystkie-artykuly-aktualnosci/wyniki-badania-gemiuspbi-za-kwiecien-2018.html (odczyt: 25.05.2018).

Levy O., Goldberg Y., *Linguistic Regularities in Sparse and Explicit Word Representations*, Proceedings of the Eighteenth Conference on Computational Natural Language Learning, Association for Computational Linguistics 2014, s. 171–180, https://aclanthology.coli.uni-saarland.de/papers/W14-1618/w14-1618.

Mikolov T., Chen K., Corrado G., Dean J., *Efficient Estimation of Word Representations in Vector Space*, 2013, arxiv.org/abs/1301.3781.

Mikolov T., Sutskever I., Chen K., Corrado G.S., Dean J., *Distributed representations of words and phrases and their compositionality. Advances in Neural Information Processing Systems*, 2013, https://arxiv.org/abs/1310.4546.

* * *

Anonymity on the Internet – identifying the gender of users based on the history of visited websites

Summary

In this article, a method of gender identification of Internet users was presented. The proposed method uses data from user profiles containing website addresses and the frequency of their visits. This approach combines the lexical analysis of the words from the Internet addresses, neural networks, mathematically sophisticated vector representation of the user profiles, and logistic regression as the main classifier. The empirical analysis was performed on the basis of 10 million profiles of Polish users, giving 82% of classification efficiency. Additional benefits from the study were the lists of the most preferred websites per gender in Poland in 2017, and the way of finding similar Internet portals, which can be used in personalized marketing as a source of savings in the form of reducing unnecessary expenses for badly targeted advertising.

Keywords: Internet, classification, user preferences, word2vec, Big Data

Zgodnie z oświadczeniem autorów, ich udział w przygotowaniu artykułu wyniósł: Łukasz Lipiński – 50%, Michał Bernardelli – 50%.