Ekonomiczne Problemy Usług nr 1/2017 (126), t. 2

ISSN: 1896-382X | www.wnus.edu.pl/epu
DOI: 10.18276/epu.2017.126/2-31 | strony: 305–313

Wiesław Wolny
Uniwersytet Ekonomiczny w Katowicach
e-mail: wieslaw.wolny@uekat.pl

Wielowymiarowa analiza mediów społecznościowych

Kod JEL: C88
Słowa kluczowe: media społecznościowe, analiza mediów społecznościowych, techniki odkrywania wiedzy
Streszczenie. Media społecznościowe w ostatnich latach przyciągają szczególną uwagę badaczy. Setki milionów ludzi spędza niezliczone godziny dzieląc się informacjami, opiniami, zdjęciami czy tworząc grupy zainteresowań. Jest to nowe, bogate źródło danych mające olbrzymi potencjał badawczy dla naukowców i do zastosowań praktycznych. W artykule zawarto koncepcję metody analizy mediów społecznościowych opartą na wielu wymiarach. Wymiary te obejmują analizę: danych tekstowych, użytkowników, sieci społecznościowych, danych geograficznych i obrazów.

Wprowadzenie

Serwisy społecznościowe w ostatniej dekadzie stały się znaczącym sposobem korzystania z internetu i tym samym bogatym źródłem wszelakiego rodzaju informacji. Informacja w mediach społecznościowych jest tworzona i publikowana przez miliony osób dziennie. Oprócz samego współdzielenia się treścią, media społecznościowe umożliwiają interakcje polegające na obserwowaniu, komentowaniu i dalszemu udostępnianiu treści.

Gigantyczne i stale rosnące zbiory danych mediów społecznościowych pochodzących z portali, takich jak Facebook, Twitter, społeczności gier komputerowych, blogów, poczty elektronicznej są często udostępniane dla badaczy. Niestety tradycyjne metody analizy danych nie sprawdzają się najlepiej w odniesieniu do mediów społecznościowych. Pomijając rozmiar, dynamizm i charakter danych typowych dla big-data, analiza ich nie jest zadaniem łatwym z powodu braku ustrukturalizowania, specyfiki używanego języka, stosowania skrótów czy częstych błędów i literówek.

Badania sieci społecznościowych od lat były prowadzone w dziedzinie nauk społecznych. Ich gwałtowny rozwój nastąpił wraz z dostępnością olbrzymich ilości danych pochodzących z portali społecznościowych.

Analiza mediów społecznościowych stała się dziedziną interdyscyplinarną, obejmującą odkrywanie wiedzy, uczenie maszynowe, eksplorację tekstu, analizę sieci społecznościowych, analizę grafiki i czasami wiele innych. Nowa dziedzina wymaga wypracowania wzorców, schematów prowadzenia badań.

1. Wymiary analizy mediów społecznościowych

Analiza mediów społecznościowych staje się popularną dziedziną badań naukowych oraz zastosowań praktycznych. Większość badań koncentruje się tylko na wybranych aspektach badania sieci społecznościowych. Najczęściej prace dotyczą analizy treści, opinii, wydźwięku oraz struktury sieci.

Współczesne media społecznościowe dają badaczom dostęp do nowych form analizy. Celowe zatem staje się stworzenie ram, w postaci wymiarów, według których analiza mediów społecznościowych może być wykonywana.

Media społecznościowe można analizować w wielu wymiarach – pierwszym z nich jest analiza danych tekstowych. Istnieje wiele podejść do analizy tekstu w mediach społecznościowych. Zwykle koncentruje się na klasyfikacji dokumentów, wykrywaniu treści czy analizie wydźwięku. Media społecznościowe dostarczają wielu informacji o współtworzących je osobach, dlatego drugim wymiarem jest analiza użytkowników. Ludzie w mediach społecznościowych wchodzą w wiele różnych relacji między sobą. Analizie tych powiązań poświęcony jest trzeci wymiar – wymiar sieci społecznościowych. Internet i media z nim związane może być również analizowany w wymiarze geograficznym, Stanowi to czwarty wymiar w przyjętej metodzie. Oprócz tekstu uczestnicy sieci społecznościowych bardzo często udostępniają zdjęcia – ich analiza może być kolejnym wymiarem badań.

Oprócz powyższych, można również identyfikować inne wymiary. Można badać odnośniki (linki hipertekstowe) umieszczone w publikowanych wpisach. W miarę potrzeb można interpretować zdarzenia na podstawie analizy szeregów czasowych.

2. Wymiar danych tekstowych

Większość mediów społecznościowych oparta jest na wymianie informacji tekstowych. Do analizy tych danych oczywiste wydaje się wiec wykorzystanie technik przetwarzania języka naturalnego (*Natural Language Processing* – NLP). Klasyczne metody przetwarzania języka naturalnego nie zawsze przynoszą dobre efekty w badaniu mediów społecznościowych. Krótkość tekstu, nieformalny język, znaczna liczba błędów w tekście powoduje, że analiza tekstu metodami NLP jest trudnym zadaniem.

Nie zawsze też potrzebne są wyniki w takiej postaci, jaką oferuje NLP. Często zamiast dokładnego zrozumienia treści wystarczą tylko łatwiejsze do uzyskania informacje, jak wydźwięk wypowiedzi, klasyfikacja tekstów, identyfikacja tematów. W badaniu mediów społecznościowych dlatego najczęściej stosuje się techniki nazywane jako eksploracja tekstu (*Text Mining*). Eksploracja tekstu zaliczana jest do metod eksploracji danych (*Data Mining*), z tą podstawową różnicą, że metody eksploracji danych skupiają się na przetwarzaniu przede wszystkim danych numerycznych, a eksploracja tekstu dotyczy danych tekstowych. Dane tekstowe przetwarzane są jednak na postać umożliwiającą ich analizę numeryczną w celu wyliczenia statystyk i odkrywania zależności w tekście.

Metody eksploracji tekstu wykształciły jednak własne, specyficzne, tylko dla tekstu, techniki przetwarzania i wizualizacji, wśród których do podstawowych można zaliczyć:

- klasyfikację dokumentów,
- ekstrakcję informacji,
- wykrywanie trentów i tematów,
- analizę wydźwięku,
- analizę emocji.

W mediach społecznościowych pojedynczy wpis dotyczy zwykle jednego konkretnego tematu, dlatego uzasadnione jest traktowanie każdego wpisu całościowo jako dokumentu i analizowanie go na tym poziomie. W taki sposób na przykład, analizuje się również e-maile w celu wykrywania spamu. Pierwszym zadaniem jest klasyfikacja tekstu. W najprostszym przypadku może polegać na wspomnianym klasyfikowaniu e-maili jako spam. W bardziej wyszukanych systemach, dokumenty można klasyfikować do określonych kategorii lub pod względem podobieństwa.

Dokumenty tekstowe mogą zawierać wiele informacji. Problemem w ich wydobyciu jest nie ustruktralizowana forma dokumentów. Dane w tekście mogą znajdować się w dowolnym miejscu i mogą nie być opisane identycznie w różnych dokumentach. W celu usunięcia informacji można jednak uruchomić specjalne procesy analizy tekstu. Istniejące systemy pozwalają np. identyfikować spółki notowane na giełdzie i dotyczące ich podstawowe informacje ekonomiczne publikowane w wiadomościach na portalu Twitter.

Analizując wpisy w mediach społecznościowych możliwa jest identyfikacja zyskujących na popularności tematów dyskusji. Pozwala to oceniać co jest aktualnie punktem zainteresowania społeczności w określonym regionie. Zwykle analizę trendów i tematów wykonuje się na poziomie państw, ale oczywiście możliwe jest zawężenie badań do wybranych rejonów czy miast. Analiza tematów uznawana jest jako istotne źródło badania opinii publicznej – dane tego typu wykorzystuje się w wielu krajach do przewidywania wyników wyborów. Analiza trendów jest również istotna dla branży medialnej – jej celem jest zidentyfikowanie najważniejszych tematów w aktualnie toczących się dyskusjach.

Wykrywanie tematów i trendów dokonuje się na podstawie statystyk występowania poszczególnych słów. Proces badawczy polega na przeglądaniu dużych zbiorów wpisów na mediach społecznościowych, identyfikacji skupień słów, grupowania ich na podstawie podobieństwa i identyfikacji wyłaniających się wzorców. Do automatycznego odkrywania tematów można wykorzystać popularny algorytm Latent Dirichlet Allocation (LDA) (Bley, Ng, Jordan, 2013). Innym dobrym narzędziem wykrywania tematów i trendów jest IBM Watson Content Analytics (Zhu i in., 2011), w którym wykorzystywane są te same technologie przetwarzania języka naturalnego, co najbardziej zaawansowany system odpowiadający na pytania IBM Watson DeepQA.

Ideą sieci społecznościowych jest dzielenie się w czasie rzeczywistym opiniami na różne tematy, dyskutowanie o bieżących wydarzeniach czy wyrażanie opinii o produktach wykorzystywanych w codziennym życiu. Dane zgromadzone w mediach społecznościowych mogą więc być cennym źródłem analizy wydźwięku wypowiedzi (*sentiment analysis*). Analiza taka może być szeroko wykorzystywana w wielu dziedzinach.

Wraz ze wzrostem popularności mediów społecznościowych analiza wydźwięku staje się polem zainteresowań wielu badaczy i praktyków, może być wykonywana na wielu poziomach. Zaczynając od poziomu dokumentów (Pang, Lee, 2004), poprzez poziom zdań (Kim, Hovy, 2004) i bardziej współcześnie – poziom wyrażeń (Wilson, Wiebe, Hoffmann, 2009) do poziomu słów (Hatzivassiloglou, McKeown, 1997).

Analiza wydźwięku obejmuje zwykle skalę dwupoziomową (wydźwięk pozytywny lub negatywny), czasem pośrodku wyodrębnia się neutralny wydźwięk wypowiedzi. Ludzkie emocje są jednak o wiele bardziej zróżnicowane. Rozwijanie badań o poszerzony zakres emocji jest nową koncepcją i obecnie niezbyt obszernie zbadaną. Socher, Pennington, Huang, Ng, Manning (2011) identyfikowali pięć typów emocji. Tromp i Pechnizkiy (2015) użyli modelu koła emocji Pluchnika i systemu regułowego do wykrywania emocji w tekście. W pracy (Wolny, 2016) do wykrywania emocji wykorzystano emotikony i znaki emocji, co pozwoliło na identyfikację kilkunastu rodzajów emocji.

3. Wymiar użytkowników

W celu przedstawienia swojej osoby w społeczności internetowej, użytkownicy w mediach społecznościowych tworzą dla siebie profile. Gdy użytkownik zakłada lub konfiguruje swój profil, dostarcza pewnych informacji o sobie, takich jak imię, nazwisko, nazwa użytkownika, hasło, adres e-mail czy numer telefonu. Użytkownik może również wprowadzić do informacji profilowej takie dane, jak krótka biografia, miejsce pobytu, adres domowej strony internetowej, datę urodzenia, zdjęcie.

Profile użytkowników są więc cennym źródłem informacji o osobach korzystających z mediów społecznościowych. Część informacji z profilu użytkownika, zwykle imię, nazwisko, nazwa użytkownika, miejsce pobytu są dostępne publicznie i mogą być

wyszukane przez wszystkich. Pozostałymi danymi dysponuje tylko właściciel serwisu społecznościowego.

Dla przykładu Twitter dostarcza w swoim API funkcję *user show*, pozwalającą wszystkim zainteresowanym uzyskać o każdym koncie użytkownika takie dane jak: imię i nazwisko, opis, zawierający zwykle dodatkowe informacje o użytkowniku, obiekty jak hasztagi, linki czy pliki medialne, które mogą prowadzić do dalszych źródeł informacji, liczba obserwujących, liczba znajomych, położenie geograficzne i język.

Cenniejsza niż funkcja zwracająca dane użytkownika jest kolejna funkcja serwisu Twitter – *follower list*, umożliwiająca dostęp do listy osób obserwujących dany profil. Funkcja ta zwraca zbliżone dane jak *user show*, lecz dla wszystkich osób obserwujących. Użycie tej funkcji dla najbardziej popularnych profili na Twitterze, gromadzących po kilka milionów osób obserwujących pozwala łatwo zgromadzić dane o milionach użytkowników, bez przekraczania limitów narzuconych przez API Twittera. Analizując te dane można bez trudu utworzyć sieci użytkowników, będące podstawą następnego wymiaru analizy.

4. Wymiar sieci społecznościowych

Sieć społecznościowa jest społeczną strukturą ludzi powiązanych bezpośrednio lub pośrednio między sobą poprzez wspólne relacje lub interesy. Analiza sieci społecznościowych jest badaniem struktury sieci społecznościowych i zachowań w nich ich członków. Do analizy sieci społecznościowych wykorzystuje się teorię grafów, które są skuteczną metodą analizy bardzo dużych zbiorów danych, czym charakteryzują się media społecznościowe.

Klasyczna analiza sieci społecznościowych skupia się na strukturze sieci. W sieciach społecznościowych węzłem grafu jest osoba (aktor), a wzajemne relacje między osobami są krawędzią grafu. Relacje mogą być wszelakiego rodzaju i przyjmować różne wartości nasilenia. W ramach sieci mogą występować również różne zachowania i działania. Po zidentyfikowaniu wszystkich osób i łączących ich relacji, do analizy sieci mogą być użyte różne miary (statystyki).

Podstawowe rodzaje sieci w mediach społecznościowych można zidentyfikować za Kumar, Morstatter i Liu (2013) jako:

- sieci przepływu informacji,
- sieci znajomych i obserwujących.

Pierwszy rodzaj sieci obrazuje osoby, które cytują, przesyłają dalej lub odpowiadają na wiadomość. Drugi rodzaj sieci oparty jest na liście znajomych i liście osób obserwujących daną osobę na portalach społecznościowych. Innym typem sieci mogą być sieci związane z wybranym wydarzeniem, np. konferencją. Wiele tego typu wydarzeń identyfikowanych jest w komunikacji sieciowej przez hasztagi lub słowa kluczowe. Hansen, Smith i Shneiderman (2011) stworzyli EventGraphs, narzędzie pozwalające

gromadzić kontakty, grupując je według przeprowadzonych dyskusji, dodania do obserwowanych i wybranych słów kluczowych lub hasztagów.

Jednym z celów analizy grafów jest identyfikacja centralnych węzłów, które w mediach społecznościowych interpretuje się jako osoby mające największe oddziaływanie na społeczność (Ghosh, Lerman, 2011).

Najpopularniejszymi narzędziami do analizy i wizualizacji sieci społecznościowych są NodeXL i Gephi. NodeXL (Hansen, 2010) jest darmowym dodatkiem do Microsoft Excel pozwalającym na przeglądanie, odkrywanie i eksplorację sieci. Gephi jest oprogramowaniem *open source* do analizy i wizualizacji sieci.

5. Wymiar geograficzny

Media społecznościowe mogą być również rozpatrywane w kategorii geograficznego rozproszenia ich użytkowników. Gromadzenie informacji lokalizacyjnych pozwala na wartościową analizę przepływu informacji czy geograficznego zasięgu sieci.

Ważnym aspektem sieci społecznościowych jest oznakowanie geograficzne wielu generowanych informacji. Większość portali, tak jak Twitter, Facebook zapisują geograficzną lokalizację użytkownika w momencie publikacji wpisu. Ponadto zdjęcia robione współczesnymi smartfonami i publikowane na mediach społecznościowych zawierają wpisane w metadane exif współrzędne geograficzne.

Osobną, niemniej bardzo znaczącą, grupę tworzą portale społecznościowe z zasady związane z logowaniem położenia geograficznego użytkowników. Portale rejestrujące aktywność fizyczną i sportową, takie jak Endomondo, Strava, portale turystyczne jak gpsis.com i wiele innych pozwalają na rejestrację przebytych tras, dostarczając tym samym wielu cennych informacji geolokalizacyjnych.

Informacje o położeniu geograficznym użytkowników mediów społecznościowych mogą być podstawą wielu analiz. Informacja, gdzie znajduje się osoba umieszczająca wpis na portalu może pozwolić lepiej zrozumieć do czego odnosi się dany wpis.

Dane geograficzne w celu lepszego wglądu, są najczęściej prezentowane w postaci wizualnej. Oczywistym sposobem ich prezentacji są mapy. Podstawowa metoda wizualizacji polega na utworzeniu mapy wpisów w mediach społecznościowych. Każdy wpis może być identyfikowany jako punkt na mapie. Zakres prezentowanych informacji może być zwiększony przez użycie różnych kolorów dla poszczególnych kategorii wpisów. Innym sposobem jest rysowanie kół o rozmiarze reprezentującym zagregowaną liczbę wpisów.

6. Wymiar grafiki

Zdjęcia są bardzo popularnym sposobem dzielenia się treścią w mediach społecznościowych. Szacuje się (Donnelly, 2016), że codziennie umieszcza się w mediach

społecznościowych 1,8 mld zdjęć a 10% wszystkich zdjęć w całej historii zostało zrobionych w ostatniej dekadzie.

W ostatnich latach zmieniła się funkcja obrazów w procesie komunikacji. Obrazy coraz częściej używane są jako główny środek wyrażania, a tekst jest jedynie krótkim, pomocniczym opisem. Powstały portale społecznościowe, takie jak Instagram, Fickr, Tumblr, Pinterest czy Snapchat, koncentrujące się na zdjęciach jako głównej formie przekazu opinii.

Z badawczego punktu widzenia, analiza obrazów jest nowym wyzwaniem w dziedzinie mediów społecznościowych. Przy analizie portali opartych na przesyłaniu grafiki, nie można już polegać na analizie tekstu w celu zrozumienia zawartości obrazów. W wielu przypadkach scenariusz jest odwrotny. Należy przeanalizować obraz, by móc zrozumieć znaczenie tekstu.

Dodatkową trudnością jest mała dostępność oprogramowania do analizy obrazów. Najlepsze rozwiązania tego typu są własnością firm Google i Facebook, ale są one w ograniczonym stopniu dostępne dla niepowiązanych z tymi firmami badaczy. Dla szerszego grona zainteresowanych dostępne jest oprogramowanie IBM Watson, które wykorzystując algorytmy *deep learning*, pozwala na zidentyfikowanie zawartości obrazów i opisanie ich w postaci tagów.

Oprogramowanie do analizy obrazów może mieć zastosowanie w marketingu do np. wykrywanie logo na zdjęciach. Podobnie analiza wieku i płci osób na zdjęciach może być wykorzystywana do identyfikacji rodzin.

Podsumowanie

Analiza mediów społecznościowych jest nowym wyzwaniem badawczym. Kompleksowa analiza mediów społecznościowych jest interdyscyplinarna i wymaga uwzględnienia wielu różnych technik badawczych. Ustalenie podstawowych obszarów badań w postaci proponowanych wymiarów pozwoli na wszechstronną, lecz ujednoliconą analizę zjawiska. Umożliwi również stworzenie platformy narzędziowej pozwalającej na prowadzenie badań w sposób powtarzalny.

Bibliografia

Blei, D.M., Ng, A.Y., Jordan, M.I. (2003). Latent Dirichlet Allocation. *The Journal of Machine Learning Research*, *3*, 993–1022.

Donelly III, J. (2016). *The future of social intelligence: image recognition and analysis*. Pobrane z: http://marketingland.com/future-social-intelligence-image-recognition-analysis-179902 (2.01.2017).

Ghosh, R., Lerman, K. (2011). Parameterized centrality metric for network analysis. *Physical Review E*, *83* (6), 66118. Pobrane z: https://doi.org/10.1103/PhysRevE.83.066118.

Hansen, D.L., Smith, M.A., Shneiderman, B. (2011). *EventGraphs: Charting Collections of Conference Connections*. 44th Hawaii International Conference on Systems Science (HICSS-44 2011), Proceedings, 4–7.01.2011. Koloa, Kauai, HI, USA (s. 1–10). Pobrane z: https://doi.org/10.1109/HICSS.2011.196.

Hansen, D., Shneiderman B., Smith M.A. (2010). *Analyzing Social Media Networks with NodeXL: Insights from a Connected World*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Hatzivassiloglou, V., McKeown, K.R. (1997). Predicting the Semantic Orientation of Adjectives. *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics* (s. 174–181). Madrid, Spain: Association for Computational Linguistics. Pobrano z: https://doi.org/10.3115/976909.979640.

Kim, S.M., Hovy, E. (2004). Determining the Sentiment of Opinions. *Proceedings of the 20th International Conference on Computational Linguistics. Geneva, Switzerland: Association for Computational Linguistics*. Pobrano z: https://doi.org/10.3115/1220355.1220555.

Kumar, S., Morstatter, F., Liu, H. (2013). *Twitter Data Analytics*. Springer.

Pang, B., Lee, L. (2004). A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts. *Proceedings of the 42Nd Annual Meeting on Association for Computational Linguistics*. Barcelona, Spain: Association for Computational Linguistics. Pobrano z: https://doi.org/10.3115/1218955.1218990.

Socher, R., Pennington, J., Huang, E.H., Ng, A.Y., Manning, C.D. (2011). Semi-supervised Recursive Autoencoders for Predicting Sentiment Distributions. *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (s. 151–161). Edinburgh, United Kingdom: Association for Computational Linguistics. Pobrano z: http://dl.acm.org/citation.cfm?id=2145432.2145450

Tromp, E., Pechenizkiy, M. (2015). Pattern-Based Emotion Classification on Social Media. W: *Advances in Social Media Analysis*, t. 602 (s. 1–20). Springer. Pobrano z: http://dx.doi.org/10.1007/978-3-319-18458-61

Wilson, T., Wiebe, J., Hoffmann, P. (2009). Recognizing Contextual Polarity: An Exploration of Features for Phrase-level Sentiment Analysis. Computational Linguistics, *35* (3), 399–433.

Wolny, W. (2016). *Emotion Analysis of Twitter Data That Use Emoticons and Emoji Ideograms*. International Conference on Information Systems Development (ISD). Pobrane z: http://aisel.aisnet.org/isd2014/proceed- ings2016/CreativitySupport/5 (1.12.2016).

Zhu, W.D., Iwai, A., Leyba, T., Magdalen, J., McNeil, K., Nasukawa, T., Redbooks, I. (2011). *IBM Content Analytics Version 2.2: Discovering Actionable Insight from Your Content*. *Vervante*. Pobrane z: https://books.- google.pl/books?id=MRnCAgAAQBAJ (1.12.2016).

MULTIDIMENTIONAL SOCIAL MEDIA ANALYSIS

Keywords: social media, social media analysis, data mining techniques
Summary. Social media has gained prominent attention in the last years. Hundreds of millions of people spending countless hours on social media to communicate, interact, share pictures and create groups of interests. Social media has become rich source of data for analysis to scientists and practitioners. Concept of multidimensional analysis of social media is presented in the article. Dimensions of analysis includes text analysis, user analysis, user networks analysis, geospatial analysis and picture analysis.

Translated by Wiesław Wolny

Cytowanie

Wolny, W. (2017). Wielowymiarowa analiza mediów społecznościowych. *Ekonomiczne Problemy Usług*, *1* (126/2), 305–313. DOI: 10.18276/epu.2017.126/2-31.