Jacek Maślankowski[1]

Automatyczna analiza treści nieustrukturyzowanej jako przykład źródła informacji dla administracji publicznej

1. Wstęp

Współczesne narzędzia analityczne pozwalają na przetwarzanie danych wielu rodzajów. Przykładem danych tego typu są dane nieustrukturyzowane, pochodzące z serwisów społecznościowych oraz komentarze zamieszczane pod artykułami prasowymi. Według typologii UNECE są one zaliczane do danych tworzonych przez ludzi. W przeciwieństwie do dwóch pozostałych rodzajów danych, generowanych przez urządzenia oraz powstających w procesach biznesowych, stanowią one znaczący odsetek analiz Big Data. Mając do dyspozycji narzędzia Big Data, analitycy są w stanie przetwarzać duże zbiory danych w sposób inny niż tradycyjny.

W niniejszym artykule opisano przetwarzanie danych Big Data, które mają charakter nieustrukturyzowany i napływają nieustannie. Analiza tych danych jest o tyle trudna, że wymaga zastosowania wielu nowatorskich metod służących do wyodrębniania najważniejszych słów kluczowych z danego komentarza czy wpisu. W kolejnym etapie należy przeprowadzić automatyczną analizę, z wykorzystaniem metod Machine Learning, która dostarczy odpowiednio przygotowane dane.

Celem głównym niniejszego artykułu jest zaprezentowanie możliwości wykorzystania narzędzi Big Data do analizy treści nieustrukturyzowanych poprzez analizę sentymentu. Jest to analiza danych jakościowych pozwalająca na klasyfikowanie treści wiadomości, np. jako pozytywnych, negatywnych lub neutralnych. Celem cząstkowym jest również zaprezentowanie wyników pozyskiwanych w taki sposób. Uzupełnieniem artykułu jest ocena wiarygodności danych przez zaprezentowanie wyzwań i zagrożeń w związku z zastosowaniem tego rodzaju narzędzi.

[1] Uniwersytet Gdański, Wydział Zarządzania.

Niniejszy artykuł został podzielony na pięć części. Po wstępie, w części drugiej znajduje się opis możliwości wykorzystania narzędzi Big Data w administracji publicznej. Część trzecia stanowi rozwinięcie części drugiej poprzez opis zastosowania analizy sentymentu w typowych działaniach podejmowanych w organizacjach. W kolejnym podrozdziale zawarto opis studium przypadku, przygotowanego na potrzeby niniejszego artykułu. Część piąta to analiza możliwości zastosowania tego rodzaju rozwiązania przez prezentację wyzwań i zagrożeń wynikających z używania go. Ostatnia, szósta część, stanowi podsumowanie oraz prezentację wniosków z przeprowadzonego badania.

Zawarte w niniejszym artykule studium przypadku odnosi się do wykorzystania narzędzi Big Data do analizy komentarzy i wpisów w mediach społecznościowych. Ze względu na reprezentacyjność tego rozwiązania zdecydowano się na wykorzystanie jako medium społecznościowego Twittera, a jako medium oferujące komentarze – typowe medium jakimi są najpopularniejsze portale w Polsce.

Hipoteza postawiona w niniejszym artykule jest następująca: zastosowanie metod *text mining* i *machine learning* pozwala uzyskać wiarygodne informacje na temat ogólnej opinii dotyczącej podejmowanych inicjatyw i zjawisk zachodzących w otoczeniu.

2. Zastosowanie Big Data w organizacjach

Typowe zastosowanie Big Data w biznesie ma na celu zwiększeie korzyści inwestorów, m.in. przez prowadzenie akcji marketingowych i ich ocenę z wykorzystaniem narzędzi do inżynierii danych (ang. *data science*)[2]. Należy zwrócić uwagę, że w wielu opracowaniach pojęcia Big Data, inżynieria danych oraz zarządzanie danymi są prezentowane jako jedna spójna całość[3]. Podobnie jak w tradycyjnych zbiorach danych, taka analiza wiąże się z wieloma problemami, takimi jak reprezentatywność danych, pomijanie istotnych zmiennych i konieczność przeprowadzania imputacji brakujących danych[4].

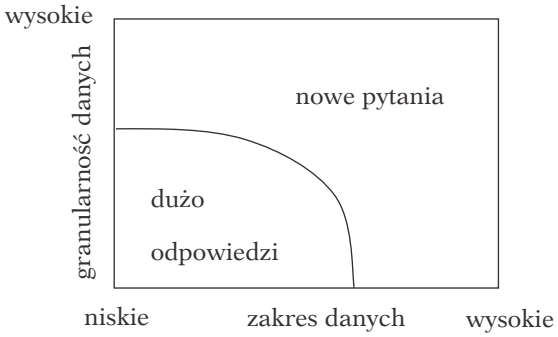
---

[2] P. Chintagunta, D. Hanssens, J. Hauser, *Marketing and Data Science: Together the Future is Ours*, "Gfk-Marketing Intelligence Review" 2016, vol. 8, no. 2, pp. 18–23.

[3] G. George, E. Osinga, D. Lavie, B. Scott, *Big Data and Data Science Methods for Management Research*, "Academy of Management Journal" 2016, no. 10.

[4] V. Bosch, *Big Data in Market Research: Why More Data Does Not Automatically Mean Better Information*, "Gfk-Marketing Intelligence Review 2016, vol. 8, no. 2, pp. 56–63.

Pomocny w zrozumieniu istoty zarządzania dużymi zbiorami danych może być rysunek 1, który przedstawia zależność pomiędzy wielkością danych a pewnością informacji.



Rysunek 1. Wielkość danych a granularność danych

Źródło: opracowanie własne na podstawie G. George, E. Osinga, D. Lavie, B. Scott, *Big Data and Data Science Methods for Management Research*, "Academy of Management Journal" 2016, no. 10.

Przedstawiony rysunek 1 wskazuje, że im bardziej szczegółowe dane posiada organizacja, tym więcej powstaje pytań, a mniej dobrych odpowiedzi. Oczekiwanym rozwiązaniem byłoby zatem przygotowanie zestawu danych, których zakres miałby granularność dedykowaną do rozwiązania danego problemu. Jest to jednak trudne do wykonania z zastosowaniem narzędzi Big Data.

Osobnym, wartym rozważenia aspektem są prawne uwarunkowania możliwości zastosowania Big Data przez jednostki administracji publicznej. W praktyce zawartość stron internetowych podlega prawnej ochronie, a ich pobieranie i przechowywanie na dyskach lokalnych może być nielegalne. Należy w szczególności zwrócić uwagę na bazy danych, które znajdują się na stronach internetowych, planowanych dla potrzeb web scrapingu, czyli pobierania danych ze stron internetowych. Wówczas należy mieć na uwadze zapisy określonego aktu prawnego, np. Ustawa o ochronie baz danych[5]. Konieczne jest zatem uzyskanie odpowiedzi na następujące pytania[6]:

- W jaki sposób organizacja przetwarza dane Big Data?
- Jak traktowane są dane osobowe, jeżeli są zbierane?

[5] Ustawa z dnia 27 lipca 2001 r. o ochronie baz danych (DzU 2001, nr 128, poz. 1402), Internetowy System Aktów Prawnych, http://isap.sejm.gov.pl/DetailsServlet?id=WDU20011281402 (3.11.2016).

[6] S. White, *6 Ethical Questions about Big Data*, "Journal Of Accountancy" 2016, vol. 222, no. 4, pp. 1–2.

- Czy organizacja ocenia ryzyko wykorzystywania określonego zbioru danych?
- Czy istnieją procedury łagodzenia ryzyka związanego z niewłaściwym wykorzystaniem danych?
- Czy pomiar ryzyka związanego z wykorzystaniem danych jest efektywnie monitorowanych?
- Czy organizacja posiada opracowane reguły udostępniania zgromadzonych danych dla podmiotów zewnętrznych?

Odpowiedź na tak postawione pytania pozwala na opracowanie przejrzystych reguł zastosowania Big Data w organizacji. Reguły te powinny zostać sformułowane w postaci dokumentu i opublikowane do użytku wewnętrznego dla pracowników organizacji.

3. Analiza sentymentu

Analiza sentymentu (ang. *sentiment analysis*) pozwala określić charakter analizowanego tekstu przez wskazanie na pozytywny lub negatywny aspekt wypowiedzi. Bardzo często analiza sentymentu jest stosowana do przetwarzania treści wpisów w mediach społecznościowych, np. na Twitterze[7]. Zastosowanie analizy sentymentu ma jednak szerszy aspekt i może dotyczyć nawet analizy danych pochodzących z rocznych raportów przedsiębiorstw[8]. Bardziej złożone analizy mogą dotyczyć zjawisk zachodzących po globalnym kryzysie ekonomicznym[9]. Jednak najczęściej ocena sentymentu jest wykorzystywana przy analizie opinii konsumentów. Jednym z takich zastosowań jest analiza wpływu komentarzy użytkowników telefonów komórkowych na ich sprzedaż[10]. Przykład analizy sentymentu, do które jako źródło danych wykorzystano portal Twitter.com, został przedstawiony na rysunku 2.

Przykład pokazany na rysunku 2 wskazuje, że ekstrakcja i klasyfikacja danych może następować na dwa sposoby – z wykorzystaniem algorytmów typu

---

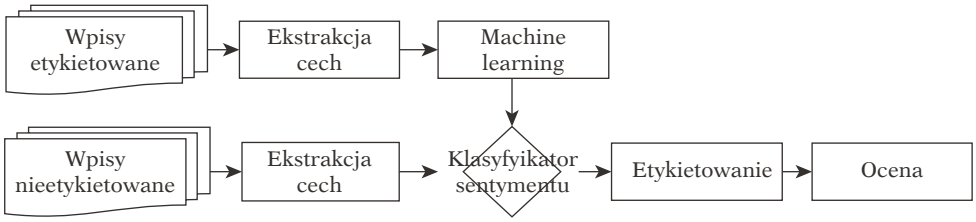
[7] S. Makarem, H. Jae, *Consumer Boycott Behavior: An Exploratory Analysis of Twitter Feeds*, "Journal Of Consumer Affairs" 2016, vol. 50, no. 1, pp. 193–223.

[8] M. Pagliarussi, M. Aguiar, F. Galdi, *Sentiment Analysis in Annual Reports from Brazilian Companies Listed at the Bm&fbovespa*, "Base" 2016, vol. 13, no. 1, pp. 53–64.

[9] E. Lozza, A. Bonanomi, C. Castiglioni, A. Bosio, *Consumer Sentiment After the global Financial Crisis*, "International Journal Of Market Research" 2016, vol. 58, no. 5, pp. 671–691.

[10] T. Liang, X. Li, C. Yang, M. Wang, *What in Consumer Reviews Affects the Sales of Mobile Apps: A Multifacet Sentiment Analysis Approach*, "International Journal Of Electronic Commerce" 2015, vol. 20, no. 2, pp. 236–260.

Machine Learning, jak również przez zautomatyzowaną klasyfikację sentymentów. W przypadku tego pierwszego rozwiązania algorytmy wykorzystują oznaczenia informacji zawartej we wpisie, np. hashtagi i inne. W drugim przypadku ekstrakcja tekstu następuje tylko na podstawie treści wpisu. Trzeba zaznaczyć, że w obu przypadkach efektem działania algorytmu jest zaklasyfikowanie treści do odpowiedniego sentymentu.



Rysunek 2. Analiza sentymentu z wykorzystaniem medium społecznościowego Twitter

Źródło: opracowanie własne na podstawie A. Giachanou, F. Crestani, *Like It or Not: A Survey of Twitter Sentiment Analysis Methods*, "ACM Computing Surveys" 2016, vol. 49, no. 2, pp. 28–41.

Należy zwrócić uwagę, że medium społecznościowe Twitter jest relatywnie łatwe do wykorzystania w analizie, gdyż pojedynczy wpis nie przekracza 140 znaków[11]. Dodatkowo dostęp do wpisów może się odbywać za pomocą interfejsów API (ang. *Application Programming Interface*), przygotowanych przez właściciela portalu[12].

Pisząc o analizie sentymentu, należy zwrócić uwagę, że istnieją również mierniki sentymentu, których zastosowanie bazuje na danych ustrukturyzowanych. Wówczas ich wykorzystanie może mieć miejsce m.in. w zakresie przewidywania koniunktury na rynku, np. odnośnie do sprzedaży nieruchomości[13].

4. Studium przypadku

W celu realizacji niniejszego studium przypadku wykorzystano typowe dla zastosowań Big Data narzędzie, jakim jest Apache Spark z językiem Python.

[11] A. Giachanou, F. Crestani, *Like It or Not: A Survey of Twitter Sentiment Analysis Methods*, "ACM Computing Surveys" 2016, vol. 49, no. 2, pp. 28–41.

[12] Developer Twitter.com, https://dev.twitter.com/overview/api (4.11.2016).

[13] G. Marcato, A. Nanda, *Information Content and Forecasting Ability of Sentiment Indicators: Case of Real Estate Market*, "Journal Of Real Estate Research", vol. 38, no. 2, pp. 165–203.

O zastosowaniu tego systemu przesądziła jego wzrastająca popularność, jak również występowanie w literaturze naukowej licznych testów, które wskazują na wysoką wydajność tego narzędzia[14]. Możliwość połączenia źródeł ustrukturyzowanych, jak i nieustrukturyzowanych w systemie plików HDFS pozwala na tworzenie rozwiązań hybrydowych, tj. integrujących te dwa rodzaje danych. Przykładem jest hybrydowa hurtownia danych, która łączy tradycyjną hurtownię z danymi przechowywanymi w HDFS[15].

Opisywany przypadek wykorzystuje język Python oraz biblioteki pozwalające na web scraping stron internetowych, jak również dostęp do mediów społecznościowych poprzez API. Z punktu widzenia Twittera istotne jest odwołanie się do opcji regionalizacyjnych. Jeżeli jednostka samorządu terytorialnego zlokalizowana jest na terenie Warszawy, należy jako punkt odniesienia przyjąć wartość wskazującą na ten obszar, przykładowo:

trendsPL = api.trends_place (23424923),

gdzie wartość w nawiasie oznacza kod geograficzny lokalizacji miejscowości.

Poniżej zaprezentowano działanie zaprojektowanego rozwiązania do web scrapingu i analizy komentarzy na anglojęzycznym portalu opisującym zjawiska zachodzące w Polsce. Zdecydowano się na analizę 121 komentarzy znajdujących się pod artykułem prasowym. Sentyment wpisów zidentyfikowano poprzez oprogramowanie przygotowane w języku Python z wykorzystaniem Machine Learning. Próbkę wyników zaprezentowano w tabeli 1.

Wykorzystane biblioteki Machine Learning pozwoliły na zaklasyfikowanie komentarzy na podstawie zestawu słów kluczowych. Ręczna analiza danych potwierdziła większość poprawnych zaklasyfikowań. Należy zwrócić uwagę, że stosowane biblioteki stale ewoluują, co oznacza, że w niedalekiej przyszłości trafność identyfikacji właściwego sentymentu dla komentarza będzie większa. W tabeli 2 przedstawiono zbiorcze wyniki analizowanych komentarzy.

Spośród 121 analizowanych komentarzy 14 miało wydźwięk pozytywny, z kolei 42 były neutralne, a 65 – miało wydźwięk negatywny. Większość komentarzy

[14] M. Zaharia, R. Xin, P. Wendell, T. Das, M. Armbrust, A. Dave, M. Xiangrui, J. Rosen, S. Venkataraman, M. Franklin, A. Ghodsi, J. Gonzalez, S. Shenker, I. Stoica, *Apache Spark: A Unified Engine for Big Data Processing*, "Communications of the ACM" 2016, vol. 59, no. 11, pp. 56–65.

[15] Y. Tian, F. Özcan, Z. Tao, R. Goncalves, H. Pirahesh, *Building a Hybrid Warehouse: Efficient Joins between Data Stored in HDFS and Enterprise Warehouse*, "ACM Transactions On Database Systems" 2016, vol. 41, no. 4, pp. 1–38.

została trafnie zdefiniowana, co zostało również potwierdzone w ramach przygotowanego rozwiązania, co zaprezentowano w tabeli 3.

Tabela 1. Efekt analizy komentarzy pod artykułem prasowym z wykorzystaniem algorytmów Machine Learning

Lp.	Kategoria	Etykieta	Prawdopodobieństwo	Treść komentarza
0	329045	negatywny	1.000	Really? E…
1	329045	negatywny	0.558	They do the…
2	329045	negatywny	1.000	A huge protest…
3	329045	negatywny	0.989	Poland has responded…
4	329045	negatywny	1.000	Right-wing po…

Źródło: opracowanie własne.

Tabela 2. Liczba analizowanych komentarzy

Lp.	Rodzaj komentarza	Liczba komentarzy
1	Negatywny	65
2	Neutralny	42
3	Pozytywny	14

Źródło: opracowanie własne.

Tabela 3. Prawdopodobieństwo właściwego sentymentu przypisanego do komentarza

Lp.	Prawdopodobieństwo	Liczba komentarzy
1	1.000	83
2	0.920	2
3	0.853	2
4	0.447	1
5	0.718	1
6	0.748	1
7	0.766	1
..	…	
121	0.588	1

Źródło: opracowanie własne.

Wśród 121 komentarzy prawdopodobieństwo poprawności zaklasyfikowania równe 1 zostało zidentyfikowane wśród 83 komentarzy. Spośród pozostałych 38 komentarzy większość miała wskazane prawdopodobieństwo na poziomie wyższym niż 0,7. Ze względu na prawdopodobieństwo błędnego zaklasyfikowania

komentarzy w środowisku produkcyjnym można rozważyć odrzucenie tych danych.

Jak widać na powyższym przykładzie, analiza danych z zastosowaniem narzędzi typu Machine Learning może znacząco usprawniać automatyczne analizy sentymentu. Oddzielnym zagadnieniem jest sprawdzenie, na ile takiego rodzaju rezultaty są wiarygodne. Przeprowadzone rozpoznanie pozwala stwierdzić, że wiarygodność tego rodzaju analiz pozwala na ich zastosowanie. Potwierdzenie tego znajduje się w odsetku poprawnie zaklasyfikowanych komentarzy. Należy jednak mieć na uwadze konieczność uwzględnienia aspektów zaprezentowanych w części piątej niniejszego artykułu.

5. Analiza możliwości wdrożenia rozwiązania

Analizując możliwość wdrożenia tego rozwiązania, należy mieć na uwadze wiele aspektów związanych z przetwarzaniem danych. Po pierwsze analiza Big Data bazuje na słowach kluczowych, które powinny być starannie wyselekcjonowane. Często narzędzia typu Machine Learning pozwalają na dostarczanie informacji nt. prawdopodobieństwa prawidłowego dopasowania danego sentymentu do analizy treści komentarza. Zostało to zaprezentowane w części czwartej niniejszego artykułu.

Rozważajac zastosowania tego rodzaju rozwiązań do analizy danych, trzeba wziąć pod uwagę następujące aspekty:

- Semantyka treści.
- Lematyzacja.
- Analiza anaforyzmów.
- Metody parsowania.
- Identyfikacja języka.
- Leksykon.

Semantyka odnosi się do prawidłowego łączenia słów kluczowych. Przykładowo występowania słowa „dobry" nie musi oznaczać pozytywnego wydźwięku komentarza. Może to być część zwrotu grzecznościowego – „dzień dobry". Ponadto słownik powinien zawierać zestawienia słów kluczowych w odniesieniu do lematyzacji, czyli możliwości zapisu słowa kluczowego na wiele sposobów, z jego licznymi odmianami. Anaforyzmy, jak również sarkazmy, są niezwykle trudne do wykrycia i również mogą prowadzić do błędnego zaklasyfikowania komentarza. Stąd też zalecane jest wykorzystanie słownika anaforyzmów. Parsowanie

danych powinno uwzględniać występowanie błędów popełnianych przez osoby tworzące komentarze czy wpisy. Ważna jest również identyfikacja języka, gdyż w niektórych mediach społecznościowych komentarze mogą występować w różnych językach, nawet gdy dotyczą jednego artykułu np. w języku angielskim. Leksykon powinien uwzględniać również szereg słów, które nie mają wpływu na identyfikację treści. Przykład braku tzw. *stop words*, czyli wyrazów niemających znaczenia dla analizy treści został zobrazowany na rysunku 3.

Rysunek 3. Przykład braku tzw. *stop words* podczas analizy treści portalu webowego
Źródło: opracowanie własne.

Na zaprezentowanej na rysunku 3 chmurze słów jednoznacznie widać, że najczęściej występujące wyrażenia w treści to „w", „i", „na" itd. Usunięcie jedynie spójników czy przyimków nie pozwoli jednak na prawidłową identyfikację treści. Zostało to zaprezentowane na rysunku 4.

Rysunek 4. Analiza treści portalu webowego po usunięciu tzw. stop words
Źródło: opracowanie własne.

W czasie prowadzenia eksperymentalnych analiz z wykorzystaniem Big Data stwierdzono, że leksykon słów kluczowych powinien być przygotowany na potrzeby konkretnego portalu webowego lub medium społecznościowego.

Taki leksykon powinien powstać dopiero po identyfikacji treści znajdującej się na danej stronie internetowej, dla konkretnego artykułu prasowego lub opisywanej inicjatywy.

Podsumowując: nieprawidłowe zastosowanie jednego z powyższych aspektów może prowadzić do powstawania błędnych wyników analiz. Ważne jest zatem skrupulatne przygotowanie procesu identyfikacji sentymentu, w zależności od źródła danych.

6. Wnioski i podsumowanie

Przedstawione w niniejszym artykule studium przypadku pozwoliło ocenić możliwość zastosowania analizy sentymentu do badania opinii nt. zjawisk zachodzących w otoczeniu. Jak wspomniano na początku artykułu, takie rozwiązania mogą być wykorzystywane przez administrację publiczną do automatycznej analizy odczuć związanych z informacjami prezentowanymi na stronach internetowych.

Jak powidziano w części piątej niniejszego artykułu, na poprawność analiz tekstu nieustrukturyzowanego ma wpływ przede wszystkim zestaw przygotowanych słów kluczowych. Pozwala to na wyeliminowanie nadmiarowych słów, które nie są istotne w analizie treści. Dostępne na rynku narzędzia typu Machine Learning, wspomagające analizy Big Data, często posiadają przygotowane leksykony słów kluczowych. Ich zastosowanie może jednak prowadzić do błędnych analiz, co wynika z niedostosowania do analizowanej dziedziny przedmiotowej.

Główny cel artykułu, jakim było zaprezentowanie możliwości wykorzystania narzędzi Big Data do analizy treści nieustrukturyzowanych poprzez analizę sentymentu, został osiągnięty. Zgodnie z przyjętymi założeniami, takie rozwiązanie może być wykorzystywane przez jednostki administracji publicznej do badania nastrojów społecznych, powiązanych z podejmowanymi inicjatywami i wydarzeniami w otoczeniu. Ważne jest przy tym spełnienie wymogów prawnych i opracowanie przejrzystych reguł monitorowania komentarzy i wpisów w mediach społecznościowych, o czym jest mowa w artykule.

Postawiona w artykule hipoteza nt. możliwości zastosowania metod text mining oraz machine learning w celu dostarczania rzetelnej informacji może zostać potwierdzona pod warunkiem dostosowania się do wytycznych zaprezentowanych w artykule.

Podsumowując: narzędzia Big Data znacznie poszerzają zakres informacji, jakie jednostki administracji publicznej, jak również inne podmioty gospodarcze, mogą pozyskać. Przy tym narzędzia te są ogólno dostępne, w znakomitej większości bezpłatne i łatwe w zastosowaniu.

Bibliografia

Bosch V., *Big Data in Market Research: Why More Data Does Not Automatically Mean Better Information*, "Gfk-Marketing Intelligence Review" 2016, vol. 8, no. 2.

Chintagunta P., Hanssens D., Hauser J., *Marketing and Data Science: Together the Future is Ours*, "Gfk-Marketing Intelligence Review" 2016, vol. 8, no. 2.

George G., Osinga E., Lavie D., Scott B., *Big Data and Data Science Methods for Management Research*, "Academy of Management Journal" 2016, no. 10.

Giachanou A., Crestani F., *Like It or Not: A Survey of Twitter Sentiment Analysis Methods*, "ACM Computing Surveys" 2016, vol. 49, no. 2.

Liang T., Li X., Yang C., Wang M., *What in Consumer Reviews Affects the Sales of Mobile Apps: A Multifacet Sentiment Analysis Approach*, "International Journal of Electronic Commerce" 2015, vol. 20, no. 2.

Lozza E., Bonanomi A., Castiglioni C., Bosio A., *Consumer Sentiment After the Global Financial Crisis*, "International Journal of Market Research" 2016, vol. 58, no. 5.

Makarem S., Jae H., *Consumer Boycott Behavior: An Exploratory Analysis of Twitter Feeds*, "Journal of Consumer Affairs" 2016, vol. 50, no. 1.

Marcato G., Nanda A., *Information Content and Forecasting Ability of Sentiment Indicators: Case of Real Estate Market*, "Journal of Real Estate Research" 2016, vol. 38, no. 2.

Pagliarussi M., Aguiar M., Galdi F., *Sentiment Analysis in Annual Reports from Brazilian Companies Listed at the bm&fbovespa*, "Base" 2016, vol. 13, no. 1.

Tian Y., Özcan F., Tao Z., Goncalves R., Pirahesh H., *Building a Hybrid Warehouse: Efficient Joins between Data Stored in HDFS and Enterprise Warehouse*, "ACM Transactions On Database Systems" 2016, vol. 41, no. 4.

White S., *6 Ethical Questions about Big Data*, "Journal of Accountancy" 2016, vol. 222, no. 4.

Zaharia M., Xin R., Wendell P., Das T., Armbrust M., Dave A., Xiangrui X., Rosen J., Venkataraman S., Franklin M., Ghodsi A., Gonzalez J., Shenker S., Stoica I., *Apache Spark: A Unified Engine for Big Data Processing*, "Communications of the ACM" 2016, vol. 59, no. 11.

Źródła sieciowe

Developer Twitter.com, https://dev.twitter.com/overview/api (4.11.2016).

Ustawa z dnia 27 lipca 2001 r. o ochronie baz danych (DzU 2001, nr 128, poz. 1402), Internetowy System Aktów Prawnych, http://isap.sejm.gov.pl/DetailsServlet?id=WDU20011281402 (2.11.2016).

* * *

Automatic Analysis of Unstructured Content as an Example of a Data Source for the Public Administration

Abstract

Organization management requires access to reliable and verified data, which allows developing a particular organizational unit by decidents. With information technology development, processing large datasets to acquire valuable information is more common. Such a data source can be social media or comments under news articles. The goal of this article is to present a case study of automatic content analysis to get a general opinion on the initiative taken by public administration units, especially self-government institutions. For this reason, a framework has been developed to allow analysing unstructured content, in which the most common form are comments. The analysis of the results taken from this system allows formulating several conclusions on Big Data tools usability as well as the reliability of the data acquired this way.

Keywords: Big Data, social media, text mining, machine learning, sentiment analysis