Z badań nad wykorzystaniem rafinacji informacji sieciowej Wybory prezydenckie i parlamentarne 2015

Włodzimierz Gogołek, Dariusz Jaruga, Krzysztof Kowalik, Piotr Celiński

Już w 2011 roku dowiedziono możliwości efektywnego wykorzystania dużych zasobów informacyjnych, nazywanych Big Data, jako źródła informacji poddających się konstruktywnej analizie ilościowej[1]. Ich znaczącą część tworzą zasoby internetu, włączając w to sieci społecznościowe. Dane tego typu są tworzone przez indywidualnych użytkowników umieszczających w sieci blogi, posty, portale, maile, strumień zapytań kierowanych do internetu, profesjonalne publikacje i inne bogate zasoby informacyjne.

Najłatwiej dostępnym zasobem informacji, określanym mianem Big Data, jest sieć/internet. Tą właśnie drogą w ciągu każdej sekundy jest przesyłane 22574 GB danych, powstaje 5700 tweetów, 55 tysięcy postów na Facebooku, a na portal YouTube dodawane są 2 godziny materiału[2]. Ten cyfrowy świat co dwa lata podwaja swoje rozmiary – w 2020 r. liczba bitów informacji wygenerowanych przez ludzkość przekroczy liczbę gwiazd we Wszechświecie[3]. Na razie, w 2014 roku, liczbę tych informacji oszacowano na 3 ZB, to jest około 40 kolumn książek z Ziemi do Słońca**.** Jednak obecnie tylko 0,5% tych zasobów jest skutecznie analizowana[4].

Przeprowadzone w Instytucie Dziennikarstwa UW wspomniane wcześniej badania[5] – w których korzystano z potencjału informacyjnego rafinacji – dotyczyły, zapewne po raz pierwszy w skali świata, problematyki pozyskiwania informacji związanych z aktywnością polityczną, w szczególności – wyborów prezydenckich i parlamentarnych. Obecnie podobne badania są prowadzone już niemal we wszystkich

[1] W. Gogołek, P. Kuczma*, Rafinacja informacji sieciowych na przykładzie wyborów parlamentarnych. Część 1. Blogi, fora, analiza sentymentów,* „Studia Medioznawcze" 2013, nr 2 (53), s. 89–109.

[2] *The Internet in real time,* http://pennystocks.la/internet-in-real-time/ [dostęp: 25.04.2015].

[3] *The digital universe of opportunities: Rich Data and the inreasing value of the Internet of things,* http://www.emc.com/leadership/digital-universe/2014iview/executive-summary.htm [dostęp: 25.04.2015].

[4] *Big Data, Bigger digital shadows, and biggest growth in the Far East,* https://www.emc.com/collateral/analyst-reports/idc-digital-universe-united-states.pdf [dostęp: 25.04.2015].

[5] W. Gogołek, P. Kuczma*, Rafinacja informacji sieciowych*…, dz. cyt.

branżach, które wykorzystują aktualne informacje w swojej działalności. Przykładem jest komercyjne narzędzie Brand24 oferujące szeroki wachlarz monitoringu opartego na zasadach rafinacji. Obejmuje on: monitoring marki – tzn. śledzenie na bieżąco, co o marce/produkcie/usłudze mówią internauci w sieci; monitoring prewencyjny (co internauci mówią, ludzkie oblicze marki); monitoring kryzysowy i monitoring sprzedażowy[6]. Podobne usługi świadczy SentiOne, firma, która udostępniła Instytutowi Dziennikarstwa UW swoje narzędzia dla testów w zakresie możliwości zbierania sentymentów dotyczących notowań spółek giełdowych. Uzyskane wyniki potwierdziły możliwość statystycznie istotnej predykcji notowań. Wyniki testów wskazały nadzwyczaj dużą korelację (r>0,8) przewidywań z rzeczywistymi notowaniami czterech spółek giełdowych (Enea SA, KGHM SA, Synthos SA i Tauron SA)[7].

Fundamentem możliwości realizacji tego typu badań były, i są obecnie w jeszcze większym zakresie, techniczne możliwość gromadzenia wspomnianych gigantycznych zasobów i narzędzi ich analizy. Mariaż technologii z potencjałem informacji nie został jeszcze zauważony jako skuteczne narzędzie pozyskiwania wtórnych informacji – są one jak rad rafinowany z rudy (zaledwie 4 g z jednej tony *rudy)*[8]. Te proporcje wydają się być adekwatną ilustracją procesów rafinacji informacji Big Data. Jej wynik stwarza nową kategorię informacji, która wcześniej nigdy nie była – i ze względów ograniczeń technologicznych – nie mogła być dostępna.

Uznano zatem za celowe potwierdzenie zasadności tego kierunku zastosowań technologii w analizie gigantycznych zasobów informacji. Rafinacja umożliwia dostrzeganie informacji wtórnych w ukrytych zasobach informacji pierwotnych (Big Data). Dane uzyskane dzięki tym analizom tworzą obraz historii, stanu, potrzeb i zachowań m.in. indywidualnych użytkowników i firm, ale także społeczności jako całości. Jednocześnie dostarczają wartościowych, wiarygodnych statystycznie informacji do analiz predykcyjnych.

Badania

Cel. Zasygnalizowany potencjał oraz realne zapotrzebowanie na aktualne, oryginalne informacje stanowiły o celu przedsięwzięcia, tzn. wskazanie głównych etapów rafinacji/ogniw łańcucha procedur/czynności składających się na proces rafinacji. Jej umiejętne zastosowanie generuje wcześniej nieznane, użyteczne informacje będące przeciwnością smogu informacyjnego przypisywanego sieci[9].

Hipoteza. Wyniki rafinacji stanowią wiarygodne informacje opisujące wybrany proces społeczny/zjawisko w czasie przeszłym, rzeczywistym, a także prognozę. Są one, po odpowiedniej obróbce, wiarygodnym źródłem opinii na temat procesu społecznego/zjawiska, np. w poszukiwaniu zagrożeń funkcjonowania firmy, oczekiwań klientów czy predykcji wyborów społecznych lub notowań spółek na giełdzie.

Założenie. Przyjęto założenie, że badania związane z rafinacją będą dotyczyć, podobnie jak w badaniach przeprowadzonych

[6] Socjomania, http://socjomania.pl/10-krokow-skutecznego-monitoringu-z-brand24 [dostęp: maj 2015].

[7] Niepublikowane prace: A. Woch*, Internetowe predykcje notowań spółek giełdowych*, ID UW, Warszawa 2015; M. Wójcikiewicz, *Analiza przydatności narzędzi Big Data w prognozowaniu notowań spółek giełdowych,* ID UW, Warszawa 2015.

[8] J.L. Marshall, *Wydobycie uranu i rafinowanie radu w St. Joachimsthal (Jáchymovie*) [w:] "Nowotwory. Journal of Oncology" 2011, Vol. 61, No. 2, p. 181–185.

[9] R. Tadeusiewicz, *Ciemna strona internetu*, wykład inauguracyjny, WSZiA w Zamościu, 16 października 1999.

w 2011 roku[10], predykcji (na podstawie danych poprzedzających dzień wyborów) wyborów prezydenckich (2015) oraz parlamentarnych (2015).

Metodologia. Jednym z ogniw procesu rafinacji jest analiza sentymentów. Jest ona rozumiana jako wyróżnianie wpisów (uniwersalne określenie paczek/fragmentów treści pozyskiwanych z Big Data) uzyskanych z sieci, które zawierają wyróżnioną *nazwę* oraz co najmniej jeden *sentyment*. Sentymentem jest słowo lub zwrot o zabarwieniu emocjonalnym. Wstępne badania dowiodły, że zasadne jest wyróżnienie trzech kategorii sentymentów: pozytywne, neutralne, negatywne[11]. Wyróżnienie słów uznanych jako sentyment (sentymenty), poza kolekcjonowaniem wpisów z sieci, jest fundamentalnym ogniwem w procedurze rafinacji opartej na sentymentach. W zależności od celu zastosowań rafinacji rolę sentymentów mogą także pełnić tematyczne konteksty, np. w odniesieniu do władz państwowych:

– merytoryczne (edukacja, finanse, gospodarka itp.);
– medialne – związane z bieżącymi wydarzeniami relacjonowanymi w mediach (np. władza, media, pieniądze, prawo)[12].

Nazwą może być dowolny termin związany z ocenianym zjawiskiem, np. ocena kondycji politycznej partii/osoby, firmy, zjawiska.

Procedura rafinacji

Mając na uwadze doświadczenia z użycia rafinacji w badaniach przebiegu wyborów prezydenckich i parlamentarnych (2011), oraz późniejsze eksperymenty związane z podobnymi badaniami, wyróżniono podstawowe ogniwa łańcucha procesu rafinacji (rysunek 1.) opartego na badaniach sentymentów. Owe ogniwa tworzą łańcuch operacji, które wraz z odpowiednim uzbrojeniem technicznym i programowym są autorskim narzędziem rafinacji. Immanentną cechą rafinacji jest możliwość uzyskiwania/wykorzystywania wyników jej stosowania w czasie rzeczywistym oraz w odniesieniu do przeszłości i przyszłości (predykcja).

Rysunek 1. Łańcuch procesu rafinacji

Źródło: opracowanie własne

[10] W. Gogołek, P. Kuczma*, Rafinacja informacji sieciowych…*, dz. cyt.

[11] V. Hatzivassiloglou, K.R. McKeown, *Predicting the semantic orientation of adjectives*, 35th Annual Meeting of the Association for Computational Linguistics, Madrid 1997, s. 174–181, http://www.anthology.aclweb.org/P/P97/P97-1023.pdf [dostęp: 30.10.2011]; P.D. Turney, *Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews*, proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia, July 2002, s. 417–424, http://acl.ldc.upenn.edu/P/ P02/P02-1053.pdf [dostęp: 29.10.2011].

[12] Dobór słów stanowiących konteksty powinien mieć swoje merytoryczne uzasadnienie m.in. w wartościach frekwencji ich występowania w rafinowanych wpisach. Zob. W. Gogołek, P. Kuczma*, Rafinacja informacji sieciowych…*, dz. cyt.

Sentymenty

Przyjętą procedurę rafinacji rozpoczyna tworzenie zbiorów nazw i sentymentów. Nazwy stanowią określenie przedmiotu badań, tutaj są nimi nazwy partii, nazwiska kandydatów na prezydenta. Sentymenty natomiast, zgodnie z podaną wcześniej definicją, mają zabarwienie emocjonalne. Zważywszy na przedmiot rafinacji, sentymenty różnią się i są dobierane stosownie do tematyki badań. Dlatego ważne jest, by po zebraniu testowej liczby wpisów dokonać weryfikacji przyjętych sentymentów w mowie potocznej (o najwyższych frekwencjach). W opisanych dalej wynikach badań do poszukiwania sentymentów związanych z kandydatami na prezydenta skorzystano z czterech zbiorów sentymentów:

1. Grupa 37 osób (studenci I roku studiów I stopnia) dokonało przeglądu zbioru tekstów (1000 wpisów) zebranych z serwisów mediowych i społecznościowych. Następnie każda z nich wybrała słowa lub wyrażenia, które negatywnie oraz pozytywnie opisywały sylwetki obu kandydatów. Powstał korpus 4650 słów i wyrażeń: Andrzej Duda – 1291 słów i wyrażeń negatywnych, 1076 pozytywnych; Bronisław Komorowski – 1134 słów i wyrażeń negatywnych, 1149 pozytywnych. Następnie dokonano analizy frekwencji wszystkich słów oraz wyrażeń określających poszczególnych kandydatów. W ten sposób wyłoniono korpus sentymentów najczęściej wskazywanych przez osoby przeglądające zbiór tekstów. W wyniku porównania sentymentów pozytywnych i negatywnych obu kandydatów wyeliminowano powtarzające się terminy, ale pozostawiono synonimy niektórych określeń, które mogą mieć duże znaczenie dla wyników (wybór oparty jest na doświadczeniu badacza i nie podlegał innej weryfikacji). Ostatecznie słowa sparowano według reguły określenie pozytywne vs. negatywne. Powstała baza 69 słów mogących wskazać sentymenty (sentymenty ST).

2. Zbiór sentymentów (sentymenty 2011), które były wykorzystane w badaniach przeprowadzonych w 2011 roku[13].

3. Zbiór sentymentów (sentymenty P) oparty na wynikach badań Walerego Pisarka[14]. Autor książki przeprowadził badania ankietowe (4873 respondentów), w których wyłoniono treści określone jako „najlepsze, najpiękniejsze i najwartościowsze" oraz treści „najgorsze, nieprzyjemne lub najszkodliwsze". W ten sposób powstała baza 54 słów sztandarowych. Respondenci wypełniali ankietę w latach 1991, 1995, 1996, 1997, a więc przed dynamicznym rozwojem sieci internetowej, i należy ich zaliczyć do pokolenia odbiorców starych mediów (prasy, radia, telewizji). Spośród słów sztandarowych wyselekcjonowano określenia, które można było wpisać w kontekst haseł toczącej się kampanii wyborczej na urząd prezydenta. Wybór został oparty na doświadczeniu badacza i nie podlegał innej weryfikacji. Następnie dokonano sprawdzenia, czy baza słów sztandarowych jest aktualna w stopniu pozwalającym na wykorzystanie ich w projekcie. W tym celu wykorzystano narzędzie Google Trends. Sprawdzono popularność terminów jako fraz wyszukiwanych przez internautów. Założono, że słowo może być użyte w przypadku, gdy Google Trends indeksuje dany termin jako poszukiwany przez użytkowników sieci. Np. serwis wskazał brak w wynikach wyszukiwania takich słów jak „zakłamanie", „dobro innych" i „dobro własne". Zastąpiono je słowami „kłamstwo" oraz „dobro".

[13] Tamże.

[14] W. Pisarek, *Polskie słowa sztandarowe i ich publiczność*, Warszawa 2002, ogólna hierarchia – tabela, s. 23–25, najlepsze – s. 26–27, najgorsze – s. 28–29. Zbiór sentymentów (sentymenty P) oparty na wynikach badań Walerego Pisarka został wykonany przez dr. Krzysztofa Kowalika z Instytutu Dziennikarstwa UW.

W kolejnym kroku słowa sztandarowe sparowano według reguły sentyment pozytywny vs. negatywny. W przypadku braku przeciwstawnego sentymentu wykorzystano *Popularny słownik synonimów i antonimów* Grzegorza Dąbkowskiego i Małgorzaty Marcjanik oraz słownik online antonimów – antonimy.net. Tą drogą powstała baza 45 słów mogących wskazać sentymenty.

4. Wykorzystano 11 sentymentów (sentymenty RP) zawartych w książce Radosława Pawelca[15].

Sentymenty, które zostały wyróżnione w podany wyżej sposób (150 pojęć/wyrazów) poddano weryfikacji frekwencyjnej. Polegała ona na obliczeniu częstotliwości występowania każdego z tych pojęć w próbce wpisów (n = 1000). Najczęściej występujące słowa stanowiły zasadniczy zbiór pojęć przyjętych jako sentymenty pozytywne i negatywne. Zasygnalizowana procedura doboru sentymentów uwzględnia autorytatywne źródła (W. Pisarek, R. Pawelec, wyniki badań z 2011 r.) oraz własne badania (wybory studentów), które były próbą uwzględnienia pojęć uznawanych przez młodych ludzi jako pozytywne i negatywne (wiosna 2015).

Kolekcjonowanie wpisów

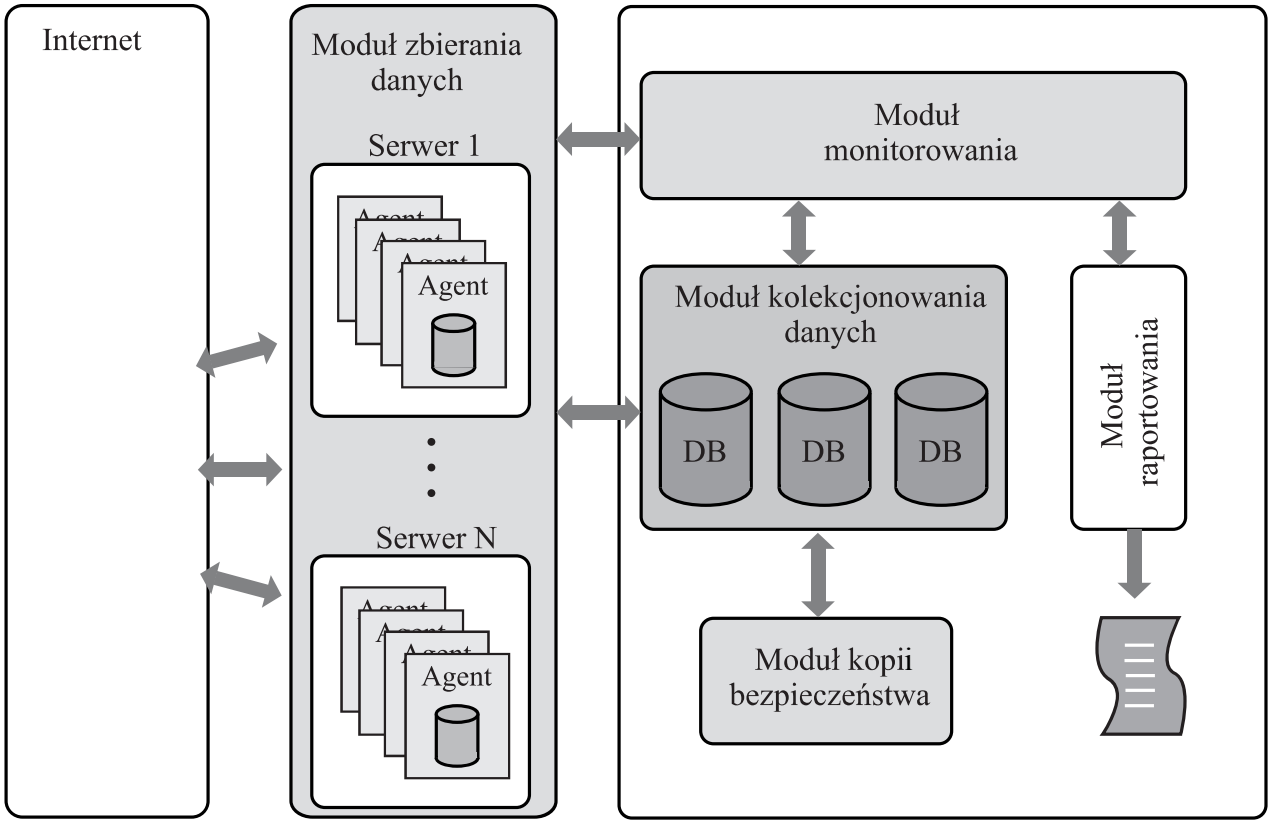
Kolekcjonowanie wpisów to kolejne ogniwo procesu rafinacji. Ta operacja jest po raz pierwszy wykonywana przez autorskiego robota nazwanego „Robot BigData”[16] (we wcześniejszych badaniach korzystano z komercyjnych robotów). Robot BigData to specjalizowany systemem teleinformatyczny do ukierunkowanego monitorowania i zbierania danych ze wskazanych serwisów internetowych. System kolekcjonuje dane udostępniane w internecie dla każdego użytkownika sieci bez konieczności autoryzacji (logowania się do danego serwisu) w sposób otwarty. Każda zarejestrowana przez robota informacja poza właściwą treścią zawiera dodatkowo źródło informacji (link) oraz datę jej publikacji albo pobrania, w zależności od zakresu danych udostępnianych przez monitorowany serwis.

System Big Data składa się z szeregu modułów (rysunek 2.), z których każdy pełni określoną funkcję. Do najważniejszych należą: moduł zbierania danych, kolekcjonowania, monitorowania i wykonywania kopii bezpieczeństwa.

Moduł zbierania danych to dedykowane oprogramowanie, które w ustalony i zdefiniowany wcześniej sposób monitoruje źródło informacji. W przypadku opublikowania nowych treści pobiera je i przekazuje do modułu kolekcjonowania. Monitorowanie i kolekcjonowanie danych odbywa się w sposób równoległy. W skład modułu zbierania danych wchodzi wiele równocześnie działających robotów (agentów), a każdy z nich w określonych i zdefiniowanych jednostkach czasu wchodzi w interakcję z monitorowanym serwisem. Częstotliwość pobierania danych jest regulowana indywidualnie dla każdego pojedynczego agenta w zakresie od 1 minuty do 365 dni. Dzięki temu monitorowanie serwisów w zależności od dynamiki zmian i ilości publikowanych artykułów w jednostce czasu może być ustawione dowolnie i stosownie do potrzeb. Dodatkowo częstotliwość skanowania może również ulegać zmianie, w zależności od pory dnia, dnia tygodnia, pory roku etc. Każdy agent, wchodząc w interakcję z monitorowanym serwisem będącym źródłem informacji, symuluje swoją pracą zachowanie człowieka przeglądającego serwisy internetowe. Zatem sposób działania robota BigData nie łamie zasad netykiety stosowanej przez internautów.

[15] R. Pawelec, *Ciemne zwierciadło. Semantyka antywartości*, Warszawa 2013.

[16] Robot oraz jego opis zostały wykonane i wykorzystane do badań przez mgr. inż. Dariusza Jarugę z Instytutu Dziennikarstwa UW.



Rysunek 2. Schemat konstrukcji systemu Big Data

Źródło: opracowanie własne

Agenci modułu zbierania danych mogą działać na jednym lub na wielu serwerach, w zależności od liczby monitorowanych serwisów. Dodatkowo, w celu zapewnienia optymalnej wydajności systemu, każdy agent modułu zbierającego posiada prywatną bazę danych, w której zapisuje postępy pracy, ograniczając tym samym ilość wymiany danych z modułem kolekcjonującym do niezbędnego minimum. Agent modułu zbierania danych pracuje w trzech trybach: produkcyjnym, konfiguracyjnym i debugowania (czyszczenia wpisów z niepotrzebnych, np. html-owych znaków). Tryb produkcyjny polega na tym, że agent informuje moduł monitorujący tylko i wyłącznie o problemach i błędach, jakie zaistniały podczas pracy w wyniku interakcji z monitorowanym serwisem. Robot BigData z założenia pracuje w trybie 7/24/365, a zbieranie informacji odbywa się w sposób ciągły.

Dane zebrane w trakcie pracy przez moduł zbierania danych są magazynowane w module kolekcjonowania danych, w skład którego wchodzi relacyjna baza danych. Informacje zawarte w bazie danych są wykorzystywane przez moduł raportujący, który generuje dane dla zewnętrznego oprogramowania do badania sentymentów w formacie wymaganym przez to oprogramowanie. Zebrane dane w module kolekcjonowania mogą być wielokrotnie wykorzystywane i pobierane, stosownie do potrzeb z określonego przedziału czasowego lub pod względem interesujących badacza słów kluczowych lub wyrażeń. Moduł raportujący potrafi wygenerować plik Excela, który z powodzeniem może być wykorzystany przez dowolne oprogramowanie trzecich firm do dalszej analizy. Modułowa budowa robota BigData pozwala na jego dalszą rozbudowę o kolejne funkcjonalności w obszarze zbierania i kolekcjono-

wania danych z różnych źródeł, w zależności od potrzeb. Obecnie robot BigData gromadzi dane udostępnione przez usługę WWW w wersji szyfrowanej (https) i nieszyfrowanej (http). Należy również zaznaczyć, że w zasadzie nie występują ograniczenia dotyczące możliwości zbierania danych z innych usług, takich jak ftp, e-mail (newslettery), API do innych systemów, np. bibliotecznych, systemów agencji prasowych itp. Taka rozbudowa jest możliwa pod warunkiem otrzymania stosownej dokumentacji oraz po wykonaniu prac programistycznych, w wyniku których powstaną dedykowani agenci modułu zbierania danych.

Robot BigData ze względu na swoją funkcjonalność znajdzie zastosowanie wszędzie tam, gdzie zachodzi konieczność zbierania dużej ilości danych tekstowych na określony temat z wybranych serwisów internetowych. Zgromadzone przez robota dane stanowią źródło informacji dla kolejnych systemów, np. do badania sentymentów, i mogą być użyteczne w zakresie predykcji przyszłych wydarzeń, trendów zjawisk społecznych, bez ograniczenia zakresu (polityka, ekonomia, zdrowie itp.).

Niezbędnym warunkiem prawidłowego zebrania danych na wskazany temat jest dobór właściwych źródeł informacji w postaci linków do serwisów internetowych. Jakość merytoryczna danych zebranych przez robota bardzo zależy od intuicji i doświadczenia badacza prowadzącego prace. Dane generowane przez system są ustandaryzowane, pobrane treści bez względu na charakter źródła zostają przekonwertowane do UTF-8, a format zapisu daty i czasu są zgodne z normą ISO 8601:2004. Ponieważ technologie internetowe i zachodzące w nim ciągłe zmiany są procesem naturalnym, robot BigData wymaga okresowych aktualizacji, których celem jest dostosowanie go do ciągle zmieniającej się rzeczywistości cyfrowego świata.

Moduł analizy treści wpisów

Metodologia analizy danych wykorzystuje wzorce wyrażeń regularnych zarówno dla nazw, jak i dla sentymentów[17]. Jednym z istotniejszych zagadnień jest odpowiednie dobranie tych wzorców. Muszą one uwzględniać wszystkie formy gramatyczne, wraz z obocznościami tematów, oraz w przypadku nazw – najpopularniejsze określenia. Na przykład dla nazwy „Platforma Obywatelska" należy uwzględnić takie określenia jak „PO", „Platformersi", „Platfusy" itp. Innym, nie mniej ważnym zagadnieniem jest dobór zestawu sentymentów pozytywnych i negatywnych.

Oprogramowanie analizuje i zlicza wystąpienia w danych wejściowych par: nazwa–sentyment (osobno dla sentymetów pozytywnych i negatywnych). Przy czym pary są poszukiwane w zadanym zakresie znaków od nazwy – zarówno lewostronnie, jak i prawostronnie. Osobno są zliczane same wystąpienia nazw bez sentymentów, co można określić jako kontekst neutralny. Zliczanie wystąpień w kontekście pozytywnym, negatywnym i neutralnym odbywa się w dwojaki sposób. Zliczanie z powtórzeniami sumuje wszystkie wystąpienia w obrębie danego rekordu (wpisu). Zliczanie bez powtórzeń zwiększa licznik o jeden, jeśli w danym rekordzie znaleziono wystąpienie. Zliczone wystąpienia zostają zsumowane dla każdej daty, dla której są dane wejściowe.

Wyniki

Wpisy gromadzone przez autorskiego robota BigData tworzą nieustannie uzupełnianą od 15 maja 2015 r. bazę rekordów odnoszących się wyborów prezydenckich i parlamentarnych. Biorąc pod uwagę datę rozpoczęcia kolekcjonowania (15 maja) – oraz początkowo niewielką, lecz każdego dnia rosnącą intensywność gromadzenia wpisów – wyniki

[17] Ten moduł został wykonany, opisany oraz zastosowany przez mgr. inż. Piotra Celińskiego.

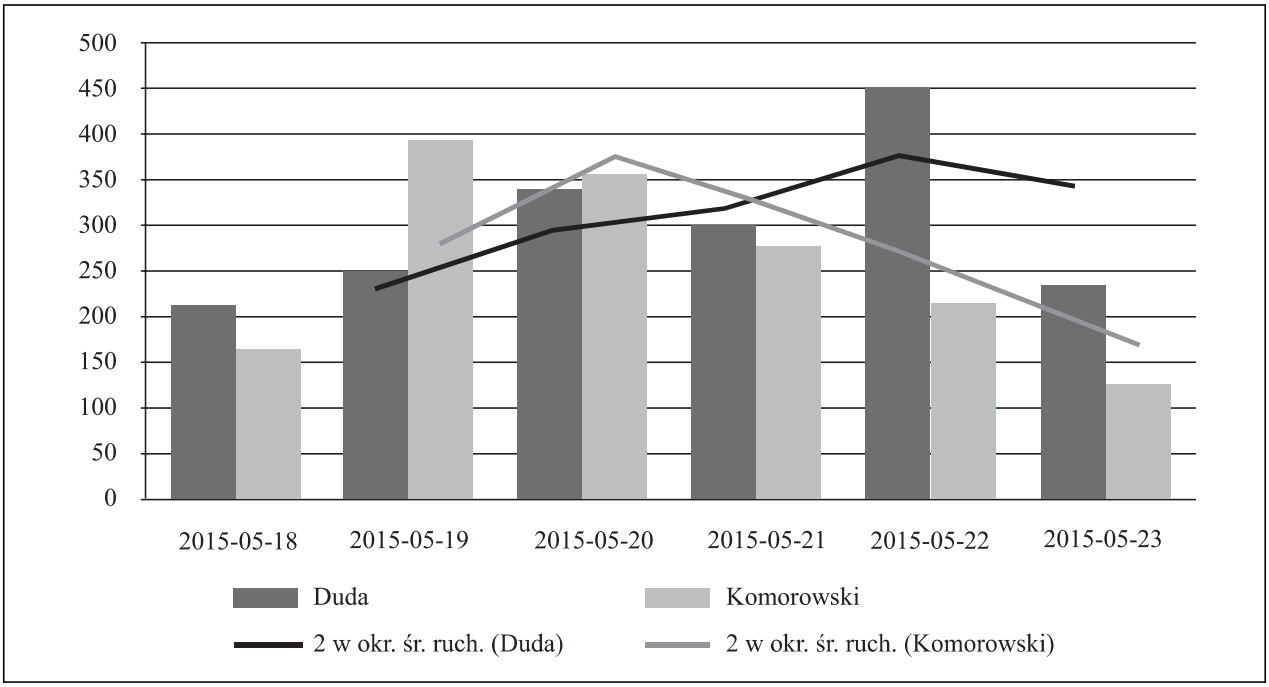
prezentowane w tym artykule są oparte na ich niewielkiej liczbie.

W odniesieniu do wyborów prezydenckich gromadzenie wpisów obejmowało okres 18–23 maja 2015 r. Ilustracją potencjału informacyjnego rafinacji przeprowadzonej na tej bazie (przy wykorzystaniu sentymentów wyróżnionych na podstawie badań Pisarka) są wartości funkcji liczb pozytywnych i negatywnych wpisów bezpośrednio przed wyborami prezydenckimi (rysunek 3.). Przedstawiona na nim wizualizacja jest jednoznaczna w odniesieniu do ostatecznych wyników wyborów prezydenckich.

Rysunek nr 3 stanowi fragment ilustracji wagi doboru stosowanych sentymentów. Sentymenty ST – dedykowane do charakteru/przedmiotu opisywanych badań – pozwoliły na uzyskanie wyników bardziej (od sentymentów P) zbliżonych do opisywanej rzeczywistości (rysunek 4.).

Wiarygodność uzyskiwanych w ten sposób danych została udowodniona w badaniach parlamentarnych 2011 r.[19]. Wymowna jest w nich także procentowa różnica (zaledwie 0,66%) pomiędzy liczbami pozytywnych sentymentów, zgromadzonymi w przeddzień wyborów Duda/Komorowski, która wynosi 2,44%, a rzeczywistą różnicą wyników kandydatów wynoszącą 3,10%.

W kontekście wiarygodności wyników uzyskiwanych z rafinacji warto podkreślić, że wartość współczynnika korelacji Pearsona pomiędzy danymi uzyskanymi z sondaży CBOS (czerwiec/lipiec) a wynikami rafinacji (rysunek 4.) wyniosła dla PIS/ZP $r = 0,97$ ($p<0,05$), a dla PO $r = 0,95$ ($p<0,05$).

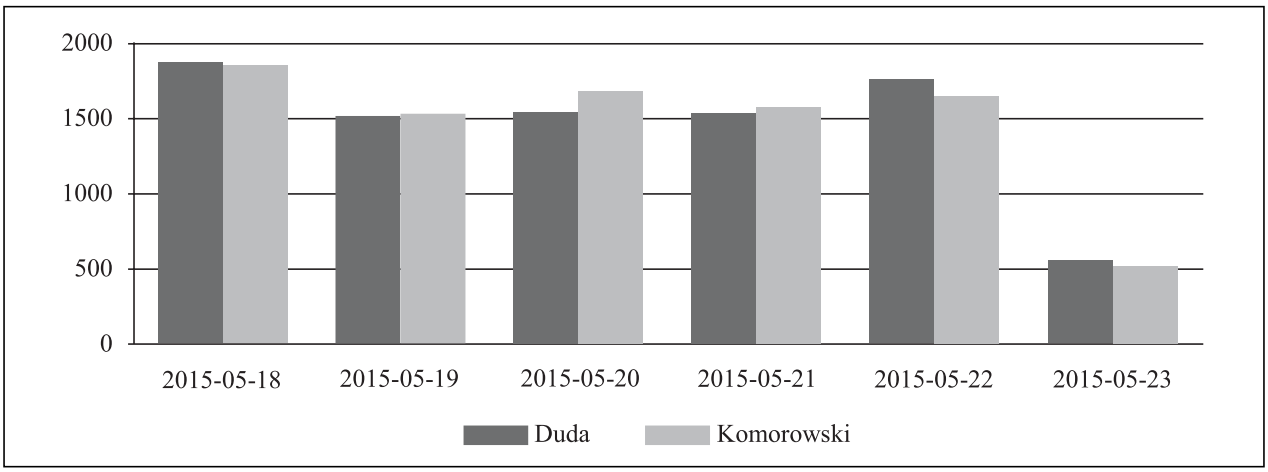


Rysunek 3. Ilustracja liczb pozytywnych sentymentów (sentymenty P)

Źródło: opracowanie własne[20]

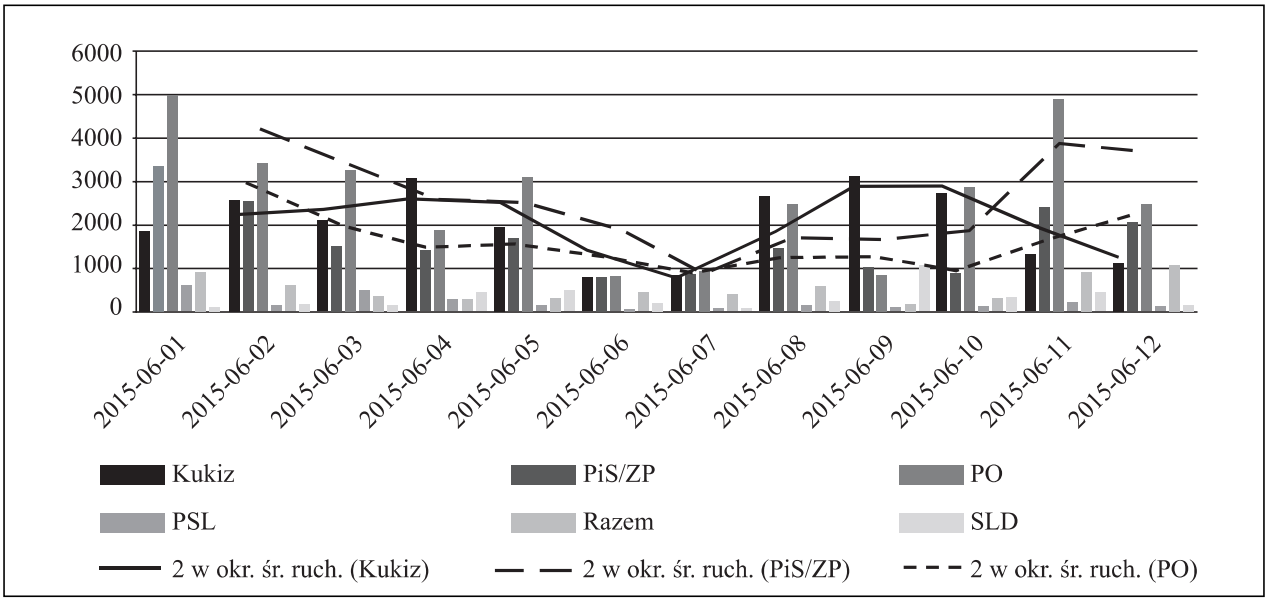
[19] W. Gogołek, P. Kuczma, *Rafinacja informacji sieciowych…*, dz. cyt.

[20] 2 w okr. śr. ruch. to linia trendu wyrównująca fluktuacje danych w celu lepszej ilustracji trendu zmian wartości zmiennej, https://support.office.com/pl-pl/article/Dodawanie-linii-trendu-i-linii-%C5%9Bredniej-do-wykresu-3-c4323b1-e377-43b9-b54b-fae160d97965?ui=pl-PL&rs=pl-PL&ad=PL [dostęp: lipiec 2015].



Rysunek 4. Ilustracja liczb pozytywnych sentymentów (sentymenty ST)

Źródło: opracowanie własne



Rysunek 5. Ilustracja liczb pozytywnych sentymentów dla wybranych partii

Źródło: opracowanie własne[21]

Dane zobrazowane w postaci wykresów (rys. 3, 4, 5) sygnalizują wstępne wyniki zastosowania przyjętej metodologii rafinacji w badaniach, które są prowadzone od maja 2015[22]. Ze względu na intencję zachowania neutralności ich pełne wyniki zostaną opublikowane dopiero po wyborach parlamentarnych.

Podsumowanie

Spektrum wartości poznawczej sygnalizowanych wyników badań tworzy nie tylko predykcja, aktualny (skala godzin), niskobudżetowy sondaż popularności osób, partii lub firm, ale także możliwość rozszerzonej analizy frekwencji sentymentów. Chodzi tu o zwrócenie uwagi

[21] 2 w okr. śr. ruch to linia trendu…, dz. cyt.

[22] Dotychczas (lipiec 2015) zgromadzono ponad 500 000 wpisów.

na ekstrema trendów i poszukiwanie przyczyn ich powstania (oraz odpowiedniej reakcji). Na przykład rysunek 3. ilustruje wybraną część majowych wyników rafinacji. Znacząca jest tam data 20 maja. Może ona wskazać czynniki (w tym znaczenie najczęściej występujących słów przyjętych jako sentymenty[23]), które spowodowały zmiany trendów. Podobnie przykładowa analiza ekstremów czerwcowych notowań najsilniejszych partii (rysunek 5.) zwraca uwagę na daty 7 i 9–10 czerwca jako znaczące. Dowodzą tego małe liczby pozytywnych sentymentów dotyczących partii. Liczby negatywnych sentymentów są także w dniu 7 czerwca mniejsze od liczb w innych dniach. Wskazuje to na wymagający dalszej analizy prawdopodobny związek liczb sentymentów ze zdarzeniami w Grecji, co oznaczałoby konieczność przeprowadzenia dodatkowych analiz korpusu zgromadzonych wpisów.

Generalizując dotychczasowe doświadczenia zastosowań rafinacji informacyjnej, najważniejszy okazuje się proces wyboru trafnych (liczonych frekwencjami) sentymentów. Uzyskiwane wyniki (trendy, ekstrema), poddane dalszej analizie, okazują się być ważnymi informacjami w ocenie przeszłego, aktualnego (liczonego godzinami) i przyszłego stanu badanego zdarzenia/procesu. Wydają się być cennym zbiorem determinant podejmowania decyzji/czynności mających wpływ na ocenę i przebieg badanego zdarzenia/procesu.

Warszawa, 28 lipca 2015

[23] Chodzi tu o znaczenie najczęściej występujących sentymentów w punktach ekstremów.