Rafinacja informacji sieciowych na przykładzie wyborów parlamentarnych.

Część 1. Blogi, fora, analiza sentymentów

Włodzimierz Gogołek, Paweł Kuczma

Korzystanie z zasobów internetu, a w szczególności sieci społecznościowych i tradycyjnych form internetowej dystrybucji informacji medialnych, staje się ważnym źródłem informacji dla badań społecznych, a szczególnie dziennikarstwa. Ów potencjał informacyjny jest pochodną komunikacyjnej siły internetu oraz potęgi dostępnych w nim zasobów informacyjnych i usługowych. Już w 2010 r. po raz pierwszy suma cyfrowych informacji wyprodukowanych na świecie w ciągu jednego roku przekroczyła jeden zeta bajt (10^{21}). Zasoby o tej skali, znane jako Big Data – ogromne nieustrukturyzowane hurtownie danych – przekroczyły krytyczną wielkość[1]. Stworzyły one nowy wymiar wartości i atrakcyjności zasobów informacyjnych do wszelkiego rodzaju badań, w tym związanych z badaniami społecznymi. Krytyczna wielkość oznacza nikłą użyteczność konwencjonalnych narzędzi do analizy tak dużych baz danych. Stanowi to uzasadnienie rozpoczęcia prac nad eksploracją/specjalistyczną analizą Big Data. Wyniki uzyskane z analizy Big Data tworzą wcześniej nieosiągalne źródła danych, których kreowanie może być postrzegane jako nowa faza rozwoju aplikacji IT (narzędzi i cyfrowych sieci wymiany informacji)[2].

Umiejętna analiza Big Data pozwala na precyzyjniejsze, w stosownym czasie, udostępnianie potrzebnych, krytycznych, a nawet wiarygodnie prognozujących informacji[3]. Pozwolą one doskonalić i rozwijać nowe generacje produktów i usług wykorzystywanych przez media.

Znaczącą część Big Data tworzą zasoby internetu, w tym sieci społecznościowe. Dane tego typu są tworzone przez i o indywidualnych użytkownikach sieci społecznościowych (blogi, posty, portale, maile czy strumień zapytań kierowanych do internetu), profesjonalne

[1] A. Beck, *Big Data Is Never Too Big When You Can Act On It*, www.clickz.com/print_article/clickz/column/2171482/act?wt.mc_ev=click&WT.tsrc=Email&utm_term=&utm_content=Print%20version&utm_campaign=05%2F02%2F12%20-%20Behavioral%20Marketing&utm_source=ClickZ%20Media&utm_medium=Email [dostęp: 22.04.2013].

[2] *Big data: The next frontier for innovation, competition, and productivity*, McKinsey Global Institute, www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation [dostęp: 22.04.2013].

[3] D. Copeland, *Harvard Researcher Uses Social Media To Predict Stock Market Volume*, http://readwrite.com/2012/02/08/harvard_researcher_uses_social_media_to_predict_st [dostęp: 20.04.2013].

publikacje i inne bogate zasoby informacyjne[4]. Interesującą częścią Big Data są zasoby ukryte w sieci (Dark Net) – Deep Web i pNet zbudowane na bazie P2P – nazywane F2F (przyjaciel do przyjaciela), np. Freenet. Zasoby te są tysiąc razy większe od dostępnych w tradycyjnej, indeksowanej przez wyszukiwarki sieci WWW[5].

Przyjęto, iż zasoby zgromadzone w Big Data tworzą informacje źródłowe, a wynik ich analizy to informacje wtórne. Proces owej analizy określany jest jako rafinacja informacji sieciowych (rafinacja).

Rafinacja

Jednym z ugruntowanych już filarów rafinacji jest *culturomics*, będąca „formą obliczeniowej leksykologii badającej ludzkie zachowania i tendencje kulturowe poprzez analizę ilościową zdigitalizowanych tekstów. Naukowcy eksplorują (*data mining*)[6] ogromne archiwa cyfrowe w celu zbadania zjawisk kulturowych poprzez ich odzwierciedlenie w języku i sposobie użycia wyrazów”[7]. Korzystanie z narzędzi *culturomics* sprawnie sygnalizuje ważne zmiany kulturalne, naukowe i historyczne. Rafinacja pozwala na dostrzeganie w ukrytych zasobach informacji pierwotnych (Big Data) – informacji wtórnych. Jest jak mikroskop umożliwiający pełniej oglądać i mierzyć rzeczy – na poziomie zarówno poszczególnych jednostek, jak i grup społecznych. Jest to rodzaj rewolucji w pomiarach. Uzyskane dzięki owym pomiarom dane tworzą obraz potrzeb i zachowań indywidualnych użytkowników, ale także społeczności jako całości.

Do rafinacji zasobów sieciowych mogą być bezpośrednio użyte takie narzędzia, jak np.: Attentio, Radian6, Sysomos, NetBase, Collective Intellect, Alterian, Google Alerts. Rafinację sieciową skutecznie przeprowadza się, wykorzystując Attentio Brand Dashboard[8]. Dowodzą tego wyniki badań dynamiki zmian obrazu informacyjnego kandydatów w wyborach prezydenckich 2010 r.[9] Innym profesjonalnym narzędziem rafinacji jest Summary of World Broadcasts (SWB) – usługa sieciowa monitorująca serwisy informacyjne. Umożliwia ona monitorowanie pełnych tekstów i streszczeń artykułów prasowych, materiałów pokonferencyjnych, materiałów telewizyjnych i radiowych oraz innych nieklasyfikowanych raportów technicznych (szarej literatury) w 130 językach[10].

---

[4] S. Stephens-Davidowitz, *Google's Crystal Ball*, http://campaignstops.blogs.nytimes.com/2012/10/20/googles-crystal-ball/ [dostęp: 20.04.2013].

[5] W. Boswell, *Five Search Engines You Can Use to Search the Deep Web,* http://websearch.about.com/od/invisibleweb/tp/deep-web-search-engines.htm [dostęp: 31.03.2012].

[6] *Data mining* – „eksploracja danych (spotyka się również określenie drążenie danych, pozyskiwanie wiedzy, wydobywanie danych, ekstrakcja danych) – jeden z etapów procesu odkrywania wiedzy z baz danych (ang. Knowledge Discovery in Databases, KDD). Idea eksploracji danych polega na wykorzystaniu szybkości komputera do znajdowania ukrytych dla człowieka (właśnie z uwagi na ograniczone możliwości czasowe) prawidłowości w danych zgromadzonych w hurtowniach danych”, por. http://pl.wikipedia.org/wiki/Eksploracja_danych [dostęp: 20.04.2013].

[7] Określenia *culturomics* użyli po raz pierwszy w końcu 2010 r. badacze z Uniwersytetu Harvarda Jean-Baptiste Michel i Erez Lieberman Aiden w artykule *Quantitative Analysis Of Culture Using Millions Of Digitized Books*, „Science” Vol. 331 (2011), nr 6014, s. 176–182, www.sciencemag.org/content/331/6014/176 [dostęp: 1.06.2011].

[8] *Attentio Brand Dashboard – monitoring mediów społecznościowych,* www.blog.mediafun.pl/attentio-brand-dashboard-monitoring-mediow-spolecznosciowych/ [dostęp: 20.04.2013], zob. też stronę firmy Attentio – http://attentio.com/.

[9] P. Kuczma, W. Gogołek, *Informacyjny potencjał sieci – na przykładzie wyborów prezydenckich 2010*, „Studia Medioznawcze” 2010, nr 4, s. 35–49.

[10] K.H. Leetaru*, Culturomics 2.0: Forecasting large-scale human behavior using global news media in time and space*, „First Monday” Vol. 16 (2011), nr 9, www.uic.edu/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/3663/3040 [dostęp: 20.04.2013].

Cel i zakres badania

Mając na uwadze potencjał Big Data, powtarzające się zapotrzebowania na bieżące informacje związane z wyborami na skalę krajową, w Instytucie Dziennikarstwa Uniwersytetu Warszawskiego w ramach jednej z prac badawczych przyjęto jako przedmiot i jednocześnie cel badań ilustrujących potencjał rafinacji sieciowej wskazanie i weryfikację narzędzi obróbki informacji umożliwiających ocenę bieżących preferencji wyborczych przed wyborami parlamentarnymi w Polsce w 2011 r. Podstawą do osiągnięcia założonego celu była ocena danych ilościowych i jakościowych oraz dynamika zmian treści ukazujących się w mediach społecznościowych oraz w sieciowych wydaniach niektórych gazet.

Równoległym celem badań było zarysowanie metodologii stanowiącej podstawowy element rafinacji sieciowej. Metodologia ta posłużyła do poszukiwania wspomnianych ocen preferencji wyborczych na podstawie informacji pozyskiwanych z sieci.

Osiągnięcie założonych celów pozwoliło wskazać sposób kreowania miarodajnego źródła danych wspomagających diagnozowanie stanu i dynamiki zmian obrazu informacyjnego aktywności komitetów wyborczych (partii politycznych) biorących udział w wyborach. Wiedza ta może stanowić wartościowe źródło informacji o przebiegu kampanii wyborczej dla mediów, zainteresowanych osób i grup społecznych.

Podobne badanie zostało przeprowadzone w 2010 r. przy okazji wyborów prezydenckich[11]. Jego wyniki w pełni potwierdziły zasadność kontynuowania ścieżki badawczej opartej na rafinacji informacji sieciowych[12].

Przyjęto następującą hipotezę: rafinacja sieci umożliwia bieżący, wiarygodny monitoring zmiennych opisujących preferencje wyborcze Polaków w okresie poprzedzającym wybory parlamentarne w 2011 r.

Hipoteza ta jest równoznaczna z twierdzeniem, że treści w sieci, szczególnie w mediach społecznościowych, są odzwierciedleniem rzeczywistych postaw użytkowników i mogą zapowiadać ich realne działania, takie jak oddanie głosu na kandydata, partię, wybór określonej odpowiedzi w referendum. Istnieje tym samym statystyczna zależność między ilościowymi miarami treści powstających w sieci a preferencjami politycznymi, których efektem jest wybór określonej opcji politycznej.

Do badania zostały zakwalifikowane komitety wyborcze powiązane z partiami/środowiskami politycznymi, których członkowie zasiadali w Sejmie RP 1 stycznia 2011 r. (w tym nowo powstałe twory polityczne obecne w sejmie związane z posłem Januszem Palikotem i Joanną Kluzik-Rostkowską), czyli: Platforma Obywatelska RP (PO), Polskie Stronnictwo Ludowe (PSL), Prawo i Sprawiedliwość (PiS), Sojusz Lewicy Demokratycznej (SLD), Polska Jest Najważniejsza (PJN)[13] i Ruch Palikota (RP)[14].

W celu dokonania ilościowej oceny krotności występowania nazw partii w tekstach zamieszczanych w sieci wyróżniono odpowiednie konteksty. Były nimi klucze (zwroty/słowa) wiązane z rządem i jego funkcjami oraz kompetencjami poszczególnych ministerstw[15], przyjęte jako konteksty merytoryczne: 1) edukacja, 2) finanse, 3) gospodarka, 4) infrastruktura, 5) kultura, 6) nauka i szkolnictwo wyższe, 7) obrona, 8) praca i polityka społeczna, 9) rolnictwo, 10) rozwój regionalny, 11) Skarb Państwa, 12) sport i turystyka, 13) sprawiedliwość, 14) sprawy wewnętrzne i administracja, 15)

[11] P. Kuczma, W. Gogołek, *Informacyjny potencjał...*

[12] Zawarte w tekście analizy i wnioski opracowane zostały przez W. Gogołka na podstawie danych źródłowych zebranych, zweryfikowanych i odpowiednio przetworzonych przez P. Kuczmę.

[13] Polska Jest Najważniejsza została zarejestrowana jako partia polityczna 17 marca 2011 r.

[14] Wcześniej Ruch Poparcia, a jako partia Ruch Palikota został zarejestrowany 1 czerwca 2011 r.

[15] Na podstawie struktury Rady Ministrów za: Postanowienie Prezydenta Rzeczypospolitej Polskiej z dnia 16 listopada 2007 r. nr 1131-50-07 o powołaniu w skład Rady Ministrów, M.P. 2007, nr 87, poz. 947.

sprawy zagraniczne, 16) środowisko oraz 17) zdrowie. Słowa opisujące kompetencje każdego z ministerstw zostały oparte na kompetencjach ministerstw zapisanych w ich statutach[16].

Drugą grupę kontekstów – konteksty medialne – stanowią te, które związane są z bieżącymi wydarzeniami relacjonowanymi w mediach. Zostały one wyłonione na podstawie formalnej analizy treści prasowych (przy użyciu programu QDA Miner v3.2 wraz z WordStat 6.0.1)[17] z największych polskich dzienników opiniotwórczych (w wersji elektronicznej)[18] o odmiennym profilu politycznym: „Gazety Wyborczej" oraz „Rzeczpospolitej". Do tej analizy wykorzystano elektroniczną wersję dzienników dostępnych za pomocą wyszukiwarki Factiva[19]. Artykuły pochodziły z okresu 1–28 lutego 2011 r., czyli z miesiąca poprzedzającego rozpoczęcie właściwego badania. Wszystkie artykuły wraz tytułami przeanalizowano pod względem ilościowym. Otrzymano w ten sposób listę 39 153 słów. Spośród nich wybrano 1000 słów, które powtarzały się statystycznie istotnie najczęściej – przynajmniej 32 razy we wszystkich analizowanych artykułach[20]. Ponieważ wśród analizowanych słów niektóre powtarzały się (np. w różnych przypadkach lub pojawiały się wyrazy bliskoznaczne w badanym zbiorze), wyodrębniono osiem grup/zbiorów słów, zwanych grupami kontekstów. W wyniku tej analizy wyłoniono następujące grupy kontekstów medialnych: 1) UE (Unia Europejska) – w tym m.in. takie słowa, jak unia, UE, europejski, prezydencja, Europa; 2) katastrofa (smoleńska) – katastrofa, smoleńska, Rosja, MAK, tragedia, tupolew, zamach; 3) władza – rząd, zarząd, władza, sejm, lider, prezydent; 4) media – media, gazeta, TVP, TVN, telewizja; 5) pieniądze – pieniądze, finanse, budżet, NBP; 6) reformy – reformy; 7) kościół – kościół; 8) prawo – prokuratura, prawo, ustawa, sąd, trybunał itp.

Uzyskane w ten sposób konteksty medialne zostały użyte do analizy merytorycznego charakteru kampanii wyborczej w 2011 r. Jej celem była próba odpowiedzi na pytanie, czy w treściach dostępnych online, intensywniej w sensie ilościowym, reprezentowane są konteksty merytoryczne czy konteksty medialne[21].

W trakcie badania zasadniczego analizowane były treści publikowane w mediach społecznościowych (fora internetowe, blogi, Facebook, Twitter), gdzie treści tworzone są przez samych użytkowników, i w serwisach informacyjnych, tworzonych przez profesjonalne redakcje. Przyjęto, iż pojedynczy wpis, rekord, fragment blogu, pobrany z sieci do dalszej analizy, nazywany będzie terminem: 'wpis'.

Rafinację danych z zasobów sieciowych przeprowadzono na zbiorach opublikowanych od 1 marca do 17 października 2011 r. Monitoring, archiwizację oraz wstępną analizę kontekstową treści publikowanych w sieci wykonano za pomocą narzędzia Attentio Brand Dashboard. Dane zostały wyróżnione na podstawie słów kluczowych (w tym przypadku – kontekstów) opisujących badane partie polityczne[22].

---

[16] Spis statutów zob. www.id.uw.edu.pl/zasoby/profile/59/Aneks_nr_2-Wykaz_statutow_ministerstw.pdf [dostęp: 23.04.2013].

[17] Programy dostępne na stronie: www.provalisresearch.com/Download/download.html [dostęp: 31.05.2010]. Używana była wersja testowa.

[18] *„Gazeta Wyborcza" i „Fakt" to najchętniej czytane dzienniki*, www.wirtualnemedia.pl/artykul/gazeta-wyborcza-i-fakt-to-najchetniej-czytane-dzienniki# [dostęp: 24.05 2010].

[19] https://han.buw.uw.edu.pl/han/ISIEM/site.securities.com/search/pub_search.html?pc=PL&sv=EMIS [dostęp: maj 2010].

[20] W związku z tym, że słowo na miejscu 1000. miało częstotliwość występowania 32, do analizy włączono wszystkie słowa z częstotliwością przynajmniej 32. Było ich w sumie 1016.

[21] Wyniki tych badań zostaną opublikowane w drugiej części opracowania *Rafinacja informacji sieciowych na przykładzie wyborów parlamentarnych. Konteksty medialne i merytoryczne*.

[22] Definicja monitorowanych słów została zamieszczona w Aneksie opublikowanym na stronie: www.id.uw.edu.pl/zasoby/profile/59/Aneks_nr_1-definicja_wyszukiwania.pdf.

Badaniu zostały poddane następujące wskaźniki pozyskane z analizy treści w sieci:
– ilość treści o danej partii,
– trendy/dynamika zmian ilości treści[23],
– ocena jakościowa, czyli wyniki analizy kontekstów, w których pojawiają się treści dotyczące partii, oraz zabarwienie treści, czyli sentyment – ilościowy podział treści na pozytywne lub negatywne.

Wskaźnik ilości treści został opracowany na podstawie liczby wszystkich wpisów/informacji w plikach zebranych przez Attentio Brands Dashboard, pochodzących ze źródeł online dotyczących danej partii oraz kontekstów. Wpisy zawierały treści uzyskane z for, blogów, Facebooka, tweetów i artykułów gazetowych.

Ocena dynamiki zmian i trendów dotyczących treści/wyszukiwania została dokonana na podstawie liczby wpisów dotyczących partii w zależności od kontekstów i sentymentów w analizowanym czasie.

Ocena jakościowa została przeprowadzona w dwóch kategoriach. Pierwszą z nich była wstępna analiza kontekstowa polegająca na pogrupowaniu pozyskanych w wyniku monitoringu treści w konteksty na podstawie listy kontek-stów merytorycznych i medialnych omówionych powyżej. Drugą była – przeprowadzona równolegle z analizą kontekstową – wstępna analiza sentymentów rozumiana jako wyróżnianie wpisów, które zawierały dowolną nazwę partii oraz słowo uznane jako „sentyment". Ze względu na wagę analizy sentymentów stanowi ona odrębną część całego badania. Wyróżnienia słów uznanych za 'sentymenty', w związku z brakiem autorytatywnej listy takich słów polskich, wykonano na bazie listy wyrażeń nasyconych emocjonalnie ANEW 2012[24]. Spośród 1031 słów z tego zbioru wybrano słowa skrajnie pozytywne i skrajnie negatywne, a wśród nich te, które w przywołanym zbiorze występowały najczęściej. Słowa te przetłumaczono następnie na język polski[25], rozszerzając, w razie potrzeby, ich znaczenie, np. przy tłumaczeniu słowa 'love' użyto zarówno formy 'miłość' – rzeczownik, jak i 'ko-

Tabela 1. Lista słów pozytywnych i negatywnych wraz z częstotliwością ich występowania we wpisach

Słowa	Częstotliwość
Słowo pozytywne	
pewny	930
wygrać, wygrana	869
cel	840
samochód	769
entuzjasta, entuzjazm	573
dziwić, zadziwiać	500
zysk, zyskiwać	407
sukces	397
okazja	341
impreza, przyjęcie	312
zabawa, zabawka	311
miłość, kochać	306
lubić	299
bogaty, bogactwo	297
Słowo negatywne	
winien, winny, wina	616
wojnę	516
trudno, trudny	396
śmierć	345
niszczyć, zniszczenia	344
katastrofa	308
najgorszy, pogarszać	279
wypadek	275
bankrut, bankructwo, upadek, upadłość, bankrutować, upadać	247

[23] Pod tym pojęciem rozumiane są relacje między ilością treści dotyczących poszczególnych kandydatów w badanym okresie.

[24] M.M. Bradley, P.J. Lang, *Affective Norms for English Words (ANEW)*. Lista wszystkich słów nie jest publicznie dostępna. Została uzyskana na specjalną prośbę autorów od jej twórców z The Center for the Study of Emotion and Attention z Uniwerytetu Florydy, http://csea.phhp.ufl.edu/media/anewmessage.html [dostęp: 2.04.2013].

[25] Tłumaczenie zbioru ANEW zostało dokonane za pomocą usługi webowej Google Translate (http://translate.google.com/).

chać' – czasownik itd. Słowa wykorzystane do analizy zawiera tabela 1.

Filtrowanie wpisów związanych z określoną partią

Pierwszy etap filtrowania

W związku z tym, że monitoring pozyskanych online treści przyniósł wyniki pokazujące ponadstandardową przewagę Platformy Obywatelskiej w liczbie wyróżnionych wpisów, została przeprowadzona dodatkowa analiza pozyskanych wpisów polegająca na zwiększeniu trafności w przypisaniu wpisów do odpowiedniej partii. W tym celu został napisany program[26], który składa się z dwóch modułów. Pierwszy konwertuje dane pozyskane z monitorowania źródeł online za pomocą aplikacji Attentio. Konwersja służy przygotowaniu pliku do etapu weryfikacji w module drugim i nie pływa na treść konwertowanego pliku. Drugi moduł weryfikuje, czy w treści i tytule wpisu znajdują się odniesienia do partii, której dany wpis został przyporządkowany. Do tego celu wykorzystano stosowne wzorce wariacji/odmian (wynikających m.in. z gramatyki polskiej) nazw partii.

Drugi etap filtrowania

Filtrowanie pierwszego etapu nie zapewniło zadawalających efektów, zwłaszcza w odniesieniu do wpisów dotyczących Platformy Obywatelskiej – w dużej części wpisy przyporządkowywane tej partii jej nie dotyczyły z powodu wielokrotnego występowania przyimka 'po'. Stanowiło to konieczności uruchomienia drugiego stopnia weryfikacji treści.

W celu precyzyjniejszego automatycznego wybierania wpisów dotyczących wyróżnionych partii – np. skrótowca partii PO i odróżniania go od przyimka 'po', opracowany został odpowiedni program analizujący dane wejściowe pod kątem wspólnego występowania wyznaczonych słów i zbioru wyrazów je określających[27]. Zliczane były wystąpienia tychże wyrazów wraz z wyznaczonymi słowami w polach tekstowych zawierających istotne z punktu widzenia analizy dane.

Do analizy występowania poszukiwanych słów program wykorzystuje wyrażenia regularne. Dane wejściowe poddawane są filtrowaniu przed rozpoczęciem analizy. Filtrowanie obejmuje usunięcie formalnych sekwencji znaków, które znalazły się w zbiorze danych uzyskanym po wstępnej analizie kontekstowej (znaczniki html, encje html, sekcje „script"), zamianę wielokrotnych białych znaków na pojedyncze spacje, zamianę wielokrotnych znaków interpunkcyjnych na pojedyncze oraz usunięcie zbędnych spacji przed znakami interpunkcyjnymi.

Program zlicza wspólne wystąpienia każdego wzorca nazwy partii z każdym wzorcem określającym sentyment/emocję (konteksty były definiowane na etapie pozyskiwania treści z sieci). Dodatkowo określono limit znaków, w obrębie którego musi znaleźć się wzorzec partii, sentymentu, aby taka para została zliczona. Przyjęta liczba 30 znaków jest połową średniej liczby znaków zdania w języku polskim. Badania sondażowe dowiodły, iż wielkość ta jest optymalna i pozwala w ten sposób wyróżnione wpisy uznać za związane z partią i przyjętym wzorcem (kontekstem, sentymentem). Wzorce kontekstów i sentymentów zliczane są dla lewostronnych i prawostronnych wystąpień względem wzorca nazwy partii.

Wzorce używanych w mediach społecznościowych nazw partii określono na podstawie najczęściej występujących (również kolokwialnych i pejoratywnych) określeń partii występujących w mediach społecznościowych. W tym celu stworzono plik wzorców nazw partii.

[26] Autorem tego programu jest Marcin Łączyński.
[27] Autorem programu jest Piotr Celiński, pgc@post.pl.

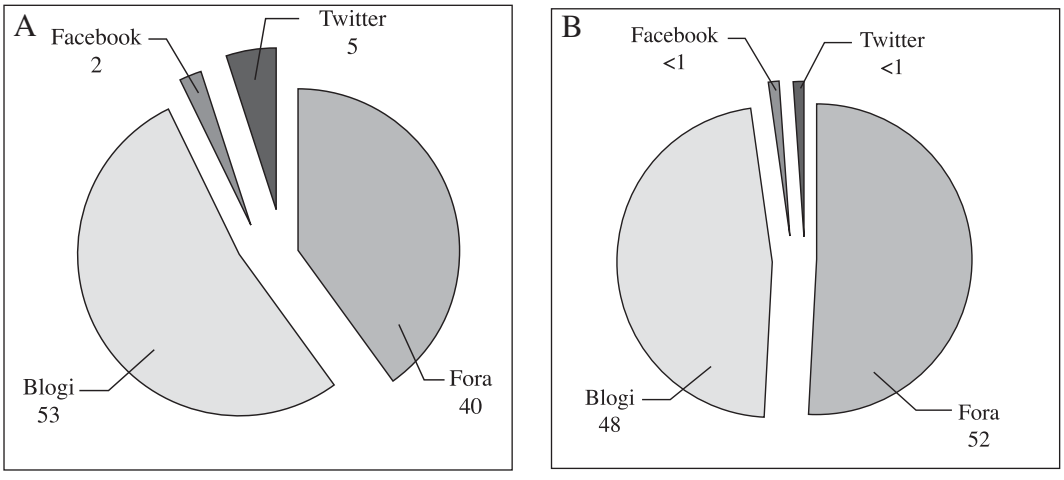
Określenie partii składa się z nazwy, która znajdzie się w pliku wyjściowym, oraz wzorca wyrażenia regularnego opisującego możliwe nazwy (i ich warianty oraz odmiany) występujące w mediach społecznościowych.

Definicja nazw partii zakłada dla każdej z partii uwzględnienie ich skrótowców, czyli PiS, PJN, PO, PSL, RP, SLD, pisanych wielkimi i małymi literami[28] oraz odmiany nazw partii, czyli np. dla partii Polskie Stronnictwo Ludowe definicja uwzględnia następujące słowa kluczowe: PSL, psl, PsL, Psl, pSl, psL, Polskie Stronnictwo Ludowe i odmiany tej nazwy przez wszystkie, występujące w języku polskim, przypadki w liczbie pojedynczej, czyli: polskie stronnictwo ludowe, polskiego stronnictwa ludowego, polskiemu stronnictwu ludowemu, polskim stronnictwem ludowym, polskim stronnictwie ludowym, a także występujące we wszystkich przypadkach liczby mnogiej sformułowanie, którym członkowie tej partii są medialnie określani, czyli: ludowcy, ludowców, ludowcom, ludowcach, ludowcami (również pisane wielką lub małą literą). Analogiczne definicje zostały stworzone dla pozostałych partii.

Dane wejściowe

Dane wejściowe do pierwszego etapu rafinacji obejmowały wszystkie dostępne, związane z partiami politycznymi w okresie: marzec – wrzesień 2011 r., zasoby informacyjne forów, blogów, Facebooka, Twittera i sieciowe portale informacyjne[29].

Za pomocą narzędzia Attentio w pierwszym etapie pozyskano z portali społecznościowych 1 418 267 wpisów, z tego 565 868 z forów, 754 850 z blogów, 23 882 z Facebooka i 73 667 z Twittera. W etapie następnym, po pierwszym etapie filtrowania treści, zostało 339 403 wpisów, z czego 175 356 z forów, 162 527 z blogów, 197 z Facebooka i 1323 z Twittera.



Wykres 1. Rozkład liczb wpisów z portali społecznościowych A – przed filtrowaniem i B – po filtracji (w proc.)

Źródło: Obliczenia własne.

[28] Wyjątek stanowi skrótowiec PO pokrywający się z przyimkiem ‘po’, co już wyżej wyjaśniono.

[29] Liczba pozyskanych wpisów z niektórych gazet online w pierwszym etapie filtrowania przekroczyła 550 000, a w następnym – 100 000. Szczegółowe dane uzyskane z gazet online i wyniki ich analizy zostaną przedstawione w drugiej części opracowania w odrębnym artykule.

Ze względu na względnie małe liczby wpisów na Facebooku i Twitterze, dane pozyskane z tych źródeł nie podlegały dalszej analizie.

W związku ze względnie dużymi wielkościami pozyskanych danych źródłowych do obliczeń związanych z wyróżnieniami wpisów – pierwszy etap rafinacji – konieczne było użycie komputera o dużej mocy obliczeniowej[30]. Do analizy wpisów wykorzystywano m.in. komputer Boreasz – IBM Power 775 udostępniony w ramach Programu Obliczeń Wielkich Wyzwań Nauki i Techniki (POWIEW) oraz serwery Instytutu Dziennikarstwa UW.

Istotnym fragmentem procedury wykorzystania danych źródłowych do dalszych badań było wskazanie zmiennych niezależnych. Stanowiły one punkt odniesienia do oceny wiarygodności uzyskanych wyników rafinacji. Przyjęto, iż owe zmienne tworzą:
– koszty poniesione przez partie na kampanię wyborczą (tabela 2),
– liczby głosów oddanych na poszczególne partie (tabela 3),
– wyniki sondaży przedwyborczych CBOS (tabela 4).

Tabela 2. Koszty kampanii (w złotych)

	PO	**PiS**	**RP**	**SLD**	**PSL**	**PJN**
Utworzenie i utrzymanie strony internetowej	60 709,57	924	2 076,98	108 100,52	41 572,90	20
Korzystanie ze środków masowego przekazu i nośników plakatów	15 881 840,10	16 224 914	572 172,52	14 334 58,14	5 018 560,83	647 485,71
Reklama w internecie (koszt usługi emisji)	2 206 623,42	1 170 403	52 438,76	1 144 647,58	321 025,43	31 284,25
Udział wydatków na internet w całości wydatków na media (w proc.)	13,89	7,21	9,16	7,99	6,4	4,83
Wykonanie materiałów wyborczych, w tym prace koncepcyjne, projektowe i wytworzenie	6 195 905,01	7 477 332	854 201,60	6 694 452,58	4 926 782,56	710 929,75
Reklama w internecie	531 144,33	344 725	29 892,90	47 859,51	28 918,65	27 727,19
Udział wydatków na internet w całości wydatków na kreacje (w proc.)	8,57	4,61	3,5	0,71	0,59	3,9
Łącznie	22 138 455	23 703 170	1 428 451	21 137 211	9 986 916	1 358 435
Łącznie wydatki na działania w internecie	2 798 477	1 516 052	84 409	1 300 608	391 517	59 031
Udział wydatków na internet w całości kosztów (w proc.)	12,64	6,4	5,91	6,15	3,92	4,35

Źródło: Komunikat Państwowej Komisji Wyborczej z dnia 13 lutego 2012 r. w sprawie sprawozdań finansowych komitetów wyborczych uczestniczących w wyborach do Sejmu Rzeczypospolitej Polskiej i do Senatu Rzeczypospolitej Polskiej, przeprowadzonych w dniu 9 października 2011 r. (przekazany do ogłoszenia w Monitorze Polskim), http://pkw.gov.pl/wybory-do-sejmu-rp-i-do-senatu-rp-2011/komunikat-panstwowej-komisji-wyborczej-z-dnia-13-lutego-2012-r.html [dostęp: 12.11.2012].

Tabela 3. Liczba głosów oddanych w wyborach parlamentarnych 2011 r. – wyniki wyborów

	PO	PiS	RP	SL	PSL	PJN
Liczba głosów	5 629 773	4 295 016	1 439 490	1 184 303	1 201 628	315 393
Liczba głosów w proc.	39,18	29,89	10,02	8,24	8,36	2,19

Źródło: Obwieszczenie Państwowej Komisji Wyborczej z dnia 11 października 2011 r. o wynikach wyborów do Sejmu Rzeczypospolitej Polskiej przeprowadzonych w dniu 9 października 2011 r. (Dz.U. 2011, nr 218, poz. 1294), http://pkw.gov.pl/g2/2011_10/0a409df0d614d3d42f854615f3ab6286.pdf [dostęp: 12.11.2012].

Tabela 4. Wyniki sondaży przedwyborczych 2011 r. przeprowadzonych przez CBOS (w proc.)

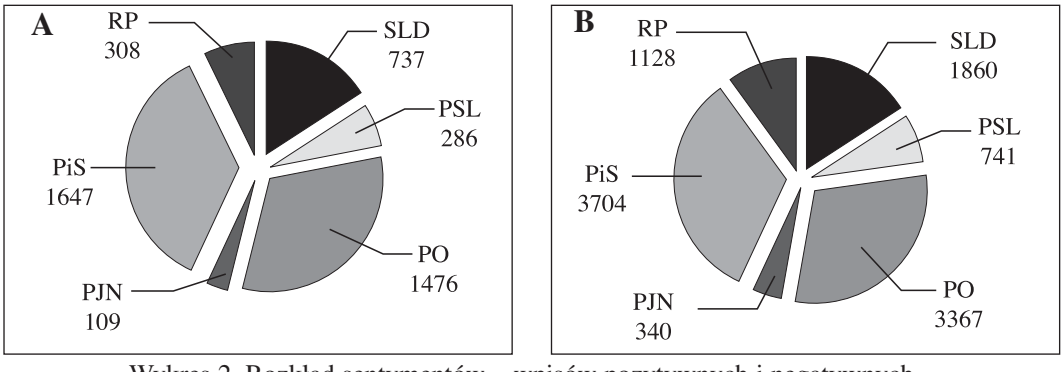
Miesiąc	PO	PiS	RP	SLD	PSL	PJN
Marzec	35	18	1	16	4	3
Kwiecień	31	23	1	12	5	2
Maj	37	21	1	12	4	1
Czerwiec	34	22	1	11	5	3
Lipiec	38	17	1	9	4	0
Sierpień	36	20	1	8	4	0
Wrzesień	37	20	2	7	6	1
Październik	34	20	7	9	6	1

Źródło: *Preferencje partyjne przed wyborami*, www.cbos.pl/SPISKOM.POL/2011/K_124_11.PDF [dostęp: 12.11.2012].

Analiza jakościowa

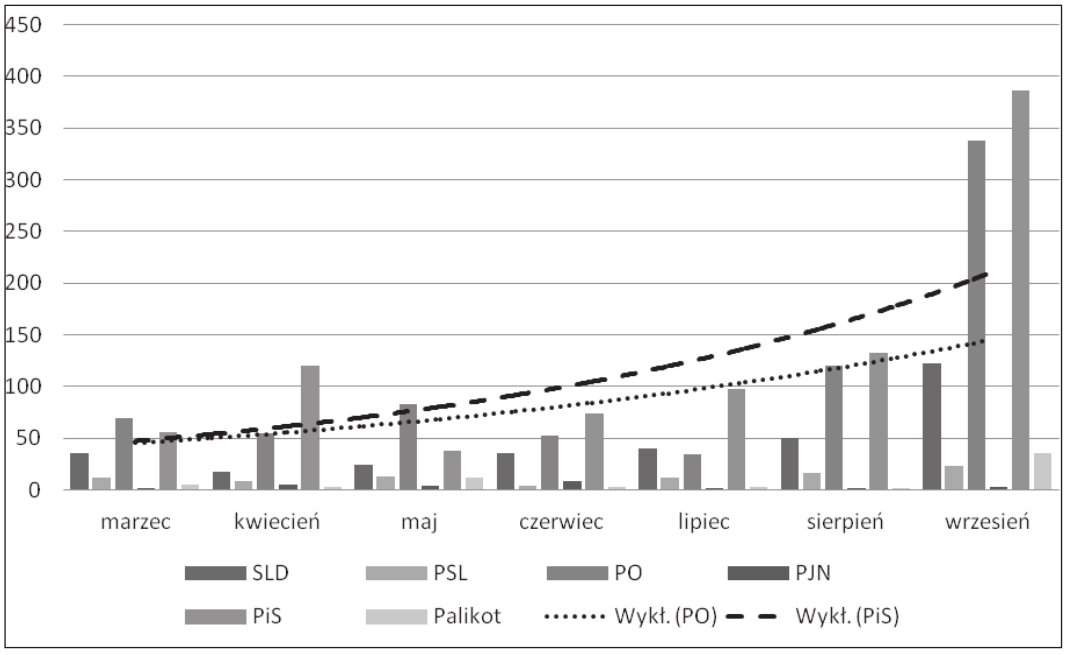
Przystępując do jakościowej analizy uzyskanych wyników, obliczono korelację Pearsona – **r** liczby głosów uzyskanych przez wszystkie partie z wynikami sondaży CBOS. Wynosi ona 0,96 ($p > 0{,}001$). Stanowi to przyjętą formę oceny wiarygodności sondaży przeprowadzanych przez CBOS i uzasadnienie założenia, iż wyniki CBOS (w poszczególnych miesiącach) stanowić będą odniesienie do dalszych badań.

Ze względu na dominujące wielkości liczby wpisów pozytywnych i negatywnych (sentymentów) dotyczących PO i PiS (fora – 62 proc., blogi – 67 proc.) dalszą analizę informacyjnej siły rafinacji przeprowadzono na przykładzie tylko tych dwóch partii (por. wykres 2).



Wykres 2. Rozkład sentymentów – wpisów pozytywnych i negatywnych – (marzec – październik) na forach (A) i blogach (B) dla wszystkich partii (w proc.)

Źródło: Obliczenia własne.



Wykres 3. Rozkład liczby wpisów negatywnych (sentymentów) na forach oraz wykładnicze krzywe trendów (wykł.) zmian tych liczb dla PO i PiS

Źródło: Obliczenia własne.

Korelacja **r** liczby głosów uzyskanych przez partie z liczbą pozytywnych wpisów na blogach wynosi 0,93 ($p > 0{,}001$) – por. wykres 5. Dowodzi to niemal pewnej zależności uzyskanych z rafinacji wyników z rzeczywistymi wynikami głosowania. Wskazuje jednocześnie na zasadność pogłębionej analizy zasygnalizowanej prawidłowości – znaczące zbieżności wyników rafinacji z oficjalnymi wynikami głosowań.

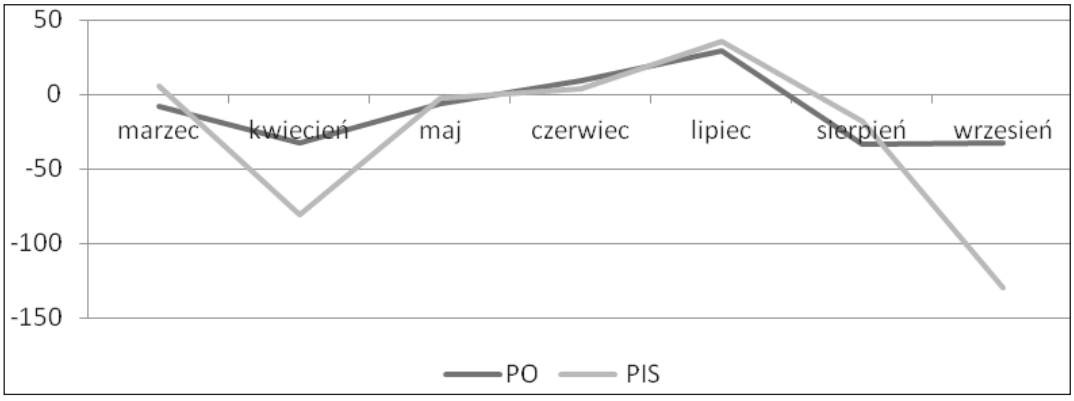
Fora

W trzecim etapie rafinacji dokonano wstępnej analizy ilościowej uzyskanych wyników – głównie na podstawie ich wizualizacji. Poza bezpośrednimi wynikami uzyskanymi z rafinacji, dodatkową formą wizualizacji są krzywe ilustrujące trend zmian liczby sentymentów – wpisów pozytywnych i negatywnych – dla wiodących w wyborach dwóch partii: PiS i PO.

Dzięki temu wyróżniono widoczne prawidłowości/zależności zmiennych uzyskanych z rafinacji wpisów. Liczba sentymentów – wpisów pozytywnych i negatywnych – na forach wskazała celowość obliczania różnic między liczbami tych wpisów. Wyraźną siłę predykcyjną ostatecznych wyników wyborów na podstawie wyników rafinacji pokazuje wykres 4.

Informacyjnej wagi uzyskanych, dzięki rafinacji, informacji dowodzą miary ilościowe zależności zmiennych badań: zmiennych niezależnych (danych źródłowych) z wynikami rafinacji (zmiennych zależnych).

W pierwszym etapie dokonano obliczeń korelacji ilościowych wyników sondaży z wynikami rafinacji. Przykładem tego jest korelacja między liczbą pozytywnych i negatywnych wpisów (sentymentów) na forach a wynikami sondaży CBOS dla wszystkich komitetów wyborczych w kolejnych miesiącach – od marca



Wykres 4. Rozkład różnic między liczbą wpisów pozytywnych a liczbą wpisów negatywnych (sentymentów) na forach dla PO i PiS

Źródło: Obliczenia własne. Wykres obrazuje bilans między wpisami negatywnymi i pozytywnymi, wartości dodatnie odzwierciedlają przewagę wpisów pozytywnych).

Tabela 5. Korelacja **r** związku wartości wyników sondaży CBOS z liczbą pozytywnych i negatywnych wpisów (sentymentów) na forach dla wszystkich partii w kolejnych miesiącach

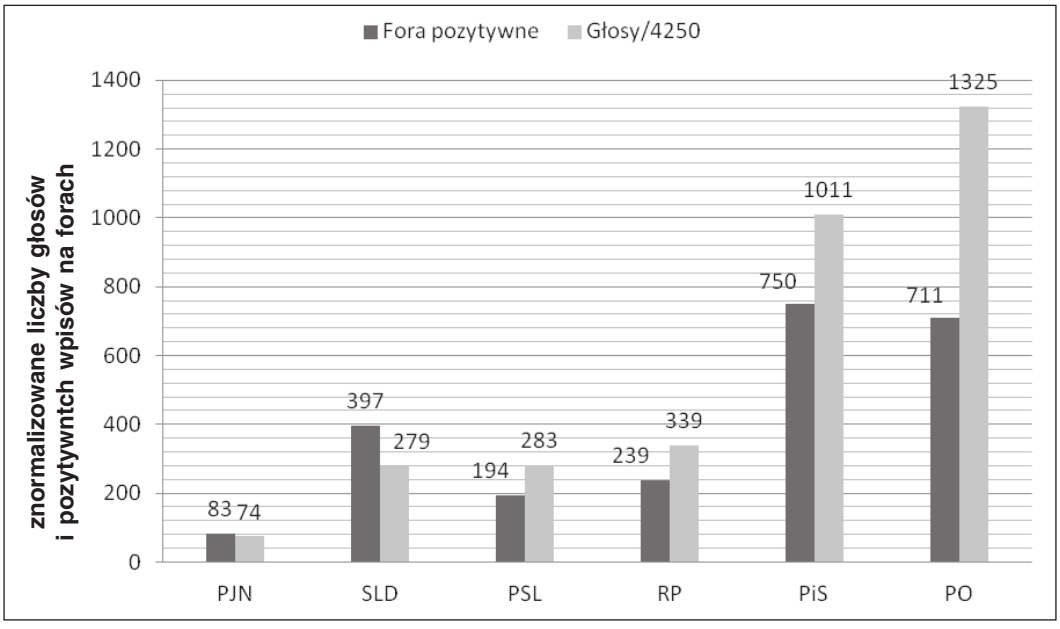
Miesiąc	r dla wpisów pozytywnych	r dla wpisów negatywnych
Marzec	0,70	0,95
Kwiecień	0,81	0,76
Maj	0,83	0,98
Czerwiec	0,89	0,84
Lipiec	0,55	0,46
Sierpień	0,88	0,93
Wrzesień	0,90	0,87
Październik	0,23	0,12

Źródło: Obliczenia własne.

do października 2011 r. Wartości tych współczynników są statystycznie znaczące ($p > 0,001$) z wyjątkiem października (ze względu na datę wyborów nie były to dane z całego miesiąca). Dowodzi to statystycznie istotnej zbieżności wyników sondaży z wynikami rafinacji. Wykazano także istotny związek między liczbą pozytywnych wpisów (sentymentów) na forach z liczbą uzyskanych przez partie głosów (wykres 5).

Współczynnik korelacji liczby pozytywnych wpisów na forach z liczbą głosów wynosi 0,93 ($p < 0,001$).

Zasygnalizowane związki: wpisów pozytywnych z sondażami i z głosami dowodzą oczekiwanej wiarygodności informacji pozyskiwanych z rafinacji i potwierdzają przyjętą hipotezę, iż rafinacja sieci umożliwia bieżący monitoring zmiennych preferencji wyborczych.



Wykres 5. Ilustracja podobieństwa proporcji (nie bezwzględnej wielkości) pozytywnych wpisów (sentymentów) na forach z liczbą uzyskanych głosów

Źródło: Obliczenia własne. W celu poprawienia efektu wizualizacji danych dokonano normalizacji danych wejściowych (liczbę głosów pomniejszono 4250 razy).

Tabela 6. Liczba pozytywnych i negatywnych wpisów (sentymentów) na forach

Miesiąc	SLD		PSL		PO		PJN		PiS		RP	
	pozytywny	negatywny	pozytywny	negatywny	pozytywny	negatywny	pozytywny	negatywny	pozytywny	negatywny	pozytywny	negatywny
Marzec	32	35	52	12	61	69	4	1	62	56	1	5
Kwiecień	8	18	3	9	22	55	2	5	39	120	5	3
Maj	68	24	4	13	77	83	2	4	35	38	5	12
Czerwiec	23	35	12	4	61	52	7	9	78	74	3	3
Lipiec	36	40	23	12	63	34	10	2	134	98	16	3
Sierpień	51	57	49	16	97	131	12	2	107	125	19	1
Wrzesień	146	122	47	23	305	338	44	3	256	386	133	36
Październik	33	9	4	3	25	3	2	0	39	0	57	6
Łącznie	397	340	194	92	711	765	83	26	750	897	239	69

Źródło: Obliczenia własne.

Blogi

Nieco większą wartość informacyjną dotyczącą liczby głosów uzyskanych przez partie od forów mają blogi.

nych z rafinacji wpisów na blogach. Można przyjąć, że blogi są istotnym źródłem informacji o liczbie głosów uzyskanych w wyborach.

Tabela 7. Liczba pozytywnych i negatywnych wpisów (sentymentów) na blogach

Miesiąc	SLD		PSL		PO		PJN		PiS		RP	
	pozytywny	negatywny	pozytywny	negatywny	pozytywny	negatywny	pozytywny	negatywny	pozytywny	negatywny	pozytywny	negatywny
Marzec	72	74	45	44	157	173	41	23	129	230	19	16
Kwiecień	121	61	71	20	202	143	26	23	144	248	14	3
Maj	152	53	37	4	295	171	3	32	152	218	12	5
Czerwiec	152	210	13	30	304	362	78	35	255	365	31	4
Lipiec	58	71	52	17	226	78	2	12	347	277	12	2
Sierpień	90	107	96	33	183	168	5	43	215	238	13	40
Wrzesień	90	57	86	1	219	87	4	12	215	267	194	201
Październik	222	270	87	105	254	345	1	0	214	190	393	169
Łącznie	957	903	487	254	1840	1527	160	180	1671	2033	688	440

Źródło: Obliczenia własne.

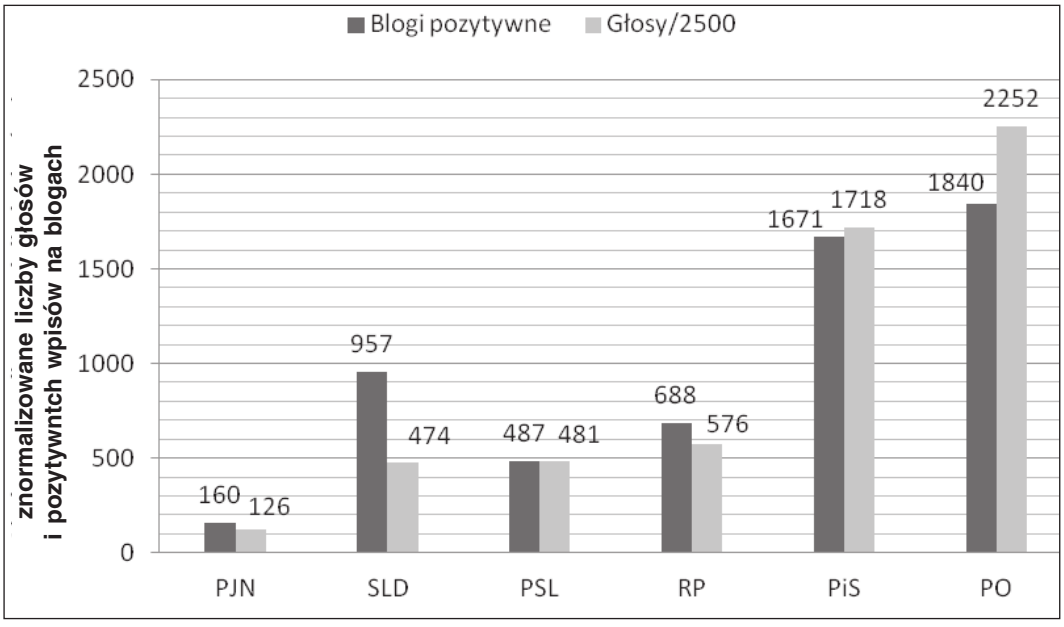
Tabela 8. Korelacja **r** związku wartości wyników sondaży CBOS z liczbą pozytywnych i negatywnych wpisów (sentymentów) na blogach dla wszystkich partii

Miesiące	r dla wpisów pozytywnych	r dla wpisów negatywnych
Marzec	0,95	0,78
Kwiecień	0,97	0,84
Maj	0,98	0,84
Czerwiec	0,95	0,93
Lipiec	0,72	0,42
Sierpień	0,84	0,79
Wrzesień	0,67	0,24
Październik	0,38	0,81

Źródło: Obliczenia własne.

O wspomnianej wartości informacyjnej świadczą przedstawione na wykresie 6 proporcje między pozytywnymi wpisami na blogach a wynikami wyborów. Współczynnik **r** korelacji między tymi zmiennym wyniósł 0,95 ($p < 0,001$). Dowodzi to o niemal pewnej wiarygodności informacji pozyskiwa-

Interesująca jest jednakowa wielkość **r** = 0,99 ($p < 0,001$) zależności liczby pozytywnych wpisów na forach i blogach oraz liczby wpisów negatywnych na forach i blogach. Dowodzi to znacznego podobieństwa miesięcznych rozkładów liczby wpisów pozytywnych i negatywnych stanowiących wynik rafinacji.



Wykres 6. Podobieństwa/różnice proporcji (niebezwzględnej wielkości) pozytywnych wpisów na blogach z liczbą uzyskanych głosów

Źródło: Obliczenia własne. W celu poprawienia efektu wizualizacji danych dokonano normalizacji danych wejściowych: liczbę głosów pomniejszono 2500 razy.

Wiarygodność wyliczonych statystyk

Wartość informacyjną wyliczonych współczynników zależności potwierdzają badania ich statystycznej istotności. Służyły temu badania związku wartości wyników sondaży CBOS z liczbą negatywnych i pozytywnych wpisów (sentymentów) na blogach dla wszystkich partii. Przyjęto, iż wyniki te można uogólnić i stanowią one także o wiarygodności rezultatów uzyskanych z rafinacji wpisów na forach.

Zważywszy na względnie małą liczbę prób/danych (tabela 8), w celu zweryfikowania istotności uzyskanych wniosków analizy statystycznej policzono wartości korelacji dla wszystkich kombinacji miesiące–partie (720 kombinacji). Oczekiwano, iż tak obliczone przypadkowe korelacje będą istotnie odmienne od zasadniczych wyników przedstawionych w tabeli 10. W kolumnie 3. tej tabeli zawarto wartości korelacji liniowej (Pearson), w kolumnie 5 korelacji rang (korelacja Spearmana, czyli korelacja liczona na rangach, tj. pozycjach w rankingu wartości) W odróżnieniu od korelacji liniowej jest to miara odporna na skalowanie – np. nie zmienia się po przejściu z oryginalnych danych np. na ich logarytmy. W kolumnach obok wartości korelacji wpisano prawdopodobieństwa (istotności statystyczne), że zaobserwowana korelacja jest większa lub równa od korelacji między wszystkimi zrandomizowanymi (przypadkowymi) danymi (720 kombinacji), czyli wektorami sondaż i blog, z których jeden ma przestawione (przepermutowane) wartości. Innymi słowy, można z pomijalnym błędem statystycznym mówić o niemal pełnej statystycznej zależności treści pozytywnych i negatywnych wpisów (sentymentów) na blogach z sondażami.

Tabela 10. Wartość korelacji Pearsona i Spearmana związku wartości wyników sondaży CBOS z liczbą negatywnych i pozytywnych wpisów na blogach dla wszystkich partii wraz z wartościami istotności

Wpisy	Miesiące	Pearson		Spearman	
		korelacja	istotność	korelacja	istotność
Pozytywne	marzec	0,948	1,000	1,000	1,000
	kwiecień	0,969	1,000	1,000	1,000
	maj	0,976	1,000	0,971	1,000
	czerwiec	0,954	0,996	0,829	0,983
	lipiec	0,720	0,926	0,943	0,999
	sierpień	0,844	0,981	0,886	0,992
	wrzesień	0,672	0,954	0,829	0,983
	październik	0,382	0,776	0,600	0,912
Negatywne	marzec	0,779	0,947	0,943	0,999
	kwiecień	0,836	0,965	0,886	0,992
	maj	0,843	0,961	0,754	0,961
	czerwiec	0,935	0,996	0,886	0,992
	lipiec	0,42	0,832	0,886	0,992
	sierpień	0,785	0,951	0,714	0,949
	wrzesień	0,237	0,683	0,429	0,822
	październik	0,806	0,986	0,943	0,999

Źródło: Obliczenia zostały wykonane przez dr. hab. Piotra Pokarowskiego z Instytutu Matematyki Stosowanej i Mechaniki Uniwersytetu Warszawskiego.

Prawidłowość ta obejmuje pozostałe wyniki mówiące o wartości korelacji wpisów z innymi zmiennymi (sondaż, liczba głosów, nakłady finansowe). Może być ona traktowana jako wskazówka celowości dalszej analizy tego typu związków w celu uzasadnionej statystycznie predykcji, np. wyborów parlamentarnych.

Nakłady na kampanię wyborczą

Analiza ilościowa kosztów poniesionych przez poszczególne partie w okresie wyborczym wskazuje na całkowity brak związku (**r** = 0,06) między kosztami utworzenia i utrzymania strony internetowej komitetu a liczbą uzyskanych głosów. Dowodzi to nadzwyczaj małej (wy-

Tabela 11. Korelacja **r** kosztów kampanii z liczbą uzyskanych głosów

Koszty kampanii	Współczynnik korelacji Pearsona
Koszty utworzenia i utrzymania strony internetowej komitetu	0,06
Wykonanie materiałów wyborczych, w tym prace koncepcyjne, prace projektowe i wytworzenie	0,61
Suma wszystkich nakładów	0,72
Korzystanie ze środków masowego przekazu i nośników plakatów	0,75
Udział wydatków na internet w całości wydatków na media	0,76
Reklama w internecie (koszt usługi emisji)	0,86
Nakłady na internet łącznie	0,89
Reklama w internecie	0,98

Źródło: Obliczenia własne.

mownej) skuteczności deklarowanych przez komitety nakładów na strony internetowe. Znacznie większe zależności liczby głosów stwierdzono z innymi nakładami.

Dane zawarte w tabeli 8 wskazują na nadzwyczajną skuteczność nakładów na reklamę w internecie, znacznie większą od nakładów na media, środki masowego przekazu i nośniki plakatów.

Zakończenie

Rafinacja zasobów Big Data umożliwia ilościową analizę – rafinację – szerokiego spektrum pierwotnej, nieustrukturyzowanej oryginalnej informacji (w omówionych wynikach badań – ponad 2 mln wpisów). Dowiedziono, iż rafinacja kreuje nową przestrzeń wartościowych źródeł informacji i otwiera nowe drogi do badań nad ich poszukiwaniem.

Badania tego typu wymagają zmian metodologii i narzędzi oraz znaczących mocy obliczeniowych wykorzystanych do eksploracji. Chodzi tu o kolekcjonowanie tematycznych informacji, ich precyzyjne filtrowanie (w omawianym przypadku odrzucono około 90 proc. wpisów) i zasadniczą analizę informacji pozyskiwanych z sieci – poszukiwanie informacji o sentymentach i ich obróbka statystyczna.

Uzyskane wyniki badań dowodzą podobieństwa, niemal identyczności uzyskiwanych dzięki rafinacji danych o poparciu dla poszczególnych partii politycznych uczestniczących w wyborach parlamentarnych 2011 r. z wynikami sondaży opinii publicznej oraz oficjalnymi wynikami ogłoszonymi przez PKW.

Dogłębna analiza rozkładu sentymentów stwarza nawet szanse na przewidywanie przyszłych zmian szacowanych wartości danych (por. wykres 4). Zmierzanie ku temu celowi – stworzenie funkcji predykcji – wymaga zaangażowania narzędzi statystycznych uwzględniających jednocześnie wiele parametrów, np. w postaci funkcji regresji wielokrotnej. Owe parametry to m.in. wiarygodnie wskazane w artykule wartości sentymentów pozyskiwanych z wpisów na forach i na blogach.

Nie mniej ważne od wartości merytorycznej uzyskanych wyników jest wielokrotnie mniejszy koszt uzyskanych, dzięki rafinacji, informacji, wobec kosztów zdobywania tych samych informacji w tradycyjny sposób – drogą sondaży.

Wyniki przedstawionych badań dowodzą, iż Big Data przestają być *terra incognita* dla nauk społecznych. Ważnym wyzwaniem jest doskonalenie metodologii rafinacji oraz opracowanie stosownych narzędzi do rafinacji informacji sieciowej, przyjaznej dla użytkowników formy dostarczania wyników, a co najważniejsze – przekonania o użyteczności tego nowego źródła informacji.