PRZEGLĄD ARCHIWALNO-HISTORYCZNY
Tom V Poznań 2018

Adrian Trzoss
(Uniwersytet im. Adama Mickiewicza w Poznaniu, Wydział Historyczny)
mgr, adrian.trzoss@amu.edu.pl

Przyczynek do badań nad metodami historii cyfrowej w świetle debaty przed EU Referendum na profilach portalu Facebook Davida Camerona i Nigela Farage'a

Dlaczego media społecznościowe?

Używanie mediów społecznościowych przez polityków wydaje się stanowić naturalną kolej rzeczy w aplikowaniu nowych technologii na rzecz docierania do wyborców i przekonywania ich do swoich racji. Sięganie po nowe rozwiązania komunikacyjne pozwalało uzyskiwać dostęp do coraz szerszych grup odbiorców czy dotrzeć do wyborców w ogóle, np. poza głównymi kanałami przekazu[2]. W XXI w. byliśmy świadkami analogicznego przełomu, kiedy w 2008 r. Barack Obama dzięki wykorzystaniu możliwości, jakie dostarcza Web 2.0, wygrał wybory i został 44. prezydentem Stanów Zjednoczonych[3]. Jak podkreślają analitycy[4], Obama w trakcie kampanii zaangażował znacz-

---

[1] Niniejszy referat jest zmodyfikowanym i rozwiniętym tekstem wystąpienia wygłoszonym podczas seminarium Instytutu Historii Nauki PAN oraz Uniwersytetu im. Adama Mickiewicza pt. *Międzykulturowy wymiar komunikacji idei w dziejach nauki* 12–13 maja 2018 r. Miałem wówczas przyjemność uczestniczyć w szerokiej dyskusji na temat historii cyfrowej i jej metod oraz ich znaczenia dla metodologii historii. W jej wyniku narodziło się wiele pytań i problemów obecnych w aktualnie dziejącym się dyskursie. Po uzupełnieniu moich wcześniejszych przemyśleń podejmuję próbę rozwiązania przynajmniej części z nich.

[2] Chociażby jak w ostatnich latach miało to miejsce w przypadku kampanii wyborczej Donalda Trumpa, który nie będąc mile widzianym w mediach głównego nurtu, komunikował się ze swoimi wyborcami poprzez portal Twitter.

[3] Oczywiście, nie mam tu na myśli absolutyzowania roli mediów społecznościowych, jednakże – jak wynika z przeprowadzonych badań – był to jeden z ważniejszych czynników determinujących tamte wybory. Por. przypis 4.

[4] *Communicator-in-Chief. How Barack Obama Used New Media Technology to Win the White House.,* pod red. J.A. Hendricks, R.E. Denton Jr., Plymouth United Kingdom 2010.

nie więcej wyborców na portalach społecznościowych niż jego oponent. Był to bodaj pierwszy taki przypadek we współczesnej historii, kiedy medium, jakim jest Internet, oraz fenomen kulturowy Web 2.0 odegrały tak znaczącą rolę w światowej polityce[5]. Od czasów kampanii Obamy zaobserwować można stały wzrost zaangażowania polityków w aktywność w mediach społecznościowych, którzy powielają jego wzorce, a także wprowadzają nowe sposoby korzystania z rozwijających się mediów cyfrowych[6].

Badania nad losami kampanii przedreferendalnej w sprawie członkostwa Wielkiej Brytanii w Unii Europejskiej toczyły się równolegle do jej przebiegu, a podczas nich zaobserwowano wiele ciekawych kwestii związanych chociażby z nastawieniem wyborców do poszczególnych tematów referendalnych czy też, już z późniejszej perspektywy, z pytaniem: jak konkretne grupy społeczne oddawały swój głos[7]. Skoro kwestia kampanii przedreferendalnej dobiegła końca, wydawać się może, iż temat ten mógłby przyciągnąć również zainteresowanie historyków. Warto byłoby przyjrzeć się, w kontekście wyników analiz socjologicznych i politologicznych, samej debacie, która towarzyszyła kampanii m.in. w swoim cyfrowym wymiarze w mediach społecznościowych. Dysponując analizami wyników głosowania, można zadać pytania o korelację pomiędzy propagandą polityczną a wynikami wyborów czy też trendami zainteresowań w debatach w poszczególnych regionach. Jest to oczywiście kwestia wyjątkowo złożona, niemniej jednak chciałbym podjąć się próby przedstawienia losów kampanii z perspektywy narracji medialnej dwóch jej głównych aktorów: premiera Wielkiej Brytanii Davida Camerona i europosła, lidera partii UKIP Nigela Farage'a. Obaj politycy aktywnie uczestniczyli w procesie przedreferendalnym, udzielając się w mediach i podczas wieców wyborczych. Tu chciałbym przyjrzeć się jednej z płaszczyzn ich zaangażowania, w której narracje te były obecne, a mianowicie mediom społecznościowym, konkretnie portalowi Facebook.

Wybór tego medium wiąże się szczególnie mocno z wyjściem na światło dzienne kontrowersji związanych z działalnością Cambridge Analytica oraz późniejszymi przesłuchaniami Marka Zuckerberga, a także śledztwa-

---

[5] W ostatnich latach takich przypadków było oczywiście więcej, np. podczas przemian politycznych na Ukrainie czy w trakcie tzw. arabskiej wiosny, por. T. Bohdanova, *Unexpected Revolution: The Role of Social Media in Ukraine's Euromaidan Uprising*, „European View" 2014, nr 13(1), s. 133–142; P.N. Howard., D. Aiden, F. Deen, H. Muzammil, M. Will, M. Marwa, *Opening Closed Regimes: What Was the Role of Social Media During the Arab Spring? Project on Information Technology and Political Islam Data Memo 2011.1*, Seattle: University of Washington, 2011.

[6] M. Broersma, T. Graham, *Social media as beat. Tweets as a news source during the 2010 British and Dutch Elections*, „Journalism Practice" 2012, vol. 6, s. 403–419.

[7] W tej kwestii całą serię analiz przeprowadził National Centre for Social Research w Wielkiej Brytanii (w skrócie NatCen). Wyniki prowadzonych przez nich badań są dostępne online: https://whatukthinks.org/eu/comment-analysis/ (dostęp 28 maja 2018 r.).

mi w sprawie możliwej ingerencji Rosjan w wewnętrzne kwestie polityczne USA oraz Wielkiej Brytanii[8]. W świetle tych wydarzeń badacze coraz bardziej zdają sobie sprawę ze znaczenia roli mediów społecznościowych w świecie, ich realnego wpływu na życie poza przestrzenią cyfrową[9]. Od dłuższego już czasu media społecznościowe wykroczyły poza swoje „cyfrowe" ramy czy to w przypadku protestów w trakcie tzw. wiosny arabskiej, kiedy stały się narzędziem do samoorganizowania się społeczeństwa przeciw władzy, czy to w kwestiach niezwiązanych z polityką, np. codziennego życia, jak marketing powszechnie używanych produktów (tu dobrze to widać w przypadku Coca-Coli, która zatrudnia influencerów[10] do promocji swoich produktów). Ponadto Facebook jest największym ogólnoświatowym medium społecznościowym gromadzącym ponad dwa miliardy aktywnych użytkowników[11], co tylko przyciąga zainteresowanie odbiorców i potencjalnych aktywistów oraz polityków.

Problemy badawcze. Historia cyfrowa i automatyzacja metod badawczych

Celem niniejszego artykułu jest nie tyle odpowiedź na to, jak wyglądała retoryka przedreferendalna Davida Camerona i Nigela Farage'a na portalu Facebook, ile raczej próba ewaluacji metod, dzięki którym taka analiza jest możliwa. Całość moich rozważań wpisuje się w nurt historii cyfrowej i badania nad źródłami *born digital*, które na gruncie polskiej historiografii wciąż są na etapie wczesnego rozwoju[12]. W moich rozważaniach chciałbym odnieść się do kilku kwestii, które pojawiły się podczas dyskusji w środowisku metodologów historii, stojących obecnie przed wyzwaniem, jakim jest refleksja nad historią cyfrową, jej obszarem badań i idącymi za tym metodami. Pierwsze

---

[8] Artykuł Guardiana nt. powiązań między politykami a organizacjami lobbującymi w trakcie EU Referendum: https://www.theguardian.com/technology/2017/may/07/the-great-british-brexit-robbery-hijacked-democracy (dostęp 28 maja 2018 r.).

[9] O ile o takim rozróżnieniu możemy jeszcze mówić. Coraz częściej podnoszone są głosy, iż obie przestrzenie ludzkiej aktywności, tj. analogowa i cyfrowa, nie tyle wzajemnie się przenikają, co stanowią swoistą jedność, por. M. Kosinski, Y. Wang, H. Lakkaraju, J. Leskovec, *Mining big data to extract patterns and predict real-life outcomes*, „Psychological Methods" 2016, nr 21(4), s. 493–506.

[10] Tzw. internetowych celebrytów, postaci, które skupiają wokół siebie liczne grono odbiorców, nadając trendy i stając się swego rodzaju *quasi*-autorytetem dla społeczności ich odbiorców.

[11] Statystyki aktywnych użytkowników Facebooka: https://www.statista.com/statistics/264810/number-of-monthly-active-facebook-users-worldwide/ (dostęp 28 maja 2018 r.).

[12] Dobrze to obrazuje przypadek historii dynamicznego rozwoju Laboratorium Cyfrowego Humanistyki, por. http://lach.edu.pl/o-laboratorium/ (dostęp 25 maja 2018 r.).

pytanie, jakie się pojawia, to czy w ogóle potrzebujemy terminu historia cyfrowa i czy wprowadzenie nowej nomenklatury z tym związanej nie jest namnażaniem bytów ponad potrzebę. Po wtóre – jeśli hipotetycznie, na potrzeby symulacji, założymy, że takowa historia jest nam potrzebna, pojawia się pytanie o to, czy stare metody badawcze są adekwatne do jej przedmiotu czy też potrzebne są nowe metody i czy całkiem je zastępują. A może jest to związek niejako hybrydowy, gdzie oba podejścia klasycznej, analogowej hermeneutyki uzupełniają się z jej cyfrową wersją? Po trzecie – pojawiło się pytanie o poziom skuteczności stosowanych zautomatyzowanych metod analizy danych źródłowych czy też źródeł cyfrowych *born digital* w ogóle. Czy metody oparte na technice programowania i szeroko pojętego *data mining*[13] (a raczej *social data science* w przypadku niniejszego referatu) nie dają nam wyników niepełnych lub w ogóle odmiennych od tych, które byśmy uzyskali tradycyjną analizą. Na koniec chciałbym wskazać na kwestie praktyczne związane ze stosowaniem metod historii cyfrowej i, jak sądzę, drzemiący w nich potencjał, który mógłby być użyteczny także w przypadku źródeł analogowych, tj. niecyfrowych *born digital*[14]. W tym celu chciałbym w kolejnych akapitach dokonać niejako rekonstrukcji procesu badawczego będącego fragmentem moich szerszych dociekań naukowych odnoszących się do analizy narracji w ramach debaty przedreferendalnej w sprawie członkostwa Wielkiej Brytanii w Unii Europejskiej.

Pytania, jakie postawiłem zebranemu zasobowi źródeł, dotyczyły kilku podstawowych kwestii, które podzieliłem na analizy ilościowe oraz jakościowe. Do analizy taksonomicznej zaliczyłem: kwestie związane ze statystykami publikowanych wpisów, ich grupowanie i korelacje tematyczne, trendy, jakie się ujawniały, oraz recepcję. W przypadku kwestii hermeneutycznych skupiam się na kwestiach związanych z analizą filologiczno-ilościową (frekwencje słów, związki korelacji między słowami) oraz znaczeniową (interpretacja poszczególnych kodów słów, metafor). W niniejszym badaniu chciałbym dokonać weryfikacji hipotezy o dominacji tematyki referendalnej w debacie politycznej w ostatnich tygodniach przed referendum. Następnie wyszczególnić, które tematy były eksponowane w narracji, po czym zestawić to z badaniami dotyczącymi preferencji wyborców oraz motywów ich wyborów podczas głosowania[15]. Oczywiście, zasób źródłowy nie jest tak duży, aby dokonać absolutyzacji wniosków i zaproponować uniwersalną teorię, jednakże zważywszy na rolę omawianych postaci w trakcie referendum, zdaje się, iż uzyskane

---

[13] R. Zafarani, M. Ali Abbasi, H. Liu, *Social Media Mining. An Introduction*, Cambridge 2014.

[14] M. Eder, *In search of the Author of Chronica Polonorum Ascribed to Gallus Anonymus: A Stylometric Reconnaissance*, „Acta Poloniae Historica" 2015, nr 112, s. 5–23.

[15] Por. przypis 7.

wyniki mogą przynajmniej wskazać kierunek dalszych dociekań. W dalszej kolejności chciałbym pochylić się nad kwestiami interpretacyjnymi, a mianowicie kwestią konstruowania narracji na profilach obu polityków, co następnie zestawiam z szerszymi danymi reprezentującymi nastawienie społeczeństwa do omawianych przez polityków problemów.

Efektem końcowym jest przedstawienie ogólnie zarysowanego poziomu skuteczności stosowanych metod i odpowiedź na powyższe zagadnienia dotyczące kampanii przedreferendalnej.

Etap pierwszy. Pozyskiwanie i selekcja danych źródłowych

Bazę źródłową dla niniejszej analizy stanowią wpisy pochodzące z oficjalnych i publicznych kont z portalu Facebook Davida Camerona oraz Nigela Farage'a. Pierwszy z nich pełnił w omawianym okresie urząd premiera Wielkiej Brytanii, drugi zaś był głównym zwolennikiem opuszczenia Unii Europejskiej przez Wielką Brytanię, a po głosowaniu okrzyknięty został przez media „człowiekiem, który wyprowadził Brytanię z Unii"[16].

Ramy czasowe, jakie proponuję, to 15 kwietnia 2016 r., kiedy Electoral Commision ogłosiła oficjalne rozpoczęcie kampanii przedreferendalnej, oraz 23 czerwca tego samego roku, czyli data referendum. Data końcowa wydaje się logicznie zasadna jako punkt kończący proces kampanii przedreferendalnej, zatem chciałbym spróbować uzasadnić pierwszą część cezury czasowej. Liczne organizacje (w tym dwie główne wyznaczone przez komisję[17]) lobbowały co prawda jeszcze przed 15 kwietnia, a sam proces legislacyjny ma swoje korzenie w przemowie królowej podczas State Opening of Parliment (27 maja 2015 r.[18]); nie sposób również pominąć kwestii dyskusji wokół samego wniosku o przeprowadzenie referendum czy kwestii roli Zjednoczonego Królestwa w Unii w ogóle. Niemniej jednak, biorąc pod uwagę konsekwencje[19], jakie idą za oficjalnym administracyjnym rozpoczęciem okresu kampa-

---

[16] Artykuł Telegraph na temat znaczenia Nigela Farage'a dla wyników EU Referendum: https://www.telegraph.co.uk/news/2016/06/24/nigel-farage-has-earned-his-place-in-history-as-the-man-who-led/ (dostęp 28 maja 2018 r.).

[17] Informacja prasowa dotycząca ogłoszenia stron kampanii przedreferendalnej: https://www.bbc.com/news/uk-politics-36038672 (dostęp 28 maja 2018 r.).

[18] Zapis transmisji State Opening of Parliment w 2015 roku na stronach Parlamentu Brytyjskiego: https://www.parliament.uk/business/news/2015/may/state-opening-of-parliament-2015/ (dostęp 28 maja 2018 r.).

[19] Chociażby jest to kwestia przyznanych funduszy państwowych na prowadzenie kampanii czy czas antenowy: https://www.electoralcommission.org.uk/find-information-by-subject/elections-and-referendums/past-elections-and-referendums/eu-referendum/grants-to-designated-lead-campaigners (dostęp 28 maja 2018 r.).

Wykres 1. Zainteresowanie EU Referendum na Wyspach Brytyjskich w wyszukiwarce Google

nii, oraz fakt, iż stanowi ona element procesu referendalnego jako swoisty kamień milowy[20], uważam tę datę za zasadną.

W wybranym przeze mnie okresie miał miejsce również jeszcze jeden, wewnętrzny proces, a mianowicie Purdah[22], czyli okres, w czasie którego rząd oraz samorząd nie mogą wnosić nowych inicjatyw legislacyjnych, mogących wpłynąć na wyniki głosowania. Czuję się zobligowany do zaznaczenia tego punktu, który następnie zostanie uwzględniony w części interpretacyjnej. Innym wydarzeniem politycznym, które może mieć znaczenie dla późniejszych wyników, były wybory lokalne w Wielkiej Brytanii 5 maja 2016 r.[23] W związku z tym politycy podróżowali po kraju, agitując, łącząc nieraz sprawy lokalne z referendalnymi (np. w kwestii rybołówstwa). Przez podróże, udział w wiecach oraz wystąpieniach medialnych, różnić się będzie także kwestia natężenia publikacji wpisów na Facebooku. Może to wprowadzić garść dodatkowych problemów natury interpretacyjnej, jak choćby decydowanie, czy dany post jest bardziej związany z debatą referendalną, czy z wyborami lokalnymi, jednakże o tym szerzej w części wniosków metodologicznych.

---

[20] To jest czas, kiedy przeprowadzane są oficjalne debaty, a za pomocą środków publicznych prowadzi się w pełni intensywną kampanię, co przekłada się na zainteresowanie tematyką referendalną wśród opinii publicznej; por. wykres 1.

[21] Na wykresie obserwujemy stopniowy wzrost zainteresowania referendum. Pierwszy mniejszy pik pokrywa się z rozpoczęciem kampanii. Wykres obrazujący miesięczne wartości względne zainteresowania (tj. mierzone w skali od 1 do 100, gdzie 1 oznacza najniższe w danym okresie i 100 analogicznie najwyższe) w wyszukiwarce Google hasła EU Referendum na Wyspach Brytyjskich w czasie od kwietnia 2015 do czerwca 2016 r. Źródło google.trends.

[22] Prawna definicja Purdah na stronach Parlamentu Brytyjskiego: https://www.parliament.uk/site-information/glossary/purdah/ (dostęp 28 maja 2018 r.). W przypadku EU Referendum Purdah trwał od 27 maja 2016 r.

[23] E. Uberoi, C. Watson, R. Keen, *Local Elections 2016.Briefing Paper,* no. CBP 7596, House of Commons Library, 2016.

Źródła zostały pozyskane[24] za pomocą API[25] portalu Facebook. Zdefiniowano następujące parametry opisowe dla obiektu, jakim jest post (tj. wpis na portalu wraz z jego częścią tekstową) z podziałem na parametry ilościowe (liczba reakcji, komentarzy oraz udostępnień), które są swego rodzaju miernikiem jego recepcji na portalu, oraz te, które można potraktować jako metadane zewnętrzne, czyli: data publikacji wpisu, adres w postaci linku, jego numer ID w portalu, typ (status tekstowy, zdjęcie, film, link). Następnie dane zostały uporządkowane i pobrane w formacie json, a później zobrazowane w postaci tabeli w formacie csv. Liczba postów w omawianym czasie dla Davida Camerona wynosi 123, a dla Nigela Farage'a 148.

Ze względu na proces selekcji danych można uznać, iż jest już ona w pewien sposób interpretacją, czyli takim modelowaniem rzeczywistości źródłowej, aby była ona możliwa w dalszej analizie – z obrazowej formy widocznej dla użytkownika (przy pomocy interfejsu graficznego portalu w WWW) do formy statystycznej przydatnej dla analityków. Pierwszy problem natury metodologicznej, jaki zaobserwowałem podczas tego etapu, to kwestia różnicy w datacji. Do parametru opisującego wpis, którym jest data publikacji, przynależy data dzienna oraz godzina. Pobrane dane opatrzone są datą zgodą z uniwersalnym czasem skoordynowanym (UTC) +0000 (tj. strefa czasowa związana z południkiem 0^{o}). Ponadto domyślnie Facebook używa dla niezalogowanych użytkowników czasu UTC-8, czyli PST[26] (co zgadza się z położeniem siedziby Facebooka). W przypadku czasu UTC (+0000) jest on stałą, na podstawie której obliczana jest godzina użytkownika (zalogowanego) w miejscu jego przebywania (tj. odpowiednio dla Polski byłoby to UTC+1) zgodnie z przyjętymi normami ISO[27]. Taki model powoduje różnice w postrzeganiu daty dla niewprawionych badaczy – inna godzina bowiem jest wyświetlana użytkownikowi, a inna widnieje w danych pobranych z serwerów Facebooka[28]. Godzina wyświetlana użytkownikowi jest automatycznie dostosowywana dla jego strefy czasowej (stąd mogą wynikać przesunięcia w dniach publikacji o jeden dzień do przodu lub do tyłu). Wskazane jest zatem nanoszenie korekty czasowej lub poprzestanie na zaznaczeniu, iż anali-

---

[24] Więcej na temat pozyskiwania źródeł z portali społecznościowych por. A. Trzoss, W. Werner, *Problemy i wyzwania związane z badaniem i archiwizacją aktywności instytucji publicznych w mediach społecznościowych*, [w:] *Toruńskie Konfrontacje Archiwalne. Pogranicza archiwistyki*, t. VI, pod red. A. Rosa, Toruń 2019 [w druku].

[25] Dokumentacja API portalu Facebook: https://developers.facebook.com/docs/graph-api (dostęp 28 maja 2018 r. v.3.0).

[26] Pacific Standard Time.

[27] Dokumentacja ISO nt. systematyzacji formatu czasów: https://www.iso.org/iso-8601-date-and-time-format.html (dostęp 28 maja 2018 r.).

[28] W przypadku pobierania z reguły jest to UTC (+0000). Dla potrzeb badawczych w algorytmach umieszcza się czasem dodatkową linijkę kodu z formatowaniem daty i godziny, według potrzeb lokalnych badaczy.

zowane dane są zgodne z czasem UTC (+0000). Wpisuje się to w szerszą gamę problemów metodologicznych, o których piszę wraz z Wiktorem Wernerem w jednej z prac[29], a mianowicie, to, co badamy i jak archiwizujemy, nie zawsze jest tym, co realnie funkcjonuje. Ma to istotne znaczenie dla sposobu zrozumienia zachodzących zjawisk oraz unikania potencjalnych błędów podczas analizy – utratę danych z interesującego dla badacza przedziału czasowego lub rozszerzenia go o dane zbędne.

Etap drugi.
Od selekcji do analizy komputerowej – zautomatyzowanej

Jak już wspomniałem selekcja danych jest już w pewnym stopniu etapem ich interpretacji. O ile samo wyodrębnienie interesującego nas zbioru źródeł jest pewnym etapem wstępnym, o tyle wyznaczenie podgrup tematycznych jest już etapem właściwym. W tym miejscu chciałbym dokonać podziału zasadniczego na posty związane z tematyką przedreferendalną oraz te niezwiązane z nią. W tym celu opracowałem następujący algorytm postępowania:

1. Przygotowanie korpusu tekstowego.
2. Oczyszczenie korpusu ze słów przystankowych, znaków interpunkcyjnych oraz innych „zanieczyszczeń" (np. znaków formatujących tekst).
3. Ustalenie częstotliwości występowania poszczególnych słów.
4. Wyznaczenie kontekstu występowania słów za pomocą kolokacji
5. (n-gramy[30]) oraz częstotliwości danych kolokacji.
6. Ustalenie kluczy – kodów, dzięki którym będzie dokonany podział postów na związane z kampanią przedreferendalną. Następnie weryfikacja ich kolokacji ze wszystkimi słowami w zestawieniu z trigramami w celu uzupełnienia ewentualnie brakujących kodów. W tym punkcie następuje klarowanie tematów związanych z referendum.
7. Selekcja za pomocą kluczy postów związanych z referendum i tych niezwiązanych. W tym miejscu zalecana jest, jeśli to możliwe, weryfikacja hermeneutyką tradycyjną (niezautomatyzowaną) otrzymanych danych. Następnie istnieje możliwość badania otrzymanych danych w celu kolejnych analiz, jak np. recepcja poszczególnych tematów, ich wzajemne korelacje, analiza sentymentów itd.

---

[29] A. Trzoss, W. Werner, *Problemy i wyzwania*.

[30] S. Bird, E. Klein, E. Loper, *Natural Language Processing with Python,* O'Reilly Media 2009, s. 202–208.

Powyższy algorytm przeniesiony został do środowiska analitycznego Jupyter[31] dzięki dystrybucji Anaconda[32] języka programowania Python[33] i jego bibliotek. Jednym z głównych rozszerzeń Pythona, które odgrywa tu zasadniczą rolę, jest moduł NLTK[34] oraz Pandas[35].

Z korpusu źródłowego eksportowana zostaje część tekstowa ze statusami wraz z numerem kolejnym statusu. Następnie tekst zostaje oczyszczony ze słów przystankowych (część z nich jest gotowa w pakiecie NLTK – są to części struktur gramatycznych; dodatkowo została uzupełniona o własne sformułowania, które zostały wprowadzone na etapie ewaluacji skuteczności oczyszczania tekstu) oraz znaków interpunkcyjnych. W dalszej kolejności taki korpus został wyeksportowany do programu analitycznego TextSTAT[36], w którym dokonano analizy listy frekwencyjnej słów. W przypadku Davida Camerona liczba słów z więcej niż jedną częstotliwością wynosi 624; analogicznie dla Nigela Farage'a 307. Dla słów z częstotliwością powyżej trzech: David Cameron – 359, Nigel Farage – 141. Zaobserwowałem, iż znaczna liczba słów pojawia się incydentalnie, co nie tylko praktycznie nie ma znaczenia w kontekście niniejszego badania (części wniosków empirycznych), ale także powoduje zaburzenia w metodyce, znacznie zaniżając i zniekształcając wszelkie obliczenia statystyczne (wnioski metodyczne). W obu przypadkach najbardziej popularnym słowem było „EU" – dla Camerona 68 powtórzeń, Farage'a 103. Pojawia się zatem pierwsze pytanie związane z tą częścią metody, a mianowicie, od jakiej liczby powtórzeń możemy uznać, iż dane są reprezentatywne, tj. są stałym, a nie losowym elementem narracji. Zaznaczam, iż są to surowe dane, gdyż nie dysponujemy na tym etapie kontekstem danych słów. Możemy zatem metodą próbkowania podjąć się próby znormalizowania układu poprzez odrzucenie wartości nieistotnych (jak jedno czy dwa powtórzenia) i wyznaczeniu zmiennych, które są znacznie powyżej odchylenia od średniej, taką liczbą byłoby dziesięć powtórzeń. Jest to to tyle zasadne, iż zależy nam na ujęciu najpopularniejszych słów jako wyznaczników tematycznych, a zatem incydentalne pojawienie się słowa nie zmienia naszych wniosków. Liczba dziesięciu powtórzeń jest zatem tą, która stanowi w tym przypadku barierę pomiędzy przypadkowością a elementem stałym narracji (zwłaszcza jeśli później dokona się grupowania słów w tematy i weryfikacji

---

[31] Dokumentacja programu: http://jupyter.org/documentation (dostęp 28 maja 2018 r.).

[32] Dokumentacja dystrybucji: https://docs.anaconda.com/ (dostęp 28 maja 2018 r.).

[33] Dokumentacja języka programowania Python: https://docs.python.org/3/index.html (dostęp 28 maja 2018 r.).

[34] Dokumentacja modułu: https://www.nltk.org/ (dostęp 28 maja 2018 r.).

[35] Dokumentacja rozszerzenia: http://pandas.pydata.org/pandas-docs/stable/ (dostęp 28 maja 2018 r.).

[36] Opis programu i jego dokumentacja: http://neon.niederlandistik.fu-berlin.de/en/textstat/ (dostęp 28 maja 2018 r.).

ich z n-gramami, można zaobserwować, iż poniżej dziesięciu powtórzeń częstotliwości słów szybko maleją, a ich związki tematyczne nie przynoszą żadnych wskazówek interpretacyjnych). Problem metodyczny, który tu chciałem podkreślić, to problem obliczeń statystycznych. Przy tak specyficznym bycie, jakim jest język naturalny, traktowanie go jako zmiennych w badaniach statystycznych może być nieco złudne. Bezgraniczne zaufanie do liczb, bez znajomości kontekstu przedmiotu analizy, powoduje, iż otrzymane dane nie tylko będą problematyczne w obliczeniach, ale ich wyniki mogą wydawać się mylące. Z drugiej strony, arbitralne narzucenie jakiejś sumy może spowodować pominięcie kilku słów, które by zmieniły nasze wyniki i interpretację. Stąd wydaje się, iż należy zachować swoisty balans między potencjalnymi możliwościami obliczeń, a ich subiektywną korektą związaną z celem badań i znajomością problematyki przedmiotu badań.

Tabela 1. Liczba powtórzeń najpopularniejszych słów dla profilu Nigela Farage'a[37]

eu	103	union	18
leave	58	want	17
i	56	brexit	14
country	46	must	14
vote	44	day	13
back	41	today	13
june	36	like	12
23rd	35	tomorrow	12
ukip	27	inside	11
borders	25	open	11
get	24	outside	11
let	22	people	11
we	21	thursday	11
control	20	believe	10
european	19	britain	10
uk	19	immigration	10
make	18	take	10
referendum	18		

[37] Opracowanie własne.

Tabela 2. Liczba powtórzeń najpopularniejszych słów dla profilu Davida Camerona[38]

eu	68	make	19
britain	67	campaign	17
europe	65	say	15
vote	58	future	14
stronger	47	every	13
remain	44	many	13
i	40	prices	13
people	37	want	13
country	35	working	13
better	31	important	12
leave	30	years	12
economy	29	economic	11
clear	28	recession	11
today	28	risk	11
jo	27	british	10
uk	26	family	10
we	26	june	10
one	25	keep	10
leaving	24	left	10
jobs	23	my	10
safer	21	public	10
referendum	20	put	10
us	20	together	10
world	20	treasury	10

Następnie należy pogrupować słowa we wstępne kategorie tematyczne. Wyszczególnione słowa są jednak pozbawione kontekstu, stąd określenie np., iż dane słowo (załóżmy pojawiające się „risk" czy „country") należy do kategorii wspólnej z „EU", byłoby obarczone ryzykiem błędu. Rzecz ta rzutuje na późniejszy podział postów na te związane z Brexitem oraz pozostałe. Stąd kolejnym krokiem było wyznaczenie kontekstów słów za pomocą metody n-gramów[39] (bi- i tri- gramów w tym przypadku[40]). Metoda ta zbiera słowa występujące w ciągu słów w danym zdaniu oraz wpisie, a następnie analizuje liczbę ich powtórzeń dla całego korpusu. Dzięki temu otrzymujemy konteksty słów i tak np. najczęściej pojawiający się bigram u Davida Camerona to „stronger", „safer" (siedemnaście powtórzeń w siedemnastu różnych postach), trigram zaś to „stronger", „safer", „better" (szesnaście powtórzeń w szesnastu różnych postach).

---

[38] Opracowanie własne.

[39] Dla n podstawia się dowolnie przybraną liczbę całkowitą.

[40] W przypadku n=4 rozproszenie i losowość danych znacząco wzrasta, stąd wybór dla n=2 oraz 3.

Tabela 3. Najpopularniejsze bigramy dla profilu Davida Camerona[41]

'stronger'	'safer'	17	17
'better'	'safer'	16	16
'stronger'	'Europe'	16	16
'Remain'	'vote'	14	13
'stronger'	'Britain'	13	13
"we're"	'stronger'	12	12
'better'	'Europe'	12	12
'EU'	'referendum'	9	9
'EU'	'leaving'	9	8
'working'	'people'	8	2
'remain'	'vote'	8	8
'European'	'Union'	8	8
'leave'	'EU'	8	7
'Europe'	'leaving'	7	7
'EU'	'remain'	7	5
'Jo'	'Cox'	6	5
'Stronger'	'Europe'	6	6
'Stronger'	'Britain'	6	6
'Bank'	'England'	6	5
'analysis'	'Treasury'	5	5
'Watch'	'Europe'	5	5
'every'	'household'	5	4
'Leave'	'campaign'	5	4
'keep'	'Britain'	5	5

Tabela 4. Najpopularniejsze trigramy dla profilu Davida Camerona[42]

'safer'	'stronger'	'better'	16	16
'Europe'	'safer'	'better'	9	9
'safer'	'Britain'	'stronger'	7	7
"we're"	'Europe'	'stronger'	6	6
'Europe'	'Britain'	'Stronger'	6	6
'safer'	"we're"	'stronger'	5	5
'Britain'	'keep'	'stronger'	4	4
'permanently'	'poorer'	'country'	3	3
'Treasury'	'analysis'	'shows'	3	3
'household'	'£4300'	'every'	3	3

[41] Opracowanie własne.
[42] Opracowanie własne.

'tribute'	'Cox'	'Jo'	3	3
'Britain'	'Europe'	'stronger'	3	3
'Watch'	'Europe'	'better'	3	3
'Polls'	'10pm'	'close'	3	3
'Union'	'reformed'	'European'	3	3
'leave'	'EU'	'vote'	3	2
'Europe'	'leaving'	'clear'	3	3

Tabela 5. Najpopularniejsze bigramy dla profilu Nigela Farage'a[43]

'EU'	'Leave'	36	36
'23rd'	'June'	35	35
'back'	'country'	23	23
'European'	'Union'	17	17
'country'	'get'	15	15
'control'	'borders'	13	13
'vote'	'Leave'	12	11
'Leave'	'Vote'	11	11
'back'	'control'	10	10
'back'	'take'	9	9
'must'	'Leave'	9	9
'want'	'country'	9	9
'Independence'	'Day'	9	9
'23rd'	'Independence'	9	9
'EU'	'outside'	8	8
'referendum'	'EU'	8	8
'make'	'June'	8	8
'EU'	'inside'	7	6
'open'	'borders'	6	6
'EU'	'open'	6	6
'European'	'Leave'	6	6
'along'	'come'	6	6
'make'	"let's"	5	5
'5th'	'May'	5	5
'independent'	'selfgoverning'	5	5
'Bus'	'Brexit'	5	5

[43] Opracowanie własne.

Tabela 6. Najpopularniejsze trigramy dla profilu Nigela Farage'a[44]

'get'	'country'	'back'	14	13
'want'	'country'	'back'	10	9
'Leave'	'vote'	'EU'	9	9
'Leave'	'must'	'EU'	9	9
'Day'	'23rd'	'Independence'	9	9
'June'	'Independence'	'23rd'	9	9
'control'	'back'	'take'	9	9
'make'	'June'	'23rd'	8	8
'back'	'borders'	'control'	6	6
'Leave'	'European'	'Union'	6	6
'Leave'	'Vote'	'EU'	6	6
'Leave'	'June'	'23rd'	5	5
'Leave'	'June'	'EU'	4	4
'EU'	'borders'	'control'	4	4
'country'	'back'	'control'	4	4
'June'	'EU'	'23rd'	4	4
'back'	'EU'	'take'	4	4
'make'	"let's"	'June'	4	4
'nation'	'selfgoverning'	'independent'	4	4
'Leave'	'vote'	'23rd'	4	4
'Leave'	'make'	'EU'	4	4
'get'	"let's"	'country'	4	4
'Leave'	'EU'	'take'	4	4
'vote'	'June'	'23rd'	4	4
'Leave'	'EU'	'control'	3	3
'UKIP'	'May'	'5th'	3	3
'Leave'	'vote'	"let's"	3	3
'Leave'	'EU'	'get'	3	3
'open'	'EU'	'borders'	3	3
'open'	'door'	'EU'	3	3
'want'	'back'	'democracy'	3	1
'politicians'	'vs'	'people'	3	1
'must'	'June'	'23rd'	3	3
'Leave'	'EU'	'start'	3	3
'LIKE'	'country'	'back'	3	3
'Leave'	'Thursday'	'EU'	3	3
'June'	'UK'	'23rd'	3	3
'UKIP'	'LIKE'	'voting'	3	2

[44] Opracowanie własne.

"let's"	'June'	'23rd'	3	3
'Leave'	'Vote'	'June'	3	3
'Bus'	'Tour'	'Brexit'	3	3

W obu profilach zauważamy, iż występuje grupa słów i kolokacji, które dominują nad innymi, stanowiąc w miarę zwarty człon, kiedy inne słowa mają kilka różnych kontekstów. W przypadku Davida Camerona narracja oscyluje wokół wpływu Unii Europejskiej na Wielką Brytanię, czyniąc ją silniejszą, lepszą i bezpieczniejszą. Połączenie haseł „vote", „remain" oraz „stronger" i „Europe" nawiązuje do hasła kampanii Britain Strongerin. Co ciekawe, równie często jak hasło „remain" występuje słowo „leave". Co może zaskoczyć, to fakt wysokiej częstotliwości hasła ekonomia przy niskiej częstotliwości jego bi- oraz tri-gramu. Na poziomie interpretacji dane te reprezentują wielość kontekstów, w których występuje to słowo. Jak zatem zweryfikować, czy nawiązania do ekonomii odnoszą się do kampanii przedreferendalnej? Z racji, iż metoda n-gramów informuje nas o najbliższym kontekście danego słowa, a niektóre wpisy są dłuższe niż jedno czy dwa zdania[45], wymagana jest dodatkowa analiza korelacji słów. Po wyznaczeniu gramów oraz słów, które z pewnością odnoszą się do retoryki referendalnej, przygotowujemy krótki algorytm (pętlę[46]), która wyszukuje wpisy zawierające słowo ekonomia oraz powyższe n-gramy i hasła związane z Brexitem. Zaczynamy od wyznaczenia liczby postów, w który pojawia się zwrot ekonomia. Przy 29 powtórzeniach słowa „economy" i jedenastu „economic" otrzymaliśmy liczbę dziesięciu wpisów (przy ewentualnym opuszczeniu wyrażenia „economic" otrzymujemy ich dziewięć, co wskazuje na dodatkowe współwystępowanie słów „economy" oraz „economic"). Następnie tworzymy w algorytmie pętlę, która wyszukuje posty zawierające frazy związane z ekonomią oraz Brexitem. Okazuje się, iż każdy post, w którym pojawia się słowo związane z ekonomią (analogiczny przypadek mamy z słowem „jobs" – pojawia się 23 razy, ale tylko w pięciu postach), został napisany w kontekście Brexitu.

Analogiczne badania dla profilu Nigela Farage'a wykazały wysoką korelację pomiędzy n-gramami i kolokacjami. Słowa związane z granicami, niepodległością (w tym kontekście bardziej niezależnością) i imigracją korelują z hasłami dotyczącymi Brexitu – „vote", „leave", „referendum", „eu".

---

[45] Przykładowo wpis dotyczący zabójstwa MP Jo Cox liczy u Davida Camerona 1126 słów (6333 znaki). Jo Cox była posłanką z Partii Pracy z okręgu Batley and Spen, która aktywnie brała udział podczas debaty dotyczącej syryjskiej wojny domowej (występując przeciwko prezydentowi Basharowi al-Assadowi) oraz wspierała w swoim okręgu grupy będące przeciwko Brexitowi.

[46] Więcej o tym jak działa pętla „for" https://pl.python.org/docs/ref/node52.html (dostęp 28 maja 2018 r.).

W następnej kolejności omówione kody posłużyły za wyznaczenie postów związanych z Brexitem na obu profilach. W przypadku Davida Camerona otrzymaliśmy liczbę 96 postów związanych z Brexitem wobec 27 z nim bezpośrednio niezwiązanych. Co wydaje się oczywiste, przyrost postów związanych z Brexitem stopniowo dominował bliżej daty referendum (z wyjątkiem zabójstwa MP Jo Cox, co spowodowało, iż część postów była poświęcona jej osobie, a sama kampania została chwilowo wstrzymana). W kwestii weryfikacji skuteczności kodów wpisy niebrexitowe zostały poddane ewaluacji tradycyjną hermeneutyką. Wpisy niezwiązane z debatą przedreferendalną dotyczyły spraw bieżących z różnych dziedzin życia: sportu, urodzin królowej, wydarzeń na świecie związanych z kataklizmami czy zamachami. Wyjątek na tym tle stanowią cztery wpisy związane z wyborami samorządowymi, a odwołujące się do kwestii związanych z gospodarką. Niemniej jednak posty te były pozbawione nawiązań do haseł „remain" czy „eu", zatem dzięki poprzedniemu etapowi ewaluacji (korelacji n-gramów i konkretnych wyrażeń) selekcja postów zgodnie z algorytmem zawieranych zwrotów dotyczących Brexitu przyniosła oczekiwane efekty i skutecznie oddzieliła posty związane z gospodarką (w przypadku wyborów samorządowych chodziło o kwestie podatków oraz inwestycji) i referendum a tymi związanymi z wyborami samorządowymi.

W kwestii wpisów Nigela Farage'a analogiczna analiza przyniosła wynik 21 postów niezwiązanych bezpośrednio (!) z Brexitem przy 127 nawiązujących to tej tematyki. Podobnie jak w przypadku Davida Camerona, pojawił się problem podczas wyborów samorządowych, jednakże tu kwestia ma się nieco odmiennie niż u premiera Wielkiej Brytanii. Nigel Farage w niektórych postach związanych z wyborami samorządowymi prowadzi narrację połączenia obu wyborów – wygrana UKIP w wyborach samorządowych miałaby być krokiem w stronę Brexitu. Tu znowu przychodzi analiza n-gramów oraz współwystępowania słów dzięki której byliśmy w stanie wskazać, które posty z kampanii samorządowej nawiązywały do Brexitu, a które bezpośrednio tego nie czyniły[47]. Posty niezwiązane z Brexitem miały po części podobny charakter co u jego adwersarza: życzenia z okazji urodzin królowej czy z okazji dnia świętego Jerzego. W przypadku selekcjonowania postów u Nigela Farage'a pojawił się problem krótkich wpisów informujących o jego wystąpieniach medialnych oraz spotkań podczas wieców. Paralelnie została przeprowadzona analiza typów tych wpisów, dzięki czemu okazało się, iż są one krótkim opisem do załączonych filmów. Szerzej rozwiązanie tej kwestii omówię w części wniosków i ewaluacji algorytmów.

---

[47] Jak już wspomniałem, problem nałożenia się dwóch wyborów w krótkim odstępie czasowym nastręcza wielu trudności. Wydaje się naturalne, iż politycy łączyli w swoich kampaniach oba wydarzenia w mniejszym bądź większym stopniu. Bardziej istotna byłaby tu kwestia przekonania do siebie wyborców w wyborach samorządowych, co mogłoby się przełożyć na analogiczne wybory referendalne.

Etap trzeci.
Interpretacja wyników i analiza komparatystyczna

Powyżej wyklarowane zostały hasła związane z narracją referendalną oraz tematyką, która się w nich wybijała, odpowiednio kwestie bezpieczeństwa, kontroli granic i imigracji oraz odzyskanie niezależnego zarządzania krajem u Nigela Farage'a; bezpieczeństwa ekonomicznego, silniejszej gospodarki oraz ryzyka ewentualnych negatywnych następstw (znowu związanych z ekonomią) opuszczenia Unii przez Wielką Brytanię – u Davida Camerona. Zaobserwowano także, jak wygląda rozłożenie frekwencji dla powyższych słów oraz ich kolokacji metodą n-gramów. W dalszej kolejności po dokonaniu selekcji wpisów otrzymałem informację o zdecydowanej dominacji tematyki referendalnej w narracji u obu polityków – 85 proc. całości postów u Nigela Farage'a oraz 78 proc. dla Davida Camerona.

W tym miejscu chciałbym dokonać zestawienia dotychczasowych przemyśleń z wspomnianymi przeze mnie wcześniej opracowaniami nastawienia wyborców podczas głosowania w referendum[48]. W badaniu styczniowym opracowanym przez sir Johna Curtice'a, profesora nauk politycznych z uniwersytetu w Strathclyde, zadano pytanie wyborcom obu stron, co sądzą o przedstawionych tematach podnoszonych w referendum.

- Stosunek do opuszczenia Unii i jego znaczenia dla gospodarki brytyjskiej: 59 proc. popierających opuszczenie uznało, iż będzie to efekt pozytywny. W przypadku chętnych za pozostaniem w Unii aż 73 proc. wyborców uważa, iż taki skutek miałby wymiar negatywny.
- Stosunek do kwestii imigracji: 65 proc. deklarujących swoje poparcie dla opuszczenia Unii uważa, iż Wielka Brytania będzie w stanie lepiej kontrolować granice i kwestie migracji po Brexicie. Analogicznie, tylko jedenaście proc. chcących pozostać w Unii uważa, iż ta kwestia nie ulegnie zmianie.
- Stosunek do kwestii ryzyka wyjścia z Unii i innych negatywnych następstw: zwolennicy opuszczenia Unii zdecydowanie (78 proc.) uważają, iż Brexit nie przyniesie negatywnych efektów i będzie to proces bezpieczny dla Brytyjczyków. Zwolennicy pozostania w strukturach unijnych są oczywiście przeciwnego zdania – aż 89 proc. z nich uważa, iż pozostanie w Unii jest bezpieczniejszym rozwiązaniem.

---

[48] W dwóch kolejnych badaniach ze stycznia oraz czerwca zostało przeanalizowane nastawienie wyborców do konkretnych tematów referendalnych, por. https://whatukthinks.org/eu/wp-content/uploads/2016/06/NatCen-EU-Referendum-Report-20.06.16-3.pdf oraz https://whatukthinks.org/eu/wp-content/uploads/2016/06/NatCen-EU-Referendum-Report-20.06.16-3.pdf (dostęp 28 maja 2018 r.).

W badaniach czerwcowych zadano kilka dodatkowych pytań:

- Stosunek do kwestii wpływu członkostwa w Unii na poczucie brytyjskiej tożsamości narodowej: 84 proc. zwolenników Brexitu uważa, iż Unia zagraża brytyjskiej tożsamości narodowej, przy czym tylko dziewiętnaście proc. zwolenników pozostania w Unii uważa podobnie.
- Stosunek do kwestii niezależności: 91 proc. zwolenników Brexitu jest zdania, iż członkostwo Zjednoczonego Królestwa w Unii ogranicza niezależność ich państwa. Ale już 35 proc. spośród zwolenników pozostania w Unii uważa tak samo.
- Stosunek do wpływu Brexitu na rynek pracy: dziewięć proc. zwolenników uważa, iż opuszczenie Unii będzie miało negatywny skutek objawiający się znacznym wzrostem bezrobocia. 46 proc. osób popierających członkostwo w Unii ma podobne obawy.
- Poczucie pewności, co wydarzy się po Brexicie: zwolennicy Unii w 31 proc. są pewni, co wydarzy się po wyjściu Wielkiej Brytanii z Unii Europejskiej (69 proc. uważa to za wielką niepewną). Po drugiej stronie sceny politycznej aż 54 proc. wyborców uważa, iż konsekwencje opuszczenia Unii są pewne do przewidzenia (46 proc. jest odmiennego zdania).

W świetle wyników referendum oraz badań opinii publicznej obserwujemy znaczny poziom skorelowania powyższych danych z przeprowadzoną przeze mnie analizą tematów, które pojawiały się u dwóch czołowych polityków w ich narracji przedreferendalnej. Nie jest celem tego artykułu odpowiedź na pytanie o taki stan rzeczy, jednakże chciałbym zwrócić uwagę na to, iż możliwe były hipotetycznie dwa scenariusze. W pierwszym politycy dokonali wcześniej sondy społecznej, aby zbadać, które tematy są istotne w kontekście członkostwa Wielkiej Brytanii w Unii Europejskiej, co następnie zostało przyjęte w ich retoryce. W drugim przypadku rozważam możliwość znacznego wpływu przekazu propagandowego na nastawienie społeczeństwa do tematyki referendalnej. Skłaniałbym się ku drugiej hipotezie, zwłaszcza jeśli spojrzy się bliżej na sposób, w jaki była kreowana narracja wokół danych tematów we wpisach polityków (były one krótkie, nacechowane emocjonalnie, powtarzające tę samą frazę, pisane prostym słownictwem). Kolejnym aspektem byłby poziom recepcji tych wpisów na portalu Facebook, co pokrywa się z badaniami dotyczącymi używania nowych mediów w kampanii przedreferendalnej i ich znaczenia na wynik głosowania[49]. To oczywiście byłoby kwestią naturalną, mającą swoje korzenie w ogólnym nurcie zwiększającej się aktywności polityków w mediach społecznościowych i wpływu tej aktywności na chociażby wybory, co jak wskazałem na początku, rozpoczęło się w kampanii prezydenckiej Baracka Obamy.

[49] *EU Referendum Analysis 2016: Media, Voters and Campaign*, pod red. D. Jackson, E. Thorsen, D. Wring, Bournemouth University 2016.

Etap czwarty. Ewaluacja metody i refleksja nad ich przydatnością w historii cyfrowej – tudzież w warsztacie historyka XXI w.

Powyższe badania miały na celu ukazanie, jak funkcjonuje opisywana przez mnie metoda, jakie daje możliwości poznawcze oraz jakie metodologiczne aspekty się z tym wiążą. Samo badanie języka naturalnego nie jest niczym nowym w nauce, a zwłaszcza w przypadku warsztatu historyka. Analiza filologiczna tekstów czy hermeneutyka w ogóle bazowały na odczytywaniu słów kluczy, metafor, znaczeń symboli zawartych w korpusach. W obecnym świecie zdominowanym przez olbrzymi przyrost danych (zwłaszcza tych cyfrowych[50]) tradycyjna analiza każdego słowa w korpusie może nie tylko być czasochłonna, ale i na swój sposób nieefektywna, zależna od wielu czynników sprowadzonych subiektywnie do kondycji badacza, uniemożliwiając czasem dostrzeżenie pewnych prawidłowości, korelacji, bardziej generalnych wzorów powiązań, nie wspominając już o tradycyjnej metodzie kartki, ołówka oraz ręcznego liczenia zmiennych i stałych. Dla historyka interesującego się badawczo przemianami zachodzącymi w społeczeństwie XXI w., gdzie media społecznościowe stały się codziennością, także tą polityczną, analizowanie dużych korpusów tekstowych staje się nieodzowną koniecznością. Sięgnięcie po metody komputerowej analizy tekstu jest kolejnym narzędziem i metodą mającą służyć wspomaganiu badań, nie tylko ich przyspieszaniu, ale, jak wyżej wspomniałem, ma pomóc nabrać nowych – metaperspektyw.

Analiza języka naturalnego przysparza wielu problemów, niejednoznaczności, co jest chyba wpisane po prostu w specyfikę języka i jego ewolucji. Daleki jestem od absolutyzacji efektów, jakie daje komputerowa analiza, nie chcąc popaść w pułapkę mitu obiektywizmu wynikającego ze stosowania metod przynależnym naukom ścisłym, a implikowanym na grunt szeroko pojętej humanistyki[51]. Pułapki te związane z bezgranicznym zaufaniem do liczb w humanistyce pojawiły się również na różnych etapach stosowanych przeze mnie algorytmów.

Na początku pracy nad źródłami *born digital* w tym referacie zaobserwowany został problem z datacją źródeł, co było spowodowane pewnymi niuansami technicznymi, które powyżej przedstawiłem. Problem ten wpisuje się w ogólną kwestię rozumienia specyfiki źródeł *born digital* – ich zmienności, hipertekstowości, płynnych granic ich formy. Problem ten, choć technicznie

---

[50] A. Radomski, *Big Data i wizualizacja. Kilka uwag o problemach i dylematach współczesnego historyka historiografii*, „Historia@Teoria" 2017, vol. 1/3.

[51] O problemie obiektywizmu i stosowaniu metod analizy języka naturalnego por. M. Eder, *Metody ścisłe w literaturoznawstwie i pułapki pozornego obiektywizmu – przykład stylometrii*, „Teksty Drugie" 2014, nr 2, s. 90–105.

wydawać się może błahy, ma istotne znaczenie dla rozumienia zachodzących procesów, odpowiedniego przypisania chronologii dla dziejących się wydarzeń. Kolejną kwestią jest to, co uwidoczniło się na etapie stosowania algorytmów do oczyszczania korpusu tekstowego oraz wyznaczenia frekwencji słów i kolokacji metodą n-gramów. Po pierwsze, gotowe zbiory słów przystankowych i znaków interpunkcyjnych okazały się niewystarczające i wymagały uzupełnienia o odgórnie i subiektywnie sporządzoną listę dodatkowych słów, które powodowały „zanieczyszczenie" korpusu. Drugim problemem metodycznym była kwestia pozbawienia słów ich kontekstu podczas tworzenia listy frekwencyjnej. Tu z pomocą przyszła omówiona wcześniej metoda n-gramów, która słowa przedstawiała w ich kolokacji. Dzięki temu byliśmy w stanie rozeznać, czy słowa związane pozornie z Brexitem rzeczywiście się do niego odnoszą (problem hasła „unia"). Trzecią kwestią była niewystarczająca ilość danych do interpretacji znaczenia słów związanych z ekonomią, które mimo dużej frekwencji w przypadku trigramów występowały w wielu różnych kontekstach, nie dając się jednoznacznie przypisać. Dopiero zastosowanie algorytmu (pętli) analizującej wpisy pod kątem współwystępowania haseł brexitowych z gospodarką pozwoliło osiągnąć oczekiwane rezultaty, jednoznacznie wskazując na powiązanie między tymi tematami. W celu weryfikacji wyników podzielone posty zostały przeanalizowane tradycyjną metodą hermeneutyczną, dzięki czemu został bliżej wyjaśniony problem z krótkimi wpisami informującymi o wystąpieniach medialnych Nigela Farage'a. Przedmiotem mojej analizy był korpus tekstowy utworzony z obiektu tekstowego wpisów na Facebooku, stąd po zastosowaniu algorytmów nie byliśmy w stanie pierwotnie dostrzec niuansu związanego z rozróżnieniem treści wpisu oraz typu wpisu: tj. postu tekstowego a postu będącego nagłówkiem dla załączonego filmu. Po tej ewaluacji dokonano grupowania postów względem ich typu oraz ponowiono analizę tekstową z uwzględnieniem treści zawartych w filmie. Weryfikacja ta pozwoliła potwierdzić skuteczność metody w przypadku analizy tekstów pod kątem ich tematyki.

W tym miejscu chciałbym wyjaśnić pewien aspekt narracji, który wykorzystałem w niniejszym artykule. W naukach humanistycznych, takie przynajmniej odnoszę wrażenie, niuanse techniczne wraz z ich nomenklaturą i technicznym opisem metod nie są czymś popularnym. Ze względu na fakt, iż niektóre z opisywanych przeze mnie metod i schemat postępowania w badaniach nie są powszechnie opisywane w literaturze polskiej[52] oraz bacząc na cele mojego artykułu, zdecydowałem się odstąpić od tego trendu i przybliżyć czytelnikowi powyższe kwestie. Mam nadzieję, iż dzięki temu nie tylko spro-

[52] Co nie oznacza, iż nie ma ich w ogóle. Zwłaszcza w ostatnich latach można zaobserwować pewien wzrost zainteresowania metodyką badań statystycznych informatycznych, co wiąże się z coraz większą interdyscyplinarnością badań.

wokuje to do szerszej dyskusji i próby falsyfikacji moich propozycji, ale również zachęci czytelników niezaznajomionych z tą tematyką do pogłębienia lektury w dziedzinie metod cyfrowych w humanistyce.

Historia cyfrowa i jej metody. Wizja dalszych badań

Kończąc, chciałbym odnieść się do postawionego przeze mnie pytania, czy potrzebujemy historii cyfrowej i jej metod. O ile w przypadku źródeł analogowych jest to sprawa opcjonalna, o tyle w kwestii źródeł *born digital*, jak mam nadzieję udało mi się to zobrazować, staje się to w świetle dużych korpusów rzeczą niemal nieodzowną. Stąd i ze względu na specyfikę przedmiotu badań, jakim są źródła cyfrowe, odpowiedź zdaje się nasuwać sama. Może pojawić się oczywiście wątpliwość o kwestie zysków ze stosowanych metod, potencjalnie niewielkich korzyści, niemniej jednak chciałbym zauważyć, iż zastosowana metoda miała na celu jedynie dokonać obserwacji tematyki referendalnej. Przygotowana baza badawcza może stać się przyczynkiem do kolejnych analiz, chociażby w kwestii recepcji treści, obiegu informacji w komentarzach, analizy sentymentu wobec konkretnych tematów referendalnych. Czy zautomatyzowane metody badań wyłączą z warsztatu przedcyfrowe metody badawcze? Jak pokazałem, metody komputerowe są elementem pomocniczym, punktem wyjścia dla dalszych tradycyjnych badań, w tym tych hermeneutycznych, które stoją na końcu każdej analizy zautomatyzowanej, weryfikując ją i interpretując jej wyniki.

Bibliografia

Opracowania

Bird S., Klein E., Loper E., *Natural Language Processing with Python*, O'Reilly Media 2009.

Binder J.M., *Alien Reading: Text Mining*, Language Standarization and the Humanities, http://dhdebates.gc.cuny.edu/debates/text/69 (dostęp 28 maja 2018 r.).

Bohdanova T., *Unexpected Revolution: The Role of Social Media in Ukraine's Euromaidan Uprising*, „European View" 2014, nr 13(1).

Broersma M., Graham T., *Social media as beat. Tweets as a news source during the 2010 British and Dutch Elections*, „Journalism Practice" 2012, vol. 6.

Communicator-in-Chief. How Barack Obama Used New Media Technology to Win the White House, pod red. J.A. Hendricks, R.E. Denton Jr., Plymouth United Kingdom 2010.

Dutton W.H., *The Oxford Handbook of Internet Studies*, Oxford University Press 2013.

Eder M., *In search of the Author of Chronica Polonorum Ascribed to Gallus Anonymus: A Stylometric Reconnaissance*, „Acta Poloniae Historica" 2015, nr 112.

Eder M., *Metody ścisłe w literaturoznawstwie i pułapki pozornego obiektywizmu – przykład stylometrii*, „Teksty Drugie" 2014, nr 2.

EU Referendum Analysis 2016: Media, Voters and Campaign, pod red. D. Jackson, E. Thorsen, D. Wring, Bournemouth University 2016.

Howard P.N., Aiden D., Deen F., Muzammil H., Will M., Marwa M., *Opening Closed Regimes: What Was the Role of Social Media During the Arab Spring? Project on Information Technology and Political Islam Data Memo 2011.1*, Seattle: University of Washington 2011.

Kosinski M., Wang Y., Lakkaraju H., Leskovec J., *Mining big data to extract patterns and predict real-life outcomes*, „Psychological Methods" 2016, nr 21(4).

Radomski A., *Big Data i wizualizacja. Kilka uwag o problemach i dylematach współczesnego historyka historiografii*, „Historia@Teoria" 2017, vol. 1/3.

Radomski A., Bomba R., *Zwrot cyfrowy w Humanistyce*, Lublin 2013.

Tenen D., *Blunt Instrumentalism: On Tools and Methods*, http://dhdebates.gc.cuny.edu/debates/text/60 (dostęp 28 maja 2018 r.).

Trzoss A., Werner W., *Problemy i wyzwania związane z badaniem i archiwizacją aktywności instytucji publicznych w mediach społecznościowych*, [w:] *Toruńskie Konfrontacje Archiwalne. Pogranicza archiwistyki*, t. VI, pod red. A. Rosa, Toruń 2019 [w druku].

Uberoi E., Watson C., Keen R., *Local Elections 2016.Briefing Paper,* no. CBP 7596, House of Commons Library 2016.

Zafarani R., Ali Abbasi M., Liu H., *Social Media Mining. An Introduction*, Cambridge 2014.

Dokumentacja programu Jupiter: http://jupyter.org/documentation (dostęp 28 maja 2018 r.).

Dokumentacja programu TextStat: http://neon.niederlandistik.fu-berlin.de/en/textstat/ (dostęp 28 maja 2018 r.).

Dokumentacja biblioteki Pandas: http://pandas.pydata.org/pandas-docs/stable/ (dostęp 28 maja 2018 r.).

Dokumentacja API portalu Facebook: https://developers.facebook.com/docs/graph-api (dostęp 28 maja 2018 r. v.3.0).

Dokumentacja programu Anaconda: https://docs.anaconda.com/ (dostęp 28 maja 2018 r.).

Dokumentacja pętli for dla Python: https://pl.python.org/docs/ref/node52.html (dostęp 28 maja 2018 r.).

Zbiór analiz NatCen dotyczący EU Referendum: https://whatukthinks.org/eu/comment-analysis/ (dostęp 28 maja 2018 r.).

Raport NatCen z 20go czerwca 2016: https://whatukthinks.org/eu/wp-content/uploads/2016/06/NatCen-EU-Referendum-Report-20.06.16-3.pdf (dostęp 28 maja 2018 r.).

Artykuł prasowy BBC o wskazaniu głównych stron lobbujących w czasie EU Referendum: https://www.bbc.com/news/uk-politics-36038672 (dostęp 28 maja 2018 r.).

Ogłoszenie Komisji Wyborczej o finansowaniu stron lobbujących: https://www.electoralcommission.org.uk/find-information-by-subject/elections-and-referendums/past-elections-and-referendums/eu-referendum/grants-to-designated-lead-campaigners (dostęp 28 maja 2018 r.).

Dokumentacja ISO dotycząca formatu daty i czasu: https://www.iso.org/iso-8601-date-and-time-format.html (dostęp 28 maja 2018 r.).

Dokumentacja biblioteki NLTK: https://www.nltk.org/ (dostęp 28 maja 2018 r.).

Definicja Purdah ze stron Parlamentu Brytyjskiego: https://www.parliament.uk/site-information/glossary/purdah/ (dostęp 28 maja 2018 r.).

Statystyka aktywnych użytkowników portalu Facebook: https://www.statista.com/statistics/264810/number-of-monthly-active-facebook-users-worldwide/ (dostęp 28 maja 2018 r.).

Artykuł Telegraph nt. znaczenia Nigela Farage'a dla wyników EU Referendum: https://www.telegraph.co.uk/news/2016/06/24/nigel-farage-has-earned-his-place-in-history-as-the-man-who-led/ (dostęp 28 maja 2018 r.).

Artykuł Guardiana nt. powiązań polityków z organizacjami lobbującymi w czasie EU Referendum: https://www.theguardian.com/technology/2017/may/07/the-great-british-brexit-robbery-hijacked-democracy (dostęp 28 maja 2018 r.).

Adrian Trzoss

Przyczynek do badań nad metodami historii cyfrowej w świetle debaty przed EU Referendum na profilach portalu Facebook Davida Camerona i Nigela Farage'a

Streszczenie

W niniejszym tekście autor omawia wykorzystanie metody analizy języka naturalnego w ramach historii cyfrowej na przykładzie debaty przed EU Referendum na portalu Facebook. Korzystając ze źródeł cyfrowych *born digital*, autor dokonuje analizy statystyk słów, kolokacji oraz kontekstów. Dzięki komputerowym metodom autor wyznaczył dominujące tematy w debacie referendalnej, słownictwo jej towarzyszące, a następnie dokonał ewaluacji stosowanych przez siebie metod. Na koniec autor omawia potencjał stosowanych metod i ich dalszą możliwość wykorzystania w historii cyfrowej zwracając również uwagę na jej ostrożne stosowanie równocześnie objaśniając niuanse techniczne ich stosowania, które mają wpływ na późniejsza interpretację otrzymanych wyników.

Słowa kluczowe: Historia Cyfrowa, EU Referendum, Analiza języka naturalnego

Introduction to research on the methods of digital history in the light of the pre-EU referendum debate on the Facebook profiles of David Cameron and Nigel Farage

Abstract

The present article discusses the use of the natural language processing method in digital history, as exemplified by the pre-EU referendum debate on the Facebook portal. Using *born digital* sources, the author analyzed collocations, n-grams, and word frequency. Computer methods allowed the author to determine the dominant subjects in the referendum debate and the vocabulary used, and subsequently evaluate the methods used in the analysis. Finally, the author discusses the potential for further use of these methods in digital history, pointing out that they must be used cautiously, and also explaining the technical details of using them, which can affect subsequent interpretation of the obtained results.

Key words: Digital history, EU Referendum, NLP