Dyskurs&Dialog 2020/3
10.5281/zenodo.4106225

Mateusz Dobrowolski, Paweł Kalisz

CZEGO NIE MÓWIĄ POLITYCY, POWIE WAM *BIG DATA*[1]

STRESZCZENIE

Artykuł ten jest opisem analizy języka, jakim posługują się politycy na Twitterze z podziałem na frakcje polityczne i wykorzystanie algorytmów uczenia maszynowego.

Słowa kluczowe: big data, twitter, analiza języka, politycy

POLITYKA W CZASACH *BIG DATA*

Jakie są pierwsze skojarzenia, które przychodzą nam do głowy, gdy myślimy o języku współczesnej polityki? Z pewnością nie jest to „wysoka jakość". Między bajki można włożyć wizję dialogu politycznego, który oparty jest na wysłuchiwaniu wzajemnych opinii i merytorycznej debacie. Językoznawcy mówią wręcz o pauperyzacji języka polityki, czyli procesie obniżania się jego jakości. Obserwując język polityków można zauważyć, że nad dyplomatyczną refleksją coraz częściej przeważają emocje i bezwzględna walka polityczna.

Jednakże jest jeszcze druga strona języka polityki, który może być także bardzo wyrachowany i ugładzony. Na co dzień politycy „komunikują się" z nami poprzez „przekazy dnia" i narracje medialne niewychodzące poza linię własnej partii politycznej. Celem takiej komunikacji może być wtedy próba stworzenia własnej prawdy politycznej – w celu autopromocji bądź uzyskania poparcia społecznego.

Jaki jest więc naprawdę język polityki? Aby dowiedzieć się więcej, nie opierajmy się jednak wyłącznie na własnych obserwacjach i emocjach, lecz spróbujmy

przeprowadzić numeryczną analizę języka polityków. W tym celu z pomocą przychodzi nam *big data*.

Jaki jest więc naprawdę język polityki? Aby dowiedzieć się więcej, nie opierajmy się jednak wyłącznie na własnych obserwacjach i emocjach, lecz spróbujmy przeprowadzić numeryczną analizę języka polityków. W tym celu z pomocą przychodzi nam *big data*.

Wraz z rozwojem Internetu i towarzyszących mu technologii ilość danych, które każdego dnia są przetwarzane na całym świecie, rośnie w ogromnym tempie. Zgodnie z www.nodegraphs.se w 2017 roku w cyfrowym świecie istniało 2.7 zettabajta danych, w 2019 było to już 4.4 ZB, natomiast predykcje na rok 2025 mówią nawet o 175 ZB. Sam zettabajt wydaje się być jednostką abstrakcyjnie dużą. Dla lepszego zrozumienia powyższych wartości warto wyjaśnić, że 1 zettabajt = 1,125,899,910,000,000 megabajtów.

Wzrost ilości danych wiąże się rozwojem szybkości łącz internetowych, a co za tym idzie – możliwością dzielenia się zawartością o coraz wyższej jakości. W dzisiejszych czasach standard HD na YouTube już nikogo nie dziwi. Innym czynnikiem mającym wpływ na ilość danych online jest zwiększająca się liczba użytkowników sieci. Pod koniec 2019 roku było to 4.6 mld użytkowników, co w porównaniu z 3.8 mld w roku 2017 daje wzrost o 21%.

Sama liczba użytkowników portali społecznościowych wzrosła w tym czasie o 33%, osiągając 3.7 mld.

Obecnie w ciągu każdej minuty wysyłanych jest 200 mln e-maili, wyszukiwanych 4.2 mln haseł w Google, postowanych 480,000 tweetów i 60,000 zdjęć na Instagramie, oglądanych 4.7 mln filmów na Youtube oraz zakładanych 400 kont na Facebooku.

Te ogromne pokłady danych stwarzają niezliczone wręcz możliwości ich analizy i okazji do karmienia algorytmów uczenia maszynowego. Przykładowo, standardem jest już rozpoznawanie twarzy na zdjęciach wrzucanych na Facebooka, podpowiadanie z dużą dokładnością słów, których chcemy użyć w pisanej właśnie wiadomości na Messengerze, czy podpowiadanie przez serwisy streamingowe, jaki *content* może nam się spodobać. To wszystko jest wynikiem przetworzenia

wielkich ilości danych przez algorytmy AI. Analiza i przetwarzanie dużych zbiorów danych doczekała się w końcu własnej, może niezbyt odkrywczej, nazwy – *big data*.

Big data to termin odnoszący się do zbiorów danych charakteryzujących się cechami opisanymi przez tzw. $5V^2$:

- Volume – duża ilość danych gromadzonych z różnych źródeł, takich jak media społecznościowe, transakcje, dane przekazywane pomiędzy urządzeniami;
- Velocity – duża szybkość przetwarzania szybko powstających danych;
- Variety – odnosi się do różnorodności formatów, w jakich dane są dostarczane;
- Veracity – weryfikacja posiadanych informacji;
- Value – wartość dla użytkownika.

Jeżeli myślimy o źródle danych do analizy języka polityków, wydaje się że naturalnym wyborem jest popularny serwis społecznościowy – twitter.com. Jest to portal, który nie wymaga szerszego przedstawienia.

Dzisiaj Twitter, ze względu na swoją popularność oraz liczbę użytkowników, wykorzystywany jest w dużej mierze nie tylko przez osoby prywatne, ale być może przede wszystkim przez kanały informacyjne i medialne, wszelkiego rodzaju przedsiębiorstwa, partie polityczne i innego rodzaju instytucje. Twitter wykorzystywany jest nie tylko do informowania o własnych działaniach, ale również do przedstawiania własnych poglądów i stanowisk oraz – w przypadku partii politycznych i ich członków – do kreowania własnego przekazu medialnego, „reklamowania" swojej działalności oraz przedstawiania szeroko rozumianej doktryny politycznej. Tysiące nieustannie dodawanych wpisów setek polityków sprawia, że mamy do czynienia z ogromnym zbiorem danych. A przecież wiadomo – w Internecie nic nie ginie. Dane te kumulują się przez lata tworząc ogromny zbiór informacji.

W naszej analizie skupimy się tylko na danych tekstowych, za to występujących w dużej ilości. Zajmiemy się przede wszystkim analizą danych oraz sprawdzimy, jakie możliwości daje analiza języka na portalu Twitter. W tym celu wykorzystamy narzędzia i biblioteki języka *Python*. Dalsze wnioski, które można wyczytać z wyników analizy (niektóre zauważalne w sposób nader oczywisty), niech pozostaną domeną politologów oraz komentatorów polskiej sceny politycznej. Przejdźmy do analizy.

PIERWSZY KROK – ZBIERANIE DANYCH

Traktując problem z perspektywy *big data* należy skupić się nie na języku używanym przez pojedynczych polityków, lecz spróbować objąć w swojej analizie jak najszerszy zbiór danych. Przenieśmy więc naszą uwagę na całe grupy (frakcje) polityczne, które ze względu na zawiązywanie partnerstw bądź koalicji ciężko nazywać tylko i wyłącznie pojedynczymi partiami politycznymi. Przyjęcie takiego założenia umożliwia nie tylko przeanalizowanie większej liczby tweetów, ale też w naszej opinii ukazuje ciekawsze i pełniejsze przedstawienie sytuacji politycznej oraz różnic (ale też podobieństw) w języku, jakim posługują się dane grupy polityków. Za najbardziej oczywisty podział grup politycznych należy uznać ten, który został ustalony w wyniku ostatnich wyborów parlamentarnych i odzwierciedla podział mandatów w Sejmie oraz Senacie RP, oraz bazuje na utworzonych tamże klubach parlamentarnych. Tym samym wyodrębnijmy pięć grup polityków wchodzących w skład następujących frakcji (w nawiasie nazwa robocza przyjęta do analiz): Prawo i Sprawiedliwość (PiS), Koalicja Obywatelska – PO, N, IP, Zieloni (KO), Klub Parlamentarny Lewica – SLD, Wiosna, Razem (LEWICA), Koalicja Polska PSL – K'15 (PSL-KUKIZ), Konfederacja Wolność i Niepodległość (KONFEDERACJA).

Gdy mamy już określone grupy polityczne, następnym etapem jest wybór kont polityków, które powinny zostać włączone do analizy. Mając na uwadze, iż wszystkie grupy polityczne starają się budować własny przekaz medialny oraz kreować taką politykę, która będzie akceptowana i popierana przez jak największe grono osób, wydaje się oczywiste, iż do analizy powinny być wybrane konta tych polityków, którzy cieszą się największą popularnością oraz którzy ze względu na swoją pozycję w danej frakcji politycznej mogą mieć wpływ na kreowanie jej języka. Dlatego też za kryteria wyboru kont przyjmijmy podejście połączone, tj. wybierzmy osoby, które z jednej strony uzyskały najwyższe wyniki podczas ostatnich wyborów parlamentarnych (nominalna ilość oddanych głosów na kandydata), ale również te, które wykazują się dużą aktywnością na portalu twitter.com – duża liczba obserwujących oraz napisanych tweetów.

Po ustaleniu listy kont, następnym krokiem w analizie jest pobranie bazy tweetów. Twitter udostępnia za darmo swoje API (*application programming interface*) jako bibliotekę Tweepy, która posiada dość bogatą dokumentację w sieci. Łącznie pobraliśmy ponad 435 tys. tweetów od przeszło 160 użytkowników[3].

Gdy mamy już utworzony duży zbiór danych, należy go „oczyścić" i przygotować. Kolejnym kluczowym procesem w analizie języka jest tzw. lematyzacja, czyli sprowadzenie słów do podstawowej formy, np. „pójdę" => „iść", „sklepu" => „sklep", „czerwonym" => „czerwony". Do tego celu możemy wykorzystać bibliotekę zawierającą słownik języka polskiego "Morfeus"[4]. Jest to niezwykle użyteczne narzędzie, pozwalające na dokładną analizę części składowych zdania. W celu uproszczenia analizy jako lematy brane są pierwsze interpretacje (najczęściej poprawne) danego słowa zwracane przez słownik.

DRUGI KROK – POLICZMY SŁOWA

Po utworzeniu bazy danych i dokonaniu lematyzacji słów można w końcu przejść do właściwej części analizy, tj. do analizy języka jakim posługują się polscy politycy na Twitterze. Na początku spróbujmy dowiedzieć się, jakie słowa występują najczęściej we wpisach danej grupy politycznej – spróbujemy więc odpowiedzieć na pytanie: **o czym mówią politycy?** W celu zaprezentowania jakie słowa występują najczęściej, wykorzystamy narzędzie *wordcloud* (chmura słów) – graficzne przedstawienie występowania najczęściej wykorzystywanych słów w danym zbiorze tekstowym. Cechą charakterystyczną tego narzędzia jest prezentowanie znaczników (tagów) jako pojedynczych słów, natomiast ich znaczenie lub częstotliwość występowania ukazywana jest za pomocą rozmiaru lub koloru czcionki. Innymi słowy, im dane słowo prezentowane jest w chmurze większą (grubszą) czcionką, tym jego występowanie w tekście jest częstsze. Narzędzie *wordcloud* umożliwia więc szybkie dostrzeżenie najbardziej znanych, najczęściej występujących terminów w celu ustalenia ich względnej ważności w danym zbiorze tekstowym.

Tak wygląda przykładowa chmura najczęściej występujących słów we wpisach polityków PiS:

PiS

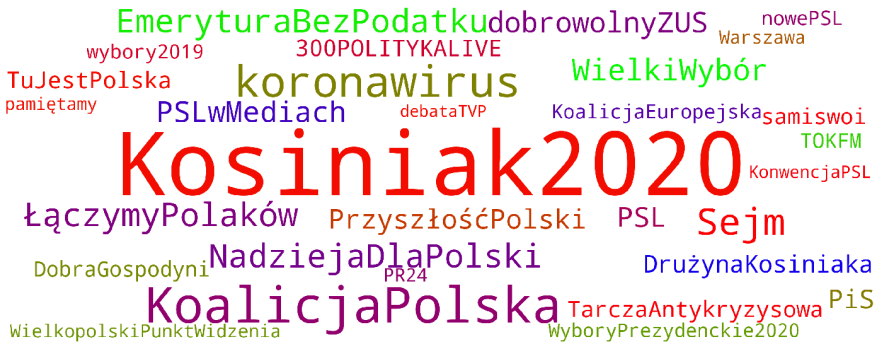
Wyniki dla innych grup politycznych są bardzo zbliżone. Nawet więcej – nie jest łatwo wskazać słowa, które byłyby charakterystyczne tylko dla danej grupy politycznej. Przeważają wyrazy o pozytywnym brzmieniu, zauważalny jest ich doniosły charakter. Jedną z zależności, którą można zauważyć, jest słowo „pis", które występuje wysoko nie w wynikach dla PiS, lecz w pozostałych grupach politycznych. Oznacza to, że politycy wszystkich frakcji chętnie odnoszą się w swoich wypowiedziach do partii rządzącej.

Powyższy model umożliwia ukazanie, jakie słowa padają najczęściej we wpisach polityków, jednakże jego wyniki są podobne dla wszystkich ugrupowań. Spróbujmy więc zrobić podobne liczenie dla „hashtagów", czyli haseł i tematów jakie politycy starają się promować w mediach. Tym razem rzućmy okiem na to, jak wyglądają wyniki dla wszystkich grup:

PiS

KO

LEWICA



PSL-KUKIZ

KONFEDERACJA

Wyniki nie odbiegają zbytnio od oczekiwań. Tematyka hashtagów obejmuje przede wszystkim hasła wyborcze wewnętrzne dla danej grupy politycznej, hasła promujące dane frakcje i ich kandydatów startujących w ostatnich wyborach parlamentarnych oraz prezydenckich. W przeważającej większości są to hasła reklamujące własną działalność (dotyczące własnego ugrupowania), a nie hasła atakujące inne grupy polityczne (o negatywnym brzmieniu). Sporadycznie pojawiają się nieliczne hasła wychodzące poza własną grupę (np. #MamyDość w wynikach KO czy #LGBT dla Lewicy). Pojawiają się nieliczne odniesienia do aktualnych wydarzeń (#koronawirus, #TarczaAntykryzysowa), lecz można domniemywać, że były one również wykorzystywane do celów politycznych. Wnioski? Można stwierdzić, że politycy wykorzystują *hashtagi* głównie do reklamy własnej osoby i swoich ugrupowań. Tematyka haseł sugeruje, iż grupy polityczne mają przeważnie „zamknięty" charakter (politycy tweetują własne hasła), a nie „otwarty" (komentowanie aktualnych tematów).

Powyżej zostały ukazane jedynie dwa przykłady wykorzystania narzędzia *wordcloud* do analizy języka na Twitterze. To oczywiście jest tylko punkt wyjścia i możliwości analizy jest o wiele więcej. Możemy na przykład zadać sobie pytanie – **o kim mówią politycy?** W tym celu sprawdzamy, jakie konta są retweetowane najczęściej, tj. wpisy z jakich kont są najczęściej udostępniane. Kolejnym pytaniem będzie: **do kogo mówią politycy?** W tym celu policzymy, do jakich kont najczęściej się odnoszono i jakim osobom najczęściej odpisywano.

W dużym skrócie wnioski można sformułować następująco: politycy najbardziej lubią mówić o sobie. Wśród najczęściej udostępnianych przez polityków

wpisów na pierwszych miejscach występują oficjalne konta danych ugrupowań (w dalszej kolejności konta najważniejszych polityków danej frakcji). Twitter wykorzystywany jest przede wszystkim do przedstawiania własnych poglądów i stanowisk. Politycy lubią też udostępniać informacje publikowane przez różne media i portale informacyjne. Tutaj też widać pewną zależność, w ramach poszczególnych ugrupowań przeważają portale informacyjne, które można w pewien sposób powiązać lub skojarzyć z daną grupą (PiS – tvp_info, RadioMaryja, wpolityce_pl, KO – Radio_TOK_FM, KONFEDERACJA – MediaNarodoweMN).

Nie ma również zaskoczeń, jeśli spojrzymy na wyniki analizy przeprowadzonej pod kątem, czyje wpisy są najczęściej komentowane. Podobnie jak przy retweetach, politycy przede wszystkim komentują lub odpowiadają na wpisy w ramach własnej grupy politycznej. Większość odpowiedzi kierowana jest do oficjalnych kont grup politycznych oraz do wpisów najważniejszych polityków danej frakcji. Po raz kolejny widać, że politycy wykorzystują Twitter do przedstawiania własnych poglądów i stanowisk oraz komunikowania się wewnątrz ugrupowań. Oprócz kont „partyjnych" występują również konta portali medialnych i informacyjnych. Ponownie widać zależność, że grupy polityczne są zamknięte, bardziej skłonne do tweetowania we własnym środowisku niż do komentowania na zewnątrz. Należy zauważyć jeden wyjątek – wyniki dla konta pisorgpl występują wysoko nie tylko dla PiS, ale też dla pozostałych grup politycznych. Można to odczytać, że politycy wszystkich frakcji chętnie odpowiadają i „dyskutują" z oficjalnym kontem PiS. Przejdźmy teraz do kolejnej części analizy, która pozwoli nam wyjść poza strumień sloganów i haseł wyborczych i odkryć to, co politycy starają się tak skrzętnie ukryć pod ugładzonymi medialnymi przekazami, czyli **jak się mówi?**

TRZECI KROK: WORD2VEC – CZYLI JAK STWORZYĆ MAPĘ SŁÓW

Techniką stosowaną do pozycjonowania słów względem siebie jest tak zwany word *embedding*. Jej początki przypadają na początek lat 2000, jednak jej rozkwit nastąpił dopiero w 2013 roku po serii publikacji zespołu Google pod przewodnictwem Tomasa Mikolova, kiedy pracowana została metoda zapisu słów za pomocą wektorów w przestrzeni wielowymiarowej. Metoda ta nosi nazwę *word2vec* i działa na zasadzie dwuwarstwowej sieci neuronowej. Zapis taki można stworzyć na dwa sposoby:

• CBoW – Continuous Bag of Words, w którym model przewiduje aktualnie uczone słowo na podstawie słów je otaczających, bez uwzględnienia kolejności;

• Skip-gram – model wykorzystuje dane słowo do przewidywania słów otaczających.

Wytrenowany model *word2vec* można wykorzystać na wiele sposobów, takich jak: sprawdzanie podobieństwa pomiędzy dwoma słowami lub grupą słów, wyszukiwanie najbardziej podobnych wyrazów oraz stworzenie wizualnej mapy słów. Warto tutaj wspomnieć, że określenie „podobne słowa" oznacza słowa, które mają największe prawdopodobieństwo do pojawienia się obok słowa zadanego w badanym tekście lub bardziej obrazowo, takie które leżą najbliżej siebie w przestrzeni wielowymiarowej. Stwarza to spore możliwości analizy tekstu pod kątem tego, jak i w jakich kontekstach dane słowo jest najczęściej używane, co jest szczególnie ciekawe w naszym przypadku, ponieważ można dostrzec różnice w usytuowaniu tych samych wyrazów w tekstach różnych środowisk politycznych.

Aby tekst zmieścił się w rozsądnych ramach postanowiliśmy sprawdzić najbliższe otoczenie tylko wybranych słów. Warto zaznaczyć, że w poniższej analizie często najwyższy stopień podobieństwa będą miały słowa bliskoznaczne oraz zawierające się w podobnej kategorii. Na przykład dla nazwiska posła lub ministra często będą zwracane nazwiska innych osób na podobnym stanowisku, dla nazw krajów – nazwy innych krajów itp. Takie wyniki w dużej mierze można pominąć, skupiając się na przymiotnikach bądź rzeczownikach pochodzących z innej kategorii.

W przypadku każdej z partii, w rankingu najczęściej występujących słów wysoko plasują się odniesienia do własnego kraju i narodu, dlatego też warto sprawdzić, w jakich kontekstach są używane.

W przypadku każdej z partii, w rankingu najczęściej występujących słów wysoko plasują się odniesienia do własnego kraju i narodu, dlatego też warto sprawdzić, w jakich kontekstach są używane. Liczby podane w tabelce oznaczają stopień podobieństwa do słowa kluczowego, przy czym maksymalną wartość 1 będzie miało samo słowo kluczowe.

Słowo kluczowe: Polska				
PiS	**KO**	**Lewica**	**Konfederacja**	**PSL-Kukiz**
kraj: 0.582	polski: 0.607	kraj: 0.564	europa: 0.483	kraj: 0.607
polski: 0.526	kraj: 0.565	europa: 0.527	kraj: 0.438	polski: 0.518
świat: 0.489	świat: 0.522	świat: 0.523	okupować: 0.427	europa: 0.481
silny: 0.453	europa: 0.505	sprawiedliwy: 0.480	świat: 0.408	świat: 0.422
ojczyzna: 0.423	strefa: 0.445	cywilizacja: 0.455	polski: 0.370	braterstwo: 0.421
wspólnota: 0.422	my: 0.437	polski: 0.455	gwarancja: 0.363	społeczeństwo: 0.395
suwerenny: 0.416	państwo: 0.397	wspólnota: 0.416	mocarstwo: 0.361	skłócić: 0.385
społecznie: 0.414	polexit: 0.377	społeczeństwo: 0.416	białoruś: 0.351	odbudować: 0.374
rozwijać: 0.412	białoruś: 0.376	różnorodny: 0.416	atomowy: 0.351	zdeterminowany: 0.364
kryzys: 0.404	warszawa: 0.375	silny: 0.406	holandia: 0.343	idea: 0.360

Odrzucając słowa bliskoznaczne, widać różnice w najbliższym otoczeniu słowa „Polska". I tak dla PiS padają takie słowa jak: „silny", „ojczyzna", „wspólnota", „suwerenny", „społecznie", „rozwijać" oraz „kryzys". Można z tego wyciągnąć wniosek, że retoryka tej partii w kontekście kraju skupia się na niezależności i społeczeństwie.

Dla KO są to: „europa", „strefa", „my", „polexit", „warszawa". Tutaj od razu rzuca się w oczy odniesienie do Europy oraz dyskusje na temat polexitu.

W przypadku Lewicy najczęściej występują takie słowa jak: „europa", „sprawiedliwy", „cywilizacja", „wspólnota", „społeczeństwo", „różnorodny", „silny". Wyniki mogą sugerować przywiązanie tej partii do wartości europejskich, różnorodności i sprawiedliwości społecznej.

Następnie dla Konfederacji są to: „europa", „okupować", „gwarancja", „mocarstwo", „atomowy". Tutaj interpretacja może nie być aż tak oczywista bez znajomości polskich realiów politycznych. Znając je jednak można stwierdzić, że partia ta nie wyraża się o Europie w dobrym świetle, mówiąc w jej kontekście o okupacji. Reszta słów może oznaczać nacisk na niezależność kraju.

Zostaje nam jeszcze PSL-KUKIZ ze słowami takimi jak: „europa", „braterstwo", „społeczeństwo", „skłócić", „odbudować", „zdeterminowany", „idea". I tutaj również przyda się nawet nieznaczna znajomość sceny politycznej, żeby stwierdzić, że partia raczej nie chce skłócić społeczeństwa, ale raczej mówi o tym, że już jest ono skłócone i ma receptę na to, jak je naprawić.

Oczywistym jest to, że same wyniki dla pojedynczych słów nie dają jednoznacznych odpowiedzi na pytanie, w jaki sposób mówi się na dany temat. Aby uzyskać szerszy obraz, warto sprawdzić też wyniki dla słów podobnych do słowa kluczowego. I tak dla słowa „Polska" może być to przymiotnik „polski". Takie zestawienie może dać szerszy obraz i nawet nie znając polskiej sceny politycznej można wyciągnąć pewne wnioski.

Słowo kluczowe: polski				
PiS	**KO**	**Lewica**	**Konfederacja**	**PSL-Kukiz**
polska: 0.526	polska: 0.607	polska: 0.455	ukraiński: 0.491	polska: 0.518
strzec: 0.459	rp: 0.457	sprawiedliwy: 0.378	obcy: 0.481	europa: 0.405
polskość: 0.452	nasz: 0.436	amerykański: 0.376	strategiczny: 0.449	przyszłość: 0.395
nasz: 0.445	włoski: 0.419	europa: 0.352	uderzać: 0.435	rozwój: 0.395
kształtować: 0.436	amerykański: 0.410	śląski: 0.348	energetyczny: 0.433	bankrutować: 0.386
ojczyzna: 0.433	naród: 0.384	cywilizacja: 0.348	litwa: 0.429	skutecznie: 0.381

polak: 0.428	zachodni: 0.382	transformacja: 0.332	futrzarski: 0.428	najeźdźca: 0.380
rzeczpospolita: 0.427	zagraniczny: 0.379	unijny: 0.332	oligarcha: 0.425	znacząco: 0.378
dbałość: 0.422	terytorialny: 0.377	kraj: 0.328	węgier: 0.421	odnowa: 0.377
stabilność: 0.414	europa: 0.374	rosja: 0.325	wileńszczyzna: 0.420	kraj: 0.376

Aby bardziej przybliżyć różnice w narracji obozów politycznych, przedstawiamy jeszcze słowa podobne do „rodzina" oraz „ksiądz". Już na pierwszy rzut oka widać odmienność wyników dla poszczególnych grup politycznych. Dla przykładu rodzina dla PiS to przede wszystkim „dziecko" i „wychowywać", podczas gdy w wynikach dla Lewicy na wysokim miejsca są takie słowa jak „partner" i „partnerka".

Słowo kluczowe: rodzina				
PiS	**KO**	**Lewica**	**Konfederacja**	**PSL-Kukiz**
dziecko: 0.500	krewny: 0.517	dziecko: 0.464	namowa: 0.581	rodzic: 0.503
wychowywać: 0.454	matka: 0.418	partner: 0.454	okołoporodowy: 0.541	karta: 0.486
społeczeństwo: 0.433	dom: 0.399	partnerka: 0.441	tradycyjny: 0.517	współczucie: 0.435
rodzic: 0.422	żona: 0.396	ofiara: 0.437	znajomy: 0.510	dziadkowie: 0.414
rodzica: 0.419	brat: 0.393	wychowywać: 0.434	rodzic: 0.508	rodzicielski: 0.412
niepełnosprawny: 0.414	ojciec: 0.389	kondolencje: 0.426	tradycja: 0.499	urlop: 0.402

godność: 0.394	bliski: 0.381	hetero: 0.414	wielodzietny: 0.497	matka: 0.400
opiec: 0.392	córka: 0.374	jednopłciowy: 0.414	łowiecki: 0.486	kondolencje: 0.395
wieś: 0.387	królik: 0.372	seks: 0.408	ulma: 0.477	przebadać: 0.389
społeczny: 0.382	par: 0.369	współczucie: 0.401	osobowy: 0.475	senior: 0.385

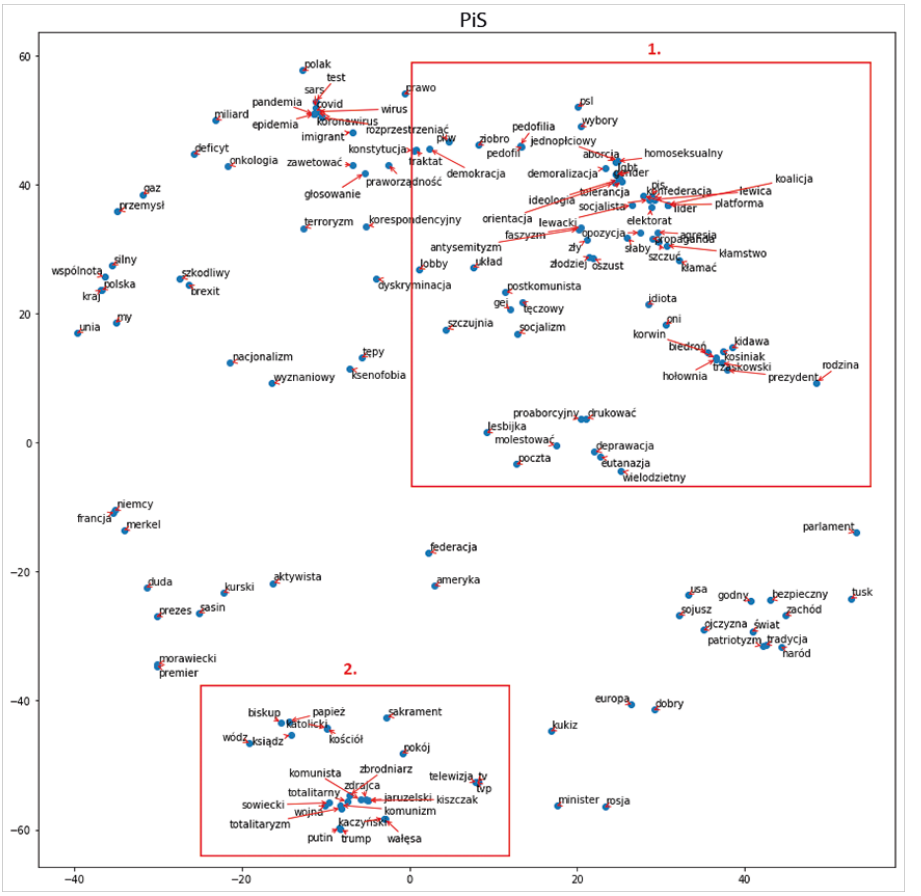
Odmienne narracje zostają jeszcze bardziej unaocznione, jeżeli przyjrzymy się wynikom dla słowa „ksiądz". Bardzo łatwo można wychwycić, które grupy polityczne mogą mieć pozytywne, a które negatywne podejście do tematów, takich jak ksiądz, religia czy kościół. Tutaj bez zaskoczenia, ugrupowania takie jak PiS i PSL-Kukiz mówią o księżach wyłącznie w pozytywnym kontekście, natomiast reszta często odnosi się do problemu pedofilii wśród księży. Dziwić może wysokie podobieństwo do słowa „tylawa" dla KO. Otóż szybko można znaleźć informacje, że jest to miejscowość, w której lokalny proboszcz dopuszczał się molestowania nieletnich.

Słowo kluczowe: ksiądz				
PiS	**KO**	**Lewica**	**Konfederacja**	**PSL-Kukiz**
popiełuszko: 0.757	tylawa: 0.785	pedofil: 0.807	pedofilia: 0.679	franciszek: 0.743
prałat: 0.740	księży: 0.687	tylawa: 0.747	pedofil: 0.674	kardynał: 0.707
biskup: 0.735	gwałciciel: 0.675	jankowski: 0.744	olga: 0.672	odejść: 0.701
proboszcz: 0.733	proboszcz: 0.670	gwałcić: 0.744	parafianin: 0.593	stefan: 0.692
blachnicki: 0.730	biskup: 0.658	ułaskawienie: 0.738	wilkosz: 0.589	pieronek: 0.691

jerzy: 0.718	zgwałcić: 0.650	sutanna: 0.735	kościół: 0.581	popiełuszko: 0.689
kapłan: 0.715	janiak: 0.647	odprawiać: 0.727	polański: 0.557	niezłomny: 0.688
suchowolec: 0.709	dziewczyn-ka: 0.629	księża: 0.721	antypedofil-ski: 0.543	marcin: 0.679
apostoł: 0.707	kuria: 0.628	jędraszewski: 0.718	zaleski: 0.542	biskup: 0.671
błogosławio-ny: 0.696	międlar: 0.628	biskupi: 0.707	pomówienie: 0.541	jerzy: 0.669

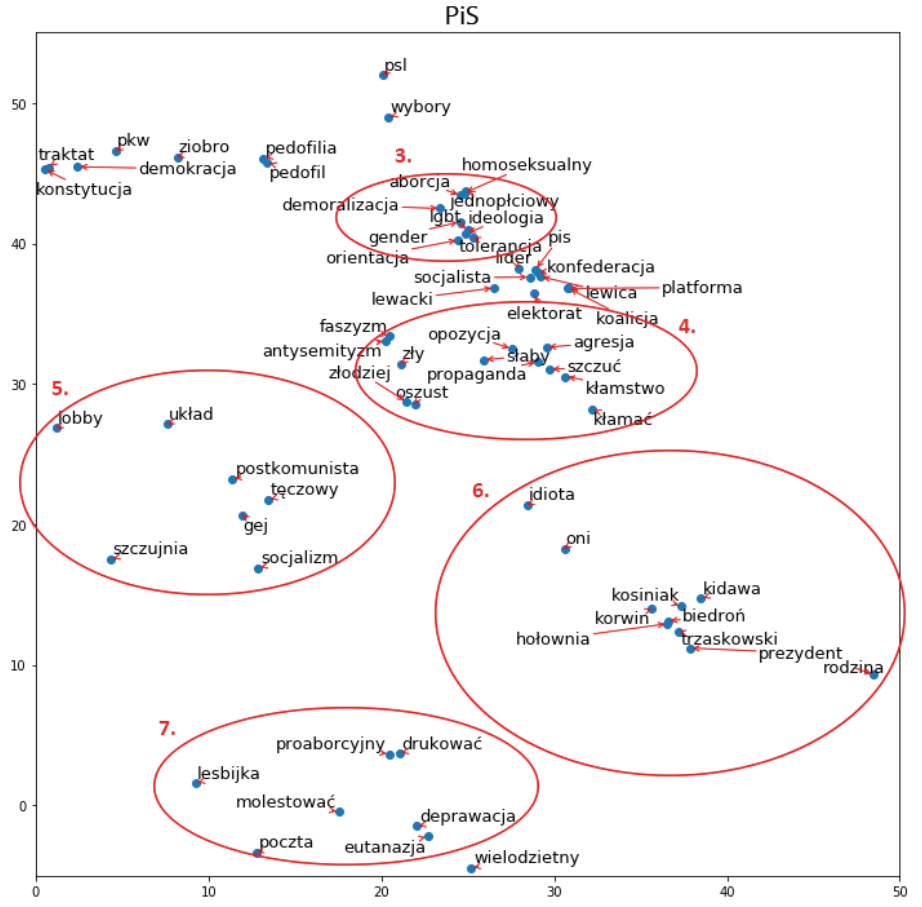
Przejdźmy teraz do już wcześniej wspomnianych map słów. Umożliwiają one spojrzenie na większy zasób słów niż zestawienia najbardziej podobnych, które wskazują tylko najbliższe otoczenie interesującego nas słowa. Aby stworzyć taką mapę konieczne jest zredukowanie liczby wymiarów do dwóch (w naszym przypadku ze stu). W tej analizie został wykorzystany algorytm t-SNE (*T-distributed Stochastic Neighbor Embedding*), który oblicza miarę podobieństwa między parami w przestrzeni o dużych wymiarach i przestrzeni o małych wymiarach, następnie próbuje zoptymalizować te dwie miary podobieństwa.

Poniżej przedstawiamy mapę wybranych słów dla aktualnie rządzącej partii PiS. Wybraliśmy tutaj słowa w pewien sposób nacechowane politycznie oraz odnoszące się do aktualnie dyskutowanych w przestrzeni publicznej tematów.

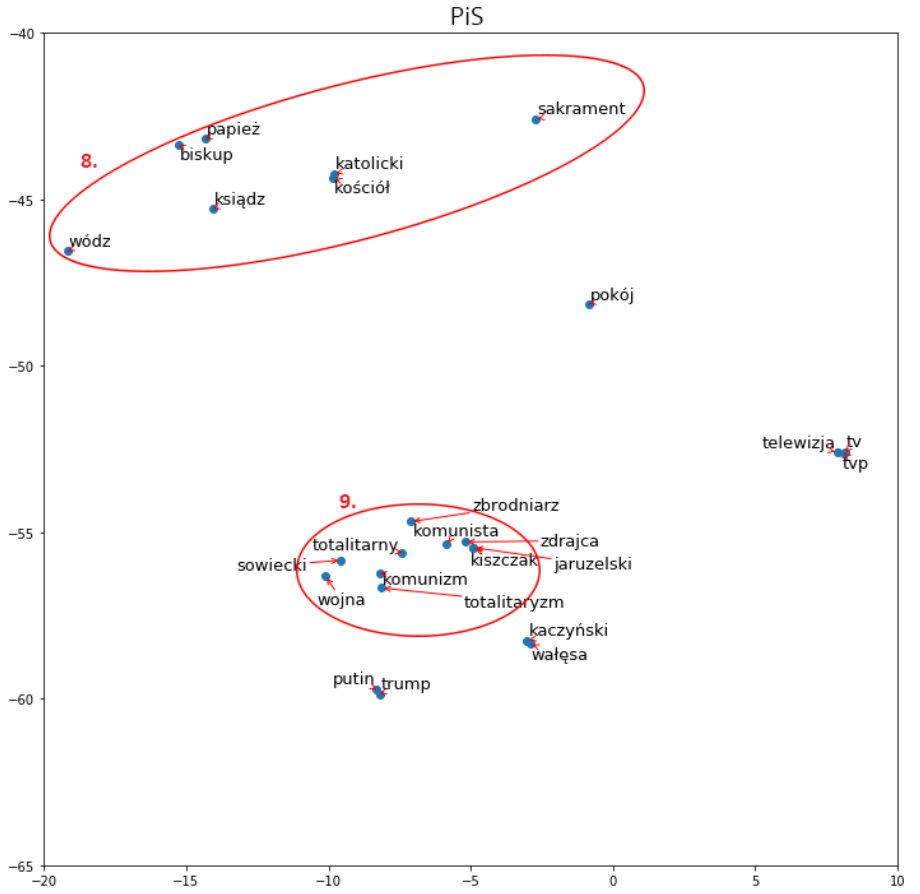


Spróbujmy teraz przyjrzeć się wycinkom powyższej mapy i wybrać kilka grup wyrazów, aby znaleźć analogiczne w mapkach innych partii i porównać narracje. Zacznijmy od wycinka numer 1. To co chyba najbardziej rzuca się w oczy, to otoczenie słowa „oni" (grupa słów nr 6), które po jednej stronie sąsiaduje ze słowem „idiota", po drugiej natomiast znajdują się nazwiska przedstawicieli opozycji oraz „prezydent". To ostatnie wcale nie musi odnosić się do głowy państwa, lecz do niektórych członków partii opozycyjnych, którzy byli lub są prezydentami miast. Nieco wyżej mamy otoczenie słowa „opozycja" (4), które również nie występuje w pozytywnych kontekstach. Zaskoczeniem może być to, że wzmianki partii o samej sobie są usytuowane niewiele wyżej, pośród nazw innych partii oraz słowa „socjalista" i „lider". To jednak, jak już wcześniej wspominaliśmy, może być wynikiem dopasowania do słów o podobnym znaczeniu, ponieważ sama nazwa

partii rządzącej może niezbyt często padać w innych kontekstach. Patrząc jeszcze wyżej na grupę słów nr 3, możemy zauważyć, że słowa takie jak: „homoseksualny", „lgbt", „gender", „aborcja", „tolerancja" są zestawione ze słowem „demoralizacja". Warto również zwrócić uwagę na grupę nr 5, która w otoczeniu słów „gej" i „tęczowy" posiada takie określenia jak: „lobby", „układ", „socjalizm", szczujnia".



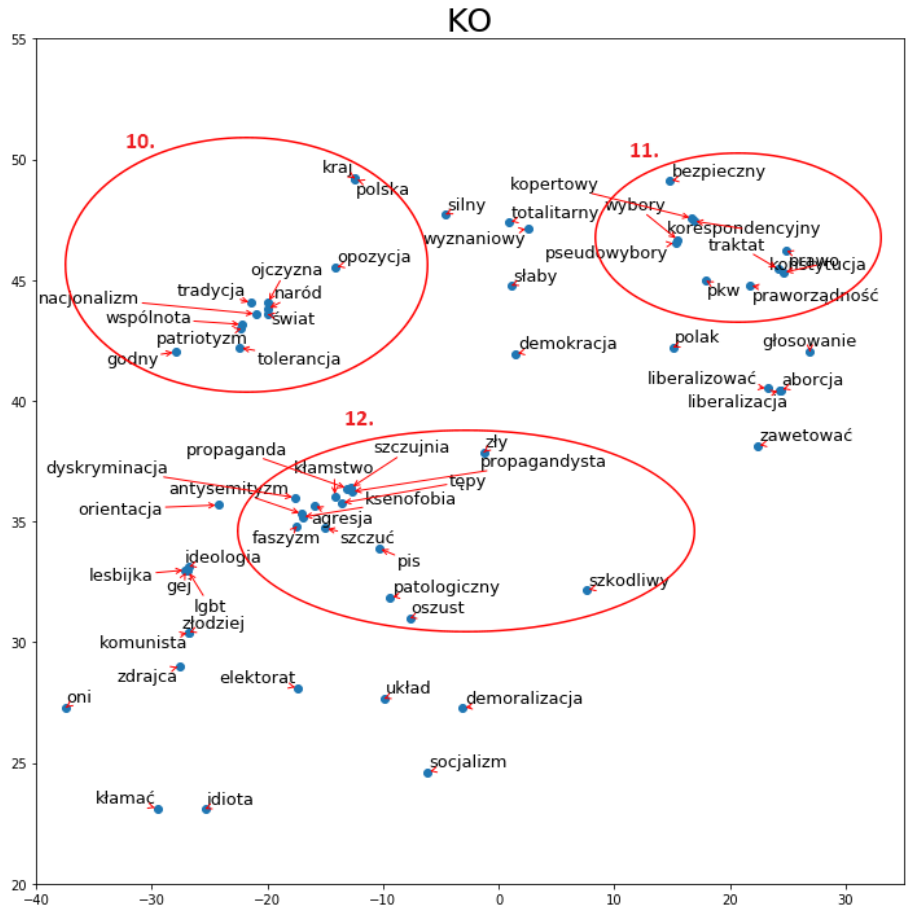
Kolejny wycinek (nr 2) przedstawia narrację partii rządzącej na temat kościoła oraz komunizmu. Jak możemy zauważyć, te dwa tematy są od siebie wyraźnie oddzielone i przedstawiane w kompletnie różnym świetle:



Teraz dla porównania przyjrzyjmy się wycinkowi z mapy dla Koalicji Obywatelskiej (pełna mapa nie została zamieszczona z powodu ograniczeń objętości). W grupie numer 12 od razu można zauważyć, że otoczenie słowa „pis" jest nacechowane negatywnie takimi wyrazami jak: „propaganda", „faszyzm", „ksenofobia". Dla kontrastu, w otoczeniu słowa „opozycja" (10) możemy znaleźć takie wyrazy jak: „ojczyzna", „tradycja", „tolerancja". To porównanie doskonale pokazuje w jaki sposób piszą o sobie przedstawiciele dwóch

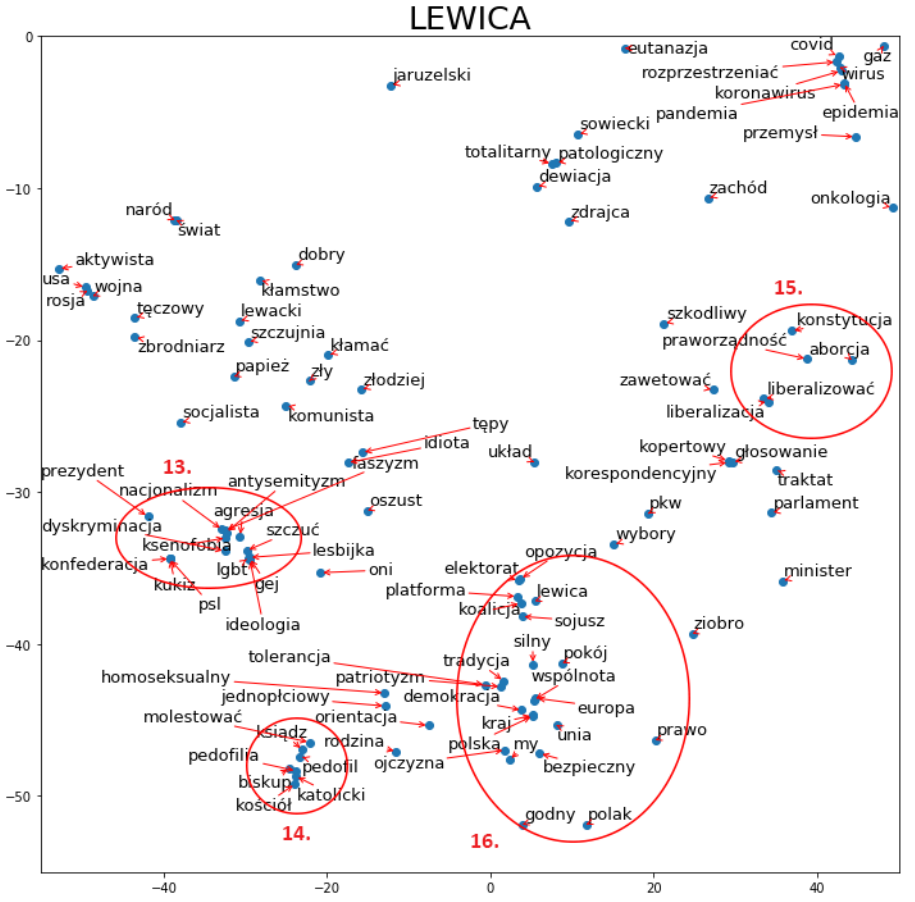
Skupmy się teraz na różnicach światopoglądowych. Najlepszym wyborem wydaje się tutaj porównanie do Lewicy. Możemy zauważyć, że narracja dotycząca kleru jest mocno odmienna od tej, którą przedstawia partia rządząca.

największych środowisk politycznych, można też wywnioskować, że raczej za sobą nie przepadają. Grupa nr 11 odnosi się do wyborów, które początkowo miały być korespondencyjne.



Skupmy się teraz na różnicach światopoglądowych. Najlepszym wyborem wydaje się tutaj porównanie do Lewicy. Możemy zauważyć, że narracja dotycząca kleru jest mocno odmienna od tej, którą przedstawia partia rządząca. Słowami leżącymi najbliżej „ksiądz" i „kościół" (14) są „molestować" czy „pedofil". Warto też przyjrzeć się otoczeniu słowa „aborcja" (15), które najczęściej jest dyskutowane w kontekście: „liberalizacja", „konstytucja", „praworządność". Dalej mamy otoczenie słowa „my" (16), które odwołuje się do wartości takich jak: wspólnota, bezpieczeństwo, demokracja oraz leży blisko słów: „sojusz", „lewica", „opozycja".

Grupa nr 13 pokazuje natomiast narrację dotyczącą ideologii.



Analizę tematyczną można prowadzić bez końca, aby więc nasz tekst nie rozrósł się ponad miarę, zdecydowaliśmy się zakończyć na pokazaniu map oaz ich wycinków tylko dla największych partii. Dla czytelników chcących bardziej zagłębić się w temat przygotowaliśmy aplikację internetową, w której samemu można sprawdzić interesujące nas słowa www.tweet2vec.com.

CO NAM MÓWI *BIG DATA*

Na koniec powróćmy do niektórych pytań, jakie zadaliśmy sobie wcześniej oraz przejrzyjmy narzędzia *big data*, które zostały wykorzystane:

1. **O czym mówią politycy?**
2. **O kim mówią politycy?**
3. **Do kogo mówią politycy?**

i wreszcie

4. **Jak mówią politycy?** – *Word2Vec*.

Próba odpowiedzi na trzy pierwsze pytania ukazała, że najczęściej wykorzystywane słowa są bardzo zbliżone dla wszystkich grup politycznych. Tematyka wpisów skupia się głównie na promocji własnych haseł oraz sloganów wyborczych. Politycy chętnie udostępniają własne przekazy partyjne i odnoszą się głównie do osób z własnego otoczenia. Wykorzystywane hashtagi sugerują, iż politycy są bardziej skłonni do promowania własnej działalności, niż do komentowania działalności innych. Wyjątkiem bądź ciekawą obserwacją jest konto pisorgpl, do którego odnoszą się wszystkie grupy polityczne (podobnie dla słowa „pis", które występuje często w każdej z grup). Wydaje się, że jeżeli politycy już muszą wychodzić poza własną grupę, to najchętniej „dyskutują" z (oraz o) aktualnie rządzącej partii.

Niemniej jednak można stwierdzić, że grupy polityczne mają bardzo zamknięty charakter i generalnie są skłonne do tweetowania jedynie wewnątrz własnego środowiska. Wyniki analizy sugerują, że politycy są nastawieni przede wszystkim na promocję własnych interesów, nie są zainteresowani komunikacją, lecz raczej przedstawianiem własnych poglądów. Twitter służy w dużej mierze jako „tuba", kanał reklamowy służący do promocji swojej osoby bądź partii. Można pokusić się o poszerzenie klasycznej definicji polityki jako sztuki rządzenia państwem, dodając również sztukę autopromocji i kreowania własnej narracji.

Analiza konotacji słów kluczowych, czyli próba odnalezienia słów występujących w podobnych kontekstach do słowa szukanego, ukazała to, czego grupy polityczne starają się nie eksponować.

Najciekawsze wyniki udało się uzyskać odpowiadając na ostatnie pytanie, które miało na celu poznanie, w jakim kontekście padają pewne istotne słowa. Analiza konotacji słów kluczowych, czyli próba odnalezienia słów występujących w podobnych kontekstach do słowa szukanego, ukazała to, czego grupy polityczne sta-

rają się nie eksponować.

Wykorzystując do analizy *big data* narzędzia *word2vec* można dowiedzieć się, co dane grupy polityczne myślą o swoich konkurentach lub w jakim kontekście piszą o takich sprawach, jak rodzina, kościół czy Polska. Nie da się ukryć, że jest to niebywale potężne narzędzie do prób określenia, co dana grupa polityczna tak naprawdę sądzi na konkretny temat, bowiem jak pisał Jan Stępień w *Zamyśleniach* – "Polityka nie lubi prawdy, zaś prawda nie lubi polityki".

W powyższej analizie została zaprezentowana jedynie część technik, które mogą stanowić punkt wyjścia do dalszej analizy języka idącej w kierunku wprowadzenia elementów uczenia maszynowego do klasyfikacji tweetów w celu stworzenia dokładniejszego opisu poszczególnych grup politycznych oraz języka jakim się posługują.

BIBLIOGRAFIA

[1] Ożóg K., *O języku współczesnej polityki*, „Polityka i Społeczeństwo" 4/2007, Rzeszów 2007.

[2] Stępień J., Szyszkowska M., *Zamyślenia*, wyd. Heliodor, Warszawa 2005.

[3] www.nodegraphs.se

[4] http://www.deepdata.pl/

[5] https://www.bbva.com

[6] Woliński M., *Morfeusz reloaded*. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk and Stelios Piperidis (editors), *Proceedings of the Ninth International Conference on Language Resources and Evaluation,* LREC 2014, pages 1106–1111, ELRA, Reykjavik 2014.

[7] Walewski Ł., *Władza w sieci. Jak nami rządzą social media*, WAM, Kraków 2020.

PRZYPISY KOŃCOWE

[1] Artykuł ten został opracowany na podstawie pracy zaliczeniowej napisanej przez jego autorów pod kierunkiem dra Bartłomieja Balcerzaka na studiach podyplomowych (nazwa kierunku: Big Data – inżynieria dużych zbiorów danych) na Polsko-Japońskiej Akademii Technik Komputerowych w Warszawie.

[2] https://www.bbva.com/en/five-vs-big-data/

[3] Pobranie bazy tweetów w dniu 22.07.2020. Analiza obejmuje tweety opublikowane do tego dnia, za początek przyjmując datę utworzenia poszczególnych kont.

[4] Morfeusz – program opracowany w Instytucie Podstaw Informatyki PAN służący do analizy morfologicznej języka polskiego. Jego biblioteka zbudowana jest na „Słowniku gramatycznym języka polskiego" (SGJP). Program często wykorzystywany jest przez aplikacje jako narzędzie do przetwarzania języka polskiego.

WHAT POLITICIANS DO NOT SAY, BIG DATA WILL TELL YOU

ENGLISH SUMMARY

This article is a description of language analysis of Polish politicians on Twitter, considering political fractions with the use of machine learning algorithms.

Keywords: big data, twitter, language analysis, politicians

Mateusz Dobrowolski

Absolwent finansów i rachunkowości, a także studiów podyplomowych z inżynierii dużych zbiorów danych – big data. W pracy zawodowej związany z administracją publiczną i finansami publicznymi. Prywatnie entuzjasta kultury i włoskiej piłki nożnej.

Paweł Kalisz

Absolwent lotnictwa i kosmonautyki oraz studiów podyplomowych big data. Inżynier w branży oil & gas zafascynowany technologiami machine learning. Od niedawna również data scientist w branży medycznej.
.