Wykorzystanie metod automatycznych do analizy treści serwisów informacyjnych

The use of automatic methods to analyze the content of information services

Kaczanowska Monika(1), Kielmer Agata(1),

(1) Wydział Elektronik i Technik Informacyjnych, Politechnika Warszawska,

Opiekun naukowy: doc. dr inż. Sławomir Kula
Kielmer Agata: akielmer@onet.eu
Kaczanowska Monika: monika.kociela@gmail.com

Artykuł badawczy:

Słowa Kluczowe: stylometria, n-gramy, transkrypcja

Streszczenie

Głównym celem niniejszej monografii jest wykazanie różnic w treści i formie przekazu wiadomości w serwisach informacyjnych. W pracy podjęto próbę wykonania analizy treści serwisów informacyjnych dwóch największych polskich stacji telewizyjnych, które mogą reprezentować odmienne opcje polityczne, wykorzystując w jak największym stopniu narzędzia automatyczne. Podjęto próbę określenia sposobu w jaki mówi się na dany temat - jakie oceny poszczególnych elementów przekazu są zasugerowane. Analizie poddano programy serwisów informacyjnych polskiej telewizji publicznej TYP1: 89 odcinków Wiadomości oraz polskiej komercyjnej stacji TVN: 81 odcinków Faktów z okresu trzech miesięcy. Treści serwisów w formie nagrań audio zostały poddane transkrypcji z wykorzystaniem programu SkryBot. Wykonano porównanie dwóch programów do transkrypcji SkryBot oraz Google Web Speech API i wykazano lepszą skuteczność polskiego programu SkryBot. Finalnie udało się określić w wyniku analizy treści główną tematykę i częstotliwość publikowania danych w serwisach informacyjnych. Wykazano różnicę w formie przekazu wyżej wymienionych programów.

1 Wstęp

Genezą poniższych analiz jest fakt, że w 2016 roku Krajowa Rada Radiofonii i Telewizji zleciła wykonanie badania, które miało ustalić czy Wiadomości w TVP realizują obowiązki i powinności nadawcy publicznego (Mrozowski i in. 2016). Głównym zadaniem było ustalenie, jak Wiadomości wypadają na tle innych serwisów informacyjnych. Analiza została wykonana przez naukowców z Uniwersytetu Humanistycznospołecznego SWPS. Zostały przeanalizowane serwisy informacyjne nadawane przez TVN, Polsat i Telewizję Polską. Analizie poddano wydania z okresu 4-11 lutego 2016 roku, czyli osiem wydań Wiadomości, Faktów i Wydarzeń. Pełen raport jest dostępny na stronie internetowej Krajowej Rady Radiofonii i Telewizji(1). Wyżej wymieniony raport stanowi o zauważalnych różnicach w przedstawianiu i doborze newsów.

[1] *http://www.krrit.gov.pl/krrit/aktualnosci/news,2252,wyniki-monitoringu-audycji-informacyjnych.html:*

Po publikacji raportu, portale internetowe z różnych środowisk zamieszczały informacje, że Wiadomości w TVP są stronnicze, a Fakty w TVN krytyczne:

- Newsweek 12 kwietnia 2016 r. „Raport KRRiT nie pozostawia złudzeń: „Wiadomości TVP prorządowe, tendencyjne i kiepskie jakościowo”[2]
- Telewizja Republika 12 kwietnia: „Wiadomości stronnicze, a „Fakty” krytyczne. Poznaliśmy raport KRRiT”[3]

Przeprowadzona w tej pracy analiza różni się od tej, która była wykonana w 2016 roku na zlecenie KRRIT. Jednak dzięki raportowi z 2016 roku istnieją podstawy do przypuszczenia, że serwisy informacyjne na różnych stacjach mogą się różnić. Dalsza analiza została wykonana na większej liczbie wydań oraz w możliwie zautomatyzowany sposób.

2 Materiał i Metody

Do analizy wykorzystano nagrania z okresu od 1 października do 31 grudnia 2017. Ze względu na trudności techniczne oraz brak umieszczania przez wydawców wszystkich wydań finalnie przeanalizowano 81 wydań Faktów i 89 wydań Wiadomości. Tezy które zostały sprawdzone:

- czy treści wiadomości różnią się między sobą?
- czy można wychwycić charakterystyczne słownictwo dla jednego lub drugiego serwisu informacyjnego?

Wybraną metodą jest analiza słów oraz fraz pojawiających się w tekście. Wykonanie tej analizy w możliwie automatyczny sposób wymaga przetworzenia nagrań na tekst, czyli wykonania transkrypcji.

Systemy automatycznego rozpoznawani mowy (ASR) mają za zadanie przetworzyć sygnał akustyczny na tekst z jak największą precyzją, wykorzystując jak najmniejsze zasoby czasu (Rykowski, 2014). Wynika z tego, że kluczowym zadaniem systemu automatycznego rozpoznawania mowy jest przypisanie dla analizowanego sygnału mowy jego najdokładniejszej transkrypcji - zapisu słownego. Zadanie to jest skomplikowane, ponieważ mowa ludzka cechuję się ogromną zmiennością sygnału wynikającą m.in. z różnorodności warunków akustycznych, każdego człowieka. Transkrypcja charakteryzuje się wysoką skutecznością, gdy mówi jedna osoba w danym momencie, a jej głos jest wyraźny, z dobrą artykulacją (Ziółko, Ziółko 2011). Ze wszystkich programów telewizyjnych ten warunek najlepiej wydawał się być spełniony przez serwisy informacyjne, gdzie większość treści przekazywana jest przez wykwalifikowanych prezenterów.

Transkrypcję wykonano polskim programem SkryBot[4], który został wybrany po przeprowadzeniu porównania z aplikacją Google Speech API. W celu oceny skuteczności programów analizie poddano takie same scenariusze tekstowe. Kolejno tekst był czytany lub odtwarzany przez kobietę i mężczyznę zgodnie z założeniami, które podlegały ocenie. Ocenie poddano dopasowanie transkrypcji do oryginalnego tekstu.

Tab.1 Rozpoznawanie mowy spontanicznej z fragmentu serwisu informacyjnego Wiadomości

	SkryBot	**Google Web Speech API**
dopasowanie transkrypcji (1 min)	85 %	85 %
dopasowanie transkrypcji (20 min)	84 %	50-60%

[2] *http://www.newsweek.pl/polska/wiadomosci-tvp-stronnicze-tendencyjne-i-prorzadowe-raport-krrit,artykuly,383800,1.html [dostęp 30.04.2018]*

[3] *http://telewizjarepublika.pl/wiadomosci-stronnicze-a-fakty-krytyczne-poznalismy-raport-krrit,32028.html [dostęp 30.04.2018]*

[4] *https://skrybot.pl/*

W przypadku sygnału mowy z serwisów informacyjnych bardzo dobre wyniki transkrypcji uzyskał program SkryBot. W przypadku porównania transkrypcji trwającej 1 minutę oraz 20 minut, zaszły tylko minimalne spadki jakości - na niekorzyść dłuższej transkrypcji - rzędu 1%. Transkrypcja krótkiego fragmentu serwisu informacyjnego Wiadomości z wykorzystaniem Google Web Speech API jest stosunkowo dobra i wynosi 85% dla fragmentów wiadomości, trwających do 1 minuty. Skuteczność drastycznie spada przy dłuższych fragmentach - do 50-60%. W transkrypcji pominięte zostały fragmenty wypowiedzi, najczęściej wtrącenia. Google Web Speech API w przypadku wtrąceń lub zbyt szybkiej mowy nie zarejestrował wszystkich słów.

Do transkrypcji wykorzystano aplikacje SkryBot ze względu na jej skuteczność w tego typu sygnałach mowy. Drugim aspektem jest również koszt, który z przypadku polskiego programu jest znacznie niższy niż produktu oferty firmy Google.

Uzyskany tekst w celu dalszej analizy został poddany:

- usunięciu stop-słów (stopwords) - są to słowa, które nie wnoszą nic do tekstu z punktu widzenia logiki i przekazu, np.: ach, aj, albo, bardzo, bez, bo itp.,
- usunięciu imion i nazwisko dziennikarzy pracujących w danych stacjach,
- lematyzacji – uwzględnienie analizy morfologicznej słów, tj. grupowanie różnych fleksyjnie form wyrazu (Allahyari, Pouriyeh 2017). Sprowadzoną słowa do form podstawowych z wykorzystaniem biblioteki Polimorfologik 2.1 [5]
- zamianie wszystkich liter na małe

Tak przygotowany tekst został poddany dalszej analizie w programie R[6] - niekomercyjnego oprogramowania wykorzystywanego do obliczeń statystycznych.

3 Wyniki i dyskusja

Pojedyncze wydania serwisów informacyjnych zostały poddane analizie z wykorzystaniem technik stylometryczych. Podstawą analiz była analiza częstotliwości n-gramów. N-gramy to ciągi słów pojawiających się w tekście. (Schonlau, Guenther 2016). Przeanalizowano 1-gramy (pojedyncze słowa), 2-gramy (bigramy), 3-gramy (trigramy).

Do grupowania wydań wykorzystano miarę Delta (Eder, Kestemont, Rybicki, 2018), a analizowaną cechą były n-gramy słów. Miara Delta została zdefiniowana przez Burrow'a (Burrows, 2002). Miara ta zależy od znormalizowanej częstotliwości występowania słów. Na jej wynik wpływa obszerność tekstu, ilości znaków lub słów w tekście.

$$\Delta_{(AB)} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{A_i-\mu_i}{\sigma_i} - \frac{B_i-\mu_i}{\sigma_i}\right| \quad (1)$$

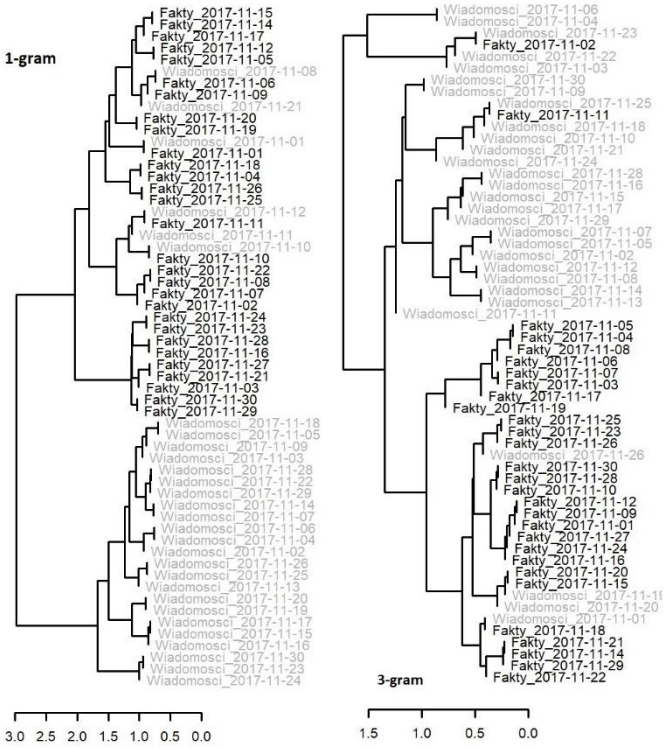
gdzie:
n – liczba najczęściej występujących n-gramów uwzględnianych w analizie;
A, B – teksty, który są porównywane;
A_i – częstotliwość danej frazy i w tekście A;
B_i – częstotliwość danej frazy i w tekście B;
μ_i – średnia częstotliwość frazy w danym tekście;
σ_i – odchylenie standardowe częstotliwości danej frazy.

Wyniki analizy skupień dla pojedynczych słów i trigramów przedstawiono na grafach – klastrach dla wydań z listopada 2017 roku (Rys.1). Wykorzystaną analizę klastrów, która umożliwia grupowanie danych obiektów, w tym przypadku słów w odpowiednie kategorie

[5] *https://github.com/morfologik/morfologik-stemming*
[6] *https://www.r-project.org/*

(Raskar, Thakore, 2011). Charakter grafów - klastrów dla wydań z października i grudnia 2017 był bardzo podobny. Analiza skupień dla pojedynczych wyrazów na podstawie miary Delta wskazuje, że większość wydań Faktów i Wiadomości pogrupowana jest razem. Wydania 1 listopada, 11 listopada, 31 grudnia, w Wigilię i Święta Bożego Narodzenia znajdują się obok siebie.

Analiza skupień dla bigramów, na podstawie miary Delta daje podobne, jednak nieco gorsze wyniki. W odróżnieniu, analizy skupień dla pojedynczych słów wydania z 1 listopada, 11 listopada i 31 grudnia nie grupują się razem. Analiza skupień dla trigramów na podstawie miary Delta pozornie wypada dobrze, jednak przy wnikliwej analizie widać, że istnieje kilka wydań Faktów i Wiadomości odstających od reszty. Analiza skupień dla 4-gramów wypada najgorzej ze wszystkich przeanalizowanych.



Rys.1 Porównanie wyników analizy skupień dla pojedynczych słów oraz trigramów

3.1 Pojedyncze słowa

Porównując listę najczęściej występujących słów (Tab.2) można zauważyć, że większość z nich jest taka sama w Faktach i w Wiadomościach, dodatkowo zdecydowanie większość z nich pojawia się codziennie. Z listy top 30 słów w Faktach, które nie pojawiają się w top 30 słów Wiadomości to: „minister", „sąd". Z listy top 30 słów w Wiadomościach, które nie pojawiają się w top 30 słów Faktach to: „sprawiedliwość", „europejski", „komisja", „kraj".

W celu sprawdzenia czy słownictwo używane w serwisach informacyjnych jest takie samo jak przeciętnie w języku polskim porównano ranking słów z listą leksemów według Jerzego Kazojcia[7].Z rankingu leksemów również, tak jak z tekstów serwisów

[7] *http://www.open-dictionaries.com/slownikfrleks.pdf [dostęp 01.04.2018]*

informacyjnych, usunięto stopwords. Zwrócono uwagę na słowa, które są wysoko w rankingach serwisów informacyjnych, a relatywnie daleko występują w rankingu leksemów. W Faktach takimi słowami były: „chcieć", „prezydent", „minister", „musieć", „sąd", „wszystek". W Wiadomościach takimi słowami były: „chcieć", „prezydent", „musieć", „sprawiedliwość", „komisja", „wszystek". Części słów w ogóle nie ma w rankingu leksemów np.: „móc", „polska", „polski", „europejski". Z analitycznego punktu widzenia to właśnie słowa, które odstają od średniej, wydają się warte uwagi, szczególne rzeczowniki, czyli np.: „prezydent", „minister", „sprawiedliwość".

Tab.2 Lista top 30 słów w Faktach i w Wiadomościach

słowo	pozycja w rankingu leksemów	Fakty		Wiadomości	
		ranking słowa	w ilu wydaniach	ranking słowa	w ilu wydaniach
mieć	44	1	100%	1	100%
rok	227	3	100%	2	99%
móc	-	2	100%	5	100%
polska	-	6	100%	3	100%
polski	-	15	99%	4	100%
chcieć	1991	5	100%	7	100%
mówić	71	7	100%	12	100%
prezydent	957	4	96%	19	96%
państwo	479	16	100%	6	100%
swój	48	12	100%	9	100%
sprawa	141	10	98%	10	99%
rząd	597	19	95%	8	97%
zostać	164	13	100%	11	100%
człowiek	15	14	100%	15	98%
czas	10	11	100%	16	100%
wielki	104	23	98%	13	100%
prawo	126	21	98%	14	98%
nowy	215	20	99%	21	100%
osoba	556	17	100%	24	100%
dobry	118	24	100%	23	100%
minister	1512	9	95%	-	-
musieć	9802	22	99%	28	100%
dzień	31	26	99%	29	97%
raz	5	25	99%	30	100%
sprawiedliwość	1683	-	-	18	94%
sąd	962	18	91%	-	-
europejski	-	-	-	22	94%
komisja	3121	-	-	25	81%
wszystek	7170	28	99%	27	99%
kraj	522	-	-	26	99%

kolorem zaznaczono słowa odstające od średniej

3.2 Bigramy

Tabele (Tab.3) zawierają najczęściej występujące bigramy w Faktach i w Wiadomościach. Częstotliwość i występowanie bigramów jest niższa niż pojedynczych słów.

Tak jak w przypadku pojedynczych słów większość bigramów z listy najczęściej występujących jest wspólna dla Faktów i Wiadomości. Biorąc pod uwagę liczbę wystąpień na pierwszych dwóch miejscach w Faktach są osoby „jarosław kaczyński" i „antoni macierewicz", w Wiadomościach nazwy partii „prawo sprawiedliwość" i „platforma obywatelski".

Tab.3 Najczęstsze bigramy w Faktach i w Wiadomościach

bigram	Fakty		Wiadomości	
	ranking bigramu	w ilu wydaniach	ranking bigramu	w ilu wydaniach
prawo → sprawiedliwość	4	62%	1	89%
rok → temu	3	74%	4	84%
unia → europejski	5	49%	3	78%
jarosław → kaczyński	1	58%	13	45%
mateusz → morawiecki	8	33%	6	43%
sąd → wysoki	9	46%	11	42%
komisja → europejski	16	32%	8	49%
andrzej → duda	7	53%	19	52%
dobry → wieczór	6	83%	26	61%
gronkiewicz → waltz	36	14%	5	45%
hanna → gronkiewicz	38	12%	9	44%
polski → rząd	40	27%	7	69%
platforma → obywatelski	86	17%	2	82%
antoni → macierewicz	2	46%	446	12%

3.3 Trigramy

Kolejnym krokiem w analizie sekwencji słów są trigramy, czyli trójki słów występujące po sobie. Tabele (Tab.4) zwierają zestawienia trigramów. Porównując zestawienia dla trigramów i bigramów widać, że ich częstotliwość spada znacząco w porównaniu z bigramami. Dla Wiadomości liczba wystąpień i udział w wydaniach jest wyższy niż dla Faktów. Jednak najczęstsze trzy 3-gramy są wspólne dla Faktów i Wiadomości, czyli „krajowy rada sądownictwo", „hanna gronkiewicz waltz", „prezydent andrzej duda".

Tab.4 Najczęstsze trigramy w Faktach i w Wiadomościach

słowo	Fakty		Wiadomości	
	ranking trigramu	w ilu wydaniach	ranking słowa	w ilu wydaniach
hanna → gronkiewicz → waltz	2	12%	1	44%
krajowy → rada → sądownictwo	1	25%	2	38%
prezydent → andrzej → duda	3	27%	3	47%
premier → beata → szydło	8	16%	4	40%
rząd → prawo → sprawiedliwość	-	-	5	35%
premier → mateusz → morawiecki	24	11%	6	20%
ustawa → sąd → wysoki	5	21%	11	18%
minister → sprawa → wewnętrzny	13	15%	10	27%
reforma → wymiar → sprawiedliwość	20	14%	9	27%
prezes → prawo → sprawiedliwość	-	-	7	27%

3.4 Analiza wybranych słów i ich kontekstów

Kilka słów poddano głębszej analizie. Słowa wytypowane przy analizie częstotliwości pojedynczych słów, te słowa to: „prezydent", „minister", „sprawiedliwość". Analiza polegała na sprawdzeniu obok jakiego innego słowa pojawia się wytypowane słowo. Wybierano co najwyżej 13 par słów i mniej więcej równy udział w obu serwisach.

Słowo „prezydent", w Faktach i w Wiadomościach pojawiło się w top 30 słowach, a zostało wytypowane do analizy, ponieważ w rankingu leksemów zajmuje dużo niższą pozycję. Tabela (Tab.5) przestawia najczęściej pojawiające się bigramy ze tym słowem. Dla 29% najczęstszych kombinacji ze słowem „prezydent" więcej kombinacji jest w Faktach niż w Wiadomościach. W Wiadomościach w 12% przypadków obok słowa „prezydent" pojawia się „andrzej". W Faktach najczęściej pojawiającym się słowem obok „prezydent" jest „warszawa", na kolejnych miejscach są „andrzej", „mieć" i „duda". Na tej podstawie można się domyślać, że w Wiadomościach zawsze pojawia się fraza „prezydent andrzej duda", a w Faktach „prezydent andrzej duda" lub „prezydent duda".

Tab.5 Najczęstsze bigramy ze słowem „prezydent"

nr	**Fakty**			**Wiadomości**		
	bigram	**n**	**%**	**bigram**	**n**	**%**
1	prezydent → warszawa	28	3,4%	prezydent → andrzej	82	11,9%
2	prezydent → andrzej	27	3,4%	prezydent → warszawa	58	8,4%
3	prezydent → mieć	22	2,8%	prezydent → stolica	35	5,1%
4	prezydent → duda	20	2,6%	prezydent → miasto	23	3,3%
5	kancelaria → prezydent	19	2,5%			
6	spotkanie → prezydent	18	2,4%			
7	prezydent → minister	15	2,0%			
8	prezydent → miasto	14	1,9%			
9	pis → prezydent	13	1,8%			
10	prezydent → prezes	13	1,8%			
11	prezydent → podpisać	12	1,7%			
12	prezydent → donald	12	1,7%			
13	kandydat → prezydent	12	1,7%			
	SUMA		29,5%	SUMA		28,7%

Kolejnym słowem wytypowanym do analizy było słowo „minister". Tak jak w przypadku słowa „prezydent" w top 40% kombinacjach dla słowa „minister" jest więcej kombinacji w Faktach niż w Wiadomościach (Tab.6). W Wiadomościach najczęściej pojawia się „minister sprawa", co jest wynikiem lematyzacji od „minister spraw", czyli bigram od ministra spraw wewnętrznych lub ministra spraw zagranicznych. Kolejni najczęściej pojawiający się ministrowie to „minister sprawiedliwości" i „minister obrony". W Faktach kolejność jest inna. Najczęściej wymienianym ministrem jest „minister zdrowia", następnie „minister obrony" i „minister sprawiedliwości". W Faktach pojawiają się wysoko w rankingu ministrowie wymienienie z nazwiska „minister macierewicz", „minister szyszko".

Tab.6 Najczęstsze bigramy ze słowem „minister"

nr	Fakty			Wiadomości		
	bigram	n	%	bigram	n	%
1	minister → zdrowie	42	6,7%	minister → sprawa	45	11,3%

2	minister → obrona	41	6,6%	minister→ sprawiedliwość	36	9,1%
3	minister → sprawiedliwość	36	5,8%	minister → obrona	25	6,3%
4	minister → macierewicz	28	4,5%	rada → minister	23	5,8%
5	minister → sprawa	21	3,4%	minister → finanse	20	5,0%
6	minister → szyszko	19	3,0%	rozmowa → minister	15	3,8%
7	minister → ziobro	18	2,9%			
8	prezydent → minister	15	2,4%			
9	minister → waszczykowski	11	1,8%			
10	mówić → minister	10	1,6%			
11	prezydencki → minister	10	1,6%			
12	minister → móc	10	1,6%			
	SUMA		41,8 %	SUMA		41,3%

Po porównaniu rankingu słów w serwisach informacyjnych i rankingu leksemów wytypowano do analizy słowo „sprawiedliwość". Popularność tego słowa w serwisach informacyjnych wynikała głównie z nazwy partii Prawo i Sprawiedliwość. Po obróbce tekstu, ta nazwa występowała jako „prawo sprawiedliwość". Pierwsze dziewięć kombinacji stanowi ponad 90% wszystkich wystąpień z tym słowem (Tab.7).

Tab.7 Najczęstsze bigramy ze słowem „sprawiedliwość"

nr	Fakty			Wiadomości		
	bigram	n	%	bigram	n	%
1	prawo → sprawiedliwość	105	41,7%	prawo → sprawiedliwość	378	56,0%
2	wymiar → sprawiedliwość	56	22,2%	wymiar → sprawiedliwość	109	16,0%
3	minister → sprawiedliwość	36	14,3%	ministerstwo → sprawiedliwość	47	7,0%
4	ministerstwo → sprawiedliwość	15	6,0%	minister → sprawiedliwość	36	5,3%
5	wiceminister → sprawiedliwość	9	3,6%	komisja → sprawiedliwość	18	2,7%
6	trybunał → sprawiedliwość	8	3,2%	sprawiedliwość → jarosław	18	2,7%
7	sprawiedliwość → mieć	7	2,8%	sprawiedliwość → mieć	17	2,5%
8	sprawiedliwość → chcieć	6	2,4%	trybunał → sprawiedliwość	14	2,1%
9	sprawiedliwość → pis	5	2,0%	resort → sprawiedliwość	13	1,9%
	SUMA		98%	SUMA		96%

4 Dyskusja i wnioski

Wykazano różnice w treściach serwisów informacyjnych nadawanych na różnych stacjach. Jednak wnioski i analiza były inne niż ta przeprowadzona na zlecenie KRRiT w 2016 roku. W tej pracy przeanalizowano więcej wydań i z zastosowaniem metod automatycznych.

Dlatego wnioski dotyczą głównie stosowanego języka, w odróżnieniu od raportu KRRiT, gdzie przeprowadzono dyskusję m.in. na temat prezentowanego światopoglądu.

Klasyfikacji tekstu do grupy Faktów lub Wiadomości można dokonać na podstawie częstotliwości pojedynczych słów lub bigramów. Dłuższe sekwencje dawały niejednoznaczne przypisania. Grupowanie wydań świątecznych (1 listopada, 11 listopada, 31 grudnia, w Wigilię i Święta Bożego Narodzenia), też lepiej wypada przy krótszych n-gramach. Wynika to ze specyfiki tekstu, w którym liczba wystąpień n-gramów znacząco spada wraz ze wzrostem n.

Częstotliwość pojawiania się pojedynczych słów oraz bigramów w Faktach i w Wiadomościach są bardzo zbliżone, co sugeruje, że treści tych serwisów są podobne. Analizując bigramy z wybranymi słowami, lista kombinacji z danym słowem prawie zawsze jest dłuższa dla Faktów niż dla Wiadomości. Na tej podstawie można stwierdzić, że język w Faktach jest bardziej chaotyczny i różnorodny, a w Wiadomościach bardziej uporządkowany i formalny.

Przykładem różniącym styl Faktów i Wiadomości jest łączenie nazw instytucji z osobami. W Faktach częściej niż w Wiadomościach nazwy stanowisk są łączone od razu z nazwiskami np. „minister szyszko" i „prezydent duda", a w Wiadomościach częściej występują pełne nazwy instytucji, a politycy wymienieni są z imienia i nazwiska np. „prezydent andrzej duda" „minister środowisko jan szyszko".

Po przeanalizowaniu wydań Faktów i Wiadomości z okresu październik – grudzień 2017 z pomocą metod automatycznych/stylometrii można stwierdzić, że style obu serwisów różnią się. W języku jakim posługą się Fakty nazwy instytucji i stanowisk i partii mogą być skracane, a w języku Wiadomości ww. nazwy nie są skracane, ale podawane pełne nazwy. Jednak jeżeli chodzi o rodzaj poruszanych tematów oraz ich częstotliwość to wydają się być bardzo podobne w obu serwisach. Niniejsza analiza różni się od raportu KRRiT, a jej celem nie było wykazanie, że Fakty są krytyczne a Wiadomości stronnicze.

5 Literatura:

Allahyari M, Pouriyeh S, (2017) A Brief Survey of Text Mining: Classification, Clustering and Extraction Techniques. [online – dostęp 15.05.2018r.] https://arxiv.org/pdf/1707.02919.pdf,

Burrows J, (2002) 'Delta': A measure of stylistic difference and a guide to likely authorship. Literary and Linguistic Computing, 17(3):267–287,

Eder M, Kestemont M, Rybicki, J, (2016) Stylometry with R: a package for computational text analysis. R Journal 8(1): 107-121.
https://journal.r-project.org/archive/2016/RJ-2016-007/index.html,

Mrozowski M, Popadiak-Kuligowska T, Ekspertyza programów informacyjnych głównych wydań TVP1 Wiadomości, TVN Fakty, Polsat Wydarzenia z okresu 4.02.2016 r. do 11.02.2016 r. Raport końcowy [online - dostęp 30 kwietnia 2018]: http://www.krrit.gov.pl/Data/Files/_public/Portals/0/komunikaty/12.04.2016/krrit_eksperyza.pdf,

Raskar S S, Thakore D M, (2011) Text Mining and Clustering Analysis. IJCSNS International Journal of Computer Science and Network Security, VOL.11 No.6: 203-207,

Rykowski J, (2014) Metody i narzędzia rozpoznawania mowy w zastosowaniach niekomercyjnych, Napędy i Sterowanie – miesięcznik naukowo-techniczny, 116-123,

Schonlau M, Guenther N, (2016) Text Mining Using N-Grams. [online – dostęp 15.05.2018 r.] https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2759033,

Ziółko B, Ziółko M, (2011) Przetwarzanie mowy, Wydawnictwo AGH, 34-35.