Wykorzystanie metod *text miningowych* do oceny treści generujących zaangażowanie na portalach społecznościowych.

Using text mining methods to evaluate content that generates engagement on social media.

Kaczanowska Monika[(1)], Kielmer Agata[(1)],
[(1)] Wydział Elektronik i Technik Informacyjnych, Politechnika Warszawska,

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

Artykuł badawczy:

Słowa Kluczowe: n-gramy, korelacja słów, lematyzacja

Streszczenie

W pracy podjęto próbę analizy tekstów, z wykorzystaniem technik *text miningowych*, zapisanych w języku polskim, które są dostępne publicznie na portalu społecznościowym Facebook. Następnie na podstawie tej analizy dokonano określenia jakie treści generują większe zaangażowanie, a jakie mniejsze. Dodatkowo podjęto się sprawdzenia jakie znaczenie w analizie tekstu w języku polskim ma lemtyzacja, czyli analiza form podstawowych słów.

1. Wstęp

Przy szybkim wzroście możliwości technicznych gromadzenia danych ustrukturyzowanych i nieustrukturyzowanych, analiza zasobów tekstowych ma coraz większe znaczenie. Zbiory danych są coraz obszerniejsze, (a procesy ich analizy często nie są poddawane automatyzacji co powoduje, że analiza jest bardzo kosztowna i czasochłonna. Stanowi to motywację dla analizowania naturalnych zasobów tekstowych narzędziami związanymi z eksploracją danych (*text mining*). Jako pierwsza zdefiniowała „*text mining*" Marti Hearst poprzez „proces mający na celu odkrycie przez komputer nowych, poprzednio nieznanych informacji z zasobów tekstowych" (Hearst 1999).

Dane podlegające przetworzeniu często są nieustrukturyzowane. Najważniejszym celem tego procesu jest odnalezienie pewnych wcześniej nieznanych danych z niestrukturalnych treści dokumentów tekstowych zapisanych w języku naturalnym. Proces polega na statystycznej analizie tekstów pod kątem występowania określonych słów kluczowych, wiązek wyrazów itd. Metoda *text mining* służy do analizy dokumentów tekstowych, szeroko rozumianych, a dokładniej do wydobywania danych z tekstu i ich dalszej obróbki. Eksploracyjna analiza tekstów obejmuje m.in. wyszukiwanie informacji, kategoryzacji tekstów, modelowanie probabilistyczne, sporządzanie streszczeń, znajdowanie grup słów o przybliżonym znaczeniu, a także automatyczne rozpoznawanie złożonych wyrażeń (Hotho i in. 2005). Analiza ta wykorzystuje również metody statystyczne i uczenie maszynowe (Kao i Poteet 2007).

Zakres zastosowań analizy *text miningowej* jest bardzo szeroki. Poniżej wymieniono tylko niektóre obszary zastosowań eksploracyjnej analizy danych tekstowych:

- pozyskiwanie informacji z dokumentów (Fan i in. 2005);

- analiza danych biznesowych;
- tłumaczenie maszynowe;
- wyszukiwanie informacji;
- przetwarzanie informacji zawartych w hurtowniach dokumentów.

Text mining jest wykorzystywany do analizy danych biznesowych. Może być narzędziem, które pozwala skuteczniej identyfikować potrzeby klientów, ich przyzwyczajenia i oczekiwania (Larose 2006; Hearst 1999). Zgłębianie danych tekstowych pozwala na wykorzystanie danych o klientach, czy też transakcjach w celu pozyskania krytycznych informacji, które można często przekształcić w przewagę konkurencyjną. Przykładem wykorzystania prostej analizy *text minigowej* jest badanie poziomu zadowolenia klienta na podstawie wypełnionych kwestionariuszy oceny (szczególnie pola np. „dodatkowe uwagi"). Dużo bardziej zaawansowane analizy pozwalają wykrywać przestępstwa bankowe, a nawet śledzić potencjalne działania terrorystyczne.

W wyniku analizy *text mining* można:

- dostarczać teksty związane z danym obszarem zainteresowania,
- opisywać zawartość tekstu w sposób przydatny do dalszego przetwarzania (przetwarzanie języka naturalnego, modelowanie statystyczne itp.),
- interpretować i analizować informację wynikową (znajdując powiązania).

Język naturalny wykorzystywany w analizie *text mining*, w tym przypadku język polski, jest trudny do przetworzenie na język zrozumiały dla komputerów (Dzieciątko i Spinczyk 2016). Człowiek potrafi odróżniać i stosować lingwistyczne wzorce tekstowe i łatwo może pokonywać przeszkody, z którymi komputery sobie nie radzą, takie jak regionalizm, slang, formy gramatyczne czy też rozumienie kontekstowe (Gładysz 2016).

Problemy mogące wystąpić w analizie wynikające z badania materiałów tekstowych w języku polskim to:

- brak wystarczających zasobów słownikowych określających sentyment, których przypisane jest znaczenie słowa jako np.: pozytywne lub negatywne.
- język polski jest językiem fleksyjnym, w związku z tym występująca deklinacja słów utrudnia wyszukiwanie związków wyrazowych.
- występują znaki diakrytyczne, często pomijane przy tworzeniu tekstów, szczególnie tych nieformalnych, co również utrudnia klasyfikację.

2. Materiały i metody

Analizy w tej pracy zostały wykonane przy wykorzystaniu pakietu R[(1)]- niekomercyjnego oprogramowania wykorzystywanego do obliczeń statystycznych.

Analizie poddano wiadomości tekstowe z wykorzystaniem następujących metod:

a. N-gramy, czyli słowa najczęściej występują po sobie, w tej analizie skupiono się na bigramach.
b. Korelacja słów, czyli sprawdzenie jakie słowa współwystępują razem w całej wiadomości.

Teksty zostały poddane wstępnej obróbce (ang. preprocessing), podczas której dane tekstowe zapisane w różnorodnych formatach zostały zaimportowane do pojedynczego zbioru, który zapewnił łatwość późniejszego odczytywania. Proces ten został zrealizowany w kilku krokach:

a. Zamieniono wielkie litery na małe.
b. Wykonano lematyzację.

Następnie tekst został poddany rozbiorowi (ang. parsing):

[1] *https://www.r-project.org/*

a. Usunięto wszystkie znaki interpunkcyjne.
b. Usunięto wszystkie znaki będące poza zbiorem liter alfabetu od a do z oraz cyfr od 0 do 9
c. W trakcie analizy zostały usunięte również słowa, które nie niosą ze sobą istotnych treści - „stopwords" czyli rodzajniki, spójniki, przyimki i inne nieistotne semantycznie części mowy (Rajaraman i Ullman 2012). Przykłady takich słów to: a, aby, ach, acz, aczkolwiek, aj, albo, ale, ależ, ani, aż.

Analizie poddano wiadomości zamieszczone na portalu społecznościowym Facebook fundacji DKMS (Bazy Dawców Komórek Macierzystych Polska), której głównym celem jest zorganizowanie i zarządzanie Ośrodkiem Dawców Szpiku i krwiotwórczych komórek macierzystych krwi obwodowej[(2)]. Wiadomości ze strony zostały pobrane z wykorzystaniem programu R i pakietu Rfacebook[(3)]. Tabela **(Tab.1**) zawiera przykłady informacji pobranych przez ten pakiet, które zostały poddane dalszej analizie. W ciągu ostatnich dwóch lat pojawiło się 787 wiadomości, czyli mniej więcej jedna wiadomość dziennie. Średnia długość wiadomości to 42 słowa. W analizie porównano posty z mniejszą i większą liczbą komentarzy. Wiadomości częściej komentowane to pierwsze 300 wiadomości pod względem liczby komentarzy (powyżej 30 komentarzy), wiadomości rzadziej komentowane to analogicznie ostatnie 300 wiadomości (poniżej 13 komentarzy). Wiadomości nie wykazujące się ani dużym, ani małym zaangażowaniem nie przypisano do żadnej z tych grup, aby ułatwić wychwycenie różnicy między grupą częściej i rzadziej komentowanych wiadomości.

Tab.1 Przykłady komentarzy umieszczonych przez DKMS wraz ze statystykami

wiadomość	data	komentarze
Dziś kilka słów od Piotra, który pewnego dnia, niedługo po swojej rejestracji w bazie dawców szpiku, otrzymał wyjątkowy telefon.... Co wydarzyło się później? Zobaczcie sami! Rejestrujcie się w bazie potencjalnych dawców! Taka osoba, chora na białaczkę, czeka na Waszą pomoc, z nadzieją oczekuje na informacje o tym, że jest dla niego szansa na ŻYCIE, to nie jest dużo, to nic nie kosztuje, ani nic nie boli. Nic się nie traci. A osoba, która czeka, każdego dnia zmaga się z chorobą. Każdego dnia boi się, że ten właśnie dzień będzie jej ostatni. Każdego dnia rodzina tej osoby martwi się o to, żeby nie zdarzyła się żadna tragedia, lub żeby się nie pogorszyło, a właśnie Ty możesz tej osobie pomóc! Możesz dać komuś szansę na życie!	2017-04-28	45
3326 ZAREJESTROWANYCH POTENCJALNYCH DAWCÓW!???? Tą piękną liczbą kończymy kolejną, wiosenną edycję #HelpersGeneration. Działania rejestracyjne koordynowali Studenccy Liderzy aż na 46 uczelniach w Polsce. KOCHANI GRATULUJEMY TEGO IMPONUJĄCEGO WYNIKU! Bez Was nie dalibyśmy rady <3<3<3 MACIE MOC! ?????? A dla tych, którzy nie zdążyli się zarejestrować - możecie zrobić to tutaj: www.dkms.pl/zostan-dawca	2017-04-27	8

[2] *https://www.facebook.com/fundacja.dkms.polska/*

[3] *https://github.com/pablobarbera/Rfacebook*

2.1. Analiza n-gramów

N-gramy to analiza występujących kolejno po sobie słów (Silge i Robinson 2017). W tym przypadku wykorzystano pary słów, czyli bigramy. Analiza polega na wskazaniu, ile razy słowo pierwsze (słowo1) występuje przed słowem drugim (słowem2), a następnie zrobienie rankingu pojawiających się par słów (słowo1→ słowo2).

2.1.1. Analiza n-gramów dla częściej i rzadziej komentowanych wiadomości przed lematyzacją

Analizy wykonano w podgrupach, czyli dla wiadomości częściej i rzadziej komentowanych. Analizowany tekst bez lemantyzacji, czyli bez sprowadzenia do form podstawowych. Porównując listę biogramów **(Tab.2**) dla tych wiadomości możliwe jest określenie, które pary słów pojawiają się w jakiej grupie wiadomości. W pracy ograniczono się do około trzydziestu najczęściej występujących bigramów.

Najczęściej pojawiające się pary słów są wspólne dla obu podgrup. Jednak w każdej z grup są pary słów, które nie występują w drugiej grupie (przykłady zaznaczono w tabelach szarym tłem). Nie oznacza to, że zaznaczone pary słów nie występują w drugiej grupie wiadomości. Najprawdopodobniej pojawiają się one, ale poza listą trzydziestu najczęściej występujących par.

W grupie częściej komentowanych wiadomości pojawiają się bigramy, które nie występują w drugiej grupie, należące do tematów np.:

- pakiet: zamów→pakiet, pakiet→rejestracyjny,
- zostanie dawcą: zgody→dawca, zostań→potencjalnym itp.

W grupie rzadziej komentowanych wiadomościach pojawiają się pary słów, które nie występują w pierwszej, należące do tematów np.:

- rejestracja: akcji→rejetracji, rejestrować→potencjalnyh,
- rozliczenie pit: rozliczeń→pit, 1→podatku.

Tab.2 Trzydzieści najczęstszych bigramów występujących w częściej i rzadziej komentowanych wiadomościach przed lematyzacją*

słowo1	**słowo2**	**liczba wystąpień w częściej komentowanych wiadomościach**	**liczba wystąpień w rzadziej komentowanych wiadomościach**
potencjalnych	dawców	56	112
komórek	macierzystych	52	50
dawcą	szpiku	46	17
nowotwory	krwi	28	42
potencjalny	dawca	40	20
potencjalnym	dawcą	39	10
pakiet	rejestracyjny	37	
dawców	szpiku	27	36
dawca	szpiku	35	11
dawcy	szpiku	35	31
rejestracji	potencjalnych	15	32
fundacji	dkms	28	19
zostań	dawcą	24	10
zgodny	dawca	19	

komórki	macierzyste	14	18
dawców	komórek		18
nowotworami	krwi		18
zgodnego	dawcy	17	
zostań	potencjalnym	17	
bazie	dawców	16	
bazie	potencjalnych	16	
bezpłatny	program		14
dawstwa	szpiku		14
rozliczeń	pit		14
dni	dawcy	13	
możesz	masz	13	
dzień	dawcy	12	13
całej	polsce		13
nowych	potencjalnych		13
niespokrewnionego	dawcy	12	
zamów	pakiet	12	
edycji	projektu		12
genetycznego	bliźniaka	11	
komuś	szansę	11	
podaruj	szansę	11	
przeszczepienie	szpiku	11	
trzymamy	kciuki	11	10
razem	możemy	10	11
1	podatku		11
akcji	rejestracji		11
naszej	fundacji		11
rejestrować	potencjalnych		11
możesz	pomóc	10	
najbliższy	weekend	10	
krwiotwórczych	komórek		10
serdecznie	dziękujemy		10

*wartości zaznaczone na szaro różnią grupy wiadomości

2.1.2. Analiza n-gramów dla częściej i rzadziej komentowanych wiadomości po lematyzacji

Tak jak w analizie wiadomości bez wykonania lematyzacji, w analizie tekstów po sprowadzeniu do form podstawowych część bigramów jest wspólna dla obu grup. Główne tematy wydają się też pozostawać te same (Błąd! Nie można odnaleźć źródła odwołania.).

W grupie częściej komentowanych wiadomości pojawiają się bigramy, które nie występują w drugiej grupie, należą do tematów np.:

- pakiet: zamówić→pakiet, pakiet→rejestracyjny,
- bliźniak: swój→bliźniak, bliźniak→genetyczny.

W grupie rzadziej komentowanych wiadomości pojawiają się pary słów, które nie występują w pierwszej grupie, należą do tematów np.:

- studenci: studencki→lider, wwwdkmspl→student,
- dawstwo: dawstwo→szpiku, idea→dawstwa.

Tab.3 Trzydzieści najczęstszych bigramów w częściej i rzadziej komentowanych wiadomościach po lematyzacji*

słowo1	**słowo2**	**liczba wystąpień w częściej komentowanych wiadomościach**	**liczba wystąpień w rzadziej komentowanych wiadomościach**
potencjalny	dawca	126	138
dawca	szpik	137	99
wwwdkmspl	zostać	87	30
zostać	dawca	84	38
nowotwór	krew	40	69
komórka	macierzysty	51	64
pakiet	rejestracyjny	46	
zgodny	dawca	41	
rejestracja	potencjalny	18	39
bliźniak	genetyczny	30	
dzień	dawca	30	26
akcja	rejestracja	17	30
zostać	potencjalny	29	
podarować	szansa	23	
dawca	komórka	15	23
móc	pomóc	22	12
komórki	macierzysty	21	
baza	dawca	20	14
baza	potencjalny	20	20
przeszczepić	szpik	20	
dawstwo	szpik		20
dobry	wiadomość	19	
trzymać	kciuk	17	19
ostry	białaczka	18	
szpik	wwwdkmspl	18	
spokrewnić	dawca	16	12
wwwdkmspl	dawca	16	
krwiotwórczy	komórka	12	15
rejestrować	potencjalny		15
studencki	lider		15
światowy	dzień		15
dawca	spokrewnić	14	
móc	mieć	14	
swój	bliźniak	14	
bezpłatny	program		14
idea	dawstwo		14
rozliczenie	pit		14
wwwdkmspl	student		14
dzień	walka		13
edycja	projekt		13
nowy	potencjalny		13
strona	wwwdkmspl		13

bitly	dawca	12	
facebook	dom	12	
zamówić	pakiet	12	
przeszczepić	komórka		12
ratować	życie		12

*wartości zaznaczone na szaro różnią grupy wiadomości

2.2. Korelacja słów

Metodą stosowaną w *text mining* jest analiza korelacji słów (Silge i Robinson 2017; Pasztyła 2005). Analiza wskazuje jakie pary słów współwystępują najczęściej w wiadomościach. W odróżnieniu od analizy bigramów w tym przypadku analizowane są wystąpienia w całej wiadomości, niezależnie od kolejności słów.

Analiza korelacji słów polega na identyfikacji słów, które najczęściej współwystępują razem. Problem może sprawiać to, że słowa które najczęściej występują razem występują bardzo często również pojedynczo. Z tego powodu badana jest korelacja między słowami, która wskazuje, jak często pojawiają się słowa razem w stosunku do tego, jak często pojawiły się osobno.

Tab.4 Poszczególne kombinacje wystąpienia słowa X oraz słowa Y

	Występuje słowo Y	Nie występuje słowo Y	Suma wystąpień
Występuje słowo X	n_{11}	n_{10}	$n_{\cdot 1}$
Nie występuje słowo X	n_{01}	n_{00}	$n_{\cdot 0}$
Suma wystąpień	$n_{\cdot 1}$	$n_{\cdot 0}$	n

Na przykład n_{11} reprezentuje liczbę dokumentów, w których pojawiają się zarówno słowo X, jak i słowo Y, n_{00} to liczba dokumentów, w którym żadne z nich się nie pojawia, oraz n_{10} oraz n_{01} określają przypadki, w których jedno pojawia się bez drugiego. Współczynnik korelacji ϕ wynosi:

$$\phi = \frac{n_{11}n_{00} - n_{10}n_{01}}{\sqrt{n_{\cdot 1}n_{\cdot 0}n_{\cdot 0}n_{\cdot 1}}} \quad (1)$$

2.2.1. Korelacja słów dla częściej i rzadziej komentowanych wiadomości przed lematyzacją

Podobnie jak w analizie bigramów, porównując listy słów współwystępujących możliwe jest ocenienie co różni grupy wiadomości, te częściej komentowane od tych rzadziej komentowanych (**Tab.5**).

W tekście oryginalnym (bez lematyzacji), w grupie częściej komentowanych wiadomości pojawiają się pary, które nie występują w drugiej grupie, należą do tematów np.:

- zwroty bezpośrednie ze słowami: podaj, prosimy, jesteś, udostępnij,
- pakiet: pakiet+rejestracyjny, pakiet+go.

W grupie rzadziej komentowanych wiadomości pojawiają się pary słów, które nie występują w pierwszej grupie, należą do tematów np.:

- helpersgeneration ze słowami: projektu, edycji, całej,
- rozliczenie pit: 1+podatku, online+pit.

Tab.5 Trzydzieści najbardziej skorelowanych słów w częściej i rzadziej komentowanych wiadomościach przed lematyzacją*

słowo1	słowo2	korelacja w częściej komentowanych wiadomościach	korelacja w rzadziej komentowanych wiadomościach
komórki	macierzyste		100%
komórek	macierzystych	97%	97%
dawców	potencjalnych	78%	88%
pakiet	rejestracyjny	85%	
1	pit		83%
dkms	fundacji	78%	39%
potencjalny	dawca		76%
dawca	potencjalny	72%	
dalej	podaj	70%	
krwi	nowotwory	67%	65%
edycji	projektu		67%
polsce	całej		67%
chorych	nowotwory		63%
zostań	dawcą	63%	
potencjalnym	dawcą	59%	
okazji	dnia		58%
prosimy	podaj	55%	
jesteś	podaj	55%	
nowotwory	chorych	52%	
dawca	zgodny	52%	
szansę	życie	52%	
prosimy	jesteś	51%	
jesteś	bazie	50%	
prosimy	przeszczepienie	49%	
udostępnij	potencjalny	48%	
projektu	helpersgeneration		48%
razem	możemy	47%	
pakiet	go	46%	
dawcą	zostań		46%
dawcy	niespokrewnionego	45%	
nowotworami	krwi		45%
podaj	pomóż	45%	
dawców	rejestracji	45%	
dawców	bazie	45%	
udostępnij	go	45%	
jesteś	przeszczepienie	44%	
krwi	chorych		44%
rejestracji	potencjalnych	44%	34%
zgodnego	dawcy	43%	
prosimy	szansą	43%	
prosimy	udostępnij	42%	
rejestracji	dawców		41%

nowotworami	dnia		41%
dzień	okazji		39%
edycji	helpersgeneration		38%
online	pit		37%
całej	helpersgeneration		37%
online	1		34%
możesz	pit		34%
online	zarejestrować		34%
nowotworami	okazji		31%
dziękujemy	3		31%
rejestracji	akcji		31%
rejestracji	akcję		29%
dzień	dnia		29%

*wartości zaznaczone na szaro różnią grupy wiadomości

2.2.2. Korelacja słów dla częściej i rzadziej komentowanych wiadomości po lematyzacji

W tekście, w którym słowa zostały sprowadzone do form podstowowych (po lematyzacji), w grupie częściej komentowanych wiadomości **(Tab.6**) pojawiają się pary, które nie występują w drugiej grupie, należą do tematów np.:

- zwroty z prośbami: podać, udostępnić,
- pakiet: pakiet+rejestracyjny, pakiet+odesłać.

W grupie rzadziej komentowanych wiadomości pojawiają się pary słów, które nie występują w pierwszej grupie, np.:

- akcja studencka ze słowami: edycji, lider, projekt, helpersgeneration, uczelnia,
- rozliczenie pit: 1+pit, przekazać+pit, 1+przekazać.

Tab.6 Trzydzieści najbardziej skorelowanych słów w częściej i rzadziej komentowanych wiadomościach po lematyzacji *

słowo1	**słowo2**	**korelacja w częściej komentowanych wiadomościach**	**korelacja w rzadziej komentowanych wiadomościach**
trzymać	kciuk	100%	89%
komórka	macierzysty	83%	98%
bliźniak	genetyczny	90%	
nowotwór	krew	74%	89%
pakiet	rejestracyjny	84%	
lider	studencki		77%
1	pit		77%
przekazać	pit		77%
życie	szansa	70%	
lider	edycja		69%
1	przekazać		68%
charytatywny	allegro		68%
studencki	edycja		67%
wypełnić	odesłać	67%	
nowotwór	chory	53%	67%

lider	projekt		67%
uczelnia	helpersgeneration		66%
światowy	walka		65%
lider	student		64%
projekt	student		64%
uczelnia	studencki		64%
potencjalny	dawca	49%	63%
czerwony	dajczerwone		63%
choroba	przeszczepić		62%
zostać	wwwdkmspl	62%	
studencki	projekt		61%
bransoletka	lilla		61%
lider	uczelnia		61%
przeszczepić	spokrewnić	60%	
edycja	projekt		59%
krew	chory		59%
lider	helpersgeneration		58%
pakiet	odesłać	58%	
dawstwo	idea		57%
edycja	student		56%
marzenie	diagnoza	55%	
prosić	udostępnić	55%	
diagnoza	choroba	55%	
studencki	student		54%
prosić	podać	53%	
rodzina	dom	53%	
uczelnia	student		53%
dawca	szpik	52%	
wiadomość	trzymać	51%	
wiadomość	kciuk	51%	
macierzysty	komórki	51%	
rejestracyjny	odesłać	50%	
udostępnić	choroba	50%	
rodzina	przeszczepić	49%	
prosić	marzenie	48%	
szansa	podarować	48%	
udostępnić	białaczka	48%	
pakiet	wypełnić	48%	
udostępnić	podać	48%	
rodzina	spokrewnić	47%	

*wartości zaznaczone na szaro różnią grupy wiadomości

3. Wyniki

Mimo częściowej automatyzacji analizy, przy wyciąganiu wniosków konieczny był duży udział oceny własnej badacza. Porównanie czym różnią się wyniki analiz i przypisanie par słów do grup było wykonane na podstawie subiektywnej oceny.

Analizy n-gramów i korelacja słów dała spójne wyniki, zarówno w przypadku analizy tekstu oryginalnego i podanego lematyzacji, czyli sprowadzenia do form podstawowych. Zebrane, najważniejsze wyniki przedstawiono poniżej w **Tab.7**.

Większym zaangażowaniem cieszyły się wiadomości z tematami o „pakiecie rejestracyjnym", „genetycznym bliźniaku" oraz prośby „podaj", „udostępnij". Wiadomości zawierające informacje o możliwości „przekazania 1% podatku" cieszyły się mniejszym zaangażowaniem, czyli były rzadziej komentowane. Tak samo wiadomości zawierające odniesienia do „studenckich projektów" oraz „akcji helpersgeneration".

Tab. 7 Podsumowanie wyników

metoda	**typ analizy**	**wiadomości częściej komentowane**	**wiadomości rzadziej komentowane**
n-gramy	bez lematyzacji	- pakiet rejestracyjny - zostanie zgodnym dawcą	- akcja rejestracji potencjalnych - rozliczenie 1 % podatku
	po lematyzacji	- pakiet rejestracyjny - swój bliźniak genetyczny	- studencki lider - idea dawstwa szpiku
korelacje słów	bez lematyzacji	- podaj, prosimy, udostępnij - pakiet rejestracyjny	- helpersgeneration - 1% podatku, pit
	po lematyzacji	- podać, udostępnij - pakiet rejestracyjny	- studencki projekt, helpersgeneration - przekazać 1%, pit

4. Dyskusja i wnioski

Głównym zadaniem metod eksploracji tekstu jest wyłuskiwanie istotnych danych, a następnie użycie ich do sporządzania prognoz i podejmowania dalszych efektywniejszych decyzji.

W wyniku powyższej analizy wysunięto następujące wnioski:

- narzędzia *text miningowe* mogą wspierać proces analizy tekstu co pozwala na zaoszczędzenie czasu oraz zasobów finansowych, które musiałyby być przeznaczone m.in na przeczytanie i eksplorowanie przez człowieka ogromnego repozytorium dokumentów tekstowych, nawet w analizie krótkich wiadomości na portalu Facebook.
- Wyniki nadal należy analizować i korygować przez człowieka, ponieważ czynnik ludzki jest nadal istotny. Maszyny wciąż nie zapewniają pełnej skuteczności identycznej z umiejętnościami ludzkimi: rozumieniem i przetwarzaniem komunikatów.
- Wykazano, że jest możliwe określenie, które słowa lub związki wyrazowe generują większe zaangażowanie. W analizowanym przypadku są to: „pakiet rejestracyjny", „bliźniak genetyczny", „podaj", „udostępnij", które wpływają na zwiększoną ilość komentarzy – w porównaniu do wiadomości zawierających informacje o „1% podatku", „akcji studenckiej" i „helpersgeneration".
- W przypadku analizy spójnych tekstów o podobnej treści, tak jak wiadomości umieszczanie przez fundację na swoim funpage na portalu Facebook, wyniki analiz wykonanych na tekstach ze sprowadzeniem słów do form podstawowych (z lematyzacją) i tekstów oryginalnych (bez lematyzacji) wskazały tematy podobne, wspólne dla grup wiadomości częściej i rzadziej komentowanych. O ile lematyzacja wydaje się lepiej określać tematy, o tyle brak lematyzacji może wskazać jak dokładnie ma wyglądać tekst. Analiza form podstawowych wskazała tematy (wymienione w punkcie poprzednim).

Analiza tekstu oryginalnego wskaże dokładne zwroty, które wydają się bardziej angażować, czyli te bezpośrednie: „podaj”, „prosimy”, „udostępnij”.

5. Literatura

Dzieciątko M, Spinczyk D (2016), Text Mining: metody, narzędzia i zastosowania, PWN Wydanie: 1

Fan W, Wallace L, Rich S (2005), Tapping into the Power of Text Mining: Communications of ACM: 4-5

Feinerer I, Hornik K, Meyer D (2008), Text Mining Infrastructure in R: Journal of Statistical Software: 23-27
https://www.jstatsoft.org/article/view/v025i05

Gładysz A (2016), Przegląd zastosowań analizy text miningowej: Autobusy: technika, eksploracja, systemy transportowe: 1742-1743

Hearst M (1999), Untangling Text Data Mining: Proceedings of ACL’99: the 37th Annual Meeting of the Association for Computational Linguistics, University of Maryland, 3-5

Hotho A, Nurnberger A, Pass G (2005), A Brief Survey of Text Mining,:.9-10
https://www.kde.cs.uni-kassel.de/hotho/pub/2005/hotho05TextMining.pdf

Kao A, Poteet S (2007), Natural Language Processing and Text Mining: Springer, 1-3
http://129.219.222.66/publish/pdf/natural_language_processing_and_text_mining.pdf

Larose D (2006), Odkrywanie wiedzy z danych, Wydawnictwo Naukowe PWN, Warszawa: 2-3

Pasztyła A (2005), Przykład badania wzorców zachowań klientów za pomocą analizy koszykowej: Data mining: poznaj siebie i swoich klientów: 55-56
https://media.statsoft.pl/_old_dnn/downloads/przyklad_badania_wzorcow_zachowan.pdf

Silge J, Robinson D (2017), Text Mining with R, a Tidy Approach, https://www.tidytextmining.com/