УДК 324(477)]:33
DOI: 10.31471/2304-7410-2019-4(56)-109-115

ЗАСТОСУВАННЯ СТАТИСТИЧНИХ МЕТОДІВ ДЛЯ АНАЛІЗУ ТЕКСТІВ ПЕРЕДВИБОРЧИХ ПРОГРАМ

М. М. Осипчук[1], І. М. Гураль[2], Л. Р. Смоловик[2]
[1]Прикарпатський національний університет імені Василя Стефаника;
76000, м. Івано-Франківськ, вул. Шевченка, 57;
e-mail: myosyp@gmail.com
[2]Івано-Франківський національний технічний університет нафти і газу;
76019, м. Івано-Франківськ, вул. Карпатська, 15;
e-mail: math@nung.edu.ua

Сучасна статистика має в своєму арсеналі методи формалізації (вимірювання) об'єктів найрізноманітнішої природи. Зокрема це стосується текстів, так званої, природної мови. В статті за допомогою статистичних методів проаналізовано тексти передвиборчих програм кандидатів на пост Президента України на виборах 2019 року. Застосовуючи метод багатовимірного шкалювання утворено набір даних, який складається з двох числових характеристик, що описують особливості розглянутих текстів програм. За допомогою кореляційного аналізу встановлено зв'язок між текстами передвиборчих програм кандидатів та офіційними результатами першого туру виборів, а також результатами загальнонаціонального екзит-полу. Застосовуючи процедури кластеризації методом Уорда, виділено чотири групи кандидатів на пост Президента України. Встановлено особливості текстів програм побудованих груп і створено хмарки ключових слів для швидкого сприйняття найбільш вживаних слів і їх розподілу за популярністю відносно один одного. Підготовку даних та всі статистичні обчислення здійснено з допомогою середовища статистичних розрахунків R.

Ключові слова: *передвиборча програма, метод багатовимірного шкалювання, кореляційний аналіз, кластерний аналіз, хмарка слів.*

Вступ

Сучасний етап розвитку людства характеризується бурхливим зростанням кількості інформації. Повноцінне і ефективне забезпечення суспільства новітньою інформацією є необхідною передумовою підвищення ефективності наукових досліджень. Однією з найбільш поширених форм зберігання інформації є інформація, представлена у вигляді текстових ресурсів на мові певної країни, тому аналіз текстів є одним з найважливіших напрямків досліджень. Аналіз текстів природної мови застосовується до широкого кола урядових, дослідницьких та бізнес-потреб.

В даний час великий інтерес у представників різних сфер діяльності викликають політичні події країни. Результати політичних виборів

мають прямий вплив на майбутнє країни і населення. Мотиви голосування визначаються багатьма факторами. Хоча в середовищі політиків існує думка, що виборці голосують не за програми і платформи, а за особистості, саме передвиборча програма є концентрованим виразом цілей, завдань і намірів кандидатів, політичних партій, виборчих блоків. Традиційні методи аналізу позицій, виражених в політичних текстах, ґрунтуються на застосуванні варіацій контент-аналізу, що вимагає експертної оцінки.

В даній статті з допомогою методів та інструментів сучасної статистики проаналізовано тексти передвиборчих програм кандидатів на пост Президента України на виборах 2019 року. Мета даної статті – вияснити чи існує зв'язок між текстами передвиборчих програм кандидатів та результатами виборів, а також, класифікувати ці програми та встановити особливості побудованих класів.

Методи та інструменти дослідження

Сучасна статистика має в своєму арсеналі методи формалізації (вимірювання) об'єктів найрізноманітнішої природи. Зокрема це стосується текстів, так званої, природної (нештучної) мови. Ми застосовуємо кількісні та сентиментальні характеристики таких текстів. Кількісні характеристики описують частоти, з якими кожне значиме слово зустрічається в тексті. Сентиментальні характеристики будуються з допомогою, так званого сентиментного аналізу (sentiment analysis). Розглядаємо тексти передвиборчих програм кандидатів на пост Президента України на виборах 2019 року.

В першу чергу необхідно підготувати тексти до аналізу. А саме, видаляємо всі незначимі (займенники, сполучники і т.д.), формальні (заголовки, підписи, посилання і т.д.) слова, видаляємо числа, знаки пунктуації, зайві пропуски і тому подібне. З одержаного набору текстів вибираємо всі слова і, спочатку підраховуємо частоти, з якими кожне слово зустрічається в кожному із текстів. Далі на основі сентиментальних категорій мови визначаємо частоти, з якими зустрічаються слова кожної категорії в цих текстах. Застосовуючи метод багатовимірного шкалювання визначаємо по одній найбільш вагомій характеристиці серед кількісних та сентиментальних характеристик. Таким чином, утворюється набір даних, який складається з двох числових характеристик (в даному дослідженні count1 та sentiment1), що описують особливості розглянутих текстів.

Маючи на меті вияснити існування зв'язку між текстами передвиборчих програм кандидатів та результатами виборів, а також, класифікувати ці програми та встановити особливості побудованих класів, застосовуємо методи кореляційного та кластерного аналізів. В кореляційному аналізі використовуємо ранговий коефіцієнт кореляції Спірмена, що продиктовано відносно невеликою кількістю спостережень (кандидатів) та тим, що цей коефіцієнт є індикатором тісноти зв'язку між характеристиками, який задається монотонною (не обов'язково лінійною)

функцією. Критерієм наявності зв'язку між характеристиками є значимість відповідного коефіцієнта кореляції. Для визначення значимості вибираємо стандартний граничний рівень значущості, рівний 0,05. В процедурах кластерного аналізу вибрано метод Уорда, який зазвичай приводить до краще структурованих кластерів.

Підготовку даних та всі статистичні обчислення здійснено з допомогою середовища статистичних розрахунків R (див. [1]). Використано також графічні можливості цього середовища. Сентиментні категорії слів визначались з допомогою морфологічного визначника для української мови UGTag [2].

Для аналізу використані наступні характеристики:

- характеристики передвиборчих програм (count1, sentiment1);
- офіційні результати першого туру виборів (Res);
- результати загальнонаціонального екзит-полу:
  - по Україні (UA);
  - на заході України (West);
  - в центрі України (Center);
  - на півдні України (Sud);
  - на сході України (Ost).

Кореляційний аналіз

В наступній таблиці наведені показники значимості коефіцієнтів кореляції в кожній парі характеристик.

	sentiment1	Res	UA	West	Center	Sud	Ost
count1	0.00396	0.00256	0.00552	0.02449	0.00972	0.00336	0.06752
sentiment1		0.03180	0.05231	0.10310	0.06609	0.20479	0.43947
Res			0.00000	0.00000	0.00000	0.00000	0.00000
UA				0.00000	0.00000	0.00000	0.00000
West					0.00000	0.00000	0.00000
Center						0.00000	0.00000
Sud							0.00000

З неї видно, що можна стверджувати про наявність тісного зв'язку між офіційними результатами першого туру голосування та результатами екзит-полу (всі значення менші 0,00001), а також, між кількісною характеристикою (count1) передвиборчих програм та всіма іншими розглянутими характеристиками, крім, можливо, результатів екзит-полу на сході України. Сентиментальна характеристика (sentiment1) передвиборчих програм пов'язана з офіційними результатами голосування та, можливо, з результатами загальнодержавного екзит-полу і екзит-полу в центрі України.

Значення коефіцієнтів кореляції Спірмена наведені в наступній таблиці.

	sentiment1	Res	UA	West	Center	Sud	Ost
count1	-0.42889	-0.46959	-0.43607	-0.35973	-0.40908	-0.45813	-0.29578
sentiment1		0.34439	0.31306	0.26494	0.29726	0.20758	0.12743
Res			0.93741	0.86868	0.91724	0.87624	0.74322
UA				0.87679	0.91345	0.87235	0.78580
West					0.78221	0.81217	0.68411
Center						0.81505	0.74401
Sud							0.71702

Можемо констатувати, що офіційні результати першого туру виборів досить сильно позитивно корелюють (коефіцієнти кореляції додатні і близькі до 0,9) з результатами екзит-полів, крім результатів екзит-полу на сході України, де оцінка коефіцієнта кореляції приблизно рівна 0,74. Крім того, зауважимо, що знаки коефіцієнтів кореляцій характеристик текстів передвиборчих програм між собою та з результатами голосування та екзит-полів не слід вважати інформативними.

Кластерний аналіз

Результатом застосування процедур кластеризації за характеристиками передвиборчих програм є наступна дендрограма.

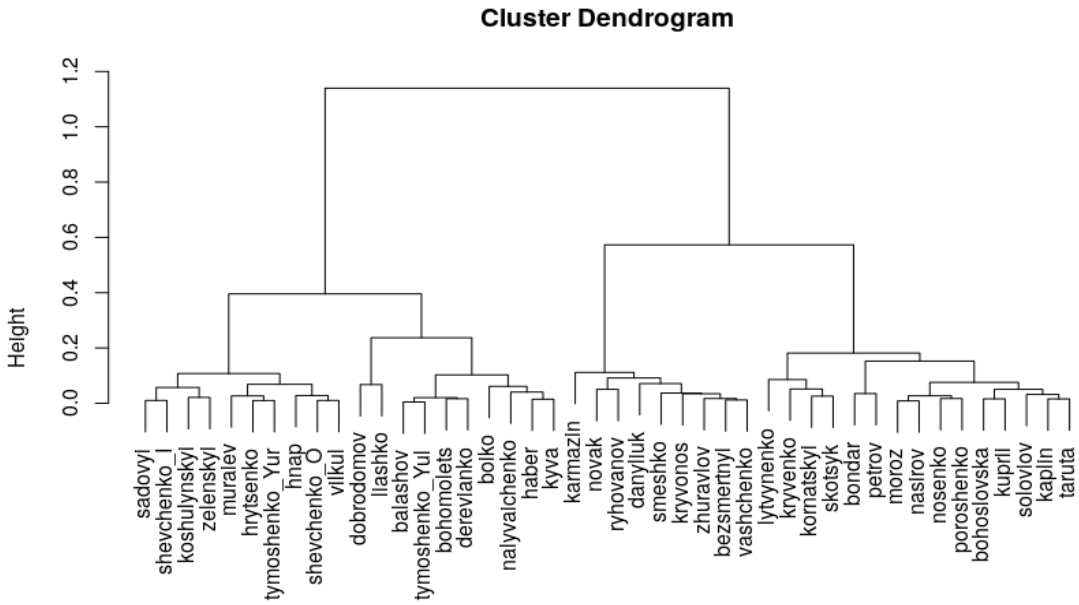

Рис. 1. Результати кластерного аналізу

З неї видно, що можна розглядати чотири групи кандидатів на пост Президента. На наступному рисунку зображені точки в системі координат count1- sentiment1, що представляють передвиборчі програми. Різними маркерами виділені програми кандидатів з різних кластерів. Якщо нумерувати їх в порядку слідування кластерів на дендрограмі, то круги відповідають першому кластеру, квадрати – другому, ромби – третьому і трикутники – четвертому.

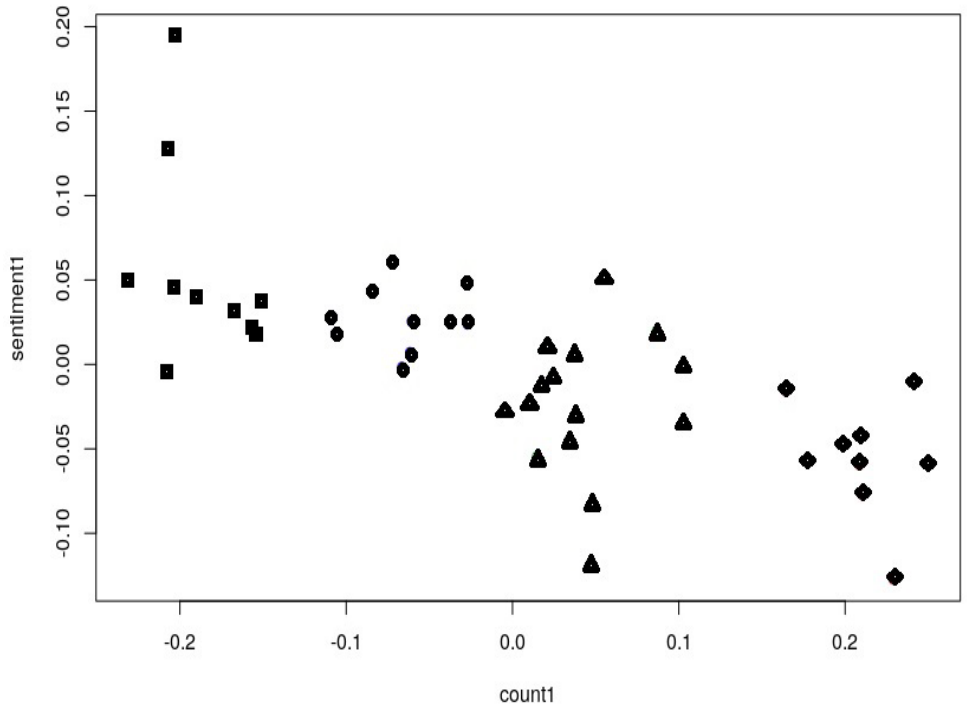

Рис. 2. Діаграма розсіювання передвиборчих програм

Цікаво вияснити, які саме ключові слова передвиборчих програм кандидатів з одержаних кластерів. Пакет word cloud із середовища R дозволяє побудувати хмарки ключових слів (див. рис. 3 – рис. 6), які є зручними для швидкого сприйняття найбільш вживаних слів і їх розподілу за популярністю відносно один одного. Більший та насиченіший шрифт слова означає, що воно частіше зустрічається в тексті (чи, в нашому випадку, текстах). Далі наводяться хмарки ключових слів в програмах кандидатів від першого до четвертого кластерів по порядку. Зразу видно, що слово «громадяни» найчастіше зустрічається в програмах кандидатів з першого кластеру і присутнє серед ключових слів в програмах кандидатів з другого кластеру. Слово «держава» найчастіше зустрічається в програмах кандидатів з 2-го, 3-го та 4-го кластерів. Слово «економіка» є серед ключових тільки в програмах кандидатів з третього кластеру і т.д.

Рис. 3. Хмарка слів в програмах кандидатів з кластеру 1

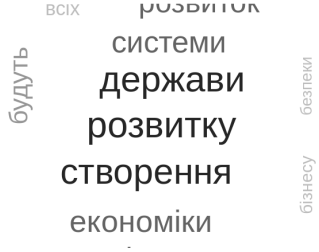

Рис. 4. Хмарка слів в програмах кандидатів з кластеру 2

Рис. 5. Хмарка слів в програмах кандидатів з кластеру 3

Рис. 6. Хмарка слів в програмах кандидатів з кластеру 4

Змістовне тлумачення одержаних хмарок слів залишаємо для зацікавленого читача.

Висновки

За результатами кореляційного аналізу можна стверджувати про наявність зв'язку між кількісною характеристикою (count1) передвиборчих програм та всіма іншими розглянутими характеристиками, крім, можливо, результатів екзит-полу на сході України. Сентиментальна характеристика (sentiment1) передвиборчих програм пов'язана з офіційними результатами голосування та, можливо, з результатами загальнодержавного екзит-полу і екзит-полу в центрі України.

За результатами процедур кластеризації щодо характеристик передвиборчих програм можна розглядати чотири групи кандидатів на пост Президента України в 2019 році.

Література

1. R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

2. UGTag – a morphological tagger for Ukrainian language. – Режим доступу: http://www.domeczek.pl/~polukr/parcor/.

Стаття надійшла до редакційної колегії 20.09.2019 р.
Рекомендовано до друку д.т.н., професором ***Мойсишиним В.М.***
д. політ. н., професором ***Монолатієм І.С.***

APPLYING STATISTICAL METHODS FOR ANALYZING TEXTS OF THE PRESIDENTIAL ELECTION PROGRAMMES

M. M. Osypchuk[1], I. M. Hural[2], L. R. Smolovyk[2]
[1]*Vasyl Stefanyk Precarpathian National University;*
76000, Ivano-Frankivsk, 57 Shevchenka Street;
e-mail: myosyp@gmail.com
[2]*Ivano-Frankivsk National Technical University of Oil and Gas;*
76019, Ivano-Frankivsk, 15 Karpatska Sreet;
e-mail: math@nung.edu.ua

Modern statistics is equipped with the methods of formalization (measurement) of the objects of different nature. This concerns in particular texts of the so called natural language. This article provides analysis (conducted by means of statistical methods) of the election programmes' texts of the candidates for Ukraine's Presidency in the 2019 election. With the method of multidimensional scaling, the data set was created that consists of two numerical characteristics that describe the peculiarities of the reviewed programmes' texts. With the correlation analysis, the correlation was established between the texts of the candidates' election programmes and the official results of the first round of the election, as well as the results of the nationwide exit poll. By applying the Ward's method cluster analysis, the four groups of the candidates for Ukraine's Presidency were outlined. Also, the peculiarities of the groups' programmes texts were identified, as well as the key words clouds were created for quick apprehension of the most frequently used words and their distribution according to popularity. Data preparation and all statistical calculations were performed with the help of the statistical calculation environment R.

Key words: *election programme, method of multidimensional scaling, correlation analysis, cluster analysis, word cloud.*