Застосування математико-статистичних методів аналізу індикативних ознак для верифікації віку учасників веб-спільнот

Галина Білущак[1], Соломія Федушко[2]

1. Кафедра вищої математики,
2. Кафедра соціальних комунікацій та інформаційної діяльності, Національний університет "Львівська політехніка", УКРАЇНА, м. Львів, вул. С. Бандери, 12,
E-mail: gbilushak@gmail.com[1], felomia@gmail.com[2]

Abstract – This paper considers the current problem of investigation of verification method of web-members' content. The set of age indicative characteristics of web-community members is formed for web-members verification. Mathematical and statistical methods in the learning sample of web-members are presented.

Ключові слова – верифікація, соціально-демографічна характеристика, вік, індикатор, індикативна ознака, математико-статистичний метод.

I. Математико-статистичний аналіз індикативних ознак

Збільшення темпів інформаційної діяльності у глобальному інформаційному просторі та наявні сучасні проблеми, які виникли на цьому ґрунті, потребують нових підходів до розроблення методів перевірки достовірності персональних даних учасників, зокрема віку учасника веб-спільнот, та інформаційного наповнення (ІН) веб-спільнот. Особливо важливим є розроблення методів перевірки коректності вказання віку користувачами соціальних комунікацій. Методами статистичних досліджень здійснено класифікацію учасників веб-форумів на основі індикативних ознак (ІО), які отримані в результаті опрацювання створеного ними ІН. Для вікової диференціації веб-учасників експертами сформовано набір вікових індикативних ознак на основі досліджень вчених та аналізу ІН веб-форумів. Набір лінгво-комунікативних індикаторів віку веб-учасника описуємо як:

$$\mathrm{LCI}\left(\mathrm{IO_i}\right) = \left(\mathrm{LCI_j}\left(\mathrm{IO_i}\right)\right)_{j=1}^{N_i^{LCI(IO)}} \tag{1}$$

Вектор маркерів визначає ІО, що в свою чергу визначає лінгво-комунікативний індикатор віку. Ця залежність задається формулою 2:

$$\mathrm{IO} = \left(\mathrm{Marker_j}\left(\mathrm{IO_i}\right)\right)_{j=1}^{N_i^{Marker}} \tag{2}$$

Порівняння основних числових характеристик розглянутих ІО у вікових групах веб-учасників (дорослі особи та підлітки) вказують на наявність статистично значущих відмінностей у більшості цих ІО. Результатом кластерного аналізу є поділ учасників веб-спільнот на два кластери: Кластер1 (41 учасник) і Кластер 2 (39 учасники). Проведено

аналіз некоректності кластеризації та проаналізовано результати кластеризації. Порівняння досліджуваних ІО для отриманих кластерів вказує на істотну відмінність середніх значень таких ІО, як Молодіжний сленг, Спрощена транслітерація, Заміна літер неалфавітними знаками, Заміна слів на основі звукової подібності. Для Кластеру 2 характерне більше середнє значення в порівнянні з Кластером 1. Оскільки, значення статистики Лямбда Уілкса 0,05255 (близьке до нуля) і значення критерію Фішера $F(23,56) = 43,901$ ($p<0,0000$) свідчать про хорошу дискримінацію, тобто класифікація проведена коректно. Також, матриця класифікації показує, що всі об'єкти класифіковані вірно.

Відповідно до стандартизованих коефіцієнтів, найвагоміший вклад в дискримінантну функцію мають такі ІО: Олбанська мова, Заміна літер неалфавітними знаками, Молодіжний сленг, Спрощена транслітерація, Надмірна кількість знаків пунктуації та спецсимволів, Фамільярна лексика. Також, значення коефіцієнтів вказують на те, що для поділу на кластери найважливішими статистично значущими є такі ІО: Олбанська мова, Заміна літер неалфавітними знаками, Молодіжний сленг, Спрощена транслітерація.

З матриці факторної структури можна зробити висновок, що структура вихідних даних в основному обумовлена такими ІО, як Молодіжний сленг, Заміна літер неалфавітними знаками, Найменування атрибутів молодіжної моди, Спрощена транслітерація.

За результатами факторного аналізу ключовими лінгво-комунікативними індикаторами для визначення віку веб-учасника є: підлітка – Сленгова варіація, Некодифіковані одиниці та невербальні засоби, Деформалізація, дорослої особи – Текстова економія. Порівняння основних числових характеристик розглянутих ІО вказують на відсутність істотних відмінностей лише в середніх значеннях таких ознаках, як Графічні смайли, Заміна літер цифрами, Складання літер. Після класифікації учасників веб-форумів у вибірку включено два контрольні спостереження (1 підліток і 1 дорослий) і проведено повторну класифікацію. При цьому задавалась однакова апріорна ймовірність ($p=0,5$) належності до кожного із кластерів обом суб'єктам. З апостеріорною ймовірністю $p=1$ обидва суб'єкти класифіковані вірно.

Висновок

Цінність математико-статистичних методів верифікації даних учасників веб-спільнот полягає у налагодження механізмів колаборативного опрацювання текстів в глобальному інформаційному просторі, механізмів керування та адміністрування віртуальними спільнотами та формуванні нових підходів до вирішення фундаментальних та прикладних проблем функціонування веб-спільнот.