УДК 81’11+32+33:004.9

Федушко Соломія Степанівна
Асистент кафедри соціальних комунікацій та інформаційної діяльності
Білущак Галина Іванівна
Доцент кафедри вищої математики

Національний університет “Львівська політехніка”, Львів

ФОРМУВАННЯ СИСТЕМИ ЛІНГВО-КОМУНІКАТИВНИХ ІНДИКАТОРІВ СОЦІАЛЬНО-ДЕМОГРАФІЧНИХ ХАРАКТЕРИСТИК WEB-УЧАСНИКІВ

Розглянуто проблему дослідження та розроблення методів верифікації інформаційного наповнення користувачів віртуальних спільнот. Першочерговим завданням у перевірці достовірності персональних даних є формування та дослідження інформаційного сліду учасника віртуальної спільноти. Створено модель інформаційного сліду учасників web-спільноти та алгоритм формування системи лінгво-комунікативних індикаторів соціально-демографічних характеристик учасників віртуальної спільноти, на основі якого побудовано схему функціонування підсистеми формування наборів лінгво-комунікативних індикаторів. Також запропоновано класифікацію гендерних, вікових індикативних ознак та індикативних ознак сфери діяльності учасника віртуальної спільноти. Застосовано статистичні методи для верифікації гендеру учасника web-спільноти.

Ключові слова: web-спільнота, контент, web-учасник, інформаційний слід, лінгво-комунікативний індикатор, соціально-демографічна характеристика, персональні дані

Расмотрена проблемав исследования и разработки методов верификации контента пользователей виртуальных сообществ. Первоочередной задачей в проверке достоверности персональных данных является исследование информационного следа участника web-сообщества. Создана модель информационного следа участников web-сообщества и алгоритм формирования системы лингво-коммуникативных индикаторов web-участников. Также предложена классификация гендерных, возрастных индикативных признаков и индикативных признаков сферы деятельности участника web-сообщества і осуществлены статистические методы верификации гендера web-участника.

Ключевые слова: web-сообщество, контент, web-участник, информационный след, лингво-коммуникативный индикатор, социально-демографическая характеристика, персональные данные

This article considers the current problem of investigation and development of the methods for content verification of virtual communities' users. The primary task of checking the authenticity of personal data is to develop and research information track of web-community member (consolidated personal data of web- member). To solving of this problem is a model of an information track of web-community member and algorithm of formation of linguistic and communicative indicators of socio- demographic characteristics of members of virtual communities. Scheme of subsystem of linguistic-communicative indicators sets formation based on algorithm of socio- demographic characteristics indicators formation of web-members is designed. Also, the classification of gender-indicative features, age-indicative features and indicative features of sphere of activity of virtual communities’ members is suggested. The indicative characteristics based on the lingo-communicative markers set up by experts. Statistical methods of the analysis of learning sample of web-members of two Ukrainian web-forums are presented. Thus, the solution of these issues is determining the topicality of this paper and the necessity of developing the computer-linguistic method of socio-demographic characteristics validation in social communications.

Keywords: web-community, content, web-member, information track, linguistic-communicative indicator, socio -demographic characteristic, personal data

Постановка проблеми

У всесвітній мережі все більше користуються попитом такі інформаційні соціальні комунікації, як web-спільноти. Отже, гостро постає проблема відстеження небажаного інформаційного наповнення цих віртуальних спільнот.

Комп'ютерно-лінгвістичний аналіз достовірності персональних даних учасника web-спільноти і контенту полягає переважно у аналізі інформаційного сліду учасника віртуальної спільноти. Проведення аналізу дає змогу верифікувати персональні дані учасника віртуальної спільноти і сформувати його соціально-демографічний портрет [1-3].

Аналіз останніх досліджень і публікацій

В дослідженнях соціальних комунікацій науковці виокремлюють багато напрямів і у зв'язку з появою нових тенденцій розвитку цієї сфери досліджень збільшується буквально щомиті.

Аналіз сучасних праць науковців виявив, що в останні роки з'явилося багато публікацій, які присвячено вивченню особливостей функціонування віртуальних спільнот (Соціальна група людей, котрі комунікують та взаємодіють через Інтернет за допомогою спеціалізованих сервісів та сайтів у WWW. Основою віртуальних спільнот є учасники та інформаційне наповнення [3]), дослідження інформаційного наповнення в Інтернеті, методів верифікації контенту і персональних даних учасників web-спільнот, способів формування мережної ідентичності web-особистості.

В українському сегменті Інтернету ці напрямки досліджень, попри важливість, науковці почали розглядати та аналізувати нещодавно, незважаючи на його доволі швидкий розвиток. Втім, проаналізувавши наявні дослідження та публікації [1-4], слід зауважити, що кількість фундаментальних досліджень у цій області невелика, проте науковий інтерес зростає і є вже доволі вагомі результати досліджень.

Мета статті

Мета статті – розглянути та проаналізувати питання питання верифікації персональних даних у мережі Інтернет.

Постановка завдання

Серед актуальних задач наукових досліджень віртуальних спільнот до організаційних задач відносимо покращання керування системою.

Підвищення ефективності функціонування віртуальних спільнот є результатом ряду задач:

– відсіювання небажаного інформаційного наповнення;

– підвищення якості інформаційного наповнення;

– фільтрація учасників за достовірністю персональними даними;

– зменшення затрат на модерування спільнотою;

– зменшення конфліктних ситуацій в спільноті;

– комп'ютерно-лінгвістичний аналіз достовірності персональних даних учасників web-спільноти;

– автоматизування комп'ютерно-лінгвістичного аналізу контенту.

Таким чином, розв'язання наведених задач вагомо впливають на функціонування інтернет-спільноти, що дає можливість спростити та пришвидшити виконання обов'язків модератора та адміністратора віртуальної спільноти і створення моделі інформаційного сліду учасника віртуальної спільноти.

Комп'ютерно-лінгвістичний аналіз для верифікації персональних даних користувачів web-спільнот на основі їх інформаційного сліду є складовою методу системної перевірки максимальної кількості даних учасника web-форуму.

На основі верифікованих даних за допомогою створених інформаційних систем формується соціально-демографічний портрет учасника віртуальної спільноти – сукупність достовірних соціально-демографічних характеристик учасника web-спільноти (ВС).

Як для будь-якого аналізу, так і для комп'ютерно-лінгвістичного аналізу необхідна база персональних даних та даних, створених учасником під час його комунікативної діяльності у спільноті.

З цією ж метою введено поняття "інформаційний слід", що окреслює всю основну інформацію, необхідну для верифікації соціально-демографічних характеристик певного учасника ВС та побудови його соцільно-демографіного портрету.

Інформаційний слід учасника web-спільноти – множина всіх даних учасника віртуальної спільноти та результати його комунікативної діяльності – створене ним інформаційне наповнення.

Інформаційний слід учасника WC є дотичним до поняття "web-особистість", введеного О. Березком [4]. У своїх дослідженнях він використовує поняття "web-особистість", як засіб персоніфікації інформаційного наповнення, створеного інтернет-користувачем у багатьох віртуальних спільнотах.

Наше дослідження спрямоване на аналіз інформаційного наповнення створеного у межах однієї спільноти певним учасником. Цей аналіз здійснюється з метою виявлення соціально-демографічних характеристик учасника віртуальної спільноти та перевірки достовірності вказаних учасником персональних даних. З цією метою і було введено поняття "інформаційний слід", яке опишемо таким чином:

$$\mathrm{InfTrack}(U_i) = \left\langle \begin{matrix} \mathrm{Content}(U_i), \\ \mathrm{PersonalData}(U_i) \end{matrix} \right\rangle.$$

Складовими інформаційного сліду є: інформаційне наповнення – Content(Ui), яке створене учасником ВС, та персональні дані – PersonalData(Ui).

Інформаційне наповнення визначається кортежем, а сааме, такими підмножинами, як дискусії, опитування та дописи:

$$\mathrm{Content}(U_i) = \langle \mathrm{Thread}(U_i), \mathrm{Poll}(U_i), \mathrm{Post}(U_i) \rangle,$$

де $\mathrm{Thread}(U_i) = \{\mathrm{Thread}_j(U_i)\}_{j=1}^{N_i^{(\mathrm{UThead})}}$ – множина дискусій, створених учасником віртуальної спільноти Ui; $N_i^{(\mathrm{UThead})}$ – кількість таких дискусій.

$\mathrm{Poll}(U_i) = \{\mathrm{Poll}_j(U_i)\}_{j=1}^{N_i^{(\mathrm{UPoll})}}$ – множина опитувань, створених учасником віртуальної спільноти Ui; $N_i^{(\mathrm{UPoll})}$ – кількість таких опитувань.

$\mathrm{Post}(U_i) = \{\mathrm{Post}_j(U_i)\}_{j=1}^{N_i^{(\mathrm{UPos})}}$ – множина дописів учасника віртуальної спільноти Ui; $N_i^{(\mathrm{UPos})}$ – кількість дописів учасника віртуальної спільноти Ui.

Комп'ютерно-лінгвістичний аналіз (КЛА) проводиться тільки тих персональних даних, які учасник web-спільнот вказав у своєму обліковому записі.

Групування персональних даних учасника web-спільнот здійснено з точки зору пріоритетності персональних даних для комп'ютерно-лінгвістичного аналізу. На основі результатів КЛА формується соціально-демографічний портрет учасника web-спільнот.

Найбільш пріоритетними для формування СДП учасника web-спільнот є обов'язкова інформація про користувача web-спільнот, менш пріоритетні – важливі дані, та все ж для повноцінного аналізу необхідний також аналіз додаткових персональних даних учасника web-спільнот.

Обсяг персональної інформації, яку користувач надає у своєму обліковому записі поділено на блоки. Такий поділ переважно не є одноманітним і залежить від типу спільноти.

Зазначимо, що вміст блоків та пріоритетність персональних даних, які надає учасник віртуальної спільноти, визначають розробники та адміністратори цієї віртуальної спільноти, орієнтуючись на тематику та призначення web-спільноти.

Загальний алгоритм формування системи лінгво-комунікативних індикаторів

Мета цього алгоритму – формування системи лінгво-комунікативних індикаторів на основі навчальної вибірки учасників web-форумів (рис. 1).

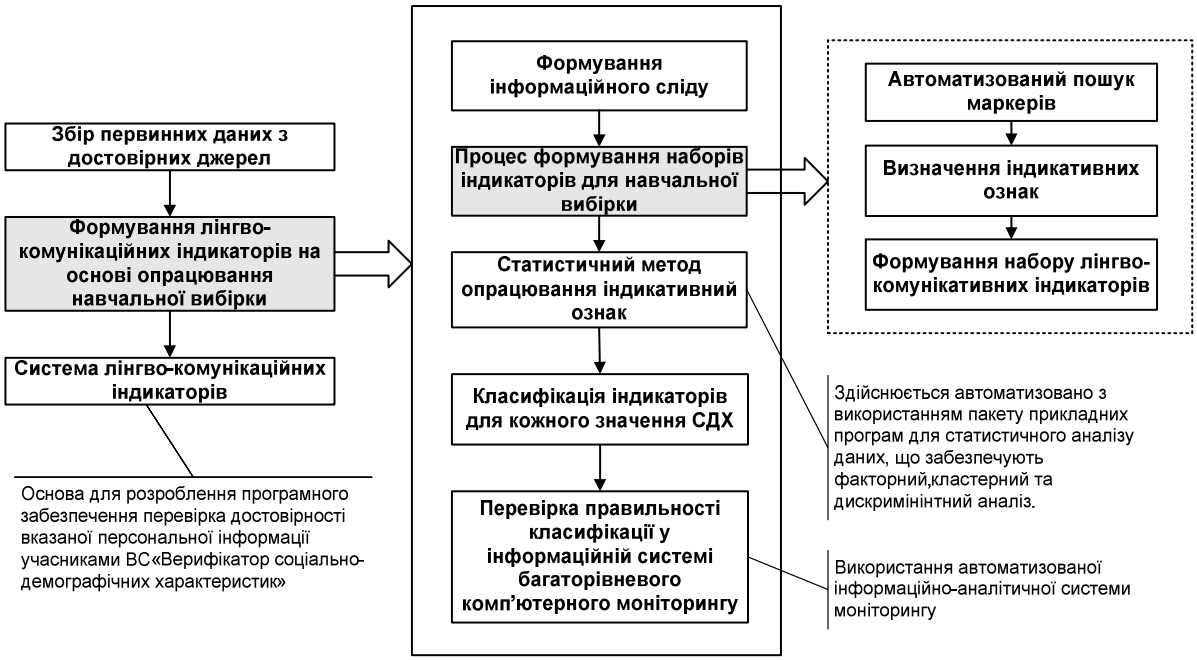

Рис. 1. Схема формування системи лінгво-комунікативних індикаторів на основі навчальної вибірки учасників web-форумів

Основні етапи підсистеми лінгво-комунікативних індикаторів [5] на основі навчальної вибірки учасників web-форумів:

1. Збирання інформаційного наповнення (контенту) учасника віртуальної спільноти.
2. Формування соціально-демографічних груп учасників web-спільноти відповідно до достовірних даних адміністрації віртуальної спільноти.
3. Формування наборів лінгвістичних та графічних маркерів соціально-демографічних характеристик учасників віртуальних спільнот та формування відповідних баз даних.
4. Формування наборів лінгво-комунікативних індикаторів на основі проаналізованих маркерів та занесення цієї інформації у базу даних індикаторів.

Графічно процес функціонування підсистеми формування лінгво-комунікативних індикаторів на основі навчальної вибірки учасників web-форумів зображено на рис. 2.

Визначення ґендерних лінгво-комунікативних індикаторів

З метою уникнення конфліктів у віртуальній спільноті модераторам та адміністраторам необхідно чітко відслідковувати ґендерну належність учасників спільноти.

Метод визначення ґендеру учасника web-спільнот використовується з метою покращення методів управління віртуальною спільнотою.

Для ґендерної диференції учасників віртуальної спільноти експертами сформовано множину ґендерних лінгвістичних ознак учасника віртуальних спільнот на основі досліджень, наукових теорій, ідеологій провідних вчених.

Перелік досліджуваних ґендерних лінгвістичних індикативних ознак учасника віртуальних спільнот наведено у табл. 1.

Для зручності аналізу множину ґендерних лінгвістичних ознак учасника віртуальної спільноти описуємо таким чином:

$$\mathrm{Gend}(U_i) = \left(\mathrm{Gend}_j(U_i)\right)_{j=1}^{N_i^{\mathrm{Gend}}},$$

де $\left(\mathrm{Gend}_j(U_i)\right)_{j=1}^{N_i^{\mathrm{Gend}}}$ – множина ґендерних лінгвістичних ознак учасника ВС; N_i^{Gend} – кількість ґендерних ознак конкретного учасника ВС.

Потрібно зауважити, що на результати досліджень суттєво впливають як контекст повідомлень, так і теми дискусій.

У дослідженні ми цей факт взяли до уваги і зробили різнопланову ґендерну вибірку повідомлень користувачів зі всіх тематичних розділів “Форуму Рідного Міста”.

У розділах “Форуму Рідного Міста” ми в однаковій мірі розглянули дискусії, які виникають у зв’язку з різноманітними інтересами та захопленнями як чоловіків, так і жінок.

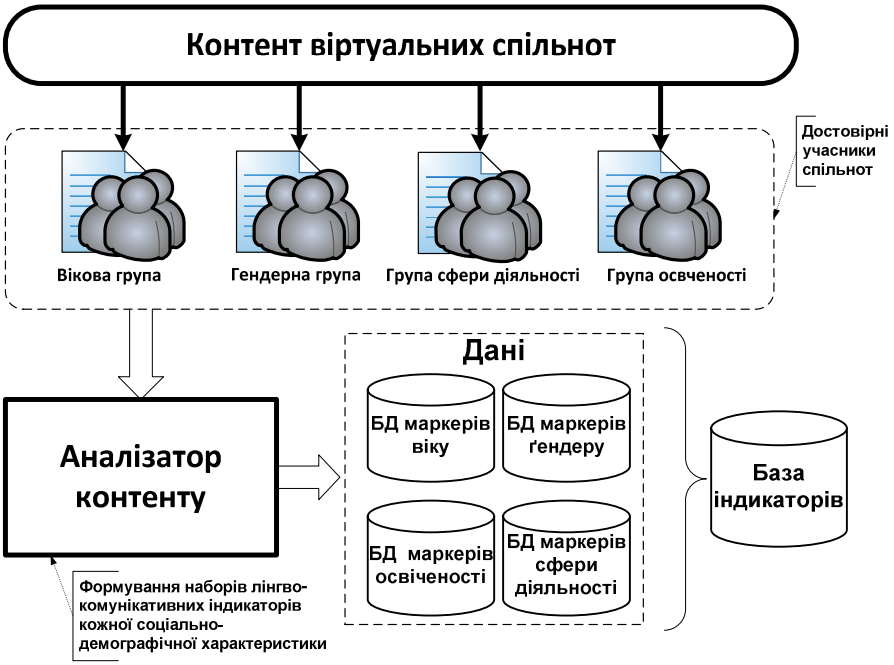

Рис. 2. Схема функціонування підсистеми формування наборів лінгво-комунікативних індикаторів

Таблиця 1

Класифікація ґендерних індикативних ознак

Лінгво-комунікативний індикатор ґендеру учасника ВС	Індикативна ознака ґендеру учасника ВС	
Емоційна складова (GENDER-A)	A(1.1) Модальні конструкції A(1.2) Згадка про емоції та почуття A(1.3) Ухиляння від відповіді A(1.4) Брак впевненості A(1.5) Аксіологічні модальні судження	A(1.6) Умовність дій A(1.7) Зменшено-пестливі форми A(1.8) Оклична інтонація A(1.9) Вираження підтримки
Культурний аспект (GENDER-B)	B(1.1) Вибачення B(1.2) Ввічливість B(1.3) Евфемізми	B(1.4) Непрямі команди і прохання B(1.5) Виправдовування
Посилання (GENDER-C)	C(1.1) Просторові посилання C(1.2) Часові посилання	C(1.3) Географічні посилання C(1.4) Посилання на кількість, величину
Вказівка та інструкція (GENDER-D)	D(1.1) Вказівка на особу, подію тощо D(1.2) Вказівка на особу, яка говорить про себе	D(1.3) Вказівка на кількох мовців
Лексичний аспект (GENDER-E)	E(1.1) Соціально-родинна лексика E(1.2) Ненормативна лексика	E(1.3) Спортивно-політична, автомобільна технологічно-інноваційна лексика
Спосіб вираження змісту (GENDER-F)	F(1.1) Протиставлення F(1.2) Згода	F(1.3) Перефразування F(1.4) Заперечення
Часові рамки (GENDER-G)	G(1.1) Згадування минулих подій	G(1.2) Обговорення поточних проблем та актуальних тем
Незмістовність (GENDER-H)	H(1.1) Словесні заповнення H(1.2) Прикметникові вислови без смислового значення.	H(1.3) Безсмислові форми
Сила, вплив, авторитетність (GENDER-I)	I(1.1) Розпорядження I(1.2) Довгі слова I(1.3) Акцентування	I(1.4) Підсилення значення I(1.5) Ствердження I(1.6) Присяги і клятви
Збагачення мови (GENDER-J)	J(1.1) Фразеологізми J(1.2) Пряме цитування	J(1.3) Гумор
Композиція (GENDER-K)	K(1.1) Безособове речення K(1.2) Запитання	K(1.3) Розділове речення K(1.4) Еліптичні речення
Конкретизація (GENDER-L)	L(1.1) Імплікація	L(1.2) Уточнення

Визначення вікових лінгво-комунікативних індикаторів

Вікову категорію вибрано для перевірки достовірності вказання віку учасників віртуальних спільнот у зв'язку з рядом важливих чинників: наявність реальних он-лайн загроз інтернет-користувачам віком від 6 до 17 років [7] (до них належать розголошення особистої інформації конфіденційного характеру, доступ до контенту, що не відповідають віковим особливостям і негативно впливають на фізичне і психологічне здоров'я дитини, он-лайн-насильство, інтернет-маркетингові злочини тощо) та необхідність відсіювання вікової групи дітей, які подали заявку чи вже стали учасником ВС призначеної тільки для повнолітніх користувачів web-простору.

Для вікової диференції учасників віртуальної спільноти сформовано множину вікових лінгвістичних ознак учасника віртуальної спільноти на основі досліджень. Для зручності аналізу множину вікових лінгвістичних ознак учасника віртуальної спільноти описуємо так:

$$\mathrm{Age}(U_i) = \left(\mathrm{Age}_j(U_i)\right)_{j=1}^{N_i^{Age}},$$

де $\left(\mathrm{Age}_j(U_i)\right)_{j=1}^{N_i^{Age}}$ – множина вікових лінгвістичних ознак учасника ВС; N_i^{Age} – кількість вікових ознак конкретного учасника web-спільнот.

Множину досліджуваних лінгвістичних ознак з прикладами наведено в табл. 2.

Таблиця 2

Класифікація вікових індикативних ознак

Лінгво-комунікативний індикатор віку учасника ВС	Індикативна ознака віку учасника ВС	
Афілічтично-агресивний стиль (AGE–A)	A(1.1) Фамільярна лексика	A(1.2) Вульгаризми
Сленгова варіація (AGE–B)	B(1.1) Олбанська мова B(1.2) Комп'ютерний сленг	B(1.3) Молодіжний сленг B(1.4) Найменування атрибутів молодіжної моди
Модуляція голосу та звукова подібність (AGE–C)	C(1.1) Символи верхнього регістру – CAPS LOCK, комбінування регістру C(1.2) Позначення вигуків	C(1.3) Заміна слів на основі звукової подібності
Текстова економія (AGE–D)	D(1.1) Усічення D(1.2) Акроніми D(1.3) Спрощена транслітерація	D(1.4) Абревіатурні скорочення D(1.5) Словоскладання
Некодифіковані одиниці та невербальні засоби (AGE–E)	E(1.1) Комбінації символів та букв E(1.2) Надмірна кількість знаків пунктуації та спецсимволів E(1.3) Заміна літер цифрами E(1.4) Заміна літер неалфавітними знаками	E(1.5) Комбінації з голосними літерами E(1.6) Складання літер E(1.7) Послідовність дужок")"
Деформалізація (AGE–F)	F(1.1) Графічні смайли	F(1.2) Фамільярні особисті імена

Надалі у роботі будемо використовувати наведені вище позначення для лінгво-комунікативних індикаторів та індикативних ознак.

Визначення лінгво-комунікативних індикаторів сфери діяльності

Для визначення сфери діяльності, інтересів та зацікавлень учасників віртуальної спільноти сформовано множину лінгвістичних ознак учасника ВС на основі досліджень науковців та експертних результатів аналізу віртуальних спільнот (табл. 3).

Таблиця 3

Класифікація індикативних ознак сфери діяльності

Лінгво-комунікативний індикатор сфери діяльності учасника ВС	Індикативна ознака сфери діяльності учасника ВС	
Фізико-математична, технічна та економічна сфера (SPHERE–A)	A(1.1) Прості речення A(1.2) Позбавлення авторського "Я" A(1.3) Логічна побудованість A(1.4) Докази істинності інформації A(1.5) Двоскладні речення з простим дієслівним присудком A(1.6) Шаблони висловлювань A(1.7) Велика кількість числових даних A(1.8) Математичні, фізичні та економічні величини	A(1.9) Назви грошових знаків A(1.10) Безособові речення із присудком, вираженим дієслівною формою на -но, -то та об'єктом - прямим додатком у формі іменника у знахідному відмінку без прийменника A(1.11) Логічність, точність, доказовість, однозначність, узагальненість, об’єктивність
Хімічна сфера (SPHERE–B)	B(1.1) Хімічні формули та позначення B(1.2) Абревіатури та акроніми B(1.3) Велика кількість апострофів, дефісів	B(1.4) Довгі слова B(1.5) Словоскладання B(1.6) Складні конструкції
Соціологічна, історична, філософська та політична сфера (SPHERE–C)	C(1.1) Логізація викладу C(1.2) Вступні слова C(1.3) Вигуки C(1.4) Вільна форма структури тексту	C(1.5) Деталізаці, конкретизація C(1.6) Повнота викладу матеріалу C(1.7) Багатозначність
Природнича сфера (SPHERE–D)	D(1.1) Складні речення з чітко вираженим складносурядним або складнопідрядним зв'язком D(1.2) Простий дієслівний присудок, виражений дієсловами теперішнього, минулого чи майбутнього часу	D(1.3) Вступні слова, вигуки, повтори та слова-звернення D(1.4) Детальний опис усіх дій

Закінчення табл. 3

Лінгво-комунікативний індикатор сфери діяльності учасника ВС	Індикативна ознака сфери діяльності учасника ВС	
Медична сфера (SPHERE–E)	E(1.1) Стійкі словосполучення з “часткОвий”, “частковИй” та “часточкОвий” E(1.2) Синоніми E(1.3) Пароніми E(1.4) Мовні засоби високого ступеня стандартизації E(1.5) Варіантивність E(1.6) Точність формулювання	E(1.7) Лексичні одиниці з латинської, грецької та давньоруської мови E(1.8) Слова з афіксами -ир-, -видний E(1.9) Медикаментозне дозування E(1.10) Вживання слів у прямому значенні E(1.11) Медичний сленг
Філологічно-педагогічна сфера (SPHERE–F)	F(1.1) Емоційно-експресивна лексика F(1.2) Образність F(1.3) Багатство мови (синоніми, антоніми, омоніми, пароніми, фразеологізми). F(1.4) Художні засоби (епітети, метафори, порівняння, символи тощо).	F(1.5) Складнопідрядні речення з чітким логічним зв'язком між компонентами F(1.6) Усталені конструкції F(1.7) Композиційність тексту. F(1.8) Запобігання повторів, багатослів'я, зайвих слів та канцеляризмів
Сфера архітектури та мистецтвознавства (SPHERE–G)	G(1.1) Авторська індивідуальна манера G(1.2) Виразна композиційна структура тексту	G(1.3) Модель терміносполук: прикметник (дієприкметник)+іменник
Сфера фізичного виховання й спорту (SPHERE–H)	H(1.1) Складні речення K(1.2) Простота H(1.3) Закличний та оціночний характер висловлювань	H(1.4) Простота викладення H(1.5) Текст без усталеної конструкції H(1.6) Сполучниковий зв'язок
Сільськогосподарська сфера (SPHERE–I)	I(1.1) Сленг фермерів I(1.2) Надто спрощена мова	I(1.3) Пряма мова I(1.4) Вживання дієслів теперішнього часу
Юридична сфера (SPHERE–J)	J(1.1) Переконливість J(1.2) Констатування фактів J(1.3) Цитування та посилання на першоджерела J(1.4) Уникнення сполучників: а, але, щоб, та ін. J(1.5) Послідовний поділ тексту із застосуванням цифрової або літерної нумерації J(1.6) Узагальнені, безособові та неозначені дієслівні форми теперішнього часу J(1.7) Дієприслівникові й дієприкметникові звороти J(1.8) Форми правових застережень J(1.10) Відсутність емоційного забарвлення	J(1.9) Достовірність, зв’язність, стислість та послідовність J(1.11) Уникнення заміни слів, зміна порядку слів, речень і частин тексту J(1.12) Системність, нейтральність, неупередженість мови J(1.13) Формалізація та уніфікація засобів вираження, орієнтована на точність та однозначність J(1.14) Уникнення окличних та питальних речень, літоти, гіперболи, метафор та літоти
Воєнна сфера (SPHERE–K)	K(1.1) Однозначність висловів K(1.2) Безособові та інфінітивні конструкції та імперативи K(1.3) Наказовий спосіб K(1.4) Порушення об’єктивного порядку слів у реченні	K(1.5) Заборона K(1.6) Військовий сленг K(1.7) Абревіатури, умовні символи та скорочення K(1.8) Кліше K(1.9) Еліптичність

Вищою атестаційною комісією (ВАК) України за погодженням з Міністерством освіти і науки України визначено такі галузі науки [6], що покладені в основу сучасної класифікації науки. Саме на цій класифікації базується поділ наук у дослідженні. Проведено комп’ютерно-лінгвістичний аналіз інформаційного наповнення багатьох форумів Укрнету. Учасник більшою або меншою мірою може належати до кількох сфер зацікавлення, оскільки поштовхом для створення контенту у web-спільноти слугують напрям отриманої освіти, професійна діяльність та коло інтересів у вільний час. Проте, результатом аналізу є визначення сфери діяльності учасника віртуальних спільнот.

Статистичні методи верифікації СДХ

Забезпечення мінімальної похибки розпізнавання та достовірної ідентифікації ґендеру учасника web-спільноти полягає у обчисленні основних статистичних характеристик навчальної вибірки (емпіричних даних, на основі яких виділяються закономірності віднесення учасників віртуальної спільноти до певної групи).

Для вирішення питання верифікації ґендеру учасників віртуальних спільнот застосовуємо комплекс методів математичної статистики [12], за допомогою якого здійснюється класифікація учасників web-форумів залежно від вікової, ґендерної та професійної категорії, а саме дискримінантний, факторний та кластерний аналіз.

Формування навчальної вибірки

Для апробації методики з учасників web-форуму «Львів. Форум Рідного Міста» [8] та web-форуму західноукраїнського рок-порталу – «Rock.Lviv.Ua» [9] сформовано навчальну вибірку найактивніших учасників цієї віртуальної спільноти. Всіх учасників рівномірно поділено на дві групи відповідно до гендерної належності. Відповідно до моделі, обрано поділ учасників віртуальної спільноти відповідно до кожної СДХ на дві групи: вік (“підлітки” і “дорослі користувачі”), ґендер (“чоловіки” та “жінки”), сфера діяльності (“прикладна” та “фундаментальна”).

Потрібно зауважити, що кожного учасника для цієї навчальної вибірки ретельно підібрано, зважаючи на повноту та достовірність інформаційного сліду та з цілком достовірними даними про ґендер учасника web-форуму. Також взято до уваги той факт, що на результати досліджень суттєво впливають, як контекст повідомлень, так і теми дискусій. Зважаючи на цей факт, основою цього дослідження є різнопланова вибірка повідомлень користувачів з усіх тематичних розділів як “Форуму Рідного Міста”, так і web-форуму західноукраїнського рок-порталу – «Rock.Lviv.Ua». Для дослідження СДХ сфер діяльності форумів проаналізовано ІН 20-ти web-спільнот. У розділах web-форумів однаковою мірою розглянуто дискусії, які виникають у зв’язку з різноманітними інтересами та захопленнями дорослих осіб та підлітків, чоловіків та жінок.

Статистичний аналіз ґендеру учасників web-спільнот (навчальна вибірка)

Дискримінантний аналіз використано для аналізу відмінностей (індикативних ознак) між ґендерними групами учасників web-форумів, кластерний аналіз – для об’єднання даних у групи так, щоб відмінності між учасниками web-спільнот кожної ґендерної групи були мінімальними, а між самими ґендерними групами – суттєвими, які впливають на досліджуваних учасників web-спільнот (рис. 3).

Обчислення здійснюємо автоматизовано з використанням пакету прикладних програм для статистичного аналізу даних, що забезпечують кластерний та дискримінантний аналіз ("STATISTICA" [10] та статистичного пакету для соціальних наук "SPSS" [11]).

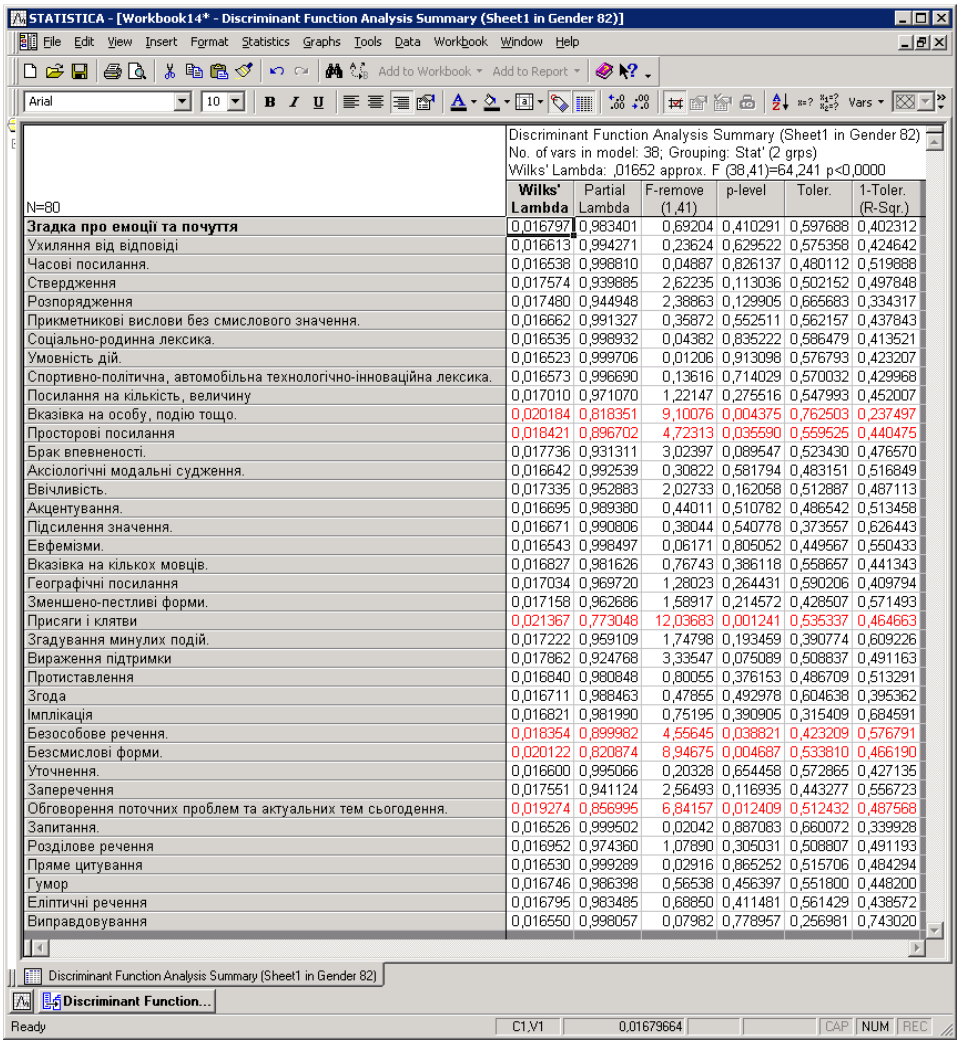

Рис. 3. Дискримінантний аналіз

У цьому дослідженні проведено статистичний аналіз індикативних ознак ґендеру учасника web-форумів (82 web-учасники). Результати порівняння основних числових характеристик розглянутих індикативних ознак у двох ґендерних групах вказують на наявність статистично значущих відмінностей у більшості з цих показників.

За результатами дискримінантного аналізу значення статистики Ламбда Уілкса (Wilk's; Lambda) 0,01652 (близьке до нуля) і значення критерію Фішера F (38,41) = 64,241 (p<0,0000) свідчать про хорошу дискримінацію, тобто класифікація проведена коректно. Також, матриця класифікації показує, що всі об'єкти класифіковані.

Поле Tolerance дозволяє виключити з моделі неінформативні змінні. Значення толерантності вказують на високу інформативність усіх індикативних ознак, при цьому статистично значущими є такі індикативні ознаки: *D(1.1) Вказівка на особу, подію тощо; C(1.1) Просторові посилання; I(1.6) Присяги і клятви; K(1.1) Безособове речення; H(1.3) Безсмислові форми; G(1.2) Обговорення поточних проблем та актуальних тем*. Це підтверджують стандартизовані коефіцієнти, які вказують на вклад індикативних ознак у значення дискримінантної функції, що є одним із підходів до визначення значущості змінної.

Результати кластерного аналізу наведені. За вибраними чинниками вибірка «Gender» поділена на два кластери – Кластер1 (чоловіки) і Кластер 2 (жінки). Оскільки ґендер учасників форуму відомий, проведено аналіз некоректності кластеризації.

Графічно процедуру кластеризації зображено на дендрограмі (рис. 4).

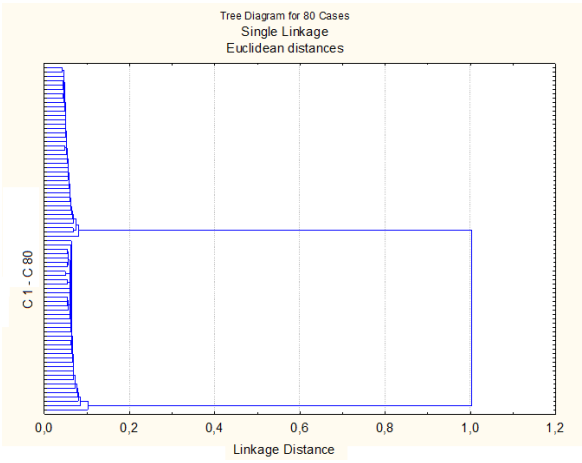

Рис. 4. Дендрограма кластеризації учасників web-форумів за ґендером

Це графічне зображення кластерів підтверджує ефективність проведеної класифікації.

Таблиця 4

Cluster Membership

Case	Stat	Cluster	Distance	Case	Stat	Cluster	Distance
1		1	3,576E-02	42		2	3,612E-02
2	m	1	4,301E-02	43	w	2	7,024E-02
3	m	1	7,114E-02	44	w	2	6,629E-02
4	m	1	6,918E-02	45	w	2	5,168E-02
5	m	1	3,804E-02	46	w	2	5,652E-02
6	m	1	5,983E-02	47	w	2	6,519E-02
7	m	1	4,521E-02	48	w	2	8,097E-02
8	m	1	3,810E-02	49	w	2	6,198E-02
9	m	1	5,757E-02	50	w	2	4,797E-02
10	m	1	4,569E-02	51	w	2	7,545E-02
11	m	1	4,130E-02	52	w	2	7,253E-02
12	m	1	5,347E-02	53	w	2	6,843E-02
13	m	1	4,848E-02	54	w	2	4,902E-02
14	m	1	5,124E-02	55	w	2	5,185E-02
15	m	1	4,247E-02	56	w	2	7,566E-02
16	m	1	4,692E-02	57	w	2	5,480E-02
17	m	1	4,915E-02	58	w	2	5,531E-02
18	m	1	5,599E-02	59	w	2	4,751E-02
19	m	1	4,623E-02	60	w	2	5,141E-02
20	m	1	5,950E-02	61	w	2	7,469E-02
21	m	1	5,144E-02	62	w	2	7,589E-02
22	m	1	6,436E-02	63	w	2	7,601E-02
23	m	1	7,721E-02	64	w	2	4,923E-02
24	m	1	7,030E-02	65	w	2	5,472E-02
25	m	1	4,537E-02	66	w	2	5,067E-02
26	m	1	5,765E-02	67	w	2	4,626E-02
27	m	1	2,734E-02	68	w	2	6,072E-02
28	m	1	5,719E-02	69	w	2	6,049E-02
29	m	1	4,328E-02	70	w	2	,111
30	m	1	4,734E-02	71	w	2	4,797E-02
31	m	1	5,463E-02	72	w	2	6,879E-02
32	m	1	6,152E-02	73	w	2	6,956E-02
33	m	1	5,548E-02	74	w	2	6,657E-02
34	m	1	6,418E-02	75	w	2	7,093E-02
35	m	1	3,849E-02	76	w	2	5,834E-02
36	m	1	3,731E-02	77	w	2	5,698E-02
37	m	1	5,959E-02	78	w	2	6,735E-02
38	m	1	8,082E-02	79	w	2	5,146E-02
39	m	1	5,026E-02	80	w	2	6,171E-02
40	m	1	5,167E-02	81	w	2	5,130E-02
41	m	1	5,368E-02	82	w	2	6,170E-02

Матриця факторної структури отриманого результату також дозволяє оцінити вклад окремих факторів у класифікацію.

Після проведеної класифікації учасників web-форумів у вибірку було 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 були класифіковані правильно.

Якщо ж одного чи кількох учасників web-спільноти під час класифікації віднесено до іншого кластеру, з огляду на адміністрування і модерування віртуальною спільнотою, є такі сценарії розвитку цієї атомарної ситуації:

– учасник навмисно неправильно вказав свій ґендер у обліковому записі, щоб цілеспрямовано проникнути у віртуальну спільноту призначену для користувачів протилежної статі з прихованими намірами;

– учасник випадково помилився у виборі свого ґендеру;

– стиль інтернет-спілкування учасника віртуальної спільноти не відповідає ґендеру учасника.

Проаналізувавши методом статистичного аналізу інформаційний слід учасника віртуальної спільноти, модератор та адміністратор має змогу верифікувати такі соціально-демографічні характеристики, як вік, сферу діяльності, освіченість та ґендер цього учасника web-спільноти, що дає можливість для ефективнішого керування спільнотою.

Висновок

Розроблення сучасного комп'ютерно-лінгвістичного підходу до верифікації даних, які надає web-користувач при реєстрації та під час участі у спільноті є актуальним питання в керуванні та модеруванні віртуальних спільнот. Зі збільшенням чисельності віртуальних спільнот та їх користувачів виникла потреба в розробці автоматизованого методу перевірки максимальної кількості даних про потенційного учасника web-спільнот.

Розроблено метод для формування інформаційного сліду учасника web-спільноти – множини всіх даних web-учасника та результати його комунікативної діяльності, аналіз якого сприяє відсіюванню небажаного контенту, підвищенню якості інформаційного наповнення, фільтрації учасників за достовірністю персональних даних, зменшення затрат на модерування спільнотою та конфліктних ситуацій в спільноті.

Розроблено алгоритм формування системи лінгво-комунікативних індикаторів соціально-демографічних характеристик учасників віртуальної спільноти, на основі якого побудовано схему функціонування підсистеми формування наборів лінгво-комунікативних індикаторів.

Також запропоновано класифікацію ґендерних, вікових індикативних ознак та індикативних ознак сфери діяльності учасника віртуальної спільноти. Для верифікації ґендеру учасника web-спільноти застосовано статистичні методи на основі аналізу навчальної вибірки учасників двох популярних українських web-форумів.

Список літератури

1. *Fedushko S. The verification of virtual community member's socio-demographic characteristics profile / S. Fedushko, O. Peleschyshyn, A. Peleschyshyn, Yu. Syerov // Advanced Computing: An International Journal (ACIJ), Vol.4, No.3, May 2013. – P. 29-38.*
2. *Fedushko S. Design of registration and validation algorithm of member's personal data/ S. Fedushko, Yu. Syerov // International Journal of Informatics and Communication Technology (IJ-ICT) Vol.2, No.2, July 2013, pp. 93-98.*
3. *Syerov Yu. The computer-linguistic analysis of socio-demographic profile of virtual community member /Yu. Syerov, A. Peleschyshyn, S. Fedushko // International Journal of Computer Science and Business Informatics (IJCSBI), Vol 4, No 1, August 2013. – P. 1-13.*
4. *Березко О.Л. Каталог Web-особистостей. Комп'ютерні системи та мережі [Текст] / О.Л. Березко // Вісник Національного університету "Львівська політехніка", № 630. – Львів: Видавництво Національного університету "Львівська політехніка", 2008. - С. 12-16.*
5. *Fedushko S. Algorithm of the cyber criminals identification / S. Fedushko, N. Bardyn // Global Journal of Engineering, Design & Technology (GJEDT), Vol. 2, No. 4(2013). – P. 56-62.*
6. *Про затвердження Переліку спеціальностей, за якими проводяться захист дисертацій на здобуття наукових ступенів кандидата наук і доктора наук, присудження наукових ступенів і присвоєння вчених звань. – Наказ Вищої атестаційної комісії України від 23 червня 2005 року N 377.*
7. *Сергєєнкова О. П. Вікова психологія : навч.посіб. / О. П. Сергєєнкова, О. А. Столярчук, О. П. Коханова, О. В. Пасєка. - К. : ЦУЛ, 2012. - 384 с.*
8. *Віртуальна спільнота «Львів. Форум Рідного Міста» [Електронний ресурс] –http://misto.ridne.net*
9. *Web-форум західноукраїнського рок-порталу – «Rock.Lviv.Ua» [Електронний ресурс] – http://rock.lviv.ua/forum*
10. *Буреева Н. Многомерный статистический анализ с использованием ППП "STATISTICA" - Нижний Новгород, 2007. - 112 с. [Електронний ресурс] – http://www.unn.ru/pages/e-library/aids/2007/57.pdf*
11. *Бююль А. SPSS: искусство обработки информации. Анализ статистических данных и восстановление скрытых закономерностей. / А.Бююль, П.Цефель – СПб.: ООО «ДиаСофтЮП», 2002, С.346-367.*
12. *Hill T. Statistics methods and applications, StatSoft. / T. Hill, P. Lewicki. Tulsa, OK, 2007. [electronic source]. - www.statsoft.com/textbook.*

References

1. *Fedushko, S., Peleschyshyn, O., Peleschyshyn, A., Syerov, Yu., (2013) The verification of virtual community member's socio-demographic characteristics profile // Advanced Computing: An International Journal (ACIJ). – Vol.4, No.3. – 29-38.*
2. *Fedushko, S., Syerov, Yu., (2013) Design of registration and validation algorithm of member's personal data // International Journal of Informatics and Communication Technology (IJ-ICT). – Vol.2, No.2. – 93-98.*
3. *Syerov, Yu., Peleschyshyn, A., Fedushko, S., (2013) The computer-linguistic analysis of socio-demographic profile of virtual community member // International Journal of Computer Science and Business Informatics (IJCSBI). – Vol 4, No 1. – 1-13.*
4. *Berezko, O. L., (2008) Catalogue of web-personalities // Lviv Polytechnic National University scientific journal "Computer systems and networks". – № 630. – 12-16.*
5. *Fedushko, S., Bardyn, N., (2013) Algorithm of the cyber criminals identification // Global Journal of Engineering, Design & Technology (GJEDT). – Vol. 2, No. 4(2013). – 56-62.*
6. *On Approval of the List of specialties for which defending of thesis for the candidate's and doctorate degrees, conferring of scientific degrees and academic statuses. The Decree of the Board of Higher Attestation Commission of Ukraine of June 23, 2005 N 377 [electronic source].–www.zakon1.rada.gov.ua/laws/show/z0713-05*
7. *Serhyeyenkova, O.P., Stolyarchuk, O.A., Kokhanova, O.P., Pasyeka, A.V., (2012) Developmental Psychology: Manual. – K. TSUL. – 384.*
8. *Virtual Community "Lviv. Forum Ridne Misto" [electronic source]. – http://misto.ridne.net*
9. *Web Forum Western rock portal – "Rock.Lviv.Ua" [electronic source] . – http://rock.lviv.ua/forum*
10. *Bureeva, N., (2007) Multivariate statistical analysis using PPP "STATISTICA" // Training method. material. – Nizhny Novgorod. – 112. [electronic source]. – http://www.unn.ru/pages/e-library/aids/2007/57.pdf*
11. *Byuyul, A., Tsefel, P. (2002) SPSS: art of information processing. Analysis of statistical data and restore hidden patterns. // St. Petersburg. LLC "DiaSoftYuP". – 346-367.*
12. *Hill, T., Lewicki, P., (2007) Statistics methods and applications, StatSoft [electronic source]. – www.statsoft.com/textbook.*

Стаття надійшла до редколегії 24.03.2014

Рецензент: д-р техн. наук, проф. А.М. Пелещишин, Національний університет "Львівська політехніка", Львів.