ПРОГРАМНО-ТЕХНІЧНІ КОМПЛЕКСИ

УДК 004.912

А.С. КОЛЕСНИК
Національний технічний університет «Харківський політехнічний інститут», Харків, Україна, e-mail: *kolesniknastya20@gmail.com*.

Н.Ф. ХАЙРОВА
Національний технічний університет «Харківський політехнічний інститут», Харків, Україна, e-mail: *nina_khajrova@yahoo.com.*

ОБҐРУНТУВАННЯ ВИКОРИСТАННЯ СТАТИСТИКИ КАППА КОЕНА В ЕКСПЕРИМЕНТАЛЬНИХ ДОСЛІДЖЕННЯХ NLP ТА TEXT MINING

Анотація. Виконано порівняння сучасних метрик оцінювання узгодженості (agreement coefficients) між результатами експериментів і експертною думкою та оцінено можливість використання цих метрик під час проведення експериментальних досліджень у галузі автоматичного оброблення текстів методами машинного навчання. Обґрунтовано вибір коефіцієнта каппа Коена як міри оцінювання узгодженості думок експертів у задачах NLP та Text Mining. Наведено приклад застосування коефіцієнта каппа Коена для оцінювання рівня узгодженості між думкою експерта і результатами ML класифікації та міри узгодженості думок експертів у випадку вирівнювання речень казахсько-російського паралельного корпусу. На підставі наведеного аналізу доведено, що завдяки зручності у використанні, простоті обчислення та високій точності результатів коефіцієнт каппа Коена є одним з найкращих статистичних методів визначення рівня узгодженості в експериментальних дослідженнях текстів.

Ключові слова: Text Mining, NLP, статистика каппа Коена, коефіцієнт узгодженості, класифікація текстів методами машинного навчання, паралельний корпус.

ВСТУП

У сучасних умовах розвитку науки гостро постає питання оцінювання результатів досліджень. Для того, щоб можна було надалі спиратись і посилатися на одержані результати експериментів, вони мають бути повністю об’єктивними та максимально точними. Ці вимоги стосуються як кількісних результатів наукової роботи, так і якісних результатів досліджень.

Оцінювання достовірності експериментів, проведених на тому чи іншому корпусі під час розв’язання задач Natural Language Prosessing (NLP) або Text Mining, також є одним з основних складних завдань, пов’язаних з автоматичним обробленням текстів. Зазвичай, результати застосування методів і моделей інформаційного пошуку, класифікації, кластеризації, морфологічного розмічування, синтаксичного парсера та інших задач оцінюють або із залученням експертів, або методом порівняння з так званим «золотим стандартом», тобто корпусом, розміченим деякими цільовими значеннями (target value), пов’язаними із завданнями Machine Learning (ML). При цьому традиційно використовують такі мет-

© А.С. Колесник, Н.Ф. Хайрова, 2022

рики як повнота, точність та F-міра. Ці метрики є досить універсальними, їх можна застосувати практично до будь-якого дослідження NLP. Проте саме через їхню універсальність не можна стверджувати, що вони є досить об'єктивними.

Під експертним оцінюванням розуміють процес отримання оцінки будь-якого явища або поняття, що ґрунтується на думці експертів, для подальшого прийняття рішення [1]. Але зазвичай самої лише індивідуальної експертної оцінки замало. Наприклад, у вузькопрофільних задачах для оцінювання роботи класифікатора необхідно використовувати декілька думок окремих експертів у цій предметній області, які є незалежними один від одного, та перевіряти рівень узгодженості їхніх думок. Ця перевірка надає змогу визначити рівень адекватності роботи експертів, достовірність та об'єктивність проведеного оцінювання. Вважають, що саме узгодженість думок експертів або експертної групи є однією з найважливіших характеристик якості результатів експерименту. Саме слабка узгодженість може бути свідченням некоректної роботи експертів і відповідно поставити під сумнів результати всього дослідження. Тому визначення великої або малої відмінності думок експертів, що допомагає у підбитті підсумків результатів експериментів, є практично обов'язковим під час оцінювання методів і моделей як NLP, так і Text Mining.

Саме через це у цьому дослідженні здійснено порівняння наявних методів та метрик оцінювання міри узгодженості експертів, та обґрунтовано можливість їхнього використання у виконанні експериментальних задач комп'ютерної та корпусної лінгвістики.

У цій роботі проведено два типи досліджень. У першому дослідженні визначено міру узгодженості між експертом та автоматичним класифікатором текстів декількома методами Machine Learning (ML), а у другому — визначено міру узгодженості думок кількох експертів щодо автоматичного розмічування паралельного корпусу. Під автоматичним розмічуванням у цьому випадку мають на увазі позначення у двох частинах двомовного корпусу семантично еквівалентних речень.

Стаття має таку структуру. У розд. 1 наведено огляд літературних джерел щодо сучасних статистичних методів оцінювання думок експертів та обґрунтовано вибір коефіцієнта каппа Коена як міри оцінювання думок у задачах NLP та Text Mining. У розд. 2 наведено опис проведених експериментів та корпусів, використаних у цих експериментах. У розд. 3 описано експериментальне дослідження застосування коефіцієнта каппа для оцінювання рівня узгодженості між експертом та автоматичним класифікатором, що використовує методи ML. У розд. 4 розглянуто застосування коефіцієнта каппа для оцінювання міри узгодженості думок експертів щодо вирівнювання речень казахсько-російського паралельного корпусу. У висновках підбито підсумок дослідження, що ґрунтується на результатах двох статистичних експериментів. Зокрема, обґрунтовано використання коефіцієнта каппа Коена як найкращої статистичної метрики оцінювання рівня узгодженості між експертами або експертною думкою та метриками ML у задачах NLP та Text Mining.

1. ОГЛЯД ЛІТЕРАТУРНИХ ДЖЕРЕЛ. СУЧАСНІ МЕТРИКИ ОЦІНЮВАННЯ УЗГОДЖЕНОСТІ ДУМОК ЕКСПЕРТІВ

Найпоширенішими статистичними методами або метриками оцінювання узгодженості думок експертів, якими можна скористатися під час проведення експериментальних досліджень у галузі комп'ютерної та корпусної лінгвістики, є такі:

— коефіцієнт варіації;
— коефіцієнт конкордації Кендала;
— коефіцієнт каппа Коена.

Коефіцієнт варіації V_j вимірюють у балах або у відсотках. Він є умовною мірою відмінностей думок відносно середньої величини оцінки групи в цілому. Він визначається для кожного порівнюваного об'єкта і характеризує ступінь узгодженості думок експертів про відносну важливість *j*-го об'єкта:

$$V_j = \frac{\sigma_j}{M_j}, \tag{1}$$

де M_j — середнє арифметичне значення величини оцінки об'єкта (в балах чи частках), а σ — середнє квадратичне відхилення оцінок, отриманих для *j*-го об'єкта.

Вважають, що чим менше значення коефіцієнта варіації, тим вище ступінь узгодженості думок експертів. Якщо коефіцієнт $V_j \le 0.30$, то ступінь узгодженості вважають задовільним. Якщо коефіцієнт варіації $V_j \le 0.20$, то ступінь узгодженості експертів вважають досить хорошим. Прийнятним результатом є значення коефіцієнта варіації не більше 0.25 [2].

Наступним коефіцієнтом оцінки узгодженості є коефіцієнт конкордації Кендала, який застосовують тоді, коли сукупність об'єктів характеризується кількома послідовностями рангів, а дослідник має встановити статистичний зв'язок між цими послідовностями [3]. Зазвичай цей коефіцієнт розраховують за формулою

$$W = \frac{12\sum_{i=1}^{n} D_i^2}{m^2 (n^3 - n)}, \tag{2}$$

де n — кількість оцінюваних об'єктів; m — кількість рангових послідовностей (кількість експертів); $D_i = d_i - \bar{d}$ — відхилення суми рангів *i*-го об'єкта від середньої суми рангів усіх об'єктів [4].

Коефіцієнт конкордації W (2) може набувати значень від 0 до 1.0. Вважають, що для $W = 1.0$ узгодженість думок експертів є повною, для $W > 0.5$ — задовільною, а для $W < 0.5$ — занизькою. Прийнятним є значення коефіцієнта конкордації не менше 0.75. Коефіцієнт рангової кореляції Кендала є ефективним і надійним способом визначення монотонних відношень між двома послідовностями даних. Однак, його застосування до цифрових даних у випадку великої кількості зв'язків дає суперечливі результати через можливе квантування [5].

Використання двох наведених коефіцієнтів (варіації та конкордації Кендала) у задачах комп'ютерної та корпусної лінгвістики може призвести до досить суперечливих результатів, особливо у тому разі, коли йдеться про колективне оцінювання та відсутність спільності поглядів. Це найчастіше виявляється під час семантичного оброблення текстів, наприклад, у випадку семантичного розмічування, вузькотематичної класифікації або кластеризації текстів та під час виконання інших задач, коли всередині експертної групи може бути висока узгодженість думок, однак узагальнені думки підгруп є протилежними.

Ефективним способом розв'язання цієї проблеми є застосування комбінації декількох методів. Зазначені коефіцієнти та різні їхні варіації ґрунтуються на початковому коефіцієнті, який у 1960 році ввів Коен і назвав його коефіцієнтом каппа. Цей коефіцієнт використовують для вимірювання скоригованої за випадковістю номінальної шкали узгодженості між двома оцінювачами. Інакше кажучи, Коен запропонував застосовувати коефіцієнт каппа як міру узгодженості між двома оцінювачами, які висловлюють свої судження за номінальною шкалою. Водночас наголошено на можливій випадковості у визначенні рівня узгодженості думок експертів [3, с. 1595], що допускає наявність випадкової узгодженості, яка впливає на результат. Саме тому це більш надійна міра, ніж простий розрахунок узгодженості за відсотками, на якому ґрунтуються два попередні коефіцієнти [6].

Традиційна формула визначення коефіцієнта каппа є такою:

$$\mathrm{K} = \frac{P_0 - P_e}{1 - P_e}, \tag{3}$$

де P_0 відображає оцінку того, наскільки спостережувана узгодженість краща за випадкову, а P_e відображає результат підрахунку максимально можливої узгодженості за винятком випадкової узгодженості.

Подібно до більшості кореляційних коефіцієнтів, коефіцієнт каппа може варіюватися від -1 до $+1$. Незважаючи на те, що значення менше нуля враховуються, Коен довів, що вони є малоймовірними [7]. Вважають, що для $\mathrm{K} = 1$ узгодженість є ідеальною (perfect agreement) і навпаки, для $\mathrm{K} = 0$ узгодженість прирівнюють до випадкової (equivalent to chance).

Останнім часом на основі традиційної статистики каппа Коена розробляють сумарні метрики узгодження. Ці метрики є або розширенням міри каппа Коена, або використовують формулу каппа з подальшим упровадженням іншої оцінювальної моделі [8], особливо тоді, коли кількість експертів є більшою за три. Наприклад, такими коефіцієнтами є каппа Флейса [9], каппа Лайта і Конджера [10, 11] та каппа Мільке [12].

Найбільш поширеним обмеженням під час вибору міри узгодженості експертів є тип досліджуваних даних. До прикладу, як міру узгодженості думок експертів під час обчислення номінальних даних зазвичай обирають коефіцієнт V Крамера та інші коефіцієнти зв'язності. Усі вони ґрунтуються на мірі зв'язку, яка відображає відносне зниження помилки. Оскільки результати експериментальних досліджень текстової інформації не є ні порядковими, ні номінальними, у цьому випадку як міру потрібно застосовувати коефіцієнт каппа Коена.

2. ОПИС ПРОВЕДЕНИХ ЕКСПЕРИМЕНТІВ

Для визначення міри узгодженості думок експертів та зіставлення результатів роботи декількох різних моделей класифікації та експерта розроблено два корпуси текстів. Перший корпус містить тексти українською мовою, отримані з новинних сайтів і присвячені кримінальній тематиці, спорту, науці, світовим новинам та економіці. Другий корпус — це паралельний казахсько-російський корпус текстів кримінальної тематики [13]. На момент написання статті процес наповнення обох корпусів тривав.

Перший корпус українських текстів розроблено на основі інформаційного вмісту двох українських сайтів актуальних новин: Главком (https://glavcom.ua) та УНІАН (https://www.unian.ua). Для дослідження використано довільну частину корпусу, що містить 6000 текстів, з яких 3000 текстів є текстами кримінальної тематики (Crime), а 3000 текстів належать різним тематичним категоріям, як-то спорт (Sport), наука (Science and IT), світові новини (World News) та економіка (Economics).

Другий, створений нами паралельний казахсько-російський корпус містить вирівняні за реченнями тексти казахською та російською мовами, що мають кримінальну забарвленість. Під паралельністю текстів розуміють їхній однаковий обсяг та повну ідентичність сенсу за кожним окремим реченням. Для збору текстів було використано чотири двомовних вебпортали Республіки Казахстан zakon.kz, caravan.kz, lenta.kz та nur.kz [14], які поряд з іншим, висвітлюють стан рівня злочинності у країні. Ці портали є білінгвістичними та містять кримінальні новини казахською та російською мовами про злочини, як-то пограбування, викрадення машин, вбивства, ДТП тощо. Саме ці тематики і визначають базовий ресурс створеного корпусу.

На основі зазначених корпусів виконано два експериментальних дослідження щодо застосування статистики каппа до розв'язання задач NLP та Text Mining. У першому дослідженні порівняно роботу кількох алгоритмів класифікації новин українськомовного корпусу з аналогічною інтелектуальною діяльністю людини. Завдання експерименту полягало в оцінюванні рівня узгодженості між автоматичним політематичним класифікатором та людиною щодо співвіднесення статті з новинного вебсайта з її тематикою, а саме кримінальною (Crime), спортивною (Sport), науковою (Science and IT), економічною (Economics) чи новинною (World News).

Завдання другого етапу дослідження полягало в оцінюванні достовірності результатів автоматичного вирівнювання створеного паралельного казахсько-російського корпусу текстів кримінальної забарвленості. Попередньо було здійснено автоматичне попарне зіставлення текстів за реченнями за допомогою розробленого алгоритму [14, с. 88]. На цьому етапі досліджено оцінку рівня узгодженості думок декількох експертів щодо повного семантичного збігу за кожним окремим реченням текстів паралельного корпусу.

В обох наведених експериментах статистика каппа ґрунтується на матриці плутанини або матриці помилок, яку часто використовують у машинному навчанні для підбиття підсумків роботи алгоритмів класифікації. Ця матриця є таблицею, в якій наведено кількість збігів думки експерта з автоматичним класифікатором або думок експертів між собою (true positive) і кількість тих випадків, коли думка експерта і автоматичного класифікатора не збігалися (false positive та false negative).

3. ОЦІНЮВАННЯ РЕЗУЛЬТАТІВ АВТОМАТИЧНОЇ ТЕМАТИЧНОЇ МУЛЬТИКЛАСИФІКАЦІЇ

Для ілюстрування матриці плутанини спостережуваної та очікуваної точності результатів автоматичної тематичної класифікації було використано три найбільш поширені методи контрольованого машинного навчання, що використовується у Text Mining, а саме, логістичну регресію, SVM (Support Vector Machine) та наївний Баєсів класифікатор.

Задачу класифікації розв'язано на невеликому фрагменті корпусу українськомовних текстів, що містить 20 екземплярів, які було класифіковано трьома методами класифікації за темами: кримінал (Crime), спорт (Sport), наука (Science and IT), світові новини (World News) та економіка (Economics). На початковому етапі отримання вихідних даних створено табл. 1, яка містить результати виконаної автоматичної класифікації та результати класифікації, здійсненої експертом-людиною.

Далі, для порівняння наявних метрик оцінювання результатів ML класифікаторів та метрики статистики Коена було пораховано традиційні *F-score* та *accuracy* для кожного класифікатора окремо. Як цільове значення класу було застосовано інтелектуальне розмічування теми відповідної новини, отримане під час здійснення скрапінгу новинного сайта. Результати, представлені на рис. 1, свідчать про те, що відповідно до значень *F-score* та *accuracy* найбільш точним класифікатором є метод опорних векторів (SVM), а найменш точним — наївний Баєсів класифікатор, що, очевидно, зумовлено невеликим розміром оброблюваного корпусу.

Для наочного порівняння методів класифікації розраховано рівень узгодженості між кожним окремим класифікатором та експертом шляхом обчислення коефіцієнта каппа Коена. Для цього результати сумісної класифікації (див. табл. 1) перетворено у спеціальні матриці або таблиці оцінки об'єктів попарного зіставлення, що групуються після збору результатів експертного оцінювання за кожним методом класифікації окремо.

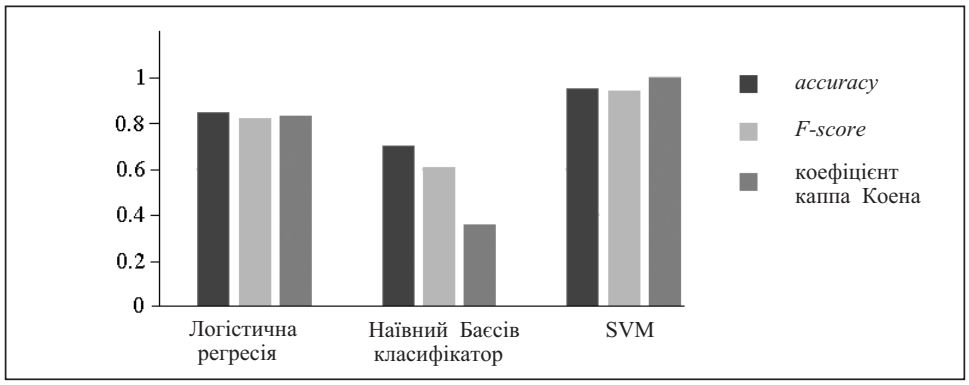

Рис. 1. Порівняння значень *accuracy*, *F-score* та коефіцієнта каппа Коена для оцінки результатів тематичної мультикласифікації текстів методами контрольованого машинного навчання

Таблиця 1. Фрагмент результатів тематичної мультикласифікації текстів, отриманої трьома методами контрольованого машинного навчання та експертним оцінюванням

Файл тексту	Логістична регресія	Наївний Баєсів класифікатор	SVM	Експерт-людина
149_uk_raw	Crime	Crime	Crime	Crime
133_uk_raw	Crime	Crime	Crime	Crime
378_uk_raw	Economics	Economics	Economics	Economics
32_uk_raw	Crime	Crime	Crime	Crime
255_uk_raw	Crime	Crime	Crime	Crime
498_uk_raw	Sport	Crime	Sport	Sport
319_uk_raw	Sport	Crime	Sport	Sport
575_uk_raw	Economics	Economics	Economics	Economics
8_uk_raw	Crime	Crime	Crime	Crime
469_uk_raw	Crime	Crime	Science and IT	Science and IT
80_uk_raw	Crime	Crime	Crime	Crime
111_uk_raw	Crime	Crime	Crime	Crime
206_uk_raw	Crime	Crime	Crime	Crime
573_uk_raw	Economics	Crime	Economics	Economics
177_uk_raw	Crime	Crime	Crime	Crime
128_uk_raw	Crime	Crime	Crime	Crime
550_uk_raw	Economics	Crime	World News	World News
416_uk_raw	Economics	Economics	Economics	Economics
182_uk_raw	Crime	Crime	Crime	Crime
316_uk_raw	Sport	Sport	Sport	Sport

Отже, для побудови матриць плутанини застосовано метод попарного зіставлення, коли об'єкти порівнюють один з одним попарно. Цей метод є засобом оцінювання та вибору рішень і широко використовується експертами під час статистичного оцінювання різних об'єктів [15]. У процесі побудови матриць здійснюється попарне порівняння кожної пари об'єктів та встановлюється якісний критерій оцінювання (збіг або його відсутність).

У табл. 2, 3 та 4 наведено матриці плутанини між експертною класифікацією та логістичною регресією, наївним Баєсовим класифікатором і методом опорних векторів відповідно.

Таблиця 2. Матриця плутанини результатів інтелектуальної експертної класифікації та логістичної регресії

Класифікатор (логістична регресія)	Інтелектуальна класифікація					
	Crime	Sport	Economics	World News	Science and IT	Разом
Crime	11	0	0	0	1	12
Sport	0	3	0	0	0	3
Economics	0	0	4	1	0	5
World News	0	0	0	0	0	0
Science and IT	0	0	0	0	0	0
Разом	11	3	4	1	1	20

Таблиця 3. Матриця плутанини результатів інтелектуальної класифікації та наївного Баєсового класифікатора

Класифікатор (наївний Басс)	Інтелектуальна класифікація					
	Crime	Sport	Economics	World News	Science and IT	Разом
Crime	11	2	1	1	1	16
Sport	0	1	0	0	0	1
Economics	0	0	3	0	0	3
World News	0	0	0	0	0	0
Science and IT	0	0	0	0	0	0
Разом	11	3	4	1	1	20

Таблиця 4. Матриця плутанини результатів інтелектуальної класифікації та SVM

Класифікатор (SVM)	Інтелектуальна класифікація					
	Crime	Sport	Economics	World News	Science and IT	Разом
Crime	11	0	0	0	0	11
Sport	0	3	0	0	0	3
Economics	0	0	4	0	0	4
World News	0	0	0	1	0	1
Science and IT	0	0	0	0	1	1
Разом	11	3	4	1	1	20

Розрахунок коефіцієнта каппа потребує обчислення спостережуваної та очікуваної точності. Спостережувана точність зазвичай є кількістю випадків, які були правильно класифіковані по всій матриці плутанини, тобто кількістю збігів думки експерта та автоматичного класифікатора. Щоб обчислити спостережувану точність, потрібно додати ті випадки, коли класифікатор і експерт зійшлися в думках (18 разів), а потім поділити на загальну кількість екземплярів (20). Інакше кажучи, спостережувана точність у випадку порівняння класифікатора «логістична регресія» з експертом становить 0.9.

Очікувана точність (випадкова узгодженість) — це точність, яку можна очікувати від будь-якого випадкового класифікатора або експерта на основі мат-

риці плутанини. У нашому прикладі 12 екземплярів були відібрані класифікатором як ті, що належать класу «Crime», а експерт у свою чергу відніс до цього класу 11 екземплярів. Інакше кажучи, очікувана точність віднесення до класу «Crime» має значення 6.6 ($11 \cdot 12 / 20 = 6.6$), до класу «Sport» — 0.45 ($3 \cdot 3 / 20 = 0.45$), до класу «Economics» — відповідно 1 ($4 \cdot 5 / 20 = 1$), до класів «World News» та «Science and IT» — 0. Отже, очікувана точність загалом — це додавання отриманих результатів з їхнім подальшим діленням на загальну кількість випадків. У результаті очікувана точність дорівнює $(6.6 + 0.45 + 1 + 0 + 0) / 20 = 0.4$.

Скориставшись формулою (3), отримуємо коефіцієнт каппа Коена оцінки рівня узгодженості результатів мультикласифікації експерта-людини та моделі логістичної регресії ML на рівні $((0.9 - 0.4) / (1 - 0.4)) = 0.83$.

Аналогічно за допомогою табл. 3 та 4 розраховано коефіцієнт каппа Коена для оцінювання наївного Баєсового класифікатора (0.36) та SVM (1).

Загальноприйнята шкала оцінки коефіцієнта каппа Коена є такою: $0.81 < K < 0.99$ — близько до ідеального рівня узгодженості (near perfect agreement); $0.61 < K < 0.80$ — значний рівень узгодженості (substantial agreement); $0.41 < K < 0.60$ — помірний рівень узгодженості (moderate agreement); $0.21 < K < 0.40$ — справедливий рівень узгодженості (fair agreement); $0.1 < K < 0.20$ — незначний рівень узгодженості (slight agreement). Спираючись на неї, можна побачити, що у першому експерименті найбільшої узгодженості досягнуто між експертом-людиною та класифікатором SVM. Тут рівень узгодженості (agreement) прирівнюється до повної та ідеальної (див. рис. 1)

4. ЗАСТОСУВАННЯ КОЕФІЦІЄНТА КАППА ДЛЯ ОЦІНЮВАННЯ РІВНЯ УЗГОДЖЕНОСТІ ДУМОК ДВОХ ЕКСПЕРТІВ

У другому експерименті досліджено міру узгодженості експертних думок за підсумками автоматичного вирівнювання речень паралельного казахсько-російського корпусу, що містить кримінальні новини.

На рис. 2 зображено головне вікно розробленої у попередніх дослідженнях програми, яке містить результати автоматичного вирівнювання речень казахсько-російського паралельного корпусу. Програму використовують для оцінювання експертом (носієм мови) правильності автоматичного вирівнювання.

Оскільки мови корпусу (казахська та російська) належать дуже різним мовним групам, у розробленні алгоритму їхньої автоматичної перевірки на семантичний збіг застосовують словниковий метод.

На першому етапі оброблення, який є морфологічним та синтаксичним обробленням, тексти корпусу поділяють на речення. На другому етапі словникового методу для кожного речення корпусу, згідно з порядком його слідуван-

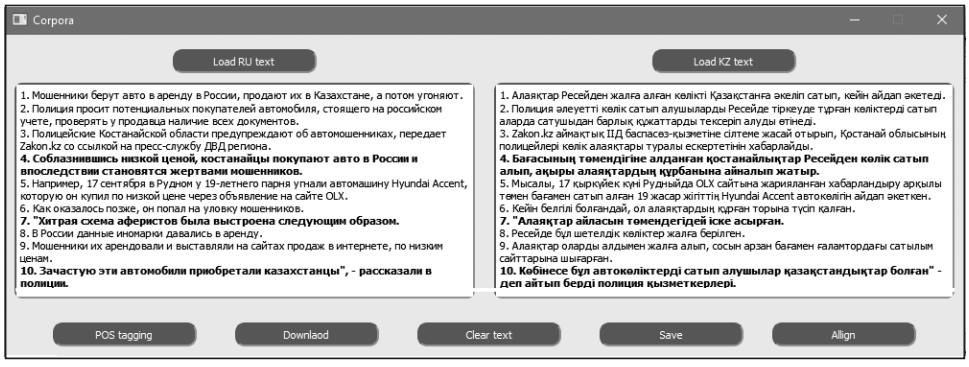

Рис. 2. Інтерфейс застосунку експертного оцінювання вирівняного паралельного корпусу

	Sentence_ID	Ru	Kz	Resul	Expert 1	Expert 2
2	1_zakon_20.07.2018_ru_raw.01	Глава государства поручил Касымову и Кожамжарову взять на контроль дело Дениса Тена.	Мемлекет басшысы Қасымов пен Қожамжаровқа Денис Теннің ісін бақылауға алуды тапсырды.	=	1	1
3	1_zakon_20.07.2018_ru_raw.02	Руководству Администрации Президента было поручено держать	Президент Әкімшілігінің Басшылығына тергеу барысын үнемі бақылауда ұстау	=	1	1
4	1_zakon_20.07.2018_ru_raw.03	генеральному прокурору Кайрату Кожамжарову и министру внутренних	қызметінің ақпаратына сүйене отырып хабарлауы бойынша, Мемлекет басшысы	=	1	1
5	1_zakon_20.07.2018_ru_raw.04	Президента было поручено держать ход расследования на постоянном	тергеу барысын үнемі бақылауда ұстау тапсырылды.	=	1	1
6	1_zakon_20.07.2018_ru_raw.05	Для расследования уголовного дела создана следственно-оперативная группа из числа наиболее опытных	Қылмыстық іс бойынша тергеу жүргізу үшін Алматы қаласы ІІМ және ІІД тәжірибелі қызметкерлерінен	=	1	1
7	1_zakon_20.07.2018_ru_raw.06	Убийцам Дениса Тена грозит пожизненное заключение.	бостандығынан айыру жазасы берілуі мүмкін.	≠	0	0
8	2_zakon_20.07.2018_ru_raw.01	За совершение убийства разыскивается Кудайбергенов Арман Бурибаевич.	Кісі өлтіргені үшін Құдайбергенов Арман Бөрібаев іздестірілуде.	≠	0	0
9	2_zakon_20.07.2018_ru_raw.02	МВД РК20 июля 2018, 11:00 Фотографию второго подозреваемого в убийстве Дениса Тена распространило	ҚР ІІМ 2018 жыл, 20 шілде 11:00 Zakon.kz ақпарат көзінің хабарлауы бойынша, ҚР ІІМ Денис Теннің өліміне	=	1	1
10	2_zakon_20.07.2018_ru_raw.03	За совершение убийства разыскивается Кудайбергенов Арман Бурибаевич, 1994 года рождения, уроженец	Кісі өлтіргені үшін Қызылорда облысының тумасы - 1994 жылғы Құдайбергенов Арман Бөрібаев іздестірілуде.	=	1	1

Рис. 3. Фрагмент таблиці з результатами експертної оцінки паралельності речень казахсько-російського корпусу

ня у тексті, здійснюють порівняння слів речення російською мовою зі словами речення казахською мовою. Для цього використовують казахсько-російський перекладний словник тематики корпусу, що містить 50000 слів. Висновок щодо семантичної еквівалентності двох речень роблять на основі збігів словникового перекладу слів російського речення зі словами відповідного казахського речення та з додатковим використанням шаблонів Pos-tagging.

Таблиця 5. Матриця прийнятих рішень

Рішення експерта 1	Рішення експерта 2	
	Так	Ні
Так	113	5
Ні	1	81

Розроблений алгоритм надав змогу автоматично вирівняти приблизно 70 % речень (у застосунку жирним шрифтом наведено речення, для якого автоматично не знайдено паралельного за перекладом речення у другому тексті). Але оскільки експерт визначив точність семантичної сумісності речень двох мов суб'єктивно, для підтвердження достовірності отриманого результату використано коефіцієнт узгодженості каппа Коена.

В експерименті взяли участь два експерти, які володіють казахською та російською мовами та є філологами за освітою. Їм було запропоновано оцінити результати роботи програми, при цьому шкала оцінювання враховувала два можливих варіанти: 0 — рішення про семантичну рівність змісту казахського та російського реченнь прийнято програмою неправильно, 1 — рішення прийнято програмою правильно. Фрагмент таблиці з результатами експертної оцінки паралельності речень корпусу, що є фрагментом матриці плутанини, наведено на рис. 3.

Наступним етапом створення статистики каппа Коена є формування матриці прийняття рішень, наведеної у табл. 5. У цій таблиці рядки відображають рішення першого експерта, а стовпці — рішення другого.

Спостережувана співвимірна узгодженість (обидва експерти сказали «так» або обидва експерти сказали «ні»): $P_0 = (113 + 81) / 200 = 0.97$, де кількість випадків, коли обидва експерти сказали «так», становить 113, а кількість випадків, коли обидва експерти сказали «ні», становить 81.

Обчислимо ймовірність випадкової узгодженості P_e. Зауважимо, що перший експерт сказав «так» 118 разів і «ні» 82 рази, а другий експерт сказав «так» 114 разів і «ні» 86 разів. Можна обчислити очікувану ймовірність того, що обидва експерти кажуть «так» випадково, за формулою $P_{\text{так}} = 118 / 200 \cdot 114 / 200 = 0.34$, а ймовірність того, що експерти кажуть «ні» випадково — за формулою

$P_{\text{ні}} = 82/200 \cdot 86/200 = 0.18$. Тоді ймовірність випадкової узгодженості $P_e = P_{\text{так}} + P_{\text{ні}} = 0.3 + 0.18 = 0.52$.

За отриманими результатами можна порахувати коефіцієнт каппа Коена: K = 0.93. Відповідно до загальноприйнятого розшифрування результатів, наведеного раніше, цей рівень узгодженості є близьким до «ідеального». Це означає, що рівень узгодженості думок двох експертів щодо семантичної сумісності речень російської та казахської мов, автоматично вирівняних у корпусі, досяг практично ідеального рівня.

ВИСНОВКИ

Результати проведеного в першому експерименті порівняння метрик *accuracy* та *F-score* з коефіцієнтом каппа Коена свідчать про те, що найвищого рівня узгодженості у результатах класифікації українськомовних текстів було досягнуто між класифікатором-людиною та методом SVM. Найменший рівень узгодженості було отримано у парі людина — наївний Баєсів метод класифікації. До того ж, наведений експеримент показує, що використання коефіцієнта каппа Коена замість традиційних *accuracy* та *F-score* надає змогу краще візуалізувати результати та більш наочно продемонструвати відмінність у застосуванні різних алгоритмів класифікації текстів.

У другому експерименті використано коефіцієнт каппа Коена. Показано, що він є ефективним для визначення рівня узгодженості думок декількох експертів щодо результатів роботи застосунку автоматичного оброблення текстів, тобто застосунку автоматичного вирівнювання речень у паралельному російсько-казахському корпусі.

Отже, проведене дослідження свідчить про те, що статистика каппа Коена, використана як інструмент спостереження міри узгодженості (agreement coefficient) між думками кількох експертів і між результатами автоматичного оброблення та цільовими значеннями, може слугувати для вимірювання надійності та достовірності проведених експериментальних досліджень NLP та Text Mining.

До того ж, ця міра є більш надійною та показовою, ніж простий розрахунок відсотка узгодженості, оскільки вона враховує можливість випадкового збігу значень. Використання метрики каппа Коена є також більш наочним і менше вводить в оману порівняно з використанням метрики точності (precision) у задачах оброблення текстів. Так, наприклад, спостережувана точність на рівні 80 % є менш разючою, якщо її порівнювати з очікуваною точністю, яка становить 75 %, а не з очікуваною точністю, що становить 50 %.

У дослідженні показано переваги використання статистики каппа не лише для оцінювання ефективності одного методу або моделі NLP чи Text Mining, а й для порівняння декількох методів між собою (як, наприклад, наведене у статті порівняння методів класифікації текстів). Ці результати доволі легко інтерпретувати та привести до потрібного наочного вигляду.

Отже, можна стверджувати, що завдяки зручності у використанні, простоті обчислення та високій точності отримуваних результатів, розрахунок коефіцієнта каппа Коена є одним з найкращих статистичних методів оцінювання рівня узгодженості під час експериментальних досліджень текстів.

СПИСОК ЛІТЕРАТУРИ

1. Lindstädt R., Proksch S.-O., Slapin J.B. When experts disagree: response aggregation and its consequences in expert surveys. *Political Science Research and Methods*. 2020. Vol. 8, Iss. 3. P. 580–588. https://doi.org/10.1017/psrm.2018.52.
2. Cohen J. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*. 1960. Vol. XX, N 1. P. 3746. https://doi.org/10.1177/001316446002000104.

3. Freitag R.M.K. Kappa statistic for judgment agreement in Sociolinguistics. *Revista de Estudos da Linguagem*. 2019. Vol. 27, N 4. P. 1591–1612. http://dx.doi.org/10.17851/2237-2083.0.0.1591-1612.
4. Franceschini F., Maisano D. Decision concordance with incomplete expert rankings in manufacturing applications. *Research in Engineering Design.* 2020. Vol. 31, Iss. 4. P. 471–490. https://doi.org/10.1007/s00163-020-00340-x.
5. Mielke P.W. Jr., Berry K.J., Johnston J.E. Unweighted and weighted kappa as measures of agreement for multiple judges. *International Journal of Management.* Vol. 26, N 2. 2009. P. 213–223.
6. Banerjee M., Capozzoli M., McSweeney L., Sinha D. Beyond Kappa: a review of interrater agreement measures. *The Canadian Journal of Statistics*. 2008. Vol. 27, Iss. 1. P. 3–23. https://doi.org/10.2307/3315487.
7. Gwet K.L. Handbook of inter-rater reliability. Gaithersburg: Advanced Analytics, LLC, 2014. 428 p.
8. Conger A.J. Integration and generalization of kappas for multiple raters. *Psychological Bulletin.* 1980. Vol. 88, Iss. 2. P. 322–328. https://doi.org/10.1037/0033-2909.88.2.322.
9. Nelson K.P., Edwards D. Measures of agreement between many raters for ordinal classifications. *Statistics in Medicine*. 2015. Vol. 34, Iss. 23. P. 3116–3132. https://doi.org/10.1002/sim.6546.
10. Ohyama T. Statistical inference of agreement coefficient between two raters with binary outcomes. *Communications in Statistics — Theory and Methods*. 2020. Vol. 49, Iss. 10. P. 2529–2539. https://doi.org/10.1080/03610926.2019.1576894.
11. Fleiss J.L. Measuring nominal scale agreement among many raters. *Psychological Bulletin*. 1971. Vol. 76, Iss. 5. P. 378–382. https://doi.org/10.1037/h0031619.
12. Light R.J. Measures of response agreement for qualitative data: Some generalizations and alternatives. *Psychological Bulletin.* 1971. Vol. 76, Iss. 5. P. 365–377. https://doi.org/10.1037/h0031643.
13. Khairova N., Kolesnyk A., Mamyrbayev O., Mukhsina K. The aligned Kazakh-Russian parallel corpus focused on the criminal theme. *Proc. 3rd International Conference on Computational Linguistics and Intelligent Systems (COLINS-2019)* (18–19 April 2019, Kharkiv, Ukraine). Kharkiv, 2019. P. 116–125.
14. Хайрова Н.Ф., Колесник А.С., Мамырбаев О.Ж., Мухсина К.Ж. Выровненный казахско-русский параллельный корпус, ориентированный на криминальную тематику. *Вестник Алматинского университета энергетики и связи.* 2020. № 1 (48). С. 84–92.
15. Nichols T.R., Wisner P.M., Cripe G., Gulabchand L. Putting the Kappa statistic to use. *The Quality Assurance Journal*. 2010. Vol. 13, Iss. 3–4. P. 57–61. https://doi.org/10.1002/qaj.481.

A.S. Kolesnyk, N.F. Khairova

JUSTIFICATION FOR THE USE OF COHEN'S KAPPA STATISTIC IN EXPERIMENTAL STUDIES OF NLP AND TEXT MINING

Abstract. Comparison of modern metrics for evaluating the agreement coefficients between the experimental results and expert opinion is made, and the possibility of using these metrics during experimental research in the field of automatic text processing using machine learning methods is estimated. The choice of Cohen's kappa coefficient as a measure of expert opinion agreement in the tasks of NLP and Text Mining is justified. An example of using Cohen's kappa coefficient for evaluating the level of agreement between the thought of an expert and the results of ML classification and measure of agreement of expert opinions in the alignment of sentences of the Kazakh–Russian parallel corpus is given. On the basis of this analysis, it is proved that the Cohen's kappa coefficient is one of the best statistical methods for determining the level of agreement in experimental studies due to its ease of use, simplicity of calculation and high accuracy of the results.

Keywords: Text Mining, NLP, Cohen's kappa statistic, agreement statistic, text classification with machine learning, parallel corpus.

Надійшла до редакції 10.09.2021