УДК 004.89
https://doi.org/10.47533/2023.1606-146X.26

Г. С. ЫБЫТАЕВА[1]*, О. Ж. МАМЫРБАЕВ[2], Н. Ф. ХАЙРОВА[3], К. Ж. МУХСИНА[2], Б. Ж. ЖҰМАЖАНОВ[2]

[1]Қ.И. Сәтбаев атындағы Қазақ ұлттық зерттеу техникалық университеті, Алматы, Қазақстан,
[2]Ақпараттық және есептеуіш технологиялар институты, Алматы, Қазақстан,
[3]«Харьков политехникалық институты» ұлттық техникалық университеті, Харьков, Украина
e-mail: ybytayeva.galiya@gmail.com, morkenj@mail.ru, nina.khairova@gmail.com, kuka_ai@mail.ru, bagasharj@mail.ru

САРАПШЫЛАР ПІКІРЛЕРІНІҢ КЕЛІСІМ ӨЛШЕМІ РЕТІНДЕ КОЭННІҢ КАППА КОЭФФИЦИЕНТІНІҢ ЕРЕКШЕЛІКТЕРІ

Эксперименттік нәтижелер мен сарапшылардың пікірі арасындағы келісім коэффициенттерін бағалауға арналған заманауи метрикалар салыстырылады және бұл метрикаларды машиналық оқыту әдістерімен мәтінді автоматты өңдеуде эксперименттік зерттеулерде пайдалану мүмкіндігі бағаланады. NLP және Text Mining есептерінде сарапшылық пікір келісімінің өлшемі ретінде Коэннің Каппа коэффициентін таңдау негізделген. Сарапшы пікірі мен ML классификациясының нәтижелері арасындағы келісім деңгейін бағалау үшін Коэннің Каппа коэффициентін қолданудың мысалы және қазақ-орыс параллель корпусының сөйлемдерін теңестіру кезінде сарапшы пікірлерінің келісім өлшемі келтірілген. Осы талдау негізінде Коэннің Каппа коэффициенті қолданудың оңайлығымен, есептеудегі қарапайымдылығымен және нәтижелердің жоғары дәлдігімен тәжірибелік зерттеулерде келісім деңгейін анықтаудың ең үздік статистикалық әдістерінің бірі екендігі дәлелденді.

Түйін сөздер: *Text Mining, NLP, Коэннің Каппа статистикасы, келісім статистикасы, мәтіндер классификациясы, параллель корпус.*

Кіріспе. Ғылым дамуының қазіргі жағдайында зерттеу нәтижелерін бағалау мәселесі өзекті болып табылады. Эксперименттердің нәтижелеріне сүйенуді және оларға сілтеме жасауды жалғастыру үшін олар толығымен объективті және мүмкіндігінше дәл болуы керек. Бұл талаптар ғылыми жұмыстың сандық нәтижелеріне де, сапалы зерттеу нәтижелеріне де қатысты.

Табиғи тілді өңдеу (NLP) немесе мәтінді өңдеу (Text Mining) мәселелерін шешуде белгілі бір корпуспен жүргізілген эксперименттердің дұрыстығын бағалау да мәтінді автоматты түрде өңдеуге қатысты негізгі күрделі міндеттердің бірі болып табылады. Әдетте, ақпаратты іздеу, жіктеу, кластерлеу, морфологиялық белгілеу, талдау және басқа тапсырмалардың әдістері мен модельдерін қолдану нәтижелері сарапшылардың қатысуымен немесе «алтын стандарт» деп аталатын, яғни машиналық оқыту (ML) тапсырмасына қатысты кейбір мақсатты мәндермен белгіленген корпуспен салыстыру арқылы бағаланады. Бұл жағдайда толықтық (completeness), дәлдік (accuracy) және F-өлшем (F-measure) сияқты метрикалар дәстүрлі түрде қолданылады. Бұл метрикалар біршама әмбебап және кез келген NLP зерттеуіне қолданылуы мүмкін.

* E-mail корреспондирующего автора: ybytayeva.galiya@gmail.com

Алайда, олардың әмбебаптығына байланысты оларды жеткілікті түрде объективті деп атауға болмайды.

Сараптамалық бағалау деп әрі қарай шешім қабылдау үшін сарапшылардың пікірі негізінде құбылысты немесе тұжырымдаманы бағалау процесі түсініледі [1]. Алайда, әдетте, тек жеке сараптамалық бағалау жеткіліксіз. Мысалы, тар бейінді міндеттерде классификатор жұмысын бағалау және олардың пікірлерінің сәйкестік деңгейін тексеру үшін берілген пәндік саладағы жекелеген сарапшылардың бірнеше тәуелсіз пікірлерін пайдалану қажет. Бұл тексеру сарапшылардың жұмысының сәйкестік деңгейін, сондай-ақ бағалаудың сенімділігі мен объективтілігін анықтауға мүмкіндік береді. Сарапшылардың немесе сарапшылық топтардың пікірлерінің сәйкес келуі эксперимент нәтижелерінің сапасының маңызды сипаттамаларының бірі болып табылады деп саналады. Бұл әлсіз келісім сарапшылардың жұмысының дұрыс еместігін көрсетуі мүмкін және сәйкесінше бүкіл зерттеудің нәтижелеріне күмән тудыруы мүмкін. Сондықтан эксперименттердің нәтижелерін жалпылауға көмектесетін сарапшылардың пікіріндегі үлкен немесе кіші айырмашылықтарды анықтау NLP сияқты әдістері мен модельдерін бағалау кезінде де, мәтінді интеллектуалды талдауда да міндетті болып табылады.

Сондықтан бұл зерттеуде сарапшылардың келісім дәрежесін бағалау үшін қол жетімді әдістер мен көрсеткіштер салыстырылады және оларды есептеу және корпустық лингвистиканың эксперименттік мәселелерінде қолдану мүмкіндігі негізделеді.

Бұл зерттеуде талдаудың екі түрі қолданылады. Біріншісі сарапшы мен мәтінді автоматты мәтіндік классификатор арасындағы келісім дәрежесін бірнеше ML әдістерімен анықтайды, ал екіншісі параллель корпусты автоматты түрде белгілеу бойынша бірнеше сарапшылар арасындағы келісім дәрежесін анықтайды. Бұл жағдайда автоматты таңбалау екі тілді корпустың екі бөлігінде семантикалық эквивалентті сөйлемдерді белгілеуді білдіреді.

Мақала келесідей құрылған. 1-бөлімде сарапшылардың пікірлерін бағалаудың заманауи статистикалық әдістері туралы әдебиеттерге шолу жасалады және NLP мен Text Mining мәселелерінде пікірлерді бағалау өлшемі ретінде Коэннің каппа коэффициентін таңдау негізделеді. 2-бөлімде жүргізілген тәжірибелер мен оларда қолданылған корпустар, сонымен қатар сарапшы мен ML әдістерін қолданатын автоматты классификатор арасындағы келісім деңгейін бағалау үшін каппа коэффициентін қолданудың эксперименттік талдауы сипатталған. 3-бөлімде қазақ-орыс параллель корпусының сөйлемдерін теңестіру туралы сарапшылардың пікірлерінің келісім дәрежесін бағалау үшін каппа коэффициентін қолдану қарастырылады. Қорытынды екі статистикалық эксперименттің нәтижелеріне негізделген зерттеуді қорытындылайды. Атап айтқанда, Коэннің каппа коэффициентін қолдану NLP және Text Mining мәселелеріндегі сарапшылар немесе сарапшылардың пікірі мен ML көрсеткіштері арасындағы келісім деңгейін бағалау үшін ең жақсы статистикалық көрсеткіш ретінде негізделген.

Әдебиетке шолу. Сарапшылардың пікірлерінің келісімін бағалаудың заманауи метрикалары. Есептеу және корпустық лингвистика саласындағы эксперименттік зерттеулерде қолдануға болатын сараптамалық пікірлердің келісімін

бағалаудың ең кең таралған статистикалық әдістері немесе метрикалары мыналар болып табылады:

– вариация коэффициенті;

– Кендалл келісімділік коэффициенті;

– Коэннің каппа коэффициенті.

V_j вариация коэффициенті әдетте пайызбен өлшенеді. Бұл жалпы топты бағалаудың орташа мәні туралы пікірлердегі айырмашылықтардың шартты көрсеткіші. Ол әрбір салыстырылатын объект үшін анықталады және *j*-ші объектінің салыстырмалы маңыздылығы туралы сарапшылардың пікірлерінің келісім дәрежесін сипаттайды:

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

мұндағы M_j – объектілерді бағалаудың орташа арифметикалық мәні, ал σ – *j*-ші объекті үшін алынған бағалаудың стандартты ауытқуы.

Вариация коэффициентінің мәні неғұрлым аз болса, сарапшылардың келісім дәрежесі соғұрлым жоғары болады деп болжанады. Егер коэффициент $V_j \le 0{,}30$ болса, онда келісім дәрежесі қанағаттанарлық болып саналады. Егер вариация коэффициенті $V_j \le 0{,}20$ болса, онда сарапшылардың келісім дәрежесі өте жақсы деп саналады. 0,25-тен аспайтын вариация коэффициентінің мәні қолайлы нәтиже болып табылады [2].

Келісімділік бағалаудың келесі коэффициенті Кендалл келісімділік коэффициенті болып табылады, ол объектілердің жиынтығы бірнеше ранг тізбегімен сипатталғанда қолданылады және зерттеуші осы реттіліктер арасында статистикалық байланыс орнатуы керек [3]. Әдетте бұл коэффициент келесі формула бойынша есептеледі:

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

мұндағы n – бағаланатын объектілердің саны; m – ранг тізбектерінің саны (сарапшылардың саны); және $D_i = d_i - \bar{d}$ – *i*-ші объект рангтарының қосындысының барлық объектілердің рангтарының орташа қосындысынан ауытқуы [4].

W келісімділік коэффициенті (1) 0-ден 1,0-ге дейінгі мәндерді қабылдай алады. Сарапшылардың пікірлерінің келісімі W = 1,0 үшін толық, W = 0,5 үшін қанағаттанарлық және W = 0,5 үшін тым төмен деп есептеледі. Кем дегенде 0,75 келісімділік коэффициенті рұқсат етілген. Кендалл рангтық корреляция коэффициенті екі деректер тізбегі арасындағы монотонды байланыстарды анықтаудың тиімді және сенімді әдісі болып табылады. Дегенмен, оның цифрлық деректерге қосылулар саны көп болған жағдайда қолданылуы мүмкін кванттауға байланысты қарама-қайшы нәтижелер береді [5].

Есептеу және корпустық лингвистика мәселелерінде осы екі коэффициентті (вариация коэффициенті және Кендалл келісімділік коэффициенті) пайдалану, әсіресе ұжымдық бағалауға және ортақ негіздердің болмауына қатысты өте қарама-қайшы нәтижелерге әкелуі мүмкін. Бұл көбінесе мәтіндерді семантикалық өңдеуде пайда болады, мысалы, семантикалық белгілеу, тар тақырып бойынша классификация немесе мәтіндерді кластерлеу жағдайында және т.б

Бұл мәселені шешудің тиімді жолы бірнеше әдістерді біріктіру болып табылады. Бұл коэффициенттер және олардың нұсқалары бастапқы коэффициентке негізделген, оны Коэн 1960 жылы енгізіп, каппа коэффициенті деп атады. Бұл коэффициент екі бағалаушы арасындағы келісімнің кездейсоқ реттелген номиналды шкаласын өлшеу үшін қолданылады. Басқаша айтқанда, Коэн номиналды шкала бойынша өз пікірлерін білдіретін екі бағалаушы арасындағы келісім өлшемі ретінде каппа коэффициентін пайдалануды ұсынды. Сонымен бірге сарапшылардың пікірлерінің келісім деңгейін анықтау мүмкіндігі [3] көрсетілген, бұл нәтижеге әсер ететін кездейсоқ келісімге мүмкіндік береді. Сондықтан да бұл бұрынғы екі коэффициент негізге алынған пайыздық келісімнің қарапайым есебіне қарағанда сенімді өлшем [6].

Каппа коэффициентін анықтаудың дәстүрлі формуласы келесідей:

$$K = \frac{P_0 - P_e}{1 - P_e}, \qquad (2)$$

мұндағы P_0 бақыланатын келісімнің кездейсоқ келісімнен қаншалықты жақсырақ екенін бағалауды көрсетеді, ал P_e кездейсоқ келісімді қоспағанда, максималды мүмкін келісімді есептеу нәтижесін көрсетеді.

Көптеген корреляция коэффициенттері сияқты каппа коэффициенті -1-ден +1-ге дейін өзгеруі мүмкін. Нөлден аз мәндер есепке алынғанымен, Коэн олардың аз ықтималды екенін дәлелдеді [7]. $K = 1$ үшін келісім мінсіз және керісінше, $K = 0$ үшін келісім мүмкіндікке эквивалентті деп есептеледі.

Жақында Коэннің дәстүрлі каппа статистикасы негізінде жалпылама келісім метрикалары әзірленді. Бұл метрикалар Коэннің каппа көрсеткішінің кеңейтілуі болып табылады немесе каппа формуласын қолданады, содан кейін басқа бағалау моделін енгізеді [8], әсіресе сарапшылар саны үштен асатын жерде. Мысалы, Флейс каппасы [9], Лайт пен Конгер каппасы [10, 11] және Миелке каппасы [12] осындай коэффициенттер болып табылады.

Сарапшылардың келісім дәрежесін таңдау кезіндегі ең көп тараған шектеу – талданатын деректер түрі. Мысалы, номиналды деректерді есептеу кезінде әдетте сарапшылар арасындағы келісім дәрежесі ретінде Крамердің V коэффициенті және басқа ассоциация коэффициенттері таңдалады. Олардың барлығы қатенің салыстырмалы төмендеуін көрсететін ассоциация дәрежесіне негізделген. Бұл жағдайда мәтіндік ақпаратты эксперименттік талдау нәтижелері реттік те, номиналды да емес болғандықтан, өлшем ретінде Коэннің каппа коэффициентін пайдалану керек.

Әдістер мен материалдар. Сарапшылардың пікірлерінің келісім дәрежесін анықтау және бірнеше әртүрлі классификациялық модельдер мен сарапшының нәтижелерін салыстыру үшін екі мәтіндік корпус әзірленді. Бірінші корпуста жаңалықтар сайттарынан алынған және қылмысқа, спортқа, ғылымға, әлем жаңалықтарына және экономикаға арналған украин тіліндегі мәтіндер бар. Екінші корпус – қылмыстық мәтіндердің параллель қазақ-орыс корпусы [13]. Мақаланы жазу кезінде екі корпусты толтыру процесі жүріп жатты.

Бірінші корпус қазіргі жаңалықтардың екі қазахстандық сайтының мазмұны негізінде әзірленді: ҚазАқпарат (https:/www.inform.kz) және Хабар 24 (https://www.24.kz). Зерттеу барысында 6000 мәтіннен тұратын корпустың бір бөлігі пайдаланылды,

оның 3000-ы қылмыс (Crime) туралы мәтіндер және 3000 мәтін спорт (Sports), ғылым (Science and IT), әлем жаңалықтары (World News) және экономика (Economics) сияқты әртүрлі тақырыптық категорияларға жатады.

Екіншісі – авторлар жасаған және қазақ пен орыс тілдерінде сөйлем деңгейінде реттелген қылмыстық маңызы бар мәтіндерден тұратын параллель қазақ-орыс корпусы. Мәтіндердің параллелизмі олардың әрбір жеке сөйлемдегі бірдей көлемдегі және мағынаның толық сәйкестігі ретінде түсініледі. Мәтіндерді жинау үшін Қазақстан Республикасының төрт қос тілді веб-порталы (zakon.kz, caravan.kz, lenta.kz және nur.kz) [14] пайдаланылды, олар басқалармен қатар елдегі қылмыстың жай-күйін қамтиды. Бұл порталдар екі тілді және тонау, көлік ұрлау, кісі өлтіру, жол-көлік оқиғасы және т.б. қылмыстар туралы қазақ және орыс тілдеріндегі қылмыстық жаңалықтарды қамтиды. Бұл тақырыптар құрылған корпустың негізгі ресурсын анықтайды.

Осы корпустардың негізінде NLP және Text Mining мәселелерін шешу үшін каппа статистикасын қолдану бойынша екі эксперименталды зерттеу жүргізілді. Бірінші зерттеу қазақ тіліндегі жаңалықтарды адамның зияткерлік белсенділігіне ұқсас жіктеуге арналған бірнеше алгоритмдердің жұмысын салыстырды. Эксперименттің міндеті автоматты политематикалық классификатор мен адам арасындағы қылмыс (Crime), спорт (Sports), ғылым (Science and IT), экономика (Economics) немесе әлем жаңалықтары (World News) тақырыптары бар жаңалықтар веб-сайтындағы мақалаға қатысты келісім деңгейін бағалау болды.

Зерттеудің екінші кезеңінің міндеті қылмыстық мәтіндердің параллель қазақ-орыс корпусының автоматты теңестіру сенімділігін бағалау болды. Бұрын мәтіндерді сөйлемдер бойынша автоматты түрде жұптық салыстыру [14] алгоритмін қолдану арқылы жүзеге асырылды. Бұл кезеңде қатарлас корпус мәтіндерінің әрбір жеке сөйлемі үшін толық мағыналық сәйкестік туралы бірнеше сарапшылардың пікірлерінің келісім деңгейін бағалау зерттелді.

Екі экспериментте де каппа статистикасы шатастыру матрицасына немесе қателік матрицасына негізделген, ол классификация алгоритмдерінің жұмысын қорытындылау үшін машиналық оқытуда жиі қолданылады. Бұл матрица автоматты классификатор-мен сарапшылық қорытындының сәйкестік санын немесе сарапшылардың бір-бірімен пікірлерін (true positive) және сарапшы мен автоматты классификатордың пікірлері сәйкес келмейтін жағдайлардың санын (false positive және false negative) көрсететін кесте болып табылады.

Автоматты тақырыптық классификатор нәтижелерінің бақыланатын және күтілетін дәлдігінің шатасу матрицасын суреттеу үшін машиналық оқытуды басқарудың үш ең кең тараған Text Mining әдісі, атап айтқанда, логистикалық регрессия, SVM (Support Vector Machine) және қарапайым Байес классификаторы пайдаланылды.

Классификация мәселесі 20 үлгіден тұратын қазақ мәтіндерінің корпусының шағын фрагментінде шешілді, олар келесі тақырыптар бойынша үш классификация әдісі бойынша жіктеледі: қылмыс (Crime), спорт (Sports), ғылым (Science and IT), әлем жаңалықтар (World News) және экономика (Economics). Бастапқы мәліметтерді алудың бастапқы кезеңінде автоматты классификация нәтижелері және адам сарапшысы жүргізген классификация нәтижелері көрсетілген 1-кесте құрылды.

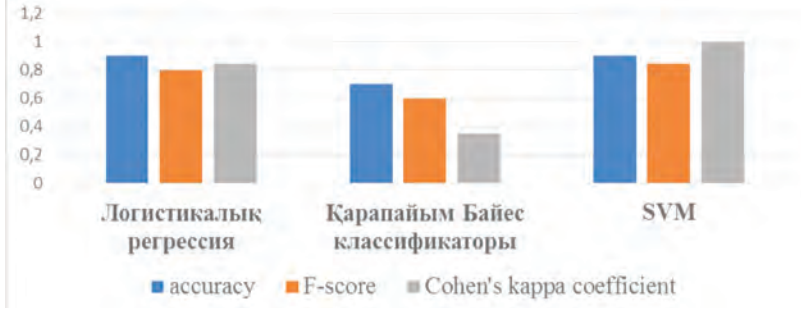
Әрі қарай, ML классификаторларының нәтижелерін бағалауға арналған қолжетімді көрсеткіштер мен Коэн статистикасының көрсеткіштерін салыстыру үшін әрбір классификатор үшін дәстүрлі *F-score* және *accuracy* бөлек есептелді. Класстың мақсатты мәні ретінде жаңалықтар сайтын скрейпинг кезінде алынған сәйкес жаңалықтар тақырыбының интеллектуалды таңбалау алынды. 1-суретте келтірілген нәтижелер *F-score* және *accuracy* мәндерінен көрініп тұрғандай, ең дәл классификатор – тірек векторлық машина (SVM), ал ең аз дәлдігі – қарапайым Байес классификаторы, бұл өңделетін корпустың шағын өлшеміне байланысты екені анық.

Классификация әдістерін салыстыруды көрсету үшін әрбір жеке классификатор мен сарапшы арасындағы келісім деңгейі Коэннің каппа коэффициентін есептеу арқылы анықталады. Бұл үшін үйлесімді классификация нәтижелері (1-кестені қараңыз) арнайы матрицаларға немесе жұппен салыстырылатын объектілерді бағалау кестелеріне түрлендіріледі, олар сарапшылық бағалау нәтижелерін жіктеудің әрбір әдісі бойынша жеке жинақтағаннан кейін топтастырылады.

Осылайша, шатастыру матрицаларын тұрғызу үшін жұптық салыстыру әдісі қолданылады, мұнда объектілер жұппен бір-бірімен салыстырылады. Бұл әдіс шешімдерді бағалау және таңдау құралы болып табылады және әртүрлі объектілерді статистикалық бағалауда мамандармен кеңінен қолданылады [15]. Матрицаларды құру процесінде объектілердің әрбір жұбы салыстырылады және сапалық бағалау критерийі белгіленеді (кездейсоқтық немесе оның болмауы).

2, 3 және 4-кестелер сәйкесінше сарапшы классификациясы мен логистикалық регрессия, қарапайым Байес классификаторы және SVM арасындағы шатасу матрицаларын ұсынады.

Каппа коэффициентін есептеу байқалатын және күтілетін дәлдікті есептеуді талап етеді. Бақыланатын дәлдік әдетте шатастыру матрицасы бойынша дұрыс жіктелген жағдайлардың саны болып табылады, яғни сарапшының пікірі мен автоматты классификатор арасындағы сәйкестіктер саны. Бақыланатын дәлдікті есептеу үшін классификатор пен сарапшы келіскен жағдайларды қосу керек (18 рет), содан кейін үлгілердің жалпы санына (20) бөлу керек. Басқаша айтқанда, «логистикалық регрессия» классификаторын сарапшымен салыстырған жағдайда байқалатын дәлдік (*accuracy*) 0,9 құрайды.



1-сурет – Машиналық оқытуды басқару әдістерімен тақырыптық көп класты мәтіндерді классификация нәтижелерін бағалау үшін *accuracy*, *F-score* және Коэннің каппа коэффициентін салыстыру.

1-кесте – Машиналық оқытуды басқарудың және сараптамалық бағалаудың үш әдісі арқылы алынған тақырыптық көп класты мәтіндерді классификация нәтижелерінің үзіндісі

Мәтіндік файл	**Логистикалық регрессия**	**Қарапайым Байес классификаторы**	**SVM**	**Адам маманы**
149_kz_raw	Crime	Crime	Crime	Crime
133_kz_raw	Crime	Crime	Crime	Crime
378_kz_raw	Economics	Economics	Economics	Economics
32_kz_raw	Crime	Crime	Crime	Crime
255_kz_raw	Crime	Crime	Crime	Crime
498_kz_raw	Sports	Crime	Sports	Sports
319_kz_raw	Sports	Crime	Sports	Sports
575_kz_raw	Economics	Economics	Economics	Economics
8_kz_raw	Crime	Crime	Crime	Crime
469_kz_raw	Crime	Crime	Science and IT	Science and IT
80_kz_raw	Crime	Crime	Crime	Crime
111_kz_raw	Crime	Crime	Crime	Crime
206_kz_raw	Crime	Crime	Crime	Crime
573_kz_raw	Economics	Crime	Economics	Economics
177_kz_raw	Crime	Crime	Crime	Crime
128_kz_raw	Crime	Crime	Crime	Crime
550_kz_raw	Economics	Crime	World news	World news
416_kz_raw	Economics	Economics	Economics	Economics
182_kz_raw	Crime	Crime	Crime	Crime
316_kz_raw	Sports	Sports	Sports	Sports

Күтілетін дәлдік (кездейсоқ келісім) – шатасу матрицасына негізделген кез келген кездейсоқ классификатордан немесе сарапшыдан күтуге болатын дәлдік. Біздің мысалда классификатор «Crime» класына жататын 12 үлгіні таңдады және сарапшы өз кезегінде осы класқа 11 үлгіні тағайындады. Басқаша айтқанда, «Crime» класына жатқызудың күтілетін дәлдігі 6,6 (11×12/20 = 6,6), «Sports» класына 0,45 (3×3/20 = 0,45), «Economics» класына 1-ге тең (4×5/20 =1), ал «World News» және «Science and IT» кластары үшін 0. Осылайша, күтілетін дәлдік тұтастай алғанда алынған нәтижелердің қосындысын және оларды одан әрі жағдайлардың жалпы санына бөлуді білдіреді. Нәтижесінде күтілетін дәлдік (6,6+0,45+1+0+0)/20=0,4 тең болады.

2-кесте – Интеллектуалды сарапшылық классификация және логистикалық регрессия нәтижелерінің шатасуы матрицасы

Классификатор (Логистикалық регрессия)	Интеллектуалды классификация					
	Crime	Sports	Economics	World news	Science and IT	Total
Crime	11	0	0	0	1	12
Sports	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
Total	11	3	4	1	1	20

3-кесте – Интеллектуалды жіктеу нәтижелері мен қарапайым Байес классификаторының шатасу матрицасы

Классификатор (Қарапайым Байес)	Интеллектуалды классификация					
	Crime	Sports	Economics	World news	Science and IT	Total
Crime	11	2	1	1	1	16
Sports	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
Total	11	3	4	1	1	20

4-кесте – Интеллектуалды классификация мен SVM нәтижелерінің шатасу матрицасы

Классификатор (SVM)	Интеллектуалды классификация					
	Crime	Sports	Economics	World news	Science and IT	Total
Crime	11	0	0	0	0	11
Sports	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
Total	11	3	4	1	1	20

((0,9 – 0,4) / (1 – 0,4)) = 0,83 деңгейінде адам сарапшысы мен логистикалық регрессияның ML моделінің көп класты классификациясы нәтижелерінің келісу деңгейін бағалау үшін (2) формуласын қолданып, Коэннің каппа коэффициентін аламыз.

Сол сияқты, 3 және 4 кестелерді пайдалана отырып, Коэннің каппа коэффициенті қарапайым Байес классификаторын (0,36) және SVM (1) бағалау үшін есептелді.

Коэннің каппа коэффициентін бағалаудың жалпы қабылданған шкаласы келесідей: 0,81< K< 0,99 – мінсіз келісімге жақын; 0,61< K< 0,80 – елеулі келісім; 0,41< K< 0,60 – қалыпты келісім; 0,21< K< 0,40 – әділ келісім; және 0,1< K< 0,20 – болмашы келісім. Оның негізінде біз бірінші экспериментте адам сарапшысы мен SVM классификаторы арасында ең үлкен келісімге қол жеткізілгенін көреміз. Мұнда келісім деңгейі толық және идеалға теңестіріледі (1-суретті қараңыз).

Нәтижелер және талқылау. Екінші эксперимент қылмыстық жаңалықтарды қамтитын параллель қазақ-орыс корпусының сөйлемдерін автоматты түрде теңестіру нәтижелері бойынша сараптамалық қорытындылардың келісім дәрежесін зерттейді.

2-суретте алдыңғы зерттеулерде жасалған бағдарламаның негізгі терезесі көрсетілген, онда қазақ-орыс параллель корпусының сөйлемдерін автоматты түрде теңестіру нәтижелері көрсетілген. Бағдарлама сарапшының (ана тілінде сөйлейтін) автоматты теңестіру дұрыстығын бағалау үшін қолданылады.

2-сурет **–** Теңестірілген параллель корпустың сараптамалық бағалау интерфейсі.

Корпус тілдері (қазақ және орыс тілдері) әртүрлі тілдік топтарға жататындықтан, олардың автоматты мағыналық сәйкестендіру алгоритмін жасау үшін сөздік әдісі қолданылады.

Морфологиялық және синтаксистік болып табылатын өңдеудің бірінші сатысында корпус мәтіндері сөйлемдерге бөлінеді. Сөздік әдісінің екінші кезеңінде корпустың әрбір сөйлемі бойынша оның мәтіндегі орын тәртібіне қарай орыс тіліндегі сөйлемдегі сөздер қазақ тіліндегі сөйлем сөздерімен салыстырылады. Ол үшін 50000 сөзден тұратын тақырыптық корпустың қазақша-орысша аударма сөздігі пайдаланылады. Екі сөйлемнің мағыналық эквиваленттілігі туралы қорытынды сөздік аудармасының орыс тіліндегі сөйлемнің сөздерімен сәйкес қазақ тіліндегі сөйлемнің сөздерімен сәйкес келуі және POS тегтеу үлгілерін қосымша қолдану негізінде жасалған.

Алгоритм сөйлемдердің шамамен 70%-ын автоматты түрде теңестіруге мүмкіндік берді (интерфейсте басқа мәтінде параллель аудармасы автоматты түрде табылмаған сөйлем қою шрифтпен бөлектелген). Дегенмен, сарапшы екі тілдегі сөйлемдердің мағыналық сәйкестігінің дұрыстығын субъективті түрде бағалағандықтан, алынған нәтиженің дұрыстығын растау үшін Коэннің каппа келісім коэффициенті пайдаланылды.

Экспериментке қазақ және орыс тілдерін білетін, мамандығы бойынша филолог екі маман қатысты. Оларға бағдарламаның нәтижелерін бағалау ұсынылды, бұл ретте бағалау шкаласы екі ықтимал нұсқаны ескерді: 0 (қазақ және орыс тіліндегі сөйлемдердің мағыналық теңдігі туралы шешім бағдарламамен қате қабылданған) және 1 (бағдарлама бойынша шешім дұрыс қабылданған). Шатасу матрицасының фрагменті болып табылатын корпустың сөйлем параллелизмін сараптамалық бағалау нәтижелерімен кестенің фрагменті 3-суретте көрсетілген.

Коэннің каппа статистикасын құрудың келесі қадамы 5-кестеде берілген шешім матрицасын құру болып табылады. Бұл кестеде жолдар бірінші сарапшының шешімін, ал бағандар екінші сарапшының шешімін көрсетеді.

Байқалған сәйкес келісім (екі сарапшы да «иә» деп жауап берді немесе екі сарапшы да «жоқ» деді) P_0 = (113 + 81) / 200 = 0,97, мұнда екі сарапшының да «иә» деп жауап берген жағдайларының саны 113 және екі сарапшы да «жоқ» деп жауап берген жағдайлар – 81.

Кездейсоқ келісім P_e ықтималдығын есептейік. Бірінші сарапшы 118 рет «иә», 82 рет «жоқ» деп жауап берсе, екінші сарапшы 114 рет «иә», 86 рет «жоқ» деп жауап бергеніне назар аударыңыз. Екі сарапшының да кездейсоқ «иә» деп айтуының күтілетін ықтималдығын $P_{иә}$ = 118 / 200 × 114 / 200 = 0,34 формуласы бойынша және сарапшылардың кездейсоқ «жоқ» деп айту ықтималдығын $P_{жоқ}$ = 82/200×86/200=0.18 формуласы бойынша есептей аламыз. Сонда кездейсоқ келісімнің ықтималдығы P_e = $P_{иә}$ + $P_{жоқ}$ = 0,3 + 0,18 = 0,52.

	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

3-сурет **–** Қазақ-орыс корпусының сөйлем параллелизміне сараптамалық бағалау нәтижелері көрсетілген кесте фрагменті.

5-кесте – Шешім қабылдау матрицасы

1-сарапшының шешімі	2-сарапшының шешімі	
	Иә	Жоқ
Иә	113	5
Жоқ	1	81

Алынған нәтижелерді пайдалана отырып, Коэннің каппа коэффициентін келесідей деп есептеуге болады: K = 0,93. Бұрын берілген нәтижелердің жалпы қабылданған түсіндірмесін ескере отырып, келісімнің бұл деңгейі «идеалдыға» жақын. Бұл екі сарапшының корпуста автоматты түрде теңестірілген орыс және қазақ тілдеріндегі сөйлемдердің мағыналық үйлесімділігі туралы пікірлерінің келісу деңгейі идеалды деңгейге дерлік жеткенін білдіреді.

Қорытынды. *Accuracy* мен *F-score* көрсеткіштерін Коэннің каппа коэффициентімен салыстыратын бірінші эксперименттің нәтижелері қазақ мәтіндерін жіктеуде адам классификаторы мен SVM әдісі арасында келісімнің ең жоғары деңгейіне қол жеткізілгенін көрсетеді. Келісімнің ең төменгі деңгейі қарапайым Бэйстің жіктеу әдісі – адам жұбында алынды. Сонымен қатар, жоғарыда келтірілген тәжірибе дәстүрлі *accuracy* пен *F-score* орнына Коэннің каппа коэффициентін пайдалану нәтижелерді жақсырақ визуализациялауға және мәтіндерді жіктеу үшін әртүрлі алгоритмдерді қолданудағы айырмашылықты айқынырақ көрсетуге мүмкіндік беретінін көрсетеді.

Екінші тәжірибеде Коэннің каппа коэффициентін қолдандық. Оның мәтінді автоматты өңдеу нәтижелері бойынша бірнеше сарапшылардың пікірлерінің келісім деңгейін анықтау, яғни параллель орыс-қазақ корпусында сөйлемдерді автоматты түрде теңестіру бағдарламасының жұмыс істеуі үшін тиімді екені көрсетілген.

Осылайша, зерттеу бірнеше сарапшылардың пікірлері арасындағы және автоматты өңдеу нәтижелері мен мақсатты мәндер арасындағы келісім коэффициентін бақылау құралы ретінде пайдаланылатын Коэннің каппа статистикасы эксперименттік NLP және Text Mining сенімділігін өлшеу үшін пайдаланылуы мүмкін деп болжайды.

Сонымен қатар, бұл өлшем келісім пайызының қарапайым есебіне қарағанда сенімдірек және индикативті болып табылады, өйткені ол мәндердің кездейсоқ сәйкестік мүмкіндігін ескереді. Коэннің каппа метрикасын қолдану мәтінді өңдеуде дәлдік көрсеткіштерін пайдаланудан гөрі айқынырақ және аз жаңылыстырады. Мысалы, 80% бақыланатын дәлдік күтілетін 50% дәлдікпен емес, күтілетін 75% дәлдікпен салыстырғанда аз таң қалдырады.

Зерттеу NLP немесе Text Mining бір әдісінің немесе моделінің тиімділігін бағалау үшін ғана емес, сонымен қатар бірнеше әдістерді бір-бірімен салыстыру үшін каппа статистикасын пайдаланудың артықшылықтарын көрсетеді (мысалы, мақалада ұсынылған классификацияны салыстыру әдістері). Бұл нәтижелерді түсіндіру және қажетті көрнекі көрініске дейін азайту өте оңай.

Осылайша, қолдану мен есептеудің жеңілдігі және нәтижелердің жоғары дәлдігі арқасында Коэннің каппа коэффициентін есептеу мәтіндердің эксперименталды талдауында келісім деңгейін бағалаудың ең үздік статистикалық әдістерінің бірі болып табылады деп айтуға болады.

Алғыс. Жұмыс Қазақстан Республикасы Ғылым және жоғары білім министрлігі Ғылым комитетінің қаржылық қолдауымен орындалды (NoAP09259309).

ӘДЕБИЕТ

1 Lindstädt R., Proksch S., Slapin J. When Experts Disagree: Response Aggregation and its Consequences in Expert Surveys / Political Science Research and Methods. 2020. – V. 8, – No. 3, – p. 580-588. doi:10.1017/psrm.2018.52

2 Cohen J. A coefficient of agreement for nominal scales Educational and Psychological Measurement. 1960. – V. XX, – No. 1, – p. 37-46. https://doi.org/10.1177/001316446002000104.

3 Freitag R.M. Ko. Kappa statistic for judgment agreement in sociolinguistics / Revista de Estudos da Linguagem. 2019. –V. 27, – No. 4, – p. 1591-1612. http://dx.doi.org/10.17851/2237-2083.0.0.1591-1612.

4 Franceschini F. and Maisano D. Decision concordance with incomplete expert rankings in manufacturing

5 Applications / Res. Eng. Design. 2020. – V. 31, – No. 4, – p. 471-490. https://doi.org/10.1007/s00163-020-00340-x.

6 Mielke Jr. P.W., Berry K.J. and Johnston J.E. Unweighted and weighted kappa as measures of agreement for multiple judges / Int. J. Manag. 2009. – V. 26, – No. 2, p. 213-223.

7 Banerjee M., Capozzoli M., McSweeney L., and Sinha D. Beyond kappa: A review of interrater agreement measures / The Canadian J. of Statistics. 2008. – V. 27, – No. 1, – p.3-23. https://doi.org/10.2307/3315487.

8 Gwet K.L. Handbook of Inter-Rater Reliability / Advanced Analytics, LLC, Gaithersburg, MD. 2014.

9 Conger A.J. Integration and generalization of kappas for multiple raters / Psychological Bulletin. 1980. – V. 88, – No. 2, – p. 322-328. https://doi.org/10.1037/0033-2909.88.2.322.

10 Nelson K.P. and Edwards D. Measures of agreement between many raters for ordinal classifications / Stat. Med. 2015.– V. 34, – No. 23, – p. 3116-3132. https://doi.org/10.1002/sim.6546.

11 Ohyama T. Statistical inference of agreement coefficient between two raters with binary outcomes / Communications in Statistics – Theory and Methods. 2020. – V. 49, – No. 10, – p. 2529-2539. https://doi.org/ 10.1080/03610926.2019.1576894.

12 Fleiss J.L. Measuring nominal scale agreement among many raters / Psychological Bulletin. 1971. – V. 76, – No. 5, – p. 378-382. https://doi.org/10.1037/h0031619.

13 Light R.J. Measures of response agreement for qualitative data: Some generalizations and alternatives / Psychological Bulletin. 1971. – V. 76, – No. 5, p. 365-377. https://doi.org/10.1037/h0031643.

14 Khairova N., Kolesnyk A., Mamyrbayev O., Mukhsina K. The aligned Kazakh-Russian parallel corpus focused on the criminal theme / Computational Linguistics and Intelligent Systems: Proc. 3rd Intern. Conf. (COLINS-2019) (Kharkiv, Ukraine, 18–19 April). 2019. – V. 1, – p. 116-125.

15 Khairova N., Kolesnik A., Mamyrbaev O., Mukhsina K. Aligned Kazakh-Russian parallel corpus, focused on the crime / Bulletin of Almaty University of Power Engineering and Telecommunications. 2020. –V. 1, – No. 48, – p. 84-92.

16 Nichols T. R., Wisner P. M., Cripe G. and Gulabchand L. Putting the Kappa statistic to use / The Quality Assurance J. 2010. – V. 13, – No. 3–4, – p. 57-61. https://doi.org/10.1002/qaj.481.

Г. С. ЫБЫТАЕВА[1], О. Ж. МАМЫРБАЕВ[2], Н. Ф. ХАЙРОВА[3], К. Ж. МУХСИНА[2], Б. Ж. ЖҰМАЖАНОВ[2]

[1] Казахский национальный исследовательский технический университет имени К. И. Сатпаева, Алматы, Казахстан,
[2] Институт информационных и вычислительных технологий, Алматы, Казахстан,
[3] Национальный технический университет «Харьковский политехнический институт», Харьков, Украина
e-mail: ybytayeva.galiya@gmail.com, morkenj@mail.ru, nina.khairova@gmail.com, kuka_ai@mail.ru, bagasharj@mail.ru

ОСОБЕННОСТИ КОЭФФИЦИЕНТА КАППА КОЭНА КАК МЕРЫ СОГЛАСИЯ МНЕНИЙ ЭКСПЕРТОВ

Сравниваются современные метрики для оценки коэффициентов согласия результатов эксперимента с мнением экспертов и оценивается возможность использования этих метрик в экспериментальных исследованиях при автоматической обработке текстов методами машинного обучения. Обоснован выбор коэффициента каппа Коэна в качестве меры согласия мнений экспертов в задачах NLP и Text Mining. Приведен пример использования коэффициента каппа Коэна для оценки уровня согласия мнения эксперта с результатами классификации ML и меры согласия мнений экспертов при выравнивании предложений казахско-русского параллельного корпуса. На основе этого анализа доказано, что коэффициент каппа Коэна является одним из лучших статистических методов определения уровня согласия в экспериментальных исследованиях благодаря простоте использования, простоте вычислений и высокой точности результатов.

Ключевые слова: *Text Mining, NLP, статистика каппа Коэна, статистика согласия, классификация текстов с помощью машинного обучения, параллельный корпус.*

G. S. YBYTAYEVA[1], O. ZH. MAMYRBAYEV[2], N. F. KHAIROVA[3], K. ZH. MUKHSINA[2], B. ZH. ZHUMAZHANOV[2]

[1] Kazakh National Research Technical University named after K.I. Satpayev, Almaty, Kazakhstan,
[2] Institute of Information and Computational Technologies, Almaty, Kazakhstan,
[3] National Technical University "Kharkov Polytechnic Institute", Kharkov, Ukraine
e-mail: ybytayeva.galiya@gmail.com, morkenj@mail.ru, nina.khairova@gmail.com, kuka_ai@mail.ru, bagasharj@mail.ru

FEATURES OF COHEN'S KAPPA COEFFICIENT AS A MEASURE OF EXPERT OPINION AGREEMENT

Modern metrics for evaluating agreement coefficients between the experimental results and expert opinion are compared, and the possibility of using these metrics in experimental research in automatic text processing by machine learning methods is assessed. The choice of Cohen's kappa coefficient as a measure of expert opinion agreement in the NLP and Text Mining problems is justified. An example of using Cohen's kappa coefficient for evaluating the level of agreement between the opinion of an expert and the results of ML classification and the measure of agreement of expert opinions in the alignment of sentences of the Kazakh-Russian parallel corpus is given. Based on this analysis, it is proved that 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, computing simplicity, and high accuracy of the results.

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