ҒТАХР 20.19.01 DOI: 10.52512/2306-5079-2021-86-2-136-144

ӘЛЕУМЕТТІК ЖЕЛІЛЕРДЕГІ ЖАЗБАЛАРЫ АРҚЫЛЫ МАШИНАЛЫҚ ОҚЫТУДЫ ҚОЛДАНЫП, АДАМДАРДЫҢ MBTI (MYERS-BRIGGS TYPE INDEX) ТИПІН АНЫҚТАУ

Ә.З. Суннатилла*, Е.С. Нурахов, А.А. Мыңжасар
Әл-Фараби атындағы Қазақ Ұлттық Университеті, Алматы, Қазақстан
e-mail*: asel.sunna@mail.ru

Бұл зерттеу адамдардың әулеметтік желілерде жариялаған мәтін негізінде психологиялық типін Myers-Briggs Type Index классификациясы бойынша анықтайтын машиналық оқыту әдістерін қолданып, классификатор жасауға бағытталған. Мақала тұлға типін анықтау тапсырмасын машиналық оқытуды пайдалану арқылы автоматтандыруды жүзеге асыруға негізделген, MBTI тұлға индикаторы арқылы жеке тұлғаның ерекшеліктерін анықтауға түсініктеме келтірілген. Машиналық оқытудың логистикалық регрессия, кездейсоқ орман және анықтамалық векторлар әдістері қолданылған, осыған ұқсас жұмыстарға әдеби талдау жасалған. Мақалада зерттеу жұмысының барысы мен әр классификатордың нәтижелері және қолданылған тәсілдердің талдауы берілген. Қазіргі карантиндік шектеулер жағдайында адамдардың онлайн жұмыс форматына ауысуына байланысты компанияларда кадрларды іріктеуде осындай зерттеулер үлкен көмегін тигізуі мүмкін, себебі зерттеу адамдардың жеке қасиеттерін әлеуметтік желідегі жазбаларына байланысты анықтауды көздейді. Бұл жұмыста қазақ тілі үшін қолдануда қарапайым, әрі көп есептеу қуаттылығын қажет етпейтін, ең тиімді машиналық оқыту алгоритмдері пайдаланылған және сәйкесінше әр әдіс үшін жұмыс нәтижелері келтірілген, келтірілген әдістердің ішінде анықтамалық векторлар әдісі арқылы қазақ тіліне арналған классификатордың дәлдігі мен сенімділігі жақсы деңгейде болды.

Түйін сөздер: *машиналық оқыту, тұлға ерекшеліктері, MBTI, әлеуметтік желілер, мәтінді өңдеу.*

Кіріспе

Жеке тұлғаны білу және түсіну арқылы көптеген артықшылықтарға қол жеткізуге болады. Технологияның қарқынды өсуімен бірге жеке тұлғаны білу автоматты түрде жүзеге асады. Психологиялық зерттеулер жеке тұлғаның кейбір ерекшеліктерінің лингвистикалық мінез-құлықпен өзара байланысын көрсетеді. Әлеуметтік желілердің қолданысына сүйене отырып, адамдардың жариялаған жаңалықтарына байланысты жеке тұлғаны болжау мүмкіндігі туып отыр.

Әр адамның әр түрлі хоббиі, қызығушылығы, жеке тұлғасы және ерекше адам ретінде қалыптастыратын басқа да сипаттамалары бар. Осы айырмашылықтардың ішінен оларды әртүрлі топтарға классификациялау мүмкіндігі туады, осылайша жарнаманың тиімділігін арттыру, маркетингтің басқа мақсаттары, жұмыс нәтижелерін өлшеу және басқа функцияларды жақсарту үшін қолдануға болады. Қазіргі уақытта адамдардың барлығы дерлік әлеуметтік желілердің ең болмағанда біреуін пайдаланады [1]. Сонымен қатар, олар әлеуметтік желілерде қызығушылығы немесе хоббиімен байланысты жазбалар жариялайды, бұл ақпарат кейінірек олардың жеке тұлғалық қасиеттерін білуге мүмкіндік береді. Сонымен қатар, қазіргі карантиндік шектеулер жағдайында адамдардың онлайн жұмыс форматына ауысуына байланысты компанияларда кадрларды іріктеуде осындай зерттеулер үлкен көмегін тигізуі мүмкін, себебі зерттеу адамдардың жеке қасиеттерін әлеуметтік желідегі жазбаларына байланысты анықтауды көздейді.

Кейбір әлеуметтік желі сайттары маркетингтің тиімділігін арттыру үшін көптеген әрекеттерді жасаса да, осы уақытқа дейін олар жеке адамдардың жариялаған ақпараттарына негізделіп болжанған тұлға ерекшеліктерін пайдалану арқылы маркетингтің тиімділігін жақсартуға бағытталған жабдықтарды жасаған жоқ. Қазіргі бар тәсілдер, әдетте қысқа мерзімді маркетингке негізделген және адамның интернеттегі іс-әрекеттеріне, cookie-файлдарға негізделген.

Жұмыста адамдарды сипаттау белгісі ретінде MBTI (Myers-Briggs type index) индикаторы таңдалынды. MBTI – Myers-Briggs индикаторы Катарин Кук Бриггс және оның қызы Изабел Бриггс Майерс әзірлеген, Карл Юнгтің психологиялық типтер теориясына негізделген жеке тұлға көрсеткіші. MBTI психологиялық тестілеу жүйесінің мәні – адамның жеке факторларының бірегей комбинацияларын өлшеу арқылы оның белгілі бір салаға бейімділігін, оның іс-әрекетінің стилін, шешімдерінің сипатын және өзіне ыңғайлы, әрі сенімді сезінуге мүмкіндік беретін басқа да

ерекшеліктерін болжауға болады. Бүгінгі күні, шетел тәжрибесіне сүйенсек, бұл индикатор жеке адамдар мен ұйымдар өздерін жақсы түсіну үшін немесе жұмыс орнының динамикасын оңтайландыру үшін қолданылатын жалпы құрал болып табылады. Ол бойынша адам 4 түрлі бинарлы шкаланың бір мәнімен бағаланады:

1. Адам санасының бағдары (E-I шкаласы): *Extrovert* – сыртқы әлеммен үнемі коммуникацияда болатын адамдар / *Introvert* – өзінің ішкі әлемімен оңашада болуды артық көретін адамдар.
2. Ақпаратты қалай қабылдауы (S-N шкаласы): *Sensing* – нақты фактілер мен тәжірибеге сүйенеді / *Intuition* – ішкі сезімдеріне назар аударатындар.
3. Шешімді қалай қабылдауы (T-F шкаласы): *Thinking* – логикаға сүйеніп, рационалды түрде шешім қабылдайтындар / *Feeling* – эмоционалды түрде адами құндылықтарға қарай шешім қабылдайтындар.
4. Өмір сүру стилі (J-P шкаласы): *Judging* – алдын-ала барлығын жоспарлап, сол бойынша әрекет етушілер / *Perceiving* – жағдайға қарай дайындықсыз әрекет ететін адамдар.

Әр адамның MBTI типі төрт санаттың жиынтығы ретінде анықталады. Мысалы, өз энергиясын көбіне басқа адамдардың ортасында болуынан алатын (E), әлемдегі ақпаратты түсіну үшін интуицияны пайдаланатын (N), рационалды шешім қабылдайтын (T) және алдын-ала жоспарлауды жөн көретін (J) адам, нәтижесінде ENTJ типіне ие болады. Сәйкес типті анықтаудың ең кең тараған тәсілі – жеке тұлғаны тестілеуден өткізу. Тест 93 сұрақтан тұрады.

Байланыс барған сайын әлеуметтік медиаға негізделген әлемде біз интернеттегі қолданушылар мен олардың жеке тұлғалары арасында тығыз байланыс бар-жоғын білуге мүдделіміз. Жеке тұлғалық қасиеттер пайдаланушылардың мінез-құлқы мен талғамын анықтауға мүмкіндік береді. Жеке тұлғаны танып білу жеке пайдаланушыға бағытталған персоналды жүйелер құру үшін маңызды ақпарат бере алады. Бұл жұмыста тұлғаны анықтау әдісі ретінде мәтіндік өңдеу тәсілдеріне ерекше назар аударылады. Негізгі мақсаты адамның MBTI типін оның әлеуметтік желідегі жазбаларының негізінде машиналық оқыту әдістері көмегімен болжау болып табылады. Алгоритм мәтіннің үзіндісін енгізілетін мәлімет ретінде қабылдап, болжанған MBTI типін шығарады (мысалы, ENTJ). Қойылған тапсырманы орындауда машиналық оқыту әдістері қарастырылған. Ең тиімді әдісті табу үшін олардың нәтижелік қателігі мен дәлдігі бойынша салыстыру және талдау жүргізілді.

Деректер Kaggle-дегі MBTI деректер жиынынан алынды. Мұнда 8675 пайдаланушыдың 45-50 ең соңғы әлеуметтік желі жазбаларының мәтіні және пайдаланушының MBTI типі ұсынылған.

Материалдар мен әдістер

Адамның тұлғасы оның талғамына және қызығушылықтарына әсер ететін шешуші фактор ретінде есептеледі. Мысалы, Раулинг және Сианкарели жеке тұлғалық ерекшеліктер және музыкалық жанр талғамы арасындағы байланысты анықтаған [2]. Бұл байланыстар тұлға туралы ақпарат бағдарламалық жасақтамаларды одан әрі дамытуға және персоналды сервис ұсыну үшін қолдануға болатындығын белгілейді. Ткальцик адамның жеке тұлғалық ерекшеліктерін рекомендациялық жүйелерде бастапқы ұсыну ауданын жақсартуға (жаңа пайдаланушыларға бастапқы ұсыну) [3], ал Ферверда музыка тыңдау жасақтамаларындағы интерфейсті музыкалық талғамға байланысты өзгертуге қолдануды ұсынды [4].

Тұлғаны анықтау үшін бірнеше модельдер жасалды. Бес факторлы модель (FFM) компьютерлік қоғамдастықта ең танымал және кең қолданылатын модель болып табылады және тұлғаны бес жалпы өлшемдерге (белгілерге) жіктейді: тәжірибеге ашықтық, адалдық, экстраверсия, келісімділік және невротизм. Алайда, кең көлемді және көп уақытты қажет ететін сауалнамаларды қолданбай, тұлғалық қасиеттерге ие болу әлі де күрделі міндет болып табылады.

Іс-әрекет мәліметтері негізінде жеке қасиеттерді қалай анықтауға болатындығы туралы зерттеулерге қазіргі кезде қызығушылық артуда. Зерттеулер көрсеткендей, жеке тұлғаны ұялы телефонды пайдалану деректерінен немесе акустикалық және визуалды белгілермен камералар мен микрофондар арқылы анықтауға мүмкіндік бар. Адамдардың өзара байланысының артуына байланысты, жақында жүргізілген зерттеулерде бейнеблогтар, Facebook-тегі іс-әрекет және профильдік суреттер сияқты мәліметтер қолданыла бастады.

Жеке тұлғаны суреттерден анықтау бойынша да зерттеулер жасалынған. Челли жеке тұлғаны анықтау үшін Facebook профиль суреттерінің мазмұнына назар аударды (мысалы, бет-әлпет, мимика,

жалғыз немесе басқалармен) [5]. Кристанидың жұмысы тұлғаны Flickr суреттерінің визуалды ерекшеліктерінен анықтауға болатындығын көрсетті [6].

MBTI классификациясының тұрақты тұлға типтеріне қатысты болжамды негізділігі туралы қазіргі кездегі пікірталастар бар. MBTI жүйесіне қарама-қарсы, психометрияда қолданылатын кең таралған тұлға типін жіктеу жүйесі - бұл Үлкен Бестік тұлғаны классификациялау жүйесі. Бұл жүйе жеке тұлғаның статистикалық бес ортогоналды өлшемдерін қарастырады: экстраверсия, келісімділік, ашықтық, саналылық және невротизм. MBTI-дан айырмашылығы, Үлкен Бес жүйесі статистикалық тұрғыдан жеке адамның өміріндегі өлшенетін белгілерге қатысты болжамды күшке ие. Алайда Пеннебакер мен Кингтің жұмыстары жеке тұлғаның бес қасиеттерінің төртеуі мен төрт MBTI өлшемдері арасындағы маңызды корреляцияны көрсетеді: ойлау/сезу келісімділікпен, бағалау/қабылдау саналылықпен, экстраверсия/интроверсия экстраверсиямен және сезім/түйсік ашықтықпен байланысы анықталды [7]. Бұл корреляциялар MBTI тұлғалық өлшемдері тұлғаның тұрақты ерекшеліктерін салыстырмалы түрде бейнелейді және жазушылық стиль мен тұрақты жеке қасиеттер арасындағы байланысты модельдеу әрекетін негіздейді.

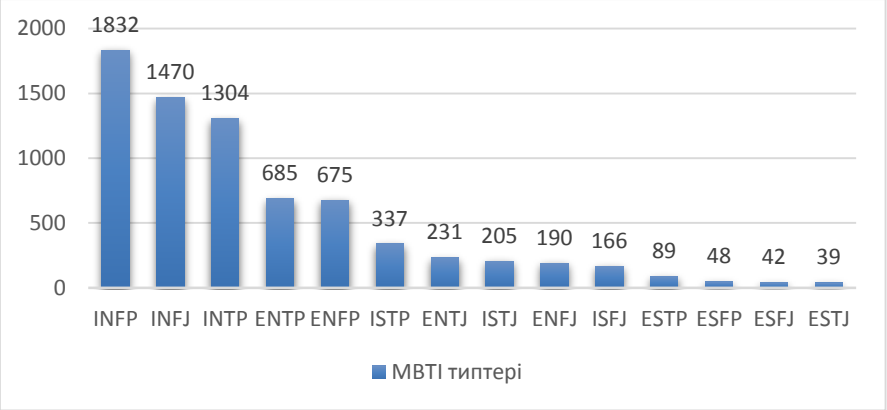
MBTI типін анықтау бойынша жұмыстар қазіргі кезде аз және мүмкіндіктері толық зерттелмеген. Дегенмен, машиналық оқыту саласында маңызды қадамдар жасалды. Машиналық оқыту саласында нейронды жүйелер MBTI типін анықтауда салыстырмалы кіші көлемдегі деректер жиынында табысты нәтижелер көрсетті [8], [9]. Гавриельску және Шампу жұмыстарында қолмен жазылған мәтіндік деректерге үш деңгейлі архитектура қолданған [10]. Бұл терең нейронды архитектуралары MBTI типін айтарлықтай дәлдікпен болжауға қабілетті деген тұжырымдаманың дәлелі. Карри Гуинн және Майк Комисин жұмысында да классикалық машиналық оқыту тәсілдерін, соның ішінде анықтамалық векторлары әдісін қолдану арқылы MBTI типін дәл анықтай алды [11]. Олардың зерттеулері адамның психологиялық ерекшеліктерін аңықтау үшін мәтіндік мәліметтердің өзі ғана жеткілікті болатынын дәлелі болып табылады.

Зерттеу барысы

Деректер жиынтығы екі бағаннан тұрады, біріншісі адамның типі, ал екіншісі бағанда сәйкес типтегі адамдар жазған мәтіндік хабарламалар жиынтығы. Бастапқы деректер жиыны ағылшын форумынан алынғандықтан, алдымен мәтіндер қазақ тіліне аударылды. Әлеуметтік желілерде адамдар мәтіндік қарым-қатынас кезінде жаргон сөздер мен белгілі бір стильді қолданады, сондықтан деректер алдын-ала өңдеуді қажет етеді. Алдын-ала өңдеу кезінде деректер жиынынан тыныс белгілері, веб-сілтемелер, сандар, хэштегтер және символдық белгілер алып тасталынды және бас әріптер кіші әріптерге ауыстырылды. Деректердегі әрбір сөз мүмкіндігінше мағыналы болуы үшін мәтіндік хабарламалар жиынтығында кездесетін «стоп-сөздер» («мен», «пен» және т.б. секілді шылаулар) және MBTI типтерінің атаулары да алып тасталынды.

Әрі қарай, NLP әдістері, соның ішінде TF-IDF-ті қолдану арқылы мәліметтердің белгілерін анықталды, осыдан кейін машиналық оқытуды белгілерге қолдану тапсырмасы қалады, алайда алдымен бізге қандай кластар бойынша жіктейтінімізді анықтап алу қажет. Бізде деректерді типтер бойынша үлестірудің екі түрлі нұсқасы бар. Бірі – толық 16 класты анықтап, солар бойынша оқыту. БҚЛ жағдайда, барлық MBTI типтері бойынша деректер жиынын 16 класқа бөлу қажет болады. Екіншісі – әр шкала бойынша бинарлы кластарға жіктеу, бұл кезде, біз мәліметтерді әр шкала бойынша екі класқа бөлеміз.

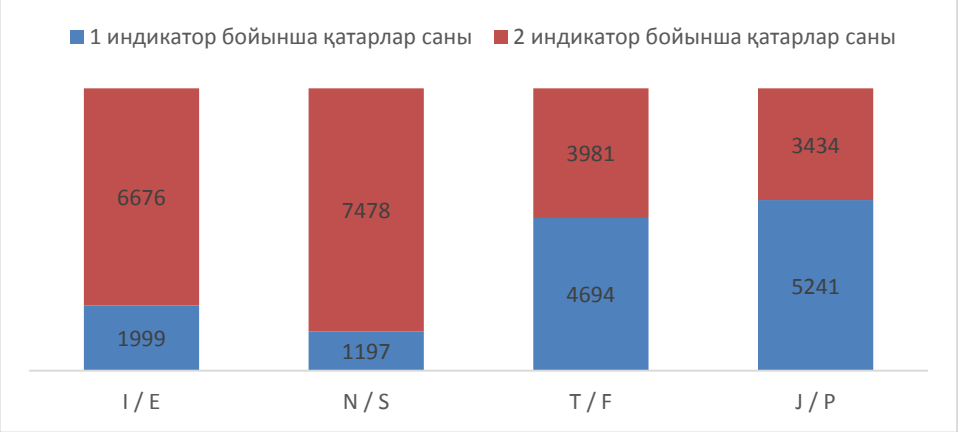
Толық 16 кластық тәсілдің бір кемшілігі – кластар арасындағы қабаттасулар көп және оларды бір-бірінен нақты ажырату қиынға соғады. Сонымен қатар, деректер жиыны барлық 16 тип бойынша біркелкі таралмаған. Мысалы, ең көп таралған INFP типі 1832 рет кездеседі, ал ең аз ISFJ тек 39 рет кездеседі. Біз осы нұсқаны пайдаланып, мәліметтерге машиналық оқытуды қолданғанда, классификатор басым типке (типтерге) сәйкес болжам жасауға бейім болады. Әлеуметтік медиа мәтіні қазірдің өзінде анық емес және әртүрлі деректерден тұратындығын ескере отырып, бұл өз кезегінде классификаторды өте ұқсас және шулы мәліметтер арасында ұсақ айырмашылықтарды іздеуге мәжбүр етеді, соңында нәтижесі нашар және тиімсіз болады. Берілген кластар іс жүзінде бір-бірінен толық тәуелсіз емес, бұл толық бөлінуді табуға тырысатын классификаторға кедергі келтіреді [12]. 1-диаграммада мәліметтер жиынындағы әр типтің кездесу жиілігі көрсетілген.



1- диаграмма. Деректер жиынында әр типтің кездесуі

Екінші тәсіл бойынша жеке тұлғаның төрт санатының әрқайсысы үшін екілік классификаторларды құру оңтайлы шешім болып табылады (мысалы, E / I, S / N, T / F, J / P), содан кейін MBTI жалпы болжамын алу үшін төрт нәтижені біріктіру ғана қажет (2-диаграмма). Бұл келесідей артықшылықтарды береді:

- Нақты айырмашылықтарды ажырата отырып, дәлдігі күрт жақсаратын мәліметтер беріледі.
- 16 класқа қарағанда екіге бөлінгенде (мысалы, E / I), әр класс үшін деректер жиыны үлкейеді.



2- диаграмма. Әр жеке класс үшін деректер жиынындағы қатарлар саны

Әрбір жеке қасиеттерге әртүрлі болжам жасау арқылы маңызды нәтиже алынады. Тағы да бір ескеру қажет фактор, бұл – әр шкаладағы индикаторлардың тәуелсіздігі. Оларды еркін түрде бөлу мүмкіншілігі болуы үшін әр шкала өзара тәуелсіз болуы керек. Мысалы, адамның экстраверт немесе интроверт болуы оның шешім қабылдауда ақылмен объективті түрде немесе эмоцияға сүйеніп қабылдауына тәуелді болмауы керек. Мұны 1-кестедегі шкала мәндері арасындағы корреляция арқылы тексеруге болады.

1-кесте. Деректер жиынындағы шкалалар арасындағы корреляция

Шкала	I/E	N/S	T/F	J/P
I/E	1.0000	-0.0458	-0.0695	0.1619
N/S	-0.0458	1.0000	-0.0809	0.0149
T/F	-0.0695	-0.0809	1.0000	-0.0046
J/P	0.1619	0.0149	-0.0046	1.0000

Деректердің барлығын екі бинарлы топтарғы бөліп алғаннан кейін оларға мәтінді алдын ала өңдеу, яғни мәтіндегі тыныс белгілері мен стоп сөздерді (шылау, одағай сияқты мағынасы жоқ сөздер) алып тасталды. Себебі мұндай сөздер бізге қажетті негізгі сөздерді анықтауда кедергі келтіруі мүмкін. Одан соң тазаланған мәтіннен TF-IDF арқылы белгілерін анықтау операциясы орындалды. Классификация тапсырмасын орындау барысында бинарлы деректердің 80%-ы оқыту жинағы үшін,

ал қалған 20%-ы тестік жинақ үшін екіге бөлінді. Сондай-ақ әр кезеңде бір қатарлар алынбас үшін деректерді бөлуде кросс-валидация пайдаланылды. Кросс-валидация – модельде қолданылатын статистикалық талдаудың тәуелсіз мәліметтер жиынтығында қаншалықты сәтті жұмыс істейтінін тексеруге арналған әдіс [13]. Әдетте кросс-валидация мақсаты болжау болып табылатын жағдайларда қолданылады және болжамды модель іс жүзінде қаншалықты жұмыс істей алатындығын бағалай алады. Бір кросс-тексеру циклі мәліметтер жиынтығын бөліктерге бөлуді, содан кейін бір бөлікке модель құруды (оқыту жиынтығы деп аталады) және басқа бөлікке үлгіні тексеруді (тест жиынтығы деп аталады) қамтиды. Нәтижелердің таралуын азайту үшін кросс-валидацияның әртүрлі циклдері әртүрлі бөлімдерде жүзеге асырылады, ал валидация нәтижелері барлық циклдерде орташа болады.

Нәтижелері

Машиналық оқыту әдістері

Тұлғалар типіне жіктеу тапсырмасын жүзеге асыру үшін машиналық оқытудың логистикалық регрессия, кездейсоқ орман әдісі, анықтамалық вектор әдістері қолданылды.

Логистикалық регрессия

Логистикалық регрессия - бұл машиналық оқытудың сызықтық классификация алгоритмі, ол болжамдық талдау алгоритмі және ықтималдық тұжырымдамасына негізделген.

Бұл әдіс көптеген белгілердің мәндеріне сәйкес белгілі бір оқиғаның ықтималдығын болжау үшін қолданылады. Ол үшін тәуелді айнымалы Y енгізіледі, ол екі мәннің біреуін ғана қабылдай алады – әдетте бұл 0 (оқиға болған жоқ) және 1 (оқиға болған) сандары және көптеген тәуелсіз айнымалылар (белгілер, предикторлар немесе регрессорлар деп те аталады) – $x_{1,} \dots, x_n$ нақты сандары, олардың негізінде тәуелді айнымалының белгілі бір мәнін қабылдау ықтималдығын есептеу қажет. Құжаттарды жіктеу жағдайында тәуелді айнымалының рөлін c_i категориясы орындайды, ал тәуелсіз айнымалылардың рөлін $d_1, \dots, d_n$ құжаттар жиынтығы атқарады.

Жалпы айтқанда, логистикалық регрессияны сигмоидты активтендіру функциясы бар бір қабатты нейрондық желі ретінде ұсынуға болады, оның салмағы логистикалық регрессия коэффициенттері болады.

Кездейсоқ орман әдісі

Кездейсоқ орман әдісі - бұл икемді, қолдануға оңай машиналық оқыту алгоритмі, тіпті гиперпараметрлерді баптаусыз да, көбінесе жақсы нәтиже береді. Бұл - сондай-ақ қарапайымдылығы және әртүрлілігімен (оны классификация және регрессиялық тапсырмалар үшін де қолдануға болады) ерекшеленетін ең көп қолданылатын алгоритмдердің бірі [14]. Қарапайым сөзбен айтқанда, кездейсоқ орман бірнеше шешім ағаштарын тұрғызады және дәлірек және тұрақты болжам жасау үшін оларды біріктіреді.

Жалпы, шешім ағашын талдау – бұл көптеген салаларда қолдануға болатын болжамды модельдеу құралы. Шешім ағаштарын алгоритмдік тәсіл арқылы салуға болады, ол мәліметтер жиынтығын жағдайларға байланысты әртүрлі жолмен бұзуы мүмкін [15]. Оларды жіктеу тапсырмалары үшін де, регрессия үшін де қолдануға болады.

Анықтамалық вектор әдісі (SVM)

Анықтамалық вектор әдісі (SVM) - өте сапалы, сенімді және реттеліп бақыланатын оқыту алгоритмдерінің бірі. Бұл әдіс логистикалық регрессия сияқты, деректер жиынтығындағы кластарды бөлетін гипержазықтықты табуға тырысады. Анықтамалық вектор алгоритмінің мақсаты - деректер нүктелерін анық жіктейтін N өлшемді кеңістіктегі гипержазықтықты табу (N - мүмкіндіктер саны). Деректер нүктелерінің екі класын бөлу үшін көптеген гипержазықтықтарды таңдауға болады. Мақсат - максималды шегі бар жазықтықты табу, яғни екі кластың деректер нүктелерінің арасындағы ең үлкен арақашықтығы бар гипержазықтықты таңдау. Шектегі қашықтықты максималды ету болашақ деректерді сенімді түрде жіктеуге болатындай етіп күшейтуді қамтамасыз етеді [16].

Гипержазықтықтар - бұл деректер нүктелерін жіктеуге көмектесетін шешім шекаралары. Гипержазықтықтың бір жағындағы деректер бірінші класқа, ал екінші жағындағылары екінші класқа тиесілі болады. Сондай-ақ, гиперпланның өлшемі ерекшеліктердің санына байланысты.

Іс жүзінде деректердің құрылымы жиі белгісіз және бөлу гипержазықтықтарын дұрыс құру өте сирек кездеседі, яғни үлгінің сызықтық бөлінуіне кепілдік беруге болмайды. Алгоритм бір класқа жататын құжаттар болуы мүмкін, бірақ іс жүзінде олар керісінше болуы керек. Мұндай деректер шығарындылар деп аталады, олар әдіс қатесін жасайды, сондықтан оларды елемеген дұрыс. Бұл сызықтық бөлінбеу мәселесінің мәні.

Нәтижелері

2-кесте. Әр әдіс бойынша болжау дәлдігінің пайыздық көрсеткіші

	I/E	N/S	F/T	J/P
Логистикалық регрессия	76.43%	86.80%	72.16%	60.29%
Кездейсоқ орман әдісі	76.14%	86.69%	67.03%	59.31%
Анықтамалық вектор әдісі	76.14%	86.69%	72.16%	60.69%

1-кестеде 2р қолданылған әдіс бойынша алынған нәтижелер классификаторлардың дәлдіктері ретінде берілген. Логистикалық регрессия және анықтамалық векторлар машиналық оқыту әдістерінің пайыздық көрсеткіштері өте жоғары мәндерді көрсетті. Мәтіннің қысқалығы мен осындай қысқаша мәтіндегі негізгі ақпаратты жинаудағы қиындықтарды ескере отырып, қол жеткізілген дәлдігіміз әсерлі көрінеді. Әртүрлі әлеуметтік сипаттағы адамдардың салыстырмалы түрде қысқа жеке әлеуметтік медиа жазбаларында кездесетін тілді қолдану тәсілдерінде үлкен айырмашылықтардың бар болуы таңғаларлық болып табылады.

Талқылау мен шектеулер

Әулеметтік желілердегі мәтіндік мәліметтер алуан түрлі формада кездеседі (SMS/жедел хабарламалар, Facebook/Instagram/форум хабарламалары, Twitter жазбалары, блог жазбалары, мақалалар және т.б.). Олардың әрқайсысы кездейсоқ және бірнеше қысқа хабарламалардан бастап ресми жазбалардың үлкен бөліктеріне дейін өздерінің жеке жазу стильдерін ұстанады. Жоғарыда айтылғандай, деректер жиынтығы жазу стилінің бір түрі ғана болып табылатын форум хабарламаларынан алынған. Модель адамның қысқа хабарламалары немесе ұзақ мақалаларынан гөрі адамның форумдарда жазған хабарламалары арқылы адамның MBTI типін дәлірек болжай алады.

Бұл зерттеуде MBTI индивидуалды типі негізінде тұлғаның типін болжау процесін автоматтандыруға арналған машиналық оқытудың логистикалық регрессия, кездейсоқ орман және анықтамалық вектор әдістері зерттелді. Мәтінді алдын ала өңдеу құралы ретінде NLTK пайдаланылды. Зерттеу бойынша шет тілдері үшін, атап айтқанда ағылшын, неміс, испан, қытай корей т.б. тілдер үшін мәтін сипатына қарап, тұлға типін анықтау мәселесі жақсы зерттелген [17], әлі күнге дейін жетілдірілуде. Алайда, қазақ тілі үшін жасалған жұмыстар аз деп айтуға болады. Бұл жұмыста қазақ тілі үшін қолдануда қарапайым, әрі көп есептеу қуаттылығын қажет етпейтін, ең тиімді машиналық оқыту алгоритмдері пайдаланған және сәйкесінше әр әдіс үшін жұмыс нәтижелері келтірілген. Бұл жұмыста келтірілген әдістердің ішінде анықтамалық векторлар әдісі арқылы қазақ тіліне арналған классификатордың дәлдігі мен сенімділігі жақсы деңгейде болды.

Қорытынды

Бұл зерттеуде MBTI индивидуалды типі негізінде тұлғаның типін болжау процесін автоматтандыруға арналған машиналық оқытудың логистикалық регрессия, кездейсоқ орман және анықтамалық вектор әдістері зерттелді. Мәтінді алдын ала өңдеу құралы ретінде NLTK пайдаланылды.

Пайдаланылған әдебиеттер

1. Стивенсон М. Введение в нейролингвистическое программирование
2. Rawlings D., Ciancarelli V. (1997) Music preference and the five-factor model of the neo personality inventory. *Psychology of Music*. P. 120–132
3. Ferwerda B., Schedl M., Tkalcic M. (2015) Personality & emotional states: Understanding users' music listening needs. UMAP 2015 Extended Proceedings
4. Ferwerda B., Schedl M. (2014) Enhancing music recommender systems with personality information and emotional states: A proposal. Proc. EMPIRE workshop.
5. Celli F., Bruni E., Lepri B. (2014) Automatic personality and interaction style recognition from Facebook profile pictures. Proceedings of the ACM International Conference on Multimedia. P. 1101–1104

6. Cristani M., Vinciarelli A., Segalin C., Perina A. (2013) Unveiling the multimedia unconscious: Implicit cognitive processes and multimedia content analysis. Proceedings of the 21st ACM international conference on Multimedia.

7. Pennebaker J.W., King L.A. (1999) Linguistic Styles: Language Use as an Individual Difference. *Personality and Social Psychology*. 77(6). P. 1296–1312

8. Hernandez R., Knight I.S. (2017) Predicting Myers-Briggs Type Indicator with Text Classification. 31st Conference on Neural Information Processing Systems, NIPS.

9. Gavrilescu M. (2015) Study on determining the Myers-Briggs personality type based on individual's handwriting. The 5th IEEE International Conference on E-Health and Bioengineering.

10. Majumder N., Poria S., Gelbukh A., Cambria E. (2017) Deep learning-based document modeling for personality detection from text. IEEE Computer Society, IEEE Intelligent Systems. https://sentic.net/deep-learning-based-personality-detection.pdf

11. Komisin M., Guinn C. (2012) Identifying personality types using document classification methods. Proceedings of the 25th International Florida Artificial Intelligence Research Society Conference, FLAIRS-25. P. 232–237

12. Ингерсолл Г.С., Мортон Т.С., Фэррис Э.Л. (2015) Обработка неструктурированных текстов. Поиск, организация и манипулирование. / Пер. с англ. Слинкин А.А. М.: ДМК Пресс. – 414 с.

13. Harrington R., Loffredo D.A. (2010) MBTI personality type and other factors that relate to preference for online versus face-to-face instruction. *The Internet and Higher Education*. Volume 13, Issues 1–2, pp. 89-95

14. Verhoeven, B., Daelemans, W., Plank, B. (2016) TwiSty: A Multilingual Twitter Stylometry Corpus for Gender and Personality Profiling. Proceedings of the 10th edition of the Language Resources and Evaluation Conference European Language Resources Association (ELRA)

15. Friedman J.H. (2001) Greedy function approximation: A gradient boosting machine. *The Annalls of Statistics*. 29(5). 1189–1232.

16. Gallo F.R., Simari G.I., Martinez M.V., Falappa M.A. (2020) Predicting user reactions to Twitter feed content based on personality type and social cues. *Future Generation Computer Systems*, volume 110, p. 918-930.

17. Bencke L., Cechinel C., Munoz R. (2020) Automated classification of social network messages into Smart Cities dimensions. *Future Generation Computer Systems*, volume 109, p. 218-237.

References

1. Stivenson M. Vvedenie v nejrolingvisticheskoe programmirovanie [Introduction to neuro-linguistic programming] [in Russian]

2. Rawlings D., Ciancarelli V. (1997) Music preference and the five-factor model of the neo personality inventory. Psychology of Music. P. 120–132

3. Ferwerda B., Schedl M., Tkalcic M. (2015) Personality & emotional states: Understanding users' music listening needs. UMAP 2015 Extended Proceedings

4. Ferwerda B., Schedl M. (2014) Enhancing music recommender systems with personality information and emotional states: A proposal. Proc. EMPIRE workshop.

5. Celli F., Bruni E., Lepri B. (2014) Automatic personality and interaction style recognition from Facebook profile pictures. Proceedings of the ACM International Conference on Multimedia. P. 1101–1104

6. Cristani M., Vinciarelli A., Segalin C., Perina A. (2013) Unveiling the multimedia unconscious: Implicit cognitive processes and multimedia content analysis. Proceedings of the 21st ACM international conference on Multimedia.

7. Pennebaker J.W., King L.A. (1999) Linguistic Styles: Language Use as an Individual Difference. Personality and Social Psychology. 77(6). P. 1296–1312

8. Hernandez R., Knight I.S. (2017) Predicting Myers-Briggs Type Indicator with Text Classification. 31st Conference on Neural Information Processing Systems, NIPS.

9. Gavrilescu M. (2015) Study on determining the Myers-Briggs personality type based on individual's handwriting. The 5th IEEE International Conference on E-Health and Bioengineering.

10. Majumder N., Poria S., Gelbukh A., Cambria E. (2017) Deep learning-based document modeling for personality detection from text. IEEE Computer Society, IEEE Intelligent Systems. https://sentic.net/deep-learning-based-personality-detection.pdf

11. Komisin M., Guinn C. (2012) Identifying personality types using document classification methods. Proceedings of the 25th International Florida Artificial Intelligence Research Society Conference, FLAIRS-25. P. 232–237

12. Ingersoll G.S., Morton T.S., Ferris E.L. (2015) Obrabotka nestrukturirovannykh tekstov. Poisk, organizacziya i manipulirovanie. / Per. s angl. Slinkin A.A. M.: DMK Press. – 414 s. [Processing of unstructured texts. Search, organization, and manipulation. / Translated from English. Slinkin A. A. M.: DMK Press - 414 p.] [in Russian]

13. Harrington R., Loffredo D.A. (2010) MBTI personality type and other factors that relate to preference for online versus face-to-face instruction. The Internet and Higher Education. Volume 13, Issues 1–2, pp. 89-95

14. Verhoeven B., Daelemans W., Plank B. (2016) TwiSty: A Multilingual Twitter Stylometry Corpus for Gender and Personality Profiling. Proceedings of the 10th edition of the Language Resources and Evaluation Conference European Language Resources Association (ELRA)

15. Friedman J.H. (2001) Greedy function approximation: A gradient boosting machine. The Annalls of Statistics. 29(5). 1189–1232.

16. Gallo F.R., Simari G.I., Martinez M.V., Falappa M.A. (2020) Predicting user reactions to Twitter feed content based on personality type and social cues. Future Generation Computer Systems, volume 110, p. 918-930.

17. Bencke L., Cechinel C., Munoz R. (2020) Automated classification of social network messages into Smart Cities dimensions. Future Generation Computer Systems, volume 109, p. 218-237.

Определение MBTI (MYERS-BRIGGS TYPE INDEX) типа человека с использованием машинного обучения на основе текста в социальных сетях

А.З. Суннатилла*, Е.С. Нурахов, А.А. Мынжасар
Казахский национальный университет им. аль-Фараби, Алматы, Казахстан
e-mail*: asel.sunna@mail.ru

Это исследование направлено на создание классификатора, используя методы машинного обучения, которые определяют психологический тип людей по классификации Myers-Briggs Type Index на основе текста, опубликованного в социальных сетях. Целью работы является автоматизация задачи определения типа человека с помощью машинного обучения, дается объяснение идентификации личностных черт с помощью индикатора личности MBTI. В машинном обучении применены методы логистической регрессии, случайного леса и опорных векторов, проведен литературный анализ аналогичных работ. В статье представлен ход исследовательской работы и результаты каждого классификатора и анализ используемых подходов. В связи с переходом людей на онлайн-формат работы в условиях нынешних карантинных ограничений подобные исследования могут оказать большую помощь в подборе персонала в компаниях, так как исследование предполагает выявление личностных качеств людей по их записям в социальных сетях. В данной работе использованы наиболее эффективные алгоритмы машинного обучения, простые в использовании для казахского языка и не требующие большой вычислительной мощности и, соответственно, приведены результаты работы для каждого метода, среди приведенных методов на хорошем уровне оказались точность и надежность классификатора для казахского языка методом опорных векторов.

Ключевые слова: *машинное обучение, черты личности, MBTI, социальные сети, обработка текста.*

Identification of MBTI (MYERS-BRIGGS TYPE INDEX) human type using text on social networks based machine learning

Assel Z. Sunnatilla*, Edil S. Nurakhov, Akniyet A. Myngzhassar
al-Farabi Kazakh National University, Almaty, Kazakhstan
e-mail*: asel.sunna@mail.ru

This study aims to create a classifier using machine learning methods that determine the psychological type of people based on the text published on social networks according to the Myers-Briggs Type Index classification. The article is based on the implementation of automation of the task of determining the personality type using machine learning, with an explanation for determining the characteristics of a person using the MBTI personality indicator. The methods of logistic regression, random forest and support vector machines were used, and a literary analysis of similar works was carried out. The article presents the progress of research work and the results of each classifier, as well as an analysis of the approaches used. In the context of the current quarantine restrictions, such studies can be of great help in the selection of personnel in companies due to the transition of people to an online format of work, since the study involves determining the personal qualities of people based on their posts in social networks. In this paper, the most effective machine learning algorithms for the Kazakh language, which are simple to use and do not require a lot of computing power, were used and, accordingly, the results of the work for each method were presented, among these methods, the accuracy and reliability of the classifier for the Kazakh language by the method of support vectors were at a good level.

Keywords: *machine learning, personality, MBTI, social network sites, text processing.*

АВТОРЛАР ТУРАЛЫ АҚПАРАТ

Суннатилла Әсел Зайниддинқызы, магистрант, әл-Фараби атындағы Қазақ ұлттық университетінің, ақпараттық технологиялар факультеті, информатика кафедрасы, Компьютерлік ғылымдар мамандығы, 2 курс. Адрес: Казақстан, Алматы, 050026, Қарасай батыр, 156; asel.sunna@mail.ru

Нурахов Еділ Сергазиевич, PhD, әл-Фараби атындағы Қазақ ұлттық университетінің, ақпараттық технологиялар факультеті, информатика кафедрасының аға оқытушысы. Адрес: Казақстан, Алматы, 050040, Тимирязева 54; eldi_mg@gmail.com

Мыңжасар Ақниет Ануарбекқызы, магистрант, әл-Фараби атындағы Қазақ ұлттық университетінің, ақпараттық технологиялар факультеті, информатика кафедрасы, Компьютерлік ғылымдар мамандығы, 2 курс. Адрес: Казақстан, Алматы, 050026, Қарасай батыр, 156; myngzhassar_akniyet@mail.ru

ИНФОРМАЦИЯ ОБ АВТОРАХ

Суннатилла Асель Зайниддинкызы – магистрант, 2 курса, специальности Компьюреные науки, кафедры информатики, факультета информационных технологий, Казахского национального университета имени аль-Фараби. Адрес: Казахстан, Алматы, 050026, Карасай батыра, 156; asel.sunna@mail.ru

Нурахов Едил Сергазиевич – PhD, старший преподаватель кафедры информатики, факультета информационных технологий, Казахского национального университета имени аль-Фараби. Адрес: Казахстан, Алматы, 050040, Тимирязева 54; eldi_mg@gmail.com

Мынжасар Акниет Ануарбеккызы – магистрант, 2 курса, специальности Компьюреные науки, кафедры информатики, факультета информационных технологий, Казахского национального университета имени аль-Фараби. Адрес: Казахстан, Алматы, 050026, Карасай батыра, 156; myngzhassar_akniyet@mail.ru

INFORMATION ABOUT THE AUTHORS

Assel Z. Sunnatilla, master's degree in Computer science, Department of Informatics, faculty of information technologies, al-Farabi Kazakh National University. Address: Kazakhstan, Almaty, 050026, Karasay batyr, 156; asel.sunna@mail.ru

Edil S. Nurakhov, PhD, senior lecturer of Computer Science Department, Faculty of Information Technology, al-Farabi Kazakh National University. Address: Kazakhstan, Almaty, 050040, Timiriyazeva 54; eldi_mg@gmail.com

Akniyet A. Myngzhassar, master's degree in Computer science, Department of Informatics, faculty of information technologies, al-Farabi Kazakh National University. Address: Almaty, Kazakhstan, 050026, Karasay batyr, 156; myngzhassar_akniyet@mail.ru

Редакцияға түсті / Поступила в редакцию / Received 10.03.2021
Жариялауға қабылданды / Принята к публикации / Accepted 29.06.2021