DOI 10.52209/1609-1825_2021_2_172 ӘОЖ 004.912

Қазақ тіліндегі мәтін стилін анықтауда табиғи тілді өңдеу (NLP) әдістері

[1]*КАЙБАСОВА Динара Женисбековна, PhD, аға оқытушы, dindgin@mail.ru,
[1]МАХАНОВА Балжан Мақсатқызы, магистрант, balzhan.makhanova@gmail.com,
[1]СҰЛЕЙМЕН Айнұр Елубайқызы, аға оқытушы, ai-box@mail.ru,
[1]Қарағанды техникалық университеті, Қазақстан, 100027, Қарағанды, Н. Назарбаев даңғылы, 56,
*автор-корреспондент.

Аңдатпа. Мақаланың мақсаты – қазақ тіліндегі мәтінді талдауда табиғи тілдік өңдеу мәселерін зерттеу. Табиғи тілдік өңдеуді жүзеге асыру үшін қолданылатын нормализацияның мүмкіншіліктеріне және атаулы ерекшеліктеріне назар аударылды. Тональдылықты талдау берілген мәтін фрагментінің көңіл-күйінің полярлығын (оң, теріс немесе бейтарап) болжауға бағытталған. Әртүрлі тілдердің тональдылығын талдаудың заманауи моделдеріне шолу жасалынды. Ұсынылған жұмыста терең оқыту, машиналық оқыту, қайталанатын нейрондық желілер (RNN), LSTM нейрондық желілері сияқты табиғи тілдерді өңдеудің әртүрлі тәсілдері мен әдістері қарастырылған. Қазақ тілінің қосымшаларын жүйелеудің заманауи жұмысы талданды.

Кілт сөздер: қазақ тіліндегі мәтін, мәтін стилі, үнділікті талдау, табиғи тілді өңдеу, терең оқыту, нейрондық желілер, жасанды интеллект.

Кіріспе

Қазіргі уақытта табиғи тілді өңдеуге байланысты әртүрлі зияткерлік және мобильді жүйелер белсенді түрде құрылуда. Өкінішке орай, қазақ тілін мәтіндік өңдеу мәселелері нашар дамыған, бұл ақпараттық технологиялардың дамуына кедергі келтіреді және келесі жағдайларға байланысты:

1. Қазақ тілінің күрделі морфологиясы бар тіл ретіндегі ерекшелігімен;

2. Осы салада қазақ тілін оқыту үшін электрондық ресурстардың болмауымен.

Сөйлем реңі адамның эмоцияларына жатады. Әр түрлі тілдерде әлеуметтік желілер арқылы өз ойларын білдіретін адамдар саны артты. Әлеуметтік желілердегі спикердің қарым-қатынасын талдаудың, жіктеудің және анықтаудың автоматты әдісін табу өте маңызды. Шынында да, бұл адамдардан тікелей кері байланыс немесе ақпарат алудың ең эмпирикалық әдісі [1]. Мысалы, бизнесте бұл компанияларға өз өнімдері немесе қызметтері туралы клиенттердің пікірлерін автоматты түрде жинауға мүмкіндік береді. Саясатта бұл шешім қабылдауға көмектесетін қоғамның бағыты мен саяси оқиғаларға реакциясы туралы қорытынды жасауға көмектеседі [2]. Барлық осы салалар қоғамдық шолулар негізінде анықталған қателіктермен жұмыс істеу үшін қоғамдық кері байланысты қажет етеді.

Көңіл-күйді талдау дегеніміз – мәтіннен белгілі бір көңіл-күйді анықтау және алу үшін ТТӨ-ні, есептеу лингвистикасы және деректерді іздеу әдістерін қолдану [3]. Оның мақсаты – мазмұны мен талдау деңгейіне (құжат деңгейі, сөйлем деңгейі, аспект деңгейі және сөз деңгейі) негізделген мәтін үзіндісінде (твит, хабарлама және т.б.) көрсетілген тональдылықты алу. Сонымен қатар, қазақ мәтініндегі көңіл-күйді талдауды қолдану өзекті болып табылады.

Материалдар мен әдістер

Тональдылықты талдау субъективті мәтіндерді екі немесе одан да көп санаттарға бөлуге бағытталған. Олардың ішіндегі ең айқын – оң және теріс. Біз бұл мәселені екілік SA (BSA) деп атаймыз. Кейбір жұмыстарға бейтарап мәтіннің үшінші класы кіреді. Біз бұл мәселені үштік SA (TSA) деп атаймыз. Соңғы нұсқа-кез-келген рейтинг жүйесі немесе рейтинг жүйесі негізінде көңіл-күйді ескеру, мысалы, 5 жұлдызды рейтинг жүйесі. Бұл MULTI-WAY SA (MWSA) ретінде белгілі [3].

Қазақ тіліндегі көңіл-күйді талдау тәсілдерінің қолжетімді зерттеулерін машиналық оқыту, лексикондық тәсіл, гибридтік немесе аралас тәсілдерге бөлуге болады.

Машиналық оқыту (МО) – көңіл-күйді талдауда жиі қолданылатын тәсіл.

Табиғи тілді өңдеу – бұл компьютерлік талдау мен табиғи тіл синтезін зерттейтін сала. Табиғи тілдік талдауды әр түрлі деңгейлерге (сатыларға) бөлуге болады. Әрбір жаңа деңгей, басқаша,

басқа төменгі деңгейлерден алынған ақпаратты пайдаланады. Келесі деңгейлер мәтінді талдауға бағытталған [4]:

- лексикалық талдау – табиғи тілдің жеке фрагменттері – лексемалардың (сөздердің) өңдеу деңгейі. Мәтінді бөлек сөздерге немесе сөйлемдерге бөлу кіреді. Бұған сонымен қатар бұрмаланған сөздерге кедергі жасау, тоқтату сөздерін (салмағы жоқ шулы сөздер) немесе эмотикон сияқты эмоционалды маңызды заттарды лексемаға айналдыру кіруі мүмкін;

- морфологиялық талдау – сөздің белгілі бір сипаттамаларын өңдеу деңгейі – грамматика (мысалы, жекеше немесе көпше – сан категориясының граммемасы). Бұл саты екі негізгі қолданбалы мәселелерді шешеді: лемматизация – сөзді қалыпқа келтіру – лемма; стемминг – сөздің түбірін табу процесі (сөздің түбірі міндетті емес);

- талдау – сөздер топтарын бөліп көрсету деңгейі және олардың арасындағы байланыс. Бұл процестің нәтижесі мәтіндегі әр сөйлемнің ағаш тәрізді иерархиясы болады;

- семантикалық талдау – мәтінді мағыналық талдау деңгейі. Бұл кезең мәтіннен мысқыл немесе иронияны бөліп көрсетуге мүмкіндік береді.

Жоғарыда айтылғандарға сүйене отырып, мәтіндік талдау дегеніміз – бұл мәтіннің үлкен фрагменттерін олардың құрамдас бөліктеріне бөлу: бірегей сөздер, жалпы тіркестер, синтаксистік заңдылықтар, кейіннен оларға статистикалық механизмдер арқылы қолданылатын болады [4].

Терең оқыту (ТО) – бұл жоғары деңгейлі деректер абстракциясын модельдеуге бағытталған машиналық оқытудың бір саласы. Бұл күрделі құрылымы бар немесе көптеген сызықты емес түрлендірулерден тұратын модельдік архитектураларды қолдану арқылы жасалады [5].

Қазақ тіліндегі көңіл-күйді талдаудың маңыздылығына қарамастан, қазіргі жасанды интеллект (ЖИ) жүйелері терең оқыту әдістеріне сүйенеді және көптеген салаларда үлкен жетістіктерге жетті. Терең оқыту (ТО), машиналық оқытудың кіші саласы (МО) мәліметтердегі жоғары деңгейлі абстракциялар үшін модель табу үшін бірнеше деңгейдегі көріністерді зерттеуге арналған алгоритмдер жиынтығына байланысты. Машиналық оқытудың әртүрлі дәстүрлі алгоритмдерінен терең оқытудың басты артықшылықтарының бірі – оның семантикалық композицияны орындау қабілетінен тұратын функцияларды өз бетінше орындау қабілеті, мәтіндік бірліктерге ұсыну векторын құру, олардың кіші компоненттерін немесе нысандарын біріктіру-тиімді және төмен өлшемді кеңістік.

Қазақ тілінің тональдылығын талдаудың тағы бір тәсілі – бұл лексикаға негізделген тәсіл. Бұл әдетте деректер белгіленбеген кезде жүзеге асырылады. Лексикондар – бұл сөзбен және оның көңіл-күйін немесе тональдылығын бағалайтын көңіл-күй сөздіктері.

Қайталанатын нейрондық желілер (RNN) – бұл мәліметтер тізбегін зерттеу үшін арнайы жасалған және негізінен мәтіндік деректерді жіктеу үшін қолданылатын терең оқытылған нейрондық желілер. Алайда, RNN деректердің ұзақ тізбегін өңдеу кезінде жоғалып кететін градиент проблемасынан зардап шегеді.

LSTM нейрондық желілері [2] осы мәселені шешу ретінде ұсынылған және сөйлеуді тану, бейне субтитрлері, музыкалық композиция сияқты көптеген өмірлік мәселелерде тиімділігін дәлелдеді. Алайда көңіл-күйді анықтау көбінесе сауалнаманың контексттік ақпаратына байланысты. Тікелей байланысқан нейрондық желілер белгілі бір дәрежеде контекстік ақпаратты ескере алмайды, сондықтан ASA-ны нашар басқарады. Осылайша, осы мақаладағы көңіл-күйді талдауға деген көзқарас тікелей және кері тәуелділіктермен жұмыс істейтін функциялар тізбегінен контекстік ақпаратты алу мүмкіндігі бар екі бағытты LSTM (BiLSTM) желісін пайдалану болып табылады. Сонымен қатар, біздің BiLSTM хронологиялық ретпен тізбекті өңдейтін тікелей LSTM және кері ретпен тізбекті өңдейтін кері LSTM көмегімен болашаққа қарауға мүмкіндік береді. Шығу тікелей және кері LSTM тиісті күйлерін біріктіреді.

Соған қарамастан, қазақ тіліндегі мәтіндерді өңдеу мәселелері іс жүзінде өте өзекті болып табылады.

Маңызды проблема – құжаттардағы нақты сөздерді жылдам табу мәселесі. Сөздерді жылдам табудың бір әдісі – құжаттардың кілт сөздерінің арасынан сөз түбірін іздеу, сәйкесінше құжатты қалауынша таңдауға мүмкіндік беру. Нормалдау (лемматизация) – ТТӨ қолдану жүйелеріндегі маңызды процестердің бірі, мысалы, ақпаратты іздеу, машиналық аударма және т.б. сөзді өзінің бастапқы негізіне келтіру.

Әр түрлі ғалымдар мен ғылыми топтар қазақ тілін қалыпқа келтірудің әртүрлі тәсілдерін талдап, қарастырды. Қазақ тілінің қосымшаларын сегментациялау бағыты бойынша жұмыстарды қарастыруға болады [1, 6], мұнда қазақ тілінің корпусындағы морфемалық құрылым талданып, аффикстің негіздері мен сегменттелуі шығарылады. Алдымен флекциялық қосымшалар Finitestate Machine (FSM) орнатылады, содан кейін флекциялық қосымшалар сегменттеледі.

Мәтіндік мәліметтерді машиналық оқыту алгоритмдерінде: сызықтық модельдер, шешім ағаштары, нейрондық желілер әдістерімен өңдеуге мүмкіндік беретіні белгілі. Мұндай үрдістерде алдымен, мәтіннің ерекшеліктері мен модельдерін тұрғызбас бұрын мәтінде қандай түрлендірулер жасау керек екенін талқылайық. Жалпылама деректер қосымшасына арналған типтік өңдеу әрекеті суретте бейнеленген қайталанатын екі кезеңді: құрастыру және орналастыру процесін жүзеге асырады. Бұл мәтінді өңдеуде машиналық оқыту үдірісінің көрінісі.

Құрастыру кезеңінде бастапқы деректер модельге ауысуға және онымен тәжірибе жасауға қо-

лайлы формаға айналады. Орналастыру кезеңінде пайдаланушыларға бағалау және болжау үшін қолданылатын модельдерді таңдау орын алады.

Мәтінді автоматты түрде жіктеу мәселесін шешудің алғашқы қадамы – құжатты трансформациялау. Бұл кезеңде символдар тізбегі түріндегі құжаттар классификациялау мәселесіне сәйкес машинада оқыту алгоритмдеріне қолайлы формаға ауыстырылады. Ондай форманы векторлық кеңістік деп атайды. Мәтіндік құжаттарды векторлық кеңістікте вектор ретінде ұсынуға арналған алгебралық модель – векторлық модель.

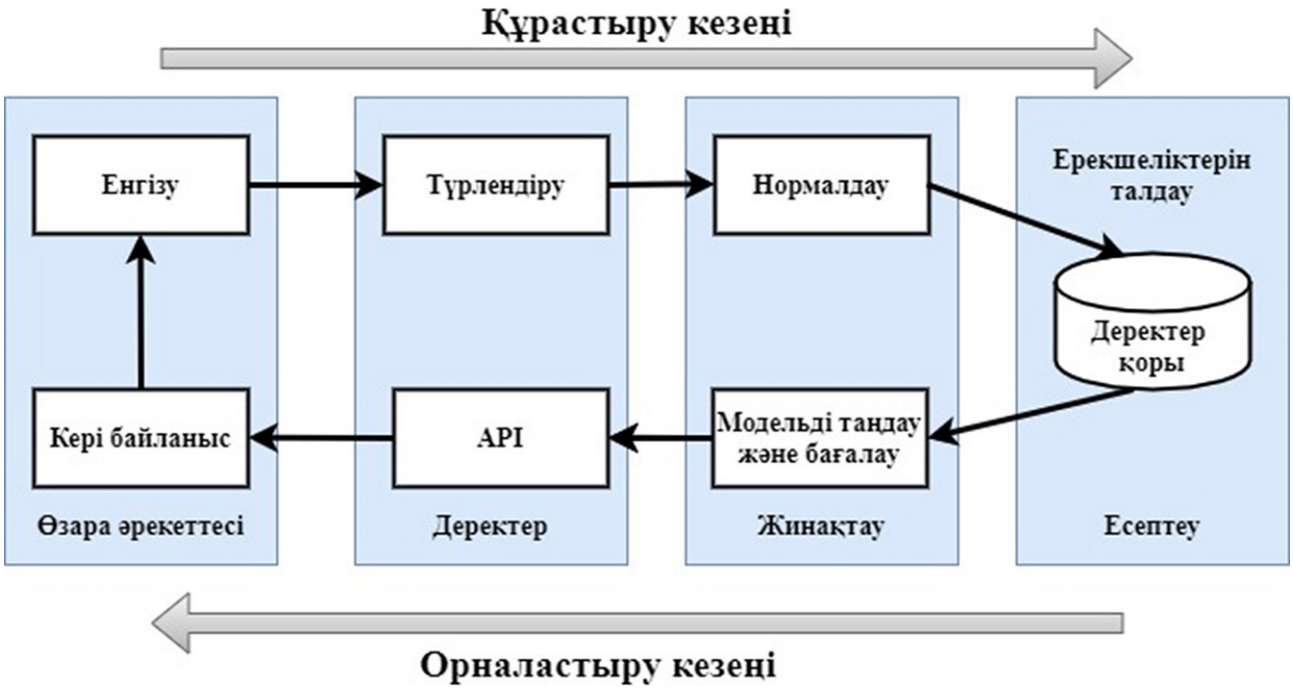
Векторлық модельде мәтін жеке сөздер немесе сөз тіркестері бола алатын терминдер жиынтығы ретінде қарастырылады. Вектордың әр компоненті жеке терминге сәйкес келеді. Компоненттің сандық мәні терминнің салмағымен сәйкес келеді, ол осы терминнің құрамындағы құжатты ұсыну үшін маңыздылығын сипаттайды. Егер құжатта термин болмаса, онда оның бұл құжаттағы салмағы нөлге тең болады.

Мәтінді векторлау үшін кестеде көрсетілген жиілік әдістері, тікелей кодтау әдісі, сөз жиілігі – кері құжат жиілігі және үлестірілген ұсыну әдісі сияқты әдістер қолданылады. Бұл әдістер болжам жасау қабілетін алу мақсатында құжаттарды алдын-ала өңделген корпустан векторлық модель кеңістігіне айналдыруға мүмкіндік береді [4].

Нәтижелер және талқылау

Қарастырылған [7] жұмыста қазақ тілі жалғауларының жіктелуі. Қазақ тілінде қосымшаларға жұрнақ пен жалғау жатады. Қосымша түрлері екі класқа бөлінеді:

1) сөздердің атаулы түбірлерінің жалғаулар жүйесі;



Мәтінді өңдеуде машиналық оқыту үдірісі

Мәтінді векторлау әдістері

Әдіс атауы	Жұмыс принципі	Қолайлылығы	Кемшіліктері
Жиілік әдісі	Сөздердің пайда болу жиілігін санау	Байес моделі	Ең жиі кездесетін сөздер әрқашан ең ақпараттылыққа ие бола бермейді
Тікелей кодтау	Сөздің болуының логикалық белгісін анықтау (0, 1)	Нейрон желісі	Барлық сөздер бірдей қашықтықта, сондықтан қалыпқа келтіру өте маңызды
Сөз жиілігі – кері құжат жиілігі	Сөздер жиілігін құжаттар бойынша қалыпқа келтіру	Жалпы мақсаттағы қосымшалар	Кең таралған сөздер құжат тақырыбы бола алмауы мүмкін
Үлестірілген ұсыну әдісі	Сөздердің ұқсастығын контекстке негізделген кодтау	Күрделі қатынастарды модельдеу	Есептеудің үлкен көлемі, қосымша құралдарды қолданбай масштабтаудың күрделілігі

2) қазақ тілінің сөздік негіздері.

Қазақ тілі сөздерінің атаулы негіздеріне жалғану жүйесінің төрт түрі бар:

1) көптік жалғаудың аяқталуы (K);
2) тәуелдік жалғаулары (T);
3) септік жалғаулары (C);
4) өздік жалғаулары (J).

Негіз (stem) (S). Мақалада формула бойынша анықталатын қосымша түрлерін орналастырудың ықтимал нұсқалары қарастырылған:

$$A_n^k = \frac{n!}{(n-k)!}.$$

Сонымен, атаулы буындары бар сөйлем мүшелері үшін жалғаулардың саны 1213 (барлық көптік нұсқалары ескеріледі), ал етістік түбірлерімен сөйлем мүшелерінің қосымша саны: етістіктер – 432, есімше жалғаулар – 1582, көсемше – 48, бұйрық рай – 240, етістер – 80 (олар жалғаулар емес, етістіктің формаларын құрайтын жұрнақтар, сондықтан біз оларды әрі қарай жұрнақ ретінде қарастырамыз). Барлығы 3565 ғана жалғау бар.

Қазақ тілінде жұрнақтардың екі түрі бар:

1) сөзжасам (TdJr);
2) қалыптастыру (TrJr).

Сөз құрушы жұрнақтар жаңа сөздерді қалыптастыруға қызмет етеді (сөздің мағынасы өзгереді).

Қалыптастырушы жұрнақтар сөз формаларын қалыптастыруға қызмет етеді. Өз кезегінде формативті жұрнақтар модификацияланған және грамматикалық жұрнақтарға бөлінеді. Барлық жұрнақтардың жалпы саны 26526 бірлікті құрайды.

Қазақ тіліне арналған стеминг алгоритмін әзірлеу. Лемматизация алгоритмдерін жасау саласында көптеген жұмыстар ұсынылған. Олардың арасында зерттелетін тілдердегі осы жұмысқа өте жақын және алгоритмнің өзін құру тәсілі – стемминг алгоритмі (stemming) бар. Жұмыста қателіктердің жеткілікті жоғары пайызының кемшіліктері қарастырылады. Қазақ тілі үшін де, сөздікті қолдану үшін де тиімді стемминг алгоритмі ұсынылған, бұл стемминг сапасын арттырады. Қазақ тіліне арналған лемматизация алгоритмі ұсынылған, онда қазақ тілінің толықсыздығы жоқ қосымшаларды жүйелеу қарастырылған. Қазақ тіліне арналған стеминг алгоритмі жарияланды. Бұл алгоритм қазақ тілінің аяқталуының ең кішкентай бөлігін ғана қамтитын стеммер Портер негізінде іске асырылған.

Стемминг – сөздердің ауыспалы бөліктерін, негізінен жалғауларды тастау. Бұл технология қарапайым және сөздіктер мен ережелер жиынтығын сақтауды қажет етпейді. Технология тіл морфологиясының ережелеріне негізделген. Стемингтің кемшілігі – қателіктердің көптігі. Стемминг ағылшын тілінде жақсы жұмыс істейді, ал орыс тілінде жақсы нәтиже бермейді [4].

Ұсынылған тәсіл қазақ тіліндегі қосымшалардың толық жүйесі негізінде қазақ тілін генерациялаудың еркін лексиконды (lexicon free) алгоритмін білдіреді.

Қорытынды

Жұмыста біз қазақ мәтінінің үндестігін талдау мәселесін қарастырдық. Қазақ тілінің үндестігін талдау жүйесінің тиімділігін бағалау үшін терең оқыту, рекурренттік нейронды желілер (RNN), LSTM нейронды желілер сияқты түрлі модельдерді пайдаланудың артықшылығы қарастырылды. Сонымен қатар, профессор У.А. Тукеевтің зерттеуіне шолу жасалды.

Осылайша, қазақ тілінің қосымшаларын жүйелеудің заманауи жұмысы зерттелді. Қазақ тілінің жалғаулары мен жұрнақтарын жіктеудің толық жүйесі әзірленді. Зат есім, сын есім, сан есім және етістікке арналған детерминистік ақырғы автоматтар қазақ тілі үшін негізгі жұрнақтар мен жалғауларды қосудың ықтимал нұсқаларын пайдалана отырып құрылған.

ӘДЕБИЕТТЕР ТІЗІМІ

1. M. HAMMAD, Mustafa et al-AWADI, «машиналық оқытуды қолдана отырып, әлеуметтік медиа шолуларының көңіл-күйін талдау», in Information Technology: 2016 жылғы жаңа буын, 131-139 бет.
2. M. Heikal, M. Torki, and N. El-Makky, «Sentiment Analysis of Arabic Tweets using Deep Learning», Procedia Comput. Sci. vol. 142, pp. 114-122, 2018, doi: 10.1016/j.procs.2018.10.466.
3. Altenbek G., Wang X.-L. Kazakh segmentation system of inflectional affixes // Proc. of the CIPS-SIGHAN Joint Conference on Chinese Language Processing (CLP 2010). Beijing, China. 2010. P. 183-190.
4. Кайбасова Д.Ж. Предварительная обработка коллекции рабочих учебных программ дисциплин для формирования корпуса текстов // Вестник КазНИТУ, № 6 (136) декабрь-2019, стр. 541-546, ISSN 1680-9211.
5. Bekmanova G., Sharipbay A., Altenbek G., Adali E., Zhetkenbay L., Kamanur U., Zulkhazhav A. A uniform morphological analyzer for the Kazakh and Turkish languages [Электронный ресурс]. URL: http://ceur-ws.org/Vol- 1975/paper3.pdf (дата обращения: 10.02.2020).
6. Федотов А.М., Тусупов Д.А., Самбетбаева М.А., Еримбетова А.С., Бакиева А.М., Идрисова А.И. Модель определения нормальной формы слова для казахского языка // Вестник Новосибирского государственного университета. Серия: Информационные технологии. 2015. Т. 13. № 1. С. 107-116.
7. Тукеев У.А., Турганбаева А. Лексикон – фри стемминг для казахского языка // Материалы международной научной конференции «Информатика и прикладная математика» («Computer science and Applied Mathematics») посвященной 25-летию Независимости Республики Казахстан и 25-летию Институт информационных и вычислительных технологий. Алматы, 2016. С. 84-88.

Методы обработки естественного языка (NLP) при определении стиля текста на казахском языке

[1]***КАЙБАСОВА Динара Женисбековна,*** *PhD, ст. преподаватель, dindgin@mail.ru,*
[1]***МАХАНОВА Балжан Мақсатқызы,*** *магистрант, balzhan.makhanova@gmail.com,*
[1]***СУЛЕЙМЕН Айнұр Елубайқызы,*** *ст. преподаватель, ai-box@mail.ru,*
[1]*Карагандинский технический университет, Казахстан, 100027, Караганда, пр. Н. Назарбаева, 56,*

**автор-корреспондент.*

Аннотация. *Целью статьи является исследование проблемы обработки естественного языка при анализе казахского текста. Было уделено внимание возможностям и номинальным характеристикам нормализации, используемой для реализации обработки естественного языка. Тональный анализ направлен на прогнозирование полярности (положительной, отрицательной или нейтральной) настроения данного фрагмента текста. Произведен обзор современных моделей тонального анализа разных языков. В предлагаемой работе рассматриваются различные подходы и методы обработки естественных языков, такие как углубленное обучение, машинное обучение, повторяющиеся нейронные сети (RNN), нейронные сети LSTM. Проанализирована современная работа по систематизации приложений казахского языка.*

Ключевые слова: *текст на казахском языке, стиль текста, анализ тональности, обработка естественного языка, глубокое обучение, нейронные сети, искусственный интеллект.*

Natural Language Processing (NLP) Methods for Determining the Style of Kazakh Text

[1]***KAIBASSOVA Dinara,*** *PhD, Senior Lecturer, dindgin@mail.ru,*
[1]***MAKHANOVA Balzhan,*** *master student, balzhan.makhanova@gmail.com,*
[1]***SULEIMEN Ainur,*** *Senior Lecturer, ai-box@mail.ru,*
[1]*Karaganda Technical University, Kazakhstan, 100027, Karaganda, N. Nazarbayev ave., 56,*

**corresponding author.*

Abstract. *The purpose of the article is to study the problem of natural language processing in the analysis of the Kazakh text. Attention was paid to the capabilities and ratings of the normalization used to implement natural language processing. Tonal analysis is aimed at predicting the polarity (positive, negative or neutral) of the mood of a given piece of text. The review of modern models of tonal analysis of different languages is made. The proposed work examines various approaches and methods of natural language processing, such as deep learning, machine learning, repetitive neural networks (RNN), LSTM neural networks. The modern work on the systematization of applications of the Kazakh language is analyzed.*

Keywords: *kazakh text, text style, tonal analysis, natural language processing, advanced learning, neural networks, artificial intelligence.*

REFERENCES

1. M. HAMMAD, Mustafa et al-AWADI, «mashinalyқ oқytudy қoldana otyryp, әleumettik media sholularynyң kөңil-kүjin taldau», in Information Technology: 2016 zhylғy zhaңa buyn, 131-139 bet.
2. M. Heikal, M. Torki, and N. El-Makky, «Sentiment Analysis of Arabic Tweets using Deep Learning», Procedia Comput. Sci. vol. 142, pp. 114-122, 2018, doi: 10.1016/j.procs.2018.10.466.
3. Altenbek G., Wang X.-L. Kazakh segmentation system of inflectional affixes // Proc. of the CIPS-SIGHAN Joint Conference on Chinese Language Processing (CLP 2010). Beijing, China. 2010. P. 183-190.
4. Kajbasova D.Zh. Predvaritel'naya obrabotka kollekcii rabochih uchebnyh programm disciplin dlya formirovaniya korpusa tekstov // Vestnik KazNITU, № 6 (136) dekabr'-2019, str. 541-546, ISSN 1680-9211.
5. Bekmanova G., Sharipbay A., Altenbek G., Adali E., Zhetkenbay L., Kamanur U., Zulkhazhav A. A uniform morphological analyzer for the Kazakh and Turkish languages [Elektronnyj resurs]. URL: http://ceur-ws.org/Vol- 1975/paper3.pdf (data obrashcheniya: 10.02.2020).
6. Fedotov A.M., Tusupov D.A., Sambetbaeva M.A., Erimbetova A.S., Bakieva A.M., Idrisova A.I. Model' opredeleniya normal'noj formy slova dlya kazahskogo yazyka // Vestnik Novosibirskogo gosudarstvennogo universiteta. Seriya: Informacionnye tekhnologii. 2015. T. 13. № 1. S. 107-116.
7. Tukeev U.A., Turganbaeva A. Leksikon – fri stemming dlya kazahskogo yazyka // Materialy mezhdunarodnoj nauchnoj konferencii «Informatika i prikladnaya matematika» («Computer science and Applied Mathematics») posvyashchennoj 25-letiyu Nezavisimosti Respubliki Kazahstan i 25-letiyu Institut informacionnyh i vychislitel'nyh tekhnologij. Almaty, 2016. S. 84-88.