МАТНЛИ МАЪЛУМОТЛАРНИ ТАҲЛИЛЛАШДА ДАСТЛАБКИ ИШЛОВ БЕРИШ МЕХАНИЗМИ

O.Babomuradov
Executive director of the Kazan Federal University branch in Jizzakh Jizzakh, Uzbekistan.E-mail: bobomuradov@gmail.com
O.Turakulov
Tashkent University of Information Technologies named after Muhammad al-Khwarizmi Tashkent, Uzbekistan. E-mail: o_xolmirzayevich@mail.ru

Аннотация. *Мазкур мақолада ижтимоий тармоқ матнли алмашинувларини таҳлил қилишда дастлабки ишлов бериш механизмини ташкил этиш воситаларининг жорий ҳолати, тадқиқот истиқбол йўналишлари ҳамда таклиф этилаётган ёндашувлар тавсифига бағишланган. Ишда маълумотларга ишлов бериш ёндашувларининг назарий таҳлили, таклиф этилаётган ечим ва натижалар келтирилган.*

Калит сузлар. *Маълумотлар, гипотезалар, Электрон, матнли ҳужжатларни, лингвистик нормаллаштириш.*

Кириш. Ҳозирги кунда маълумотлар уммонида турли тоифа ва турларга тегишли бўлган маълумотлар ҳажми жадал суръатлар билан ошиб бормоқда. Маълумотлар ҳажми жуда ҳам катта бўлганлиги сабабли, улар ичидан фойдалунувчи ўзига керакли бўлган ахборотларни ажратиб олиши масаласи мураккаблашиб бормоқда. Инсоният ўзига керакли ахборотларни излаш ва уларни ажратиб олиши учун маълумотларни қайта ишлаши, уларни таҳлил қилиши, аниқроғи маълумотлардан зарурий парчаларини ажратиб олиши лозим бўлмоқда. Масаланинг бундай қўйилиши маълумотларни таҳлил қилишнинг анъанавий усуллари, асосан, маълумотлар ҳақидаги олдиндан мавжуд бўлган гипотезаларни текширишга қаратилган ёндашувларидан кўра, интеллектуал таҳлил ёрдамида маълумотлар тузилмасини, маълумотлар орасидаги илгари маълум бўлмаган боғлиқликлар ва қонуниятларни аниқлашни амалга ошириш мақсадга мувофиқ бўлишини кўрсатмоқда.

Маълумотлар жамланмаси уларни фойдаланиш мақсадларига кўра ҳамда сақланиш турига қараб турлича бўлади. Турли кўринишдаги маълумотларга ишлов бериш учун турлича ёндашувлар талаб этилади.

Бир хилдаги ёндашув бир турда бошқача ва яна бир туркумда бошқача ишлов натижаларини кўрсатиши мумкин. Айниқса, ҳозирги кундаги маълумотларни жуда катта ҳажмларга эга бўлиши уларга ишлов беришдаги қийинчиликларни келтириб чиқармоқда.

Электрон матнли ҳужжатларни таснифлаш, таҳлиллаш ёки башоратлаш масалаларини ечиш самарадорлигини ошириш биргина таснифлаш, таҳлиллаш ёки башоратлаш механизмини яхшилаш, уни такомиллаштиришнинг ўзи етарли бўлмаслиги турли тадқиқотчилар томонидан асослаб берилган. Тузилмаланмаган матнларни таҳлил қилишда сифатни оширишга асосий эътибор қаратилса, ҳужжатни таснифлаш аниқлигига таъсир кўрсатиши мумкин. Хужжатларга дастлабки ишлов бериш босқичини мувафақиятли амалга ошириш сифатининг ошиши таҳлиллаш ёки таснифлаш (олдиндан белгиланган синфлар асосида)га кетувчи вақтни камайтириш ва сифатини ошириш имконияти яратилади. Матнга дастлабки ишлов беришнинг асосий 2 усулга ажратилади: ҳусусиятларни ажратиш (FE) ва ҳусусиятларни танлаш (FS).

Дастлабки ишлов беришнинг назарий асосланиши. Ҳусусиятларга ажратиш йўналиши асосан морфологик, синтактик ҳамда семантик таҳлилларга бўлинади. Морфологик таҳлил матнли ҳужжатда мавжуд алоҳида сўзлар билан токенлаштириш, ўгириш – тўхтатиш ва сўзларнинг бирикмалари билан ишлайди. Токенлаштиришда матнли ҳужжатдаги сўзлар кетма-кетлиги сифатида қаралиб, тиниш белгилар олиб ташлаш орқали сўзлар ажратиб олинади. Матнда учрайдиган ажратиш, тўхталиш сўзлари олиб ташланади ва калит сўзлар ажратилади. Бу жараённинг қўлланилиши ҳужжатда сўзлар сонини камайишига, матнга ишлов бериш самарадорлигини оширилишига олиб келади. Бир қатор сўзларни лингвистик нормаллаштириш орқали, яъни ўзак сўзни ажратиб олиш орқали “stemming-word”ни амалга оширишни назарда тутади. Ёндашув орқали сўзлар таҳлилида асосий мазмундаги ўзак сўзларни ажратиб олиш масаласи ечилади. Бу орқали ҳам сўзлар портфелини камайтиришга эришиш, ўқитиш жараёнини енгиллаштириш имконияти яратилади. Бундай амални бажариш учун “қўпол куч”, $suf-x-striping$, $afx\ x-remove$, n-gramm каби турли алгоритмлардан фойдаланиш мумкин [1-4].

Матндаги гапдан мантиқий маънони ҳосил қилиб олиш учун жумла грамматик қоидаларга бўйсиндирилган бўлиши керак. Синтактик таҳлил боғлиқликларни ифодалайди ва тугунларда ушбу таиий тилнинг сўзлари жойлашади ва матн қисми ҳамда грамматик

характеристикаси берилади. Морфологик таҳлилда матн элементларининг нутқ қисмлари ҳамда морфологик характеристикаси асосида таҳлиллаш амалга оширилса, синтактик таҳлил матн элементлари орасидаги боғлиқликларни семантик қирра кўринишида аниқланади, у ерда аргумент, атрибут, коньюнкция, дизюнкция, х.клар акс этган бўлади. Синтактик таҳлил грамматик тузилмани аниқлашда, табиий тил матнининг от, феъл, ясовчи қўшимчалар, тиниш белгилари каби нутқ қисмларидан ташкил топган бўлади. Синтактик таҳлиллаш нутқ қисмини аниқлаш (POS tugging) ҳамда таҳлиллаш қисмларидан иборат бўлади, бу матндан мантиқий маънони ажратиб олиш учун қўлланилади [6].

POSларни белгилаш жараёни грамматик қоидалар асосида берилган сўзнинг контекстдаги боғлиқлиги ифодалаш учун ишлатилади. Таҳлиллаш жараёни амалга оширилаётганда сўзларнинг лексик тегишлилик синфи аниқ бўлса, таҳлиллаш жараёни осонлаштирилади [Yoshida 2007]. Кўпгина манбаларда POSларни белгилаш турли ёндашувларда амалга оширилган[6,7]. Бундай ёндашувлар орасида энг истиқболлиларидан HMM(яширин марков модели) ҳисобланади. Чунки бу модел тушуниш ва қўллашга осон ҳисобланиб, кириш кетма-кетлигини ҳосил қилиш учун қоидалар шакллантирилади[8]. Ҳусусиятларни ажратиш жумланинг грамматик тузилишини ўрганишда қарор дарахти кўринишдаги моделдан фойдаланиш ҳам кенг тадқиқ қилинган. Бу ёндашув грамматик таҳлилда юқоридан пастга ёки пастдан юқорига кўринишларидан фойдаланилади[9,10].

Тарқатилаётган табиий тил матни мазмуни тушунилган ҳолда қабул қилинади, буни автоматик тарзда шаклланувчи механизм ишлаб чиқилиши талаб этилади[11]. Калит сўзларни аниқлаган матнни таснифлаш учун WordNet-Affect ёки SentiWordNet матнларидаги кайфият (ҳис-туйғу)ларни аниқлашга йўналтирилган механизмга эга [10,11]. Бироқ бу каби тизимлар катта МБ талаб қилади.

Матнли ҳужжатларни таснифлашда дастлабки ишлов бериш процедураларидан бири ҳусусиятларини танлаш бўлиб, унинг ёрдамида матндан аҳамиятсиз ва ортиқча маълумотларни олиб ташлашни амалга ошириш мумкин. Бунинг учун сўзлар тўплами ҳосил қилиниб, ҳужжатдаги сўз аҳамияти маълум бир бирликда белгиланади [xua 2009]. Хусусият векторларини муддатли частота (IDF) кўринишдаги усуллари таклиф қилинган [5].

TF ҳужжат тўпламидан матнларга сўзларнинг учраш частотасини аниқлайди. Матнли ҳужжатларни уларда учрайдиган сўзлар гуруҳи

частотаси билан тегишли гуруҳини (синфини) аниқлаш мумкин бўлади. IDF эса матнли ҳужжатда сўзларнинг кам учраш кесимлари қаралади. IFIDF механизми юқоридаги икки воситанинг комбинациясини ҳисобга олган ҳолда берилган сўз частотаси ва мослигини аниқлашга йўналтирилади [12].

Бир қатор манбаларда тадқиқотчилар матнларга дастлабки ишлов бериш учун мосликларни (ўхшашликларни) ўлчовини ҳисобга олиш усулларидан фойдаланишган [10,13]. Ўхшашликни ўлчашни грамматик бир бирига яқин бўлган (тузилиши жиҳатидан) ҳалқа мазмунан яқин бўлган сўз (атама)лар олинади, уларнинг миқдори бўйича аниқлаш амалга оширилади. Бунга сўзларнинг соҳаларга (калит сўзларнинг предмет соҳани ифодалайдиган салмоғи) тегишлилик даражаси ҳисобга олинади. Масалан, "ўқувчи", "ўқитувчи", "ўқитиш" каби сўз (атама)лар таълим жараёнининг асосий идентификатор сўзларидан саналади ва бу атамаларнинг бир бирига яқинлик даражаси катта. Бироқ бу ҳам матнда ҳосил қилинаётган мазмун ҳис-туйғуни аниқ ифода қилолмаслиги мумкин. Чунки баъзи жумлалар яқин бўлгани билан қарама-қарши фикрларни англатиши мумкин. "Ўқувчи" билан "ўқитувчи" таълим тизимида икки томонни эгалловчи субъектлар ҳисобланади. Мисол тариқасида "Мен ўйлайманки бу йил бизларга муваффақиятсизлик келтирмайди", мазкур жумла муваффақият келтиришидан дарак беради. Бироқ "муваффақиятсизлик" сўзи салбий ҳис-туйғу ифодаси ҳисобланади (75% салмоқ билан). Сўзларнинг бундай мазмунан ажратиш имкониятини яратиш учун етарли даражада лингвистик (табиий тил) корпус билан таъминланиши зарур бўлади. Шу билан бирга ўхшашликларни ўлчаш усулларини қўллашда алгоритмни катта ҳажмли маълумотлар билан тажрибавий тадқиқот ўтказиб синов қилиш талаб этилади [14,15].

Тадқиқотларда FSнинг латент-семантик индекслаш (LSI) ва тасоддифий хариталаш (RM) каби усуллар мавжуд бўлиб, LSI семантик ажратишни қўллаган ҳолда лексик мосликни таъминлашга интилади, RM эса катта ҳужжатлар тўплами таркибидан матнларнинг яқинлик харитасини ишлаб чиқади.

[16] ижтимоий тармоқларда матнларни таҳлил қилиш учун энг кўп тарқалган таснифлаш ҳамда мета маълумот (матндан)ни ҳосил қилиш орқали (кластеризация) ёндашувларни тадқиқ қилган.

Эътибор билан кузатадиган бўлсак матнли ҳужжатларга дастлабки ишлов бериш орқали таснифлаш масаласи веб – муҳитда матнли ҳужжатларни таҳлиллаш масаласини долзарблигини кўрсатади. Иш

туфайли навбатдаги босқич таҳлилларимиз, айнан, веб-муҳитда матнли ҳужжатларни таҳлиллаш масаласини веб – муҳитда матнли ҳужжатларни таҳлиллаш масаласини долзарблигини кўрсатади. Иш туфайли навбатдаги босқич таҳлилларимиз, айнан, веб-муҳитда матнли ҳужжатларни таҳлиллаш масаласини ечишга йўналтирилади. Чунки веб 2.0 нинг ривожланиши фойдаланувчиларни нафақат фойдаланиш мақоми билан чекланишга, балким, ўзларининг ушбу тармоқ ривожига ҳисса қўшиш имкониятини яратди [1,2]. Web 2.0 да фойдаланувчилар ҳамда турли махсус ташкилотлар ҳосил қилаётган маълумотлар сегменти бозорда реклама хизмати, маркетинг хизматини ривожлантиришда асосий омил ҳиосбланади [3,4]. Бироқ мазкур маълумотларнинг хилма-хиллиги унга қўшимча равишда ишлов беришни талаб этади [5,6]. Бунинг учун катта массивли маълумотларга ишлов бериш, саралаш ва зарурий маълумотларни ажратиб олиш муҳим вазифалардан ҳисобланади [7]. Бу кўринишдаги ярим автомат маълумотларни саралаш, ажратиш ва қонуниятларни (мазмунли) юзага чиқариш фикрларини (ижтимоий тармоқларда) ўрганиш учун муҳим механизм ҳисобланади [3].

Махсус тайёрланмаган "ҳом" маълумотларни қайта ишлашда (таҳлиллашда) бир қатор муаммолар юзага келиши мумкин[8]. Ижтимоий гуруҳлар фикрларини ўрганишда пайдо бўладиган бир қатор қийинчиликлар ечим топишни талаб этади. Уларни қуйидаги кўринишда гуруҳлаштирилади:

1. Реал ишлаб турган тизимлар, тажрибавий тадқиқот ўтказилувчиларига нисбатан кўпроқ табиий тилларга эга[9,10]. Табиий тилларни кўпроқ олиниши социумнинг билдирадиган фикрларини, ҳис-туйғуларини аниқлашда ҳатолик катталашиб кетади, чунки табиий тил тузилмасига нисбатан усул ва моделларда маълум ўзгартириш ёки параметрлаштириш амалга оширилиши лозим, акс ҳолда интерпретация нотўғри бўлади.

2. Бир қатор ижтимоий тармоқлар қараладиган бўлса, дастурий платформа ёки интерфейс ўзида асосий маълумотлар билан бир қаторда реклама, қўшимча маълумотлар каби чет маълумотлар мавжуд. Мазкур чет элементлар асосий контексига ҳалақит бўлиб ҳизмат қилади ва таснифлаш аниқлигини пасайтириш мумкин [11].

3. Фойдаланувчилар томонидан яратилаётган контент турли қичқартма ёки тизим томонидан таклиф этиладиган белгилар (смайликлар, эмодзилар, gif лар) ҳам қўшилиши мумкин. Бу ҳолатда ҳам контент эгаси ёки эгаларининг ҳис-туйғуларини акс эттирувчи

асосий маълумотлар йўқолиб қолиши мумкин [12]. Асосий эътиборни мазкур кўринишдаги муаммоларни бартараф этишга қаратиш устувор вазифа сифатида эътироф этилади ҳамда ечим берувчи усул, модел ва алгоритмлар устида тадқиқот кучайтириш лозим бўлади. Умуман олганда матнлардан мазмун (фикр)ни чиқариб олиш катта масалалардан ҳисобланади ва қуйидаги қисм масалаларига бўлинади [13]:

1) Матнда ифодаланган моҳият (ҳис-туйғу) таснифи социум вакилларининг руҳий ҳолатлари таснифлаш масаласи [14,15];

2) Фикрларнинг асосий хусусиятларига қараб қидириш ҳис-туйғуларни ифодаловчи жумлалардан келиб чиққан ҳолда аниқлаш масаласи [16];

3) Олдиндан мавжуд бўлган шаблонлар асосида таққослашларни амалга ошириш масаласи [17].

Кўп ҳолларда аниқлиги юқори бўлишини таъминлаш учун тил корпуслари қўлланилади бироқ бу мураккаблик келтириб чиқаради [18-20]. Луғатга асосланган ҳолда масала ечимини ишлаб чиқиш, таққосламаларнинг кўпайиши ҳамда NP-тўла масалага келиш эҳтимолининг ошишига олиб келиши мумкин [14,16,21,22].

Ҳужжатлардан мазмунан (ҳис-туйғу) аниқланишни амалга ошириш учун бир қанча алгоритмлар мавжуд, буларга SVM, чизиқли регрессия, сунъий нерон тармоқлар, LDA-ажратиш, генетик алгоритмларни киритиш мумкин [23-27,29-32].

Матнларни интеллектуал таҳлиллаш корхона ёки тармоқдаги тузилмалаштирилмаган табиий тил матнидан ташкил топган базадан фойдаланади [17]. Таҳлиллашнинг икки усулига таянган механизми мавжуд бўлиб, улардан ёпиқ луғатли база сўз ва иборалари маълум предмет соҳани қамраб олган масалаларни ҳал этади [18]. Иккинчи механизм очиқ луғат жамланмали бўлиб, унинг учун табиий тил турининг фарқи йўқ ҳолда таҳлиллаш амалга оширилади, бироқ бу ёндашувни ҳам NLP(табиий тил матнини тушуниш) масаласига олиб бороймайди [19]. NLP билан боғлиқ тадқиқотлар[20]да баён этилган. Очиқ луғатли таҳлиллаш масаласини амалиётга қўллаш масаласи [21] ҳамда ижтимоий масала ҳисобланган зарурий номзодларни танлаш масалаларида кенг қўлланилди [22,23,24,25].

Ҳар икки тур масала ҳам табиий тил матнини таҳлиллаш каби ўта муҳим муаммони ҳал этишга йўналтирилади.

Юқорида келтирилган таҳлил асосида табиий тил матнини таҳлиллашда ёки таснифлашда дастлабки ишлов бериш

механизмининг ўрни беқиёс эканлигини кўришимиз мумкин. Мазкур тадқиқот ишида матнли ҳужжатларни таснифлаш учун дастлабки ишлов бериш масаласи ечимларига бағишланади.

ФОЙДАЛАНИЛГАН АДАБИЁТЛАР РЎЙХАТИ:

1. Гмурман В. Е. Теория вероятностей и математическая статистика. — Москва : Высшая школа, 2013. — 479 с.
2. Вапник В. Н., Стерин А. М. Об упорядоченной минимизации суммарного риска в задаче распознавания образов // Автоматика и телемеханика. — 1978. — № 10. — С. 83—92.
3. Епрев А.С. Автоматическая классификация текстовых документов // Математические структуры и моделирование / Под ред. А.К. Гуца. – Омск: "Омское книжное издательство", 2010. – Вып. 21. – С. 65-81. – [Электрон ресурс]. URL: http://msm.univer.omsk.su/sbornik/jrn21/sbornik_n21.pdf
4. Kunneman F., Bosch A. van den. Event detection in Twitter: A machine-learning approach based on term pivoting // Proceedings of the 26th Benelux Conference on Artificial Intelligence / Grootjen, F., Otworowska, M., Kwisthout, J. (ed.). – Nijmegen, 2014. – P. 65-72. – [Электрон ресурс]. URL: http://antalvandenbosch.ruhosting.nl/papers/event-detection-twitter.pdf
5. Sebastiani F. Machine Learning in Automated Text Categorization // ACM Computing Surveys (CSUR). – New York, 2002. – Vol. 34, No. 1. – P. 1-47. – [Электрон ресурс]. URL: http://nmis.isti.cnr.it/sebastiani/Publications/ACMCS02.pdf
6. Hastie T., Tibshirani R., Friedman J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. – Springer, 2009. – 745 p.
7. Rish I. An empirical study of the naive Bayes classifier // IJCAI 2001 workshop on empirical methods in artificial intelligence. – IBM New York, 2001. – Vol. 3, Issue 22. – P. 41-46. – [Электрон ресурс]. URL: http://www.research.ibm.com/people/r/rish/papers/RC22230.pdf
8. Тузовский А. Формирование семантических метаданных для объектов управления знаниями // Известия Томского политехнического университета. — 2007. — Т. 310. — С. 108—112.
9. Amaravadi C. S. Knowledge Management for Administrative Knowledge //Expert Systems. | 2005. | 25(2). | Pp. 53{61.

10. Kuznetsov S., Poelmans J. Knowledge representation and processing with formal concept analysis // Wiley interdisciplinary reviews: Data mining and knowledge discovery. — 2013. — № 3. — С. 200—215.

11. Roussopoulos N. Conceptual Modeling: Past, Present and the Continuum of the Future // Conceptual Modeling: Foundations and Applications. 2009. | Pp. 139{152.

12. Hutchins J. ALPAC: The (In)Famous Report // Readings in machine translation. 2003. Vol. 14. P. 131–135.

13. Маннинг К. Д., Рагхаван П., Шютце Х. Введение в информационный поиск. : Пер. с англ. / Под ред. П. И. Браславского, Д. А. Клюшина, И. В. Сегаловича. М.: ООО «И.Д. Вильямс», 2011. 528 с.

14. Лукашевич Н. В. Тезаурусы в задачах информационного поиска. М.: Изд-во Московского университета, 2011. 512 с.

15. Deliyanni A., Kowalski R. A. Logic and Semantic Networks // Communications of the ACM. 1979. Vol. 22, no. 3. P. 184–192.

16. Корепанова А. А., Абрамов М. В., Тулупьева Т. В. Идентификация аккаунтов пользователей в социальных сетях «вконтакте» и «одноклассники» // Семнадцатая национальная конференция по искусственному интеллекту с международным участием КИИ-2019, сборник научных трудов – 2019. – С. 153.]

17. О. Ж. Бабомурадов and Л. Б. Бобоев, "Ўзбек тилидаги матнли ҳужжатларни таснифлашнинг тасодифий ўрмон усули," in *Олий таълим тизимида масофали таълимни жорий этишнинг техник-дастурий ва услубий таъминотини такомиллаштириш истиқболлари, Республика илмий-амалий конференсияси, Қарши, 28 май 2021*, 2021, pp. 112–115.

18. О. Ж. Бабомурадов, Л. Б. Бобоев, "Таснифлашни баҳолаш ўлчовлари," in *Инновацион ёндашувлар илм-фан тараққиёти калити сифатида: ечимлар ва истиқболлар, ЎзМУ Жиззах филиали, Республика миқиёсидаги илмий-техник анжумани*, 2020, pp. 146–153.

19. О.Ж.Бабомурадов, Л. Б. Бобоев, Х. Т. Дусанов, "Матнннинг кетма-кетлик модели," in *Инновацион ёндашувлар илм-фан тараққиёти калити сифатида: ечимлар ва истиқболлар, ЎзМУ Жиззах филиали, Республика миқиёсидаги илмий-техник анжумани*, 2020, pp. 192–197.

20. О. Ж. Бабомурадов, Н. С. Маматов, Л. Б. Бобоев, Б. И. Отахонова, "Text documents classification in Uzbek language," *International journal of recent technology and engineering*, vol. 8, no. 2, pp. 3787–3789, 2019.

21. Y. Du, J. Liu, W. Ke, and X. Gong, "Hierarchy construction and text classification based on the relaxation strategy and least information model," *Expert Systems with Applications*, vol. 100, pp. 157–164, 2018.

22. Гришеленок Д. А., Ковель А. А. Использование результатов математического планирования эксперимента при формировании обучающей выборки нейросети //Известия высших учебных заведений. Приборостроение. – 2011. – Т. 54. – №. 4. – С. 51-54].

23. [Sabuj M.S., Afrin Z., Hasan K.M.A. (2017) Opinion Mining Using Support Vector Machine with Web Based Diverse Data. / Pattern Recognition and Machine Intelligence. PReMI 2017. Lecture Notes in Computer Science, vol 10597. Springer, pp 673-678.

24. Filippov A., Moshkin V., Yarushkina N. (2019) Development of a Software for the Semantic Analysis of Social Media Content. // Recent Research in Control Engineering and Decision Making. ICIT 2019. Studies in Systems, Decision and Control, vol 199. Springer, Cham pp 421-432;

25. Хомский Н. Три модели описании языка//Кибернитеческий сборник.-1961.-Вын.2.-с.81-92.

26. Филмор Ч. Дело о падеже// Новое в зарубежной лингвистике. Вып. X.-М.: Лингвистическая семантика, Прогресс, 1981,-с.369-495.

Мельчух И.А. Опыт теорий лингвистических моделей «смысл-текст».-М.: Наука, 1974,-314с.

27. Yarushkina N. G., Moshkin V. S., Andreev I. A. The sentimentanalysis algorithm of social networks text resources based on ontology //Информационные технологии и нанотехнологии (ИТНТ-2020). – 2020. – pp. 226-232.