МАТН БЕЛГИЛАРИ ЎЛЧОВИНИ ҚИСҚАРТИРИШ ҲАМДА ТАБИИЙ ТИЛ МАТНИНИ ТАҲЛИЛЛАШ ЁНДАШУВИ

O.Babomuradov
Executive director of the Kazan Federal University branch in Jizzakh Jizzakh, Uzbekistan.
O.Turakulov
Tashkent University of Information Technologies named after Muhammad al-Khwarizmi Tashkent, Uzbekistan.

Аннотация: *Мазкур мақолада матн белгилари ўлчовини қисқартириш ҳамда табиий тил матнини таҳлиллаш ёндашуви хакида суз боради.*

Калит сузлар: *Матнлар, ёндашув, матрица, компонентлар, ҳотира, мураккаблик даражаси.*

Матнларни таснифлашда белгилар фазосини қисқартириш муҳим аҳамият касб этади [26], [27]. Бу борада турли ҳил ҳолатларни юзага келтирувчи ёндашувларни кўриб чиқиш мумкин [28], [29]. Терминга асосланган вектор моделларидаги матнлар кетма-кетлиги белгилар фазоси ҳилма ҳил бўлади. Мазкур ҳолатда ҳотира ва вақт ресурсининг ортиқча сарфига дуч келинади. Буни бир нечта ёндашув ёрдамида ҳал этиш мумкин. Ёндашувларни танлаш ҳолатларнинг мураккаблигига қараб танланади. Булардан юиринчиси асосий компонентлар таҳлили бўлиб, у ўзаро боғлиқ бўлмаган қисмларни аниқлашга қаратилган, янги ўзаро боғлиқ бўлмаган ўзгарувчини фарқини топиш ва уни сақлаб қолишга интилтириш ажратишни амалга оширишга олиб келади. Уни қуйидаги кўринишда амалга ошириш мумкин. Уни қуйидаги кўринишда формаллаштириш мумкин, маълумотлар тўплами $x^{(i)}; i = 1,\ldots,m$ ва ҳар бир $i(n\ \ m)$ учун $x^{(i)} \in\ ^{n}$ берилган бўлсин. X матрицанинг j - устуни вектор бўлиб, j - ўзгарувчининг устида кузатувлар x_j ҳисобланади [29], [30].

Асосий компонентлар таҳлили ўқитиш алгоритмини ишлатишдан олдин маълумотлар тўпламининг ўлчамини камайтириш учун дастлабки ишлов бериш воситаси сифатида ишлатилиши мумкин, унда $x^{(i)}s$ кириш сифатида берилади. Асосий компонентлар таҳлили халақитларни камайтириш алгоритми сифатида қўлланилиб, ортиқчалик муаммосини олдини олишда ёрдам бериши мумкин. Асосий компонентларни таҳлил қилиш ядро усули ёрдамида чизиқли асосий компонентлар таҳлилини чизиқли бўлмаган ҳолатга умумлаштирадиган яна бир ўлчамни камайтириш усули ҳисобланади [31].

Тузилаётган луғатнинг мураккаблик даражасига қараб мустақил компонентлар таҳлили Ж. Ҳераулт томонидан киритилган. Бу усул кейинроқ С. Жуттен ва Ж.

Ҳераулт томонидан янада ривожлантирилган. Мустақил компонентлар таҳлили кузатилган маълумотлар чизиқли трансформация сифатида ифодаланган статистик моделлаштириш усули ҳисобланади.

Чизиқли дискриминант таҳлил маълумотларни таснифлаш ва ўлчамини камайтириш учун кенг қўлланиладиган усул ҳисобланади. Чизиқли дискриминант таҳлил, айниқса, синф ичидаги частоталар тенг бўлмаган ва уларнинг ишлаши тасодифий генерация қилинган синов маълумотлар бўйича баҳоланган ҳолларда жуда фойдали саналади. Синфга боғлиқ ва синфга боғлиқ бўлмаган трансформация чизиқли дискриминант таҳлилга иккита ёндашув бўлиб, унда синф дисперсияси ва синф дисперсиясининг нисбати ўртасида ва умумий дисперсия ва синф фарқи нисбати ўртасида мос равишда қўлланилади.

Манфий бўлмаган матрица факторизацияси ёки манфий бўлмаган матрицанинг яқинлашиши матн ва кетма-кетликни таҳлил қилиш каби жуда юқори ўлчамли маълумотлар учун жуда кучли усул эканлигини кўрсатган. Ушбу усул ўлчамларни қисқартиришнинг истиқболли усули ҳисобланади.

Ушбу усулларга асосланган ўлчамни камайтириш қуйидаги 5 босқични ўз ичига олади:

Дастлабки ишлов беришдан сўнг термин индексини ажратиб олиш ва матнни тозалаш каби белгиларни ажратиш, m белгиларга эга n ҳужжатлар мавжуд бўлади;

N ҳужжат яратиш $\left(d \in \{d_1, d_2, \ldots, d_n\}\right)$, бу ерда $a_{ij} = L_{ij} \times G_i$ вектори L_{ij} j ҳужжатдаги i - терминнинг локал вазнини билдиради ва G_i - бу i - ҳужжат учун глобал вазнлар;

Усулни барча ҳужжатлардаги барча терминларга бирма-бир қўллаш;

Китилган ҳужжат векторини r ўлчамли фазога проекциялаш;

Худди шу трансформациядан фойдаланиб, r - ўлчамли фазога синов тўпламини хариталаш.

Матнли ҳужжатларни таснифлаш табиий тил билан боғлиқ бўлган ҳолатларда бир қатор ёндашувлар мавжуд бўлиб, уларнинг табиий тил матнини синтаксис таҳлиллашни амалга оширишда семантик моделлаштириш орқали семантик таҳлиллагични ишлаб чиқиш қараб ўтилади.

Табиий тил матнини таҳлиллаш тил гуруҳи ва тузилишига боғлиқ. Бунда матннинг синтаксис тузилмасини ўрганиш орқали семантикасини ҳосил қилиш, яъни мантиқий боғлиқликларни таҳлиллаш муҳим омил ҳисобланади.

Аксарият назарияларда (80-йилларда қўлланилган) бирикмалар ташкил этивчи қисмларни (wellformedness) ифодаловчи чеклов қоидалари (licensing ruls)да грамматикани тавсифлашга ўтилди. Тилни бундай тавсифлаш усулида тил синтаксиси берилмайди, турли чекловлар бир-бири билан боғланмаган ҳолатда бўлади. Шу қаторда таҳлил орқали барча чекловларни қаноатлантирадиган ифодалашни топишга ҳаракат қилинади, бу ерда мумкин бўлган конструкция варианти параллел равишда

қурилади. Ушбу йўналишда лингвистик конструкция хусусиятини ифодаловчи қоидалар ушбу конструкцияларни тавсифлашда чекловлар учун умумийроқ бўлган тамойиллардан фойдаланиш ўринли бўлади. Ҳусусан бундай ёндашув қурилган қоидаларни аниқ бир конструкцияга боғланмаган ҳолда лексик бирликларнинг грамматик хусусиятларини ифодалаш имконини беради.

Синтаксис қоидани қўллашнинг икки усули мавжуд: пастдан юқорига ва юқоридан пастга усуллари. Биринчи ҳолатда ўнг томонни тавсифловчи тузилмасини чап томонни ифодаловчи белги билан алмаштирувчи қоида қўлланилади. Иккинчи ҳолатда эса берилган гапни дастлабки S белгисидан бошланиши исботланади. Кўпинча пастдан юқорига таҳлилида қоидани бир неча усулда қўллаш имконияти пайдо бўлади.

Синтаксис таҳлилда муқобил танловни амалга оширишнинг икки стандарт қоидасини қўллаш мумкин: “Энига” қидирув ҳамда “чуқур” қидирув. Биринчи ҳолатда барча мумкин бўлган муқобиллар эслаб қолинади ва уларнинг ҳар бири параллел очилади. Агар муқобиллар мувофақиятли чиқмаса ушбу муқобиллар имкон тўпламидан ўчирилади. “Чуқур” қидирувда муқобиллардан биттаси таққослама сифатида олинади, мувофақиятсиз чиқса дастлабки нуқтадан таҳлиллаш қайта бошланади. Юқоридан пастга ёндашувли таҳлиллашни қўллаш грамматик бўлмаган муқобилларни ҳосил қилиш имконини беради. Бошқа томондан юқоридан-пастга таҳлиллаш ушбу гапга тўғри келмайдиган ажратиш фаразини генерациялашга тўсқинлик қилади.

Ушбу муқобиллар устуворликларини қисман ажратиш натижаларини мужассамлаштирувчи жадвал ёрдамида таҳлиллаш имкони яратилади. Бирор сабабга кўра таҳлил боши берк кўчага кириб қолса, оҳирги фойдаланилган қоида нуқтасига қайтарилади ва бошқа қоидадан фойдаланишга ҳаракат қилинади. Бироқ аввалги қоида асосида тўлдирилган жадвал сақлаб қолинади ва ажратишнинг жорий босқичида зарурат бўйича қўлланилиши мумкин бўлади. Бу билан ҳар-бир фойдаланилган қоида ёрдамида шакллантирилган жадвал (ёки муқобил)нинг устувор жиҳатлари (элементлари)дан фойдаланилади. Бундан бир ёндашувда ёмон натижа берган муқобил бошқасини қўллаётганда яхшироқ натижа бериши мумкинлигини асослайди. Шу мақсадда фаразлар ҳам, текширув натижалари ҳам эслаб қолинади. Бу ёндашув схемали таҳлил дейилади. Уни биринчи бўлиб Мартин Кей Powerful Parser тизимида таклиф қилган [32].

Семантик моделлардан фойдаланиш. Ҳозирда табиий тилдаги матнни таҳлиллаш, қайсидир маънода жумлалар мазмуни асосида генерациялай олувчи лингвистик таҳлиллагич моделлари ишлаб чиқилган. Шу қаторда жараённи моделлаштиришга ёндашувлар ҳам турлича бўлиб, уларни таҳлилда қўлланилаётган воситалари, яъни “тушиниш” даражаси, билимларни ифодалаш ҳажми ва услублари каби лингвистик таҳлиллагич модели сифатини белгилаб берувчи элементлари

асослайди. Уларнинг баъзилари матнларни таҳлиллашга асосланган ифодалашни шакллантирган аниқ тизимларни ташкил этган [33].

Таҳлиллаш масаласига дастлабки матнни мазмунига кўра ажратиш ва ушбу мазмунни тизимнинг ички тилида баён этиш киради. Тизим тилига ўтказиш дастлабки матнни тизим билимига айлантириш тушинилади. Бу билим ўзида ҳам мазмунни ҳам бўлакларни ва улар орасидаги муносабатлар шаклида мужассамлашади. Таҳлилнинг ушбу параметри матн моҳиятини чуқурроқ “англаш”га олиб келади.

Лингвистик таҳлилнинг мавжуд моделларида мазмунни ажратиш ва тасвирлашнинг қуйидаги усулларини ажратиш мумкин: компонентали таҳлил; концептуал тармоқ; образ бўйича мазмунни идентификациялаш; интеграл ёндашув.

Дастлабки матнни формаллаштиришга илк уринишлар компонентали таҳлиллаш ёндашувида амалга оширилган. Унда табиий тил семантикаси тузилмалаштирилмаган семантик кўпайтмалар тўпламидаги якуний атамалар орқали ифодаланиши белгиланган. Сўзларни қараб чиқишда сўзларни алоҳида гуруҳларда таҳлиллаш орқали натижалар олинади. Кейинчалик ушбу ёндашув Ч.Филморнинг “Семантик роллар”и моделида ўз аксини топди [34].

Иккинчи синфдаги моделларга матн мазмуни концептуал тармоқ кўринишида берилувчи моделлар киради. Бу моделлар ёрдамида матннинг концептуал боғлиқликлари топилади [36]. Концептуал тармоқ квази граф бўлиб, унда бинар муносабатлар билан бир қаторда тернар ва кварнар муносабатлар қаралади, қирралари эса нафақат учлари балки бошқа қирраларни ҳам бирлаштиради.

Навбатдаги тур модел “Семантик афзаллик” бўлиб, унда мазмунни идентификациялаш намуналар бўйича амалга оширилади. Моделнинг ўзига хос ҳусусияти уларда морфологик блок ва синтаксис таҳлил мавжуд бўлмайди, бу эса уларнинг камчилиги ҳисобланади. Камчилик ўзини матндаги семантик боғлиқликни аниқ белгилаш учун зарур бўлган сўзларнинг қийматини чуқур таҳлилини таъминлаб беролмайди.

Бу моделда (Уилкс) матн қуйидаги моҳиятлари билан характерланади: сўз мазмуни, хабарлар, матн фрагментлари ва семантик мувофиқлиги билан [35].

Яна бир ёндашувда таҳлиллаш учун намуналар жадвал усули ишлатилади. У гапда учрайдиган калит сўзларни таҳлиллашга асосланган.

Етарли даражада морфологик, синтаксис ва муаммовий таҳлиллар жараёни инобатга олинган моделлар сирасига тилни интеграл ёндашувли тавсифлашга асосланган моделларини киритилади. Бундай моделларга “Мазмун-матн” ва контекст фрагментлаш моделини мисол сифатида келтириш мумкин.

“Мазмун-матн” модели ўзида матнларни кўп даражали мазмуний транслятори ва аксинча жараённи амалга оширади [37]. Тўртта асосий даража ажратилади: фонетик, морфологик, синтаксис ва муаммовий. Уларнинг ҳар-бири муаммовийдан ташқари икки бошқа юза ва чуқур даражаларга бўлинади. Ушбу модел матнни тўлиқ мазмунини тушиниш зарур бўлган тизимларда қўлланилиши мумкин. Бироқ

“Мазмун-матн” моделини тўлиқ қўллаш учун юз минглаб луғат, морфологик ва лексик бирликларнинг жуда катта миқдордаги жуфтликларнинг индивидуал ҳусусиятларини ҳисобга олиш керак бўлади.

Контекст фрагментлаш моделининг асосини уч даражали тизим ташкил этади: лингвистик модел, гапларга ишлов беришнинг базавий механизмлари ҳамда лингвистик билимларни фрагментлаш. Моделда жуда чуқур синтаксис таҳлил бир вақтнинг ўзида топиб олинаётган синтаксис муносабатларни семантикга ўтказиш билан амалга оширилади.

Шундай қилиб, турли ёндашувли ва йўналишли лингвистик процессор моделлари қуйидаги имкониятларни амалга оширади:

берилган матндаги билимларни ҳосил қилиш ва мазмуннинг берилган қийматлари асосида табиий тилнинг тўғри гапларини шакллантириш;

ушбу гаплар бирикмаларини ўзгартириш;

уларнинг боғлиқлигини баҳолаш ва бошқа масалаларни амалга ошириш.

СЕМАНТИК ТАҲЛИЛЛАГИЧ АЛГОРИТМИ

Бу масалаларни ечишнинг бош воситаси баён қилишни ёзиш учун маҳсус семантик тил ҳамда табиий ва семантик тилдаги гапларнинг ўзаро мувофиқлигини аниқлаш механизми ҳисобланади. Таҳлил қилиш жараёни ишлаб чиқилган моделларнинг устиворлиги билан бир қаторда камчиликлари ҳам мавжудлигини кўрсатди. Таҳлил натижаларидан келиб чиққан ҳолда семантик таҳлиллагичнинг умумлашган алгоритми ҳамда ролли тузилмани аниқлаш ва аргументлаш алгоритмларини асослаймиз.

Семантик таҳлиллагичнинг умумлашган алгоритми. Семантик таҳлил алгоритми беш босқичда амалга оширилади:

предикат сўз (ПС) ва уларга мос келувчи луғат мақолаларини қидириш;

ПС аргументларини қидириш;

бошқа аргументлардан мустақил ҳолда ПС луғат мақолаларининг ҳар бирига мос топилган аргументларнинг семантик ролини белгилаш;

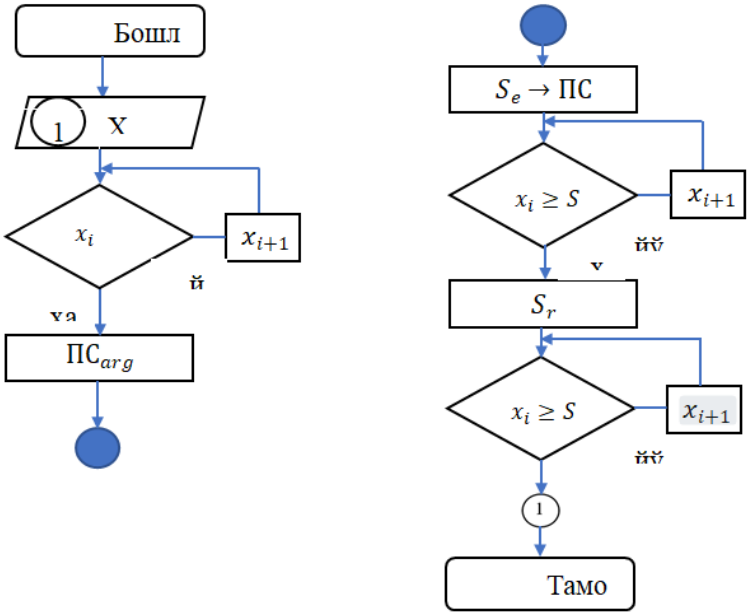
ролларни тақсимлашни оптималлаштириш орқали берилган луғат қисми учун аргументлари бўйича семантик ролларни энг мақбул тақсимотини танлаш;

энг мақбул ПС луғат қисми ва унга мос келувчи семантик роллар жамламасини танлаш.

Биринчи босқичда ПСни қидириш учун гапда морфосинтаксис шаблонлар ишлатилади. Асосан шартлар сўзлашув қисмига ва синтаксис тармоқ жойига жойлаштирилади. Перидикат сўз - бош бўлакни қидиришда матн гапларида семантик луғатдаги ПС леммадан фойдаланади. Шундан сўнг берилган луғат қисми учун ҳар бир аргументга семантик рол жамланмаси белгиланади ва уларнинг вазни аниқланади. Унда семантик роллар аргументлар орасида шундай тақсимланадики, берилган ПС учун ҳар бир семантик аргумент биттадан ортиқ рол ҳосил қилмаслиги таъминланади. Белгиларни таққослаш орқали эвристик асосда семантик рол

аргументига 0 дан 1 гача вазн қиймати берилади ва баҳоланади. Шундай қилиб, натижада ПС луғат қисмига мосроғи учун яхшироқ қамраб олувчи семантик аргументли семантик роллар жамланмаси шакллантирилади.

Гапнинг ролли тузилмасини аниқлаш алгоритми 1-расмда келтирилган:



1-расм. Гапнинг ролли тузилмасини аниқлаш алгоритми.

Гапнинг ролли тузилмасини аниқлашда, аввало, гапнинг морфосинтаксис тузилмаси (X) берилган дастлабки маълумотлар асосида шакллантирилади ва барча ПСлар бўйича таққослама амалга оширилиб, ПС аргументи ва унинг вазни аниқланади. Аниқланган ПС аргументи асосида ПС учун семантик луғат қисми танланади ва ПС луғат қисми рўйхати шакллантирилади. Навбатдаги шарт орқали барча танланган қисмлар асосида аргументларнинг семантик роли (Sr) белгиланади. Шу қаторда ушбу вазнлар ёрдамида ролларни аргументлар бўйича тақсимлаш амалга оширилади. Шундан сўнг ПС бўйича қайта таққосланиш бажарилиб, ПС учун энг мақбул луғат қисми танланади ҳамда гапнинг ролли тузилмаси ҳосил қилинади.

Тажрибавий тадқиқот натижалари. Ўзбекистон Республикаси Олий таълим, фан ва инновациялар вазирлиги (собиқ Олий ва ўрта махсус таълим вазирлиги) ҳузуридаги Таълим муассасаларида электрон таълимни жорий этиш марказида жорий этилган бўлиб, у ўзида бир қанча ахборот ресурсларини мужассамлаштирган ахборот тизимлари мажмуасини ҳосил қилади.

Мазкур тизим ўз таркибига ўндан ортиқ қисм тизимларни мужассамлаштирган, уларнинг тадқиқот доирасида бир қаторларидан ахборот-ресурс манбалари, таҳлиллаш объектлари сифатида фойдаланилди:

www.edu.uz- вазирликнинг асосий веб сайти бўлиб, ўзида маълумот берувчи вазифани бажаради. Шу билан бирга барча зарурий ресурслар билан боғланиш имконини берувчи мурожаатларни сақлайди;

my.edu.uz - Олий таълим тизимида интерактив хизматлар ва ахборот тизимлари портали;

vazir.edu.uz - вазирнинг электрон қабулхонаси бўлиб, мазкур восита ёрдамида вазирга мурожаат қилиш мумкин;

билимни баҳолаш тест тизими ёрдамида турли фанлардан тест назоратларни ўтказиш учун база шакллантирилган;

taklif.edu.uz - Ўзбекистон Республикаси Олий ва профессионал таълим тизимига оид таклифлар портали;

Ўзбекистонда очилган ҳорижий ОТМлар ва факультетлар мониторинг тизими;

- Иш жойидан маълумотнома олиш сервиси;
- Ўқиш жойидан маълумотнома олиш сервиси;
- Раҳбар қабулига ёзилиш сервиси;
- Бўш иш ўринлари сервиси.

Ишлаб чиқилган алгоритмларга асосланган дастурий восита Ўзбекистон Республикаси Олий ва ўрта махсус таълим вазирлиги ҳузуридаги Таълим муассасаларида электрон таълимни жорий этиш марказида масофавий ва электрон таълимни ташкил этиш жараёнида фойдаланиш учун тақдим этилган. Дастурий восита матнли ҳужжатларга дастлабки ишлов бериш ва нормаллаш орқали матнли ҳужжатларни, постларни ва билдирилган фикр-мулоҳазаларни ҳиссий таълуқлилигига қараб юқори аниқликда таснифлашни таъминлаган. Ўзбек тилидаги матнли ҳужжатларни таҳлил қилиш самарадорлиги 10-12% ни ташкил этган.

“Жиззах овози” Жиззах шаҳар ҳокимлиги ва халқ депутатлари кенгаши газетасининг ижтимоий тармоқдаги электрон ахборот-ресурсларидаги турли кўринишдаги постлар, янгиликлар ва турли хабарлар, уларга билдирилган фикр-мулоҳазаларнинг ҳиссий таълуқлилигини таснифлаш масалаларини ҳал этишда жорий қилинган. Ишлаб чиқилган таснифлаш механизмининг қўлланилиши ёзишмалар таҳлилига кетадиган вақтни 50% га қисқартириш имконини берган ҳолда, иш самарадорлигини 12-15%га оширди.

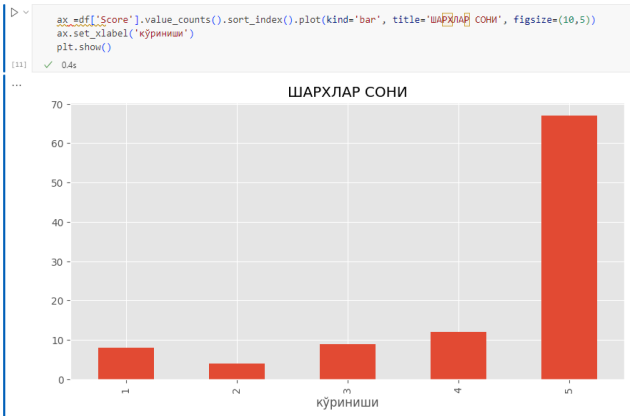
Оила ва хотин-қизлар илмий-тадқиқот институтининг Маҳалла ва оила нашриётининг электрон контентларини таҳлил қилишда қўлланди. Қўллаш натижасида ўзбек тилидаги матнларни таснифлаш орқали таҳлиллаш механизми самарадорлик кўрсаткичи 15-17%ни ташкил этди.

Тажрибавий тадқиқотлар учун “Жиззах овози” газетасининг расмий ахборот манбаидан олинган ёзишмаларнинг ҳиссий таълуқлилиги 2000дан зиёд ўзбек тилидаги постлар таҳлил қилинди.

Амалга оширилган тажрибавий тадқиқотларда дастлабки ишлов бериш орқали эришилган натижалар постларнинг киритилиши ва уларни ажратиш масалаларида қўлланилди. Натижалар қуйидаги кўринишда акслантирилди.

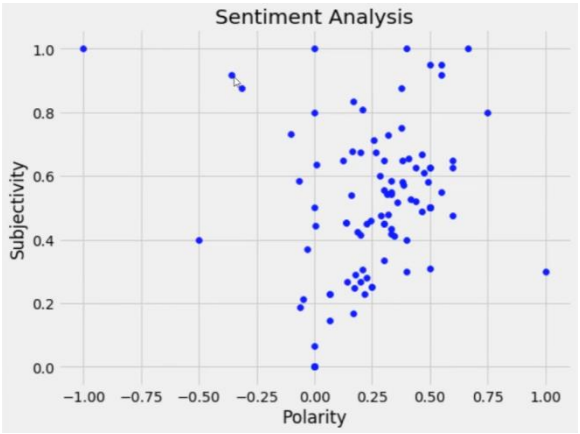
```
sia.polarity_scores('Bu eng yomon narsa')
[244]
{'neg': 0.451, 'neu': 0.549, 'pos': 0.0, 'compound': -0.6249}
```

2-расм. Киритилган кичик изоҳлар таҳлили.



3-расм. Дастлабки 100та шарҳнинг 5 баллик тизимдаги баҳоси.

Ҳосил қилинган матнли ёзишмалар базасида турли қутбли сўзлар массиви ҳосил қилинди. Ушбу сўзлар базаси боғлиқлиги сентиминтал-таҳлили амалга оширилганда қуйидаги кўринишдаги натижа ҳосил бўлди (4):



4-расм. Сентиминтал таҳлиллашда субъектив қутбланиш боғлиқликлари.

Бу ерда субъективлик чегарасини ПС бўйича эксперт белгилаб беради. Қутблилик эса 0 қийматдан юқори бўлиши ижобийликка олиб боради. Ижобий ҳиссиётлар ПС бўйича эксперт чегараси ҳамда ижобий қутб чегараси кесишмаларида аниқланади.

Хулоса. Ўтказилган назарий тадқиқотлар матнли маълумотларни таҳлил қилишда дастлабки ишлов бериш ёндашувлари таҳлилларига асосланган ҳолда адаптив кўринишга эга бўлган ҳамда ўзбек тилидаги матнли маълумотларга ишлов беришга йўналтирилган механизмни ишлаб чиқиш, уларнинг самарадорлигини ошириш бўйича қилинган ишларда ўз аксини топган. Маълумотларни таҳлиллаш механизмини амалга оширишда гап тузилмаларини аниқлашга йўналтирилган дастлабки ишлов бериш компонентаси ишлаб чиқилган ҳамда тажрибавий тадқиқот учун олинган ўзбек тилидаги матнли маълумотларни (ёзишмаларни) таҳлил қилишда олинган натижаларнинг ижобийлиги билан асосланган.

ФОЙДАЛАНИЛГАН АДАБИЁТЛАР РЎЙХАТИ:

1. Xin-She Yang Introduction to Algorithms for Data Mining and Machine Learning// Copyright © 2019 Elsevier Inc. All rights reserved. Academic Press, ISBN: 978-0-12-817216-2, 171p.

2. Hemlata Sahu, Shalini Shrma, Seema Gondhalakar A Brief Overview on Data Mining Survey, International Journal of Computer Technology and Electronics Engineering (IJCTEE), 2013, Volume 1, Issue 3; P. IndiraPriya, Dr. D.K. Ghosh A Survey on Different Clustering Algorithms in Data Mining Technique, International Journal of Modern Engineering Research (IJMER) www.ijmer.com Vol.3, Issue.1, Jan-Feb. 2013 pp-267-274.

3. M. A. Deshmukh, Prof. R. A. Gulhane Importance of Clustering in Data Mining, International Journal of Scientific & Engineering Research, Volume 7, Issue 2, February-2016

4. Jaro M. A. Advances in record linkage methodology as applied to the 1985 census of Tampa Florida // Journal of the American Statistical Association.1989. | 84 (406). | Pp. 414{420. | DOI: 10.1080/01621459. 989.10478785.

5. Рассел С. Искусственный интеллект. Современный подход [Текст] / С. Рассел, П. Норвиг, 2-е изд.: Пер. с англ. – М.: Издательский дом «Вильямс», 2006. – 1408 с.

6. Feldman R. The text mining handbook: advanced approaches in analyzing unstructured data [Текст] / R. Feldman, J. Sanger. – Cambridge University Press, 2007. – 410 p.

7. Moyotl-Hernandez E. An Analysis on Frequency of Terms for Text Categorization [Текст] / E. Moyotl-Hernandez, H. Jimenez-Salazar // Procesamiento del lenguaje natural. – 2004. – Vol. 33. – P. 141-146.

8. Moyotl-Hernandez E. Some Tests in Text Categorization using Term Selection by DTP [Текст] / E. Moyotl-Hernandez, H. Jimenez-Salazar // Proceedings of the Fifth Mexican International Conference on Computer Science ENC'04. – Colima. – 2004. – P. 161-167.

9. Большакова Е., Лукашевич Н., Нокель М. Извлечение однословных терминов из текстовых коллекций на основе методов машинного обучения // Информационные технологии. — 2013. — С. 31—37

10. Usama F., Smyth P., Piatetsky–Shapiro G. From Data Mining to Knowledge Discovery in Databases // Arti_cal intelligence Magazine. | 1996. |17(3). | Pp. 34-54.