ТОШКЕНТ АХБОРОТ ТЕХНОЛОГИЯЛАРИ УНИВЕРСИТЕТИ ҲУЗУРИДАГИ ИЛМИЙ ДАРАЖАЛАР БЕРУВЧИ DSc.13/30.12.2019.T.07.01 РАҚАМЛИ ИЛМИЙ КЕНГАШ АСОСИДАГИ БИР МАРТАЛИК ИЛМИЙ КЕНГАШ

ТОШКЕНТ АХБОРОТ ТЕХНОЛОГИЯЛАРИ УНИВЕРСИТЕТИ ҲУЗУРИДАГИ АХБОРОТ-КОММУНИКАЦИЯ ТЕХНОЛОГИЯЛАРИ ИЛМИЙ-ИННОВАЦИОН МАРКАЗИ

БОБОЕВ ЛОЧИНБЕК БОЙМУРОТОВИЧ

МАТНЛИ ҲУЖЖАТЛАРНИ ТАСНИФЛАШ АЛГОРИТМЛАРИ

05.01.03 – Информатиканинг назарий асослари

ТЕХНИКА ФАНЛАРИ БЎЙИЧА ФАЛСАФА ДОКТОРИ (PhD) ДИССЕРТАЦИЯСИ АВТОРЕФЕРАТИ

Тошкент – 2021

УДК: 004.92

Техника фанлари бўйича фалсафа доктори (PhD) диссертацияси автореферати мундарижаси

Оглавление автореферата диссертации доктора философии (PhD) по техническим наукам

Contents of Dissertation Abstract of the Doctor of Philosophy (PhD) on Technical Sciences

ТОШКЕНТ АХБОРОТ ТЕХНОЛОГИЯЛАРИ УНИВЕРСИТЕТИ ҲУЗУРИДАГИ ИЛМИЙ ДАРАЖАЛАР БЕРУВЧИ DSc.13/30.12.2019.T.07.01 РАҚАМЛИ ИЛМИЙ КЕНГАШ АСОСИДАГИ БИР МАРТАЛИК ИЛМИЙ КЕНГАШ

ТОШКЕНТ АХБОРОТ ТЕХНОЛОГИЯЛАРИ УНИВЕРСИТЕТИ ҲУЗУРИДАГИ АХБОРОТ-КОММУНИКАЦИЯ ТЕХНОЛОГИЯЛАРИ ИЛМИЙ-ИННОВАЦИОН МАРКАЗИ

БОБОЕВ ЛОЧИНБЕК БОЙМУРОТОВИЧ

МАТНЛИ ҲУЖЖАТЛАРНИ ТАСНИФЛАШ АЛГОРИТМЛАРИ

05.01.03 – Информатиканинг назарий асослари

ТЕХНИКА ФАНЛАРИ БЎЙИЧА ФАЛСАФА ДОКТОРИ (PhD) ДИССЕРТАЦИЯСИ АВТОРЕФЕРАТИ

Тошкент – 2021

Техника фанлари бўйича фалсафа доктори (PhD) диссертацияси мавзуси Ўзбекистон Республикаси Вазирлар Маҳкамаси ҳузуридаги Олий аттестация комиссиясида B2019.4.PhD/T1419 рақам билан рўйхатга олинган.

Диссертация Тошкент ахборот технологиялари университети ҳузуридаги ахборот-коммуникация технологиялари илмий-инновацион марказида бажарилган.

Илмий раҳбар: **Бабомурадов Озод Жўраевич**
техника фанлари доктори

Расмий оппонентлар: **Мирзаев Номаз**
техника фанлари доктори
Юлдашев Зафар Бахтиярович
техника фанлари бўйича фалсафа доктори

Етакчи ташкилот: **"UNICON.UZ" - Фан-техника ва маркетинг тадқиқотлари маркази**

Диссертация ҳимояси Тошкент ахборот технологиялари университети ҳузуридаги DSc.13/30.12.2019.T.07.01 рақамли Илмий кенгашнинг 2021 йил «10» сентябрь соат 14⁰⁰ даги мажлисида бўлиб ўтади. (Манзил: 100202, Тошкент шаҳри, Амир Темур кўчаси, 108-уй. Тел.: (99871) 238-64-43, факс: (99871) 238-65-52, e-mail: tuit@tuit.uz).

Диссертация билан Тошкент ахборот технологиялари университети Ахборот-ресурс марказида танишиш мумкин (219 рақам билан рўйхатга олинган.). (Манзил: 100202, Тошкент шаҳри, Амир Темур кўчаси, 108-ўй. Тел.: (99871) 238-65-44).

Диссертация автореферати 2021 йил «20» августа да тарқатилди.
(2021 йил «23» июля даги 1 рақамли реестр баённомаси.)

Р.Х. Хамдамов
Илмий даражалар берувчи илмий кенгаш раиси, техника фанлари доктори, профессор

Ф.М. Нуралиев
Илмий даражалар берувчи илмий кенгаш илмий котиби, техника фанлари доктори, доцент

М.А. Исмаилов
Илмий даражалар берувчи илмий кенгаш қошидаги илмий семинар раиси, т.ф.д., профессор

КИРИШ (фалсафа доктори (PhD) диссертацияси аннотацияси)

Диссертация мавзусининг долзарблиги ва зарурати. Жаҳонда матнли ҳужжатларни таснифлаш тизимларини ишлаб чиқишга алоҳида эътибор қаратилмоқда. Ушбу соҳада матнли маълумотларга дастлабки ишлов бериш, таснифлаш ва таҳлиллаш усуллари ва алгоритмларини ишлаб чиқиш, такомиллаштириш ва амалга ошириш долзарб масалалардан бири ҳисобланади. Бундан ташқари дунёнинг ривожланган мамлакатлари, жумладан, Хитой, Россия, АҚШ, Англия, Германия, Ҳиндистон, Франция ва бошқа давлатларда матнли маълумотларга дастлабки ишлов бериш ҳамда таснифлаш йўналишларининг назарий ва амалий масалаларини ечишга катта эътибор қаратилмоқда.

Жаҳонда матнли маълумотларга ишлов бериш ва матнни таснифлаш усул ва алгоритмларини такомиллаштириш, ишлаб чиқиш ҳамда ҳисоблаш алгоритмларини яратишга йўналтирилган илмий тадқиқот ишлари олиб борилмоқда. Бу борада, жумладан, матнли маълумотларни интеллектуал таҳлилини амалга ошириш учун ҳужжатларни таснифлаш сифатини ошириш ва самарали воситасини ишлаб чиқиш, матнларни интеллектуал таҳлилашни амалга ошириш учун зарурий таснифлаш ва башоратлаш алгоритмларини ишлаб чиқиш, такомиллаштириш ҳамда матн моделлари ва автоматлаштирилган таснифлаш тизимларини яратиш муҳим вазифалардан ҳисобланади.

Республикамизда мазкур йўналишда ўзбек тилидаги матнли ҳужжатларга дастлабки ишлов бериш, интеллектуал таҳлил қилиш ва таснифлашга мўлжалланган дастурий воситаларни ишлаб чиқишга алоҳида эътибор қаратилмоқда. 2017-2021 йилларда Ўзбекистон Республикасини янада ривожлантириш бўйича Ҳаракатлар стратегиясида иқтисодиёт, ижтимоий соҳа ва бошқарув тизимига ахборот-коммуникация технологияларини жорий этиш, идоралараро ҳужжат алмашинув тизимини такомиллаштириш, давлат хизматларини кўрсатиш сифатини ошириш[1], «Электрон ҳукумат» тизимини янада ривожлантириш бўйича йўл харитасида тасдиқланган вазифаларни амалга ошириш, матнли маълумотлар (ҳужжатлар)ни таснифлаш ва башоратлаш орқали электрон мурожаатларни автоматлаштирилган таҳлилини амалга оширувчи бошқариш тизимларини яратиш жамиятни ахборотлаштиришда муҳим аҳамият касб этади. Бу борада, матнли ҳужжатларни таснифлаш ва башоратлаш тизимлари самарадорлигини оширишни инобатга олган ҳолда ривожлантириш, электрон мурожаатларни тўғри таснифлаш асосида манзилли кўриб чиқилишини таъминлаш орқали фуқаро мурожаатларини ҳамда турли кўринишдаги матнли маълумотларни самарали таҳлил қилиш тизимларида қўллаш асосий вазифалардан ҳисобланади.

[1] Ўзбекистон Республикаси Президентининг 2017 йил 7 февралдаги ПФ-4947-сон «Ўзбекистон Республикасини янада ривожлантириш бўйича Ҳаракатлар стратегияси тўғрисида”ги Фармони

2015 йил 18 ноябрда қабул қилинган «Электрон ҳукумат тўғрисида» ги Қонун, Ўзбекистон Республикаси Президентининг 2017 йил 7 февралдаги ПФ-4947-сон «Ўзбекистон Республикасини янада ривожлантириш бўйича Ҳаракатлар стратегияси тўғрисида»ги Фармони, 2021 йил 17 февралдаги ПҚ-4996-сонли «Сунъий интеллект технологияларини жадал жорий этиш учун шарт-шароитлар яратиш чора-тадбирлари тўғрисида»ги Қарори ҳамда мазкур фаолиятга тегишли бошқа меъёрий-ҳуқуқий ҳужжатларда белгиланган вазифаларни амалга оширишга ушбу диссертация тадқиқоти муайян даражада хизмат қилади.

Тадқиқотнинг республика фан ва технологиялари ривожланишининг устувор йўналишларига мослиги. Мазкур тадқиқот республика фан ва технологиялар ривожланишининг IV. «Ахборотлаштириш ва ахборот коммуникация технологияларини ривожлантириш» устувор йўналишига мос равишда бажарилган.

Муаммонинг ўрганилганлик даражаси. Матнли маълумотларга ишлов бериш ва таснифлаш модел, усул ва алгоритмларини ишлаб чиқиш ҳамда такомиллаштириш масалаларини ечиш ва уларни амалиётга жорий этиш бўйича хорижий олимлардан С.Агарвал, Т.Миколов, З.Янг ва бошқаларнинг илмий ишлари диққатга сазовор.

Ўзбекистонда таниб олиш ва матнли маълумотларни интеллектуал таҳлил қилишнинг назарий асосларини ривожлантиришга М.М.Камилов, Т.Ф.Бекмуратов, Ш.Х.Фозилов, С.С.Содиқов, Н.С.Маматов, О.Ж.Бабомурадов ва бошқалар ўзларининг ҳиссаларини қўшиб келмоқдалар.

Ҳозирги кунда матнли маълумотларга дастлабки ишлов бериш орқали матнли ҳужжатларни таснифлаш сифати ва самарадорлигини ошириш, бошқарувга йўналтирилган ахборот тизимларини ишлаб чиқиш ва матнларга ишлов бериш технологиялари жадал суръатлар билан ривожланаётган йўналиш ҳисобланади. Бироқ, бундай технология асосида матнли ҳужжатларни таснифлаш ва таҳлил қилиш автоматлаштирилган тизимларини яратишда вужудга келадиган муаммолар ҳозирги кунгача етарли даражада ҳал этилмаган. Бундан ташқари матнли маълумотларни таснифлаш ва таҳлил қилишнинг ишончли, мустақил ва юқори тезликни таъминловчи усул ва алгоритмларини ишлаб чиқиш муаммоси етарли даражада ўрганилмаган.

Диссертация тадқиқотининг диссертация бажарилган олий таълим ва илмий-тадқиқот муассасасининг илмий-тадқиқот ишлари режалари билан боғлиқлиги. Диссертация тадқиқоти Тошкент ахборот технологиялари университети илмий-тадқиқот ишлари режасининг БВ-М-Ф4-09 «Норавшан муҳитда мураккаб тузилмали ахборотларга ишлов бериш, таниб олиш ва башоратлашнинг баҳоларни ҳисоблаш алгоритмлари мантиқий-эвристик синфига асосланган интеллектуал тизимлари» (2017 2020), ҳамда ПЗ-201906202 «Ўзбек тилидаги матнли ахборотларни таҳлил қилиш ва таснифлаш тизимини ишлаб чиқиш» (2019-2021) мавзудаги лойиҳалари доирасида бажарилган.

Тадқиқотнинг мақсади матнли ҳужжатларни таснифлаш алгоритмлари ва дастурий мажмуасини ишлаб чиқишдан иборат.

Тадқиқотнинг вазифалари:

матнли маълумотларга ишлов бериш ва таснифлаш борасида амалга оширилган илмий ишларни таҳлил қилиш;

матнли маълумотларга дастлабки ишлов бериш усул ва алгоритмларини тадқиқ қилиш;

ўзбек тилидаги матнли ҳужжатларни таснифлаш ёндашувини ишлаб чиқиш;

ўзбек тилидаги матнли ҳужжатларни таснифлашнинг чуқур ўқитишга асосланган алгоритмини ишлаб чиқиш;

мавжуд ва таклиф этилган алгоритмлар асосида матнли ҳужжатларни таснифлаш дастурий мажмуасини ишлаб чиқиш.

Тадқиқотнинг объекти сифатида электрон матнли ҳужжатларни таснифлаш жараёни қаралган.

Тадқиқотнинг предметини матнли маълумотларни таҳлиллаш ва таснифлаш усул ва алгоритмлари ҳамда дастурий таъминоти ташкил этади.

Тадқиқотнинг усуллари. Назарий тадқиқотларда тизимли таҳлил, имитацион моделлаштириш, эҳтимоллик назарияси, математик статистика, дискрет математика, матнларга ишлов бериш ва таснифлаш усулларидан фойдаланилган.

Тадқиқотнинг илмий янгилиги қуйидагилардан иборат:

ўзбек тилидаги матнли ҳужжатларни автоматлаштирилган таснифлаш механизмини амалга ошириш ва такомиллаштиришда эҳтимолли ва шажарали моделлари қурилган;

ахборот ресурсларида мавжуд ўзбек тилидаги матнли ҳужжатларни таснифлаш тизимини амалга оширишда белгилар фазосини шакллантиришга хизмат қилувчи маълумотларга дастлабки ишлов беришнинг кетма-кетликка асосланган CBOW(Continuous Bag of Words) модели такомиллаштирилган;

гиперпараметрларни созлаш орқали чуқур ўқитишга асосланган ўзбек тилидаги матнли ҳужжатларни таснифлаш алгоритми модификацияси ишлаб чиқилган;

матнли ҳужжатларда матн элементларини тўғри шакллантириш ҳамда таснифлаш механизми натижадорлигини яхшиловчи ўзбек тилидаги матнли ҳужжатларни таснифлашнинг CNN+RNN(BLSTM) гибрид алгоритми ишлаб чиқилган.

Тадқиқотнинг амалий натижалари қуйидагилардан иборат:

матнли маълумотларга дастлабки ишлов бериш ва таснифлаш алгоритмлари ишлаб чиқилган;

мавжуд ва ишлаб чиқилган усул ва алгоритмларга асосланган «MT 1.0» ўзбек тилидаги матнли ҳужжатларни танифлаш дастурий мажмуаси ишлаб чиқилган.

Тадқиқот натижаларининг ишончлилиги матнли маълумотларга ишлов бериш ва таснифлаш алгоритмларни ишлаб чиқишда математик аппаратининг тўғри қўлланилиши ҳамда тажрибавий тадқиқотларнинг ижобий натижалари билан тасдиқланади.

Тадқиқот натижаларининг илмий ва амалий аҳамияти. Тадқиқот натижаларининг илмий аҳамияти матнли маълумотларга дастлабки ишлов бериш ҳамда матнли ҳужжатларни таснифлашнинг назарий асосларининг истиқболли ривожланишига ишлаб чиқилган алгоритмларнинг ҳисса қўшиши билан изоҳланади.

Тадқиқот натижаларининг амалий аҳамияти ишлаб чиқилган дастурий мажмуа матнли ҳужжатларни таснифлаш автоматлаштирилган тизимларини яратишда қўлланилиши билан изоҳланади.

Тадқиқот натижаларининг жорий қилиниши. Матнли ҳужжатларни таснифлаш масаласини ечиш билан боғлиқ бўлган мавжуд ҳамда таклиф этилган усул ва алгоритмлари асосида яратилган дастурий таъминот асосида:

ўзбек тилидаги матнли ҳужжатларни таснифлашда дастлабки ишлов беришнинг кетма-кетликка асосланган CBOW модели ва таснифлашнинг CNN+RNN(BLSTM) гибрид алгоритми асосида яратилган дастурий восита «Darakchi Inform Service» МЧЖ да фойдаланиш учун жорий этилган (Ахборот технологиялари ва коммуникацияларни ривожлантириш вазирлигининг 2021 йил 18 январдаги 33-8/395-сон маълумотномалари). Натижада ишлаб чиқилган дастурий восита янгиликларни рукнларга таснифлашда 90% дан юқори аниқликда ишламоқда. Бу янгиликларни таснифлаш учун кетадиган вақтни 50% га қисқартириш имконини берди;

ўзбек тилидаги матнли ҳужжатларни автоматлаштирилган таснифлашнинг эҳтимолли, шажарали ва параметрсиз моделлари асосида яратилган дастурий мажмуа Ўзбекистон Республикаси Халқ таълими вазирлиги ҳузуридаги Халқ таълими соҳасида ахборот коммуникация технологияларини ривожлантириш марказига жорий қилинган (Ахборот технологиялари ва коммуникацияларни ривожлантириш вазирлигининг 2021 йил 18 январдаги 33-8/395-сон маълумотномалари). Натижалар матнли ҳужжатларга дастлабки ишлов бериш ва нормаллаш орқали ҳужжатларни таснифлашга кетадиган вақтни 20-25% га қисқартириш имконини берди;

гиперпараметрларни созлаш орқали чуқур ўқитишга асосланган ўзбек тилидаги матнли ҳужжатларни таснифлаш алгоритми асосида яратилган 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 жорий этилган (Ахборот технологиялари ва коммуникацияларни ривожлантириш вазирлигининг 2021 йил 18 январдаги 33-8/395-сон маълумотномалари). Натижада матнли ҳужжатларга дастлабки ишлов бериш ва нормаллаш орқали ҳужжатларни юқори аниқликда таснифлашни таъминламоқда. Бу ўзбек тилидаги матнли ҳужжатларни таҳлил қилиш самарадорлигини 15-20% га ошириш имконини берди.

Тадқиқот натижаларининг апробацияси. Мазкур тадқиқот натижалари 7 та халқаро ва 14 та республика илмий-амалий анжуманларида маъруза қилинган ва муҳокамадан ўтказилган.

Тадқиқот натижаларининг эълон қилинганлиги. Тадқиқот мавзуси бўйича жами 29 та илмий иш чоп этилган бўлиб, шулардан Ўзбекистон

Республикаси Олий аттестация комиссияси томонидан тавсия этилган илмий нашрларда 6 та мақола, жумладан 1 таси хорижий ва 5 таси республика журналларида нашр қилинган ҳамда 1 та ЭҲМ учун яратилган дастурий воситаларини қайд қилиш гувохномалари олинган.

Диссертациянинг тузилиши ва ҳажми. Диссертация кириш, тўртта боб, хулоса, фойдаланилган адабиётлар рўйхати ва иловалардан иборат. Диссертациянинг ҳажми 117 бетни ташкил этади.

ДИССЕРТАЦИЯНИНГ АСОСИЙ МАЗМУНИ

Кириш қисмида диссертация мавзусининг долзарблиги ва зарурияти асосланган бўлиб, тадқиқотнинг Ўзбекистон Республикаси фан ва технологиялари тараққиётининг устувор йўналишларига мослиги кўрсатилган. Тадқиқотнинг мақсад ва вазифалари белгилаб олинган ҳамда тадқиқот объекти ва предмети аниқланган, олинган натижаларнинг ишончлилиги асослаб берилган, уларнинг назарий ва амалий аҳамияти кўрсатилган, тадқиқот натижаларини амалиётга жорий қилиш ҳолати, нашр этилган ишлар ва диссертация тузилиши бўйича маълумотлар келтирилган.

Диссертациянинг **«Маълумотларга ишлов бериш тизимларида матнни таснифлашнинг назарий асослари»** номланган биринчи бобида матнли ҳужжатларни таснифлашнинг назарий асосланиши, муаммонинг ҳозирги ҳолатини таҳлил қилиш натижалари, ҳал қилиш ёндашувлари келтирилган бўлиб, таҳлил қилинган ҳар бир ёндашувнинг ўзига ҳос ижобий томонлари ва учраган камчиликлари акс эттирилган. Бундан ташқари, матнли маълумотларни таснифлаш йўналишида мавжуд бўлган асосий муаммолар шакллантириб олинган. Матнли ҳужжатларга дастлабки ишлов бериш ва таснифлашнинг маълум усул, модел ва алгоритмларини тадқиқ қилиш асосида мазкур йўналишдаги тадқиқот ишларини ривожлантириш муаммолари, уларни ҳал этиш йўлларига аниқлик киритилган ҳолда илмий тадқиқот ишнинг мақсади ва вазифалари шакллантирилган.

1.1-параграфда маълумотлар интеллектуал таҳлилининг асосий масалаларидан бўлган такрорланувчи белгиларни ажратиш, кластер таҳлили, четлашишни таҳлил қилиш ва таснифлаш масалаларига тўхталиб ўтилган бўлиб, олиб борилган таҳлилий тадқиқот асосида матнли маълумотларга ишлов беришда маълумотларни интеллектуал таҳлиллаш усул ва воситаларининг ўзига ҳос жиҳатлари очиб берилган.

Маълумотларга ишлов беришда ўқитиш ва автоматик таснифлаш масалаларини ечишдаги эришилган натижалар таҳлил этилганда уларнинг ўзига ҳос жиҳатлари ажратиб берилган бўлиб, уларни қай ҳолатларда қўлланилиши мақсадга мувофиқлиги кўрсатилган. Анъанавий усулларни қўлланилиши катта массивли матнли маълумотларга ишлов беришда катта ҳатоликлар келтириб чиқариши, мураккароқ тузилмага эга бўлган чуқур ўқитишга асосланган ёндашувларда эса ўқув танлама миқдорининг катта бўлиши ҳамда параметрлаштиришга талабнинг юқорилиги катта камчилиги эканлиги ойдинлаштирилган.

1.2-параграфда матнни таснифлаш борасида дунё тажрибаси, амалга оширилган тадқиқотлар, олинган натижалар тадқиқотнинг асосий йўналишини белгилаб олиш мақсадида қилинган таҳлили келтирилган. Ўтказилган таҳлилий тадқиқот асосида матнли ҳужжатларни таснифлаш борасида олиб борилаётган тадқиқотларнинг асосий йўналишлари ва эришилаётган натижалар қараб ўтилди. Таҳлил асосида матнларни таснифлаш соҳасида ривожланишнинг асосий тенденцияси аниқланган ва тадқиқот мобайнида эътибор бериладиган асосий жиҳатлар белгилаб олинган. Назарий ишланмаларга асосланган ҳолда ишлаб чиқилган турли кўринишдаги тизимлар таҳлили тадқиқот ишининг 1.3-параграфида баён этилган. Унда табиий тил тури ва таркибини ҳисобга олмаган ҳолда матнларни таснифлаш тизимларини қандай архитектура ҳамда таркибий элементлар билан амалга оширилиши баёни келтирилган. Мазкур йўналишдаги ютуқ ва камчиликлар, уларни келтириб чиқарувчи асосий омиллар, тадқиқот доирасида ишлаб чиқиладиган тизимни қуришда инобатга олиниши учун белгилаб қўйилган.

Амалга оширилган назарий ва амалий ишланмалар таҳлилидан келиб чиққан ҳолда предмет соҳа бўйича амалга оширилган аналитик таҳлил натижаларидан келиб чиққан ҳолда тадқиқот масаласининг қўйилиши 1.4-параграфда акс эттирилган.

Диссертациянинг «**Матнли маълумотларни таснифлаш жараёни**» деб номланувчи иккинчи боби тўртта параграфдан иборат бўлиб, матнли маълумотларга дастлабки ишлов бериш босқичлари, матн белгилари ўлчовини қисқартириш ёндашувлари, матнли ҳужжатларни таснифлашнинг анъанавий моделлари ва таснифлашни баҳолаш ўлчовларига бағишланган.

2.1-параграфда матнли маълумотларни таснифлаш масаласини ечишда асосий рол ўйнайдиган маълумотларга дастлабки ишлов беришнинг бир қатор ёндашувлари қараб ўтилган бўлиб, уларнинг асосида қўйилаётган матнли маълумотларни таснифлаш масаласини ечиш самарадорлигини ошириш учун тадбиқ этиш ёндашувлари таклиф этилган. Бунинг учун мазкур қисмда платформага асосланган ажратиш ва ўгириш, бўлакларни қайта ишлаш, нормаллаштириш, TF-IDF модели, Word2Vec моделлари тадқиқ қилиниб, дастлабки ишлов беришнинг мос самарали ёндашуви қўлланилди.

2.2-параграфда матнларни таснифлашни самарали амалга оширишга хизмат қилувчи матн белгиларини камайтириш ёндашувлари қараб ўтилган бўлиб, уларни турли ҳолатларда қўлланиши бўйича таклифлар берилган. Шулардан анъанавий усулларга асосланган ўлчамни камайтириш ёндашуви беш босқични ўз ичига олади ва у қуйидагича:

1) Дастлабки ишлов бериш термин индексини ажратиб олишни ва матнни тозалаш амалга оширилади, натижада m белгиларга эга n ҳужжатлар ҳосил бўлади;
2) 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;

3) Усулни барча ҳужжатлардаги барча терминларга бирма-бир қўллаш;
4) Киритилган ҳужжат векторини r ўлчамли фазога проекциялаш;
5) Худди шу трансформациядан фойдаланиб, r - ўлчамли фазога синов тўпламини акслантириш.

2.3-параграфда мавжуд матнли ҳужжатларни таснифлаш алгоритмлари ўзбек тилидаги матнни таснифлаш учун қўлланилиши баён қилинган бўлиб, турли ҳолатларда матнни таснифлаш учун соддадан мураккабга йўналтирилган алгоритмлар тадқиқ этилган. Матнни таснифлашда фойдаланиладиган Роккио алгоритми, ансамбл ўқитиш алгоритмларининг иккита машҳур усулига мурожаат қилинган: boosting ва bagging, мантиқий регрессия, Содда Байес ва k-яқин қўшнилар каби ёндашувлар қўлланилиши қараб чиқилган. Шу билан бирга таснифлашда SVM, қарор дарахти ва тасодифий ўрмонлар каби алгоритмларни ҳужжатларни таснифлаш учун тезлиги ва аниқлиги юқорилиги тадқиқотда аниқланди. Мазкур ёндашувлар тадқиқот давомида ўзбек тилидаги матнли маълумотларни таснифлаш учун қўллаб кўрилди. 2.4-параграфда тадқиқот давомида таснифлаш сифатини аниқлаш ва такомиллаштиришни яхшилаш мақсадида таснифлашни баҳолаш ўлчовларини ўз ичига олган мезонлар тадқиқ қилинган. Бу матнли ҳужжатларни таснифлашда битта моделдан фойдаланишнинг ўзи етарли эмаслигини асослади. Тадқиқотнинг мазкур қисмида таснифлашни баҳолаш ўлчовлари асосида таснифлаш жараёнининг турли босқичида турли механизмлардан фойдаланиш ўринли эканлигини исботлаган.

Диссертациянинг **«Матнли ҳужжатларни таснифлашнинг гибрид алгоритмини ишлаб чиқиш»** деб номланувчи учинчи боби матннинг кетма-кетлик моделини қуриш, чуқур ўқитишга асосланган матнни таснифлаш моделлари тадқиқ қилинган, матнли ҳужжатларни таснифлаш гибрид алгоритмларини ишлаб чиқишга бағишланган.

Диссертациянинг 3.1-параграфида матнларни кетма-кетлик моделига асосланган алгоритми тавсифланган.

Матнннинг j-кирувчи сўзи учун p-ўлчовли жойлаштирилиши бутун корпус бўйлаб қаралиши $\bar{u}_j = (u_{j1}, u_{j2}, \ldots, u_{jp})$ ни беради ва $\bar{h} = (h_1 \ldots h_p)$ кирувчи матнннинг ўзига хос жойлашувини таъминлайди. Яширин қатлам қуйидаги натижани ҳосил қилади:

$$h_q = \sum_{i=1}^{m} \left[\sum_{j=1}^{d} u_{jq} x_{ij} \right] \qquad \forall q \in \{1 \ldots p\} \tag{1}$$

Ифода кўпайтмаси учун аҳамиятсиз бўлсада, аксарият ҳолларда ифоданинг ўнг томонида m нинг аниқланиш соҳасидан фойдаланилади. Ушбу муносабатни вектор шаклида ҳам ёзиш мумкин:

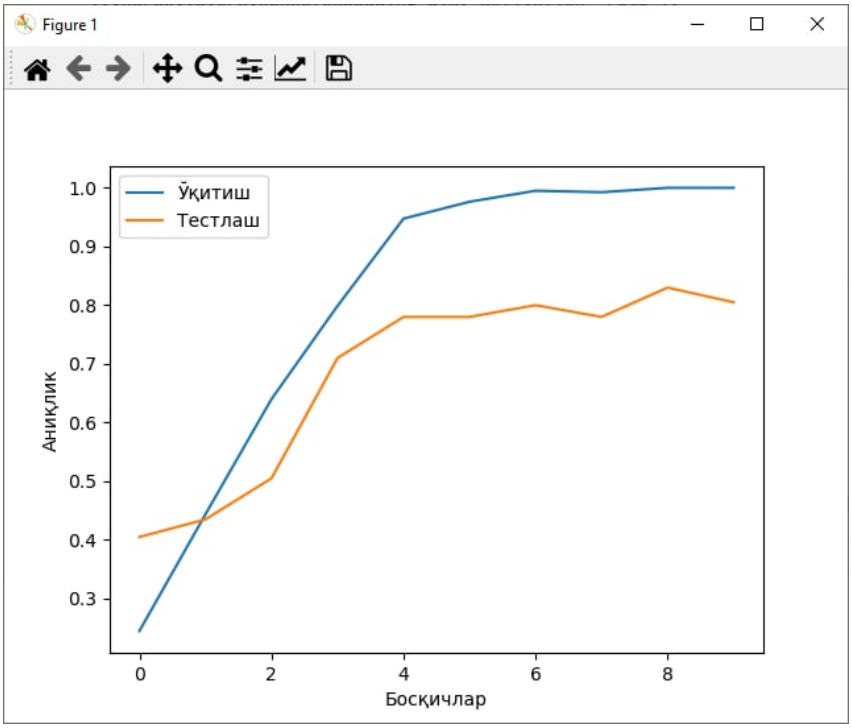
$$\bar{h} = \sum_{i=1}^{m} \sum_{j=1}^{d} \bar{u}_j x_{ij} \tag{2}$$

Жойлаштириш орқали $(h_1 \ldots h_p)$ мақсад сўзнинг ҳар-бир d-натижадан биттаси бўлишини тахмин қилиш учун *softmax* функциясидан фойдаланилади. Чиқувчи қатлам вазнлари $p \times d$ матрица орқали аниқланиб, $V = \left[v_{qj} \right]$ билан ифодаланади. *Softmax* функсияси битта қиймат 1 қолганлари 0 бўлган натижаларда y_j нинг $P(w \mid w_1 \ldots w_m)$ - эҳтимоллиги қуйидагича ҳисобланади:

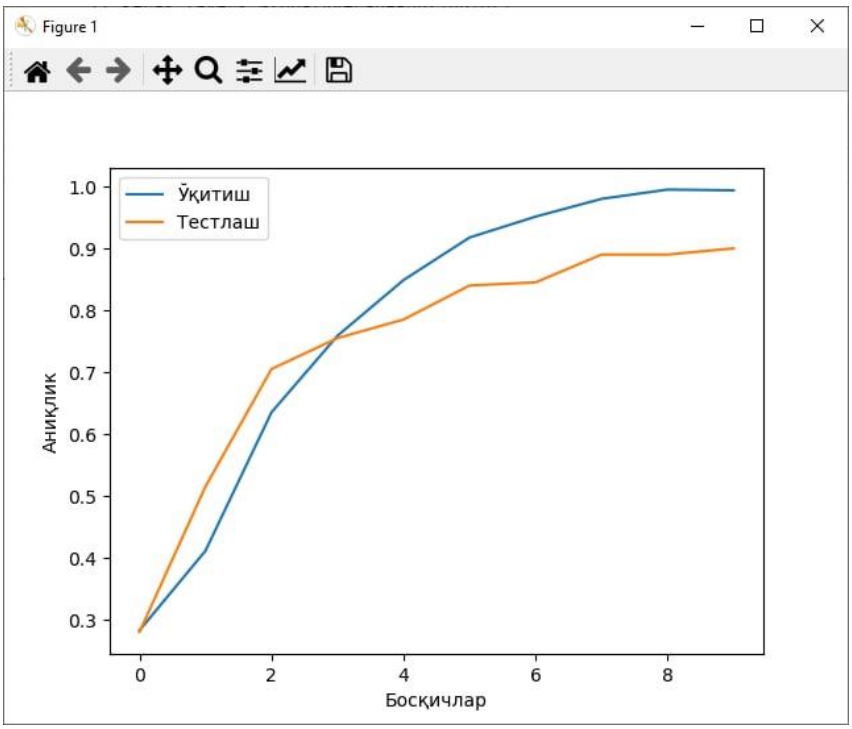
$$\hat{y}_j = P(y_j = 1 \mid w_1 \ldots w_m) = \frac{\exp\left(\sum_{q=1}^{p} h_q v_{qj} \right)}{\sum_{k=1}^{d} \exp\left(\sum_{q=1}^{p} h_q v_{qk} \right)} \quad (4)$$

3.2-параграфда кўп қатламли нейрон тармоқлари, такрорланувчи нейрон тармоқлари (RNN) ва конволюцион нейрон тармоқлари (CNN) модел ва архитектураси қараб ўтилган бўлиб, ушбу ёндашувлар асосида ўзбек тилидаги матнларни таснифлаш алгоритмларининг ишлаш самарадорлиги тажрибавий тадқиқотлар асосида асослаб берилган.

Олиб борилган тадқиқотлар натижасида чуқур ўқитишнинг RNN ва CNN усуллари мос равишда 80% ва 90% аниқлик билан ишлаши аниқланди (1 ва 2-расмлар).



1-расм. RNN усулида олинган натижалар



2-расм. CNN усулида олинган натижалар

3.3-параграфда чуқур ўқитишга асосланган ёндашувлар негизида гибрид алгоритм ишлаб чиқилган бўлиб, уни ўзбек тилидаги матнли ҳужжатларга тадбиқ этиш усули асосланган. Ушбу алгоритм 8 қадамда амалга оширилган:

1-қадам: $D=\{w_1,w_2,\ldots,w_n\}$ кириш ҳужжати дастлабки ишлов бериш босқичидан ўтказилади.

2-қадам: Ҳужжатдаги ҳар бир w_i сўз, $v_i\in\mathbb{R}^d$ - сўзлар векторига трансформация қилинади,.

3-қадам: $w_{i:i+k-1}=\{w_i,w_{i+1},\ldots,w_{i+k-1}\}$ сўзлар кетма-кетлиги учун мос равишда $v_{i:i+k-1}=\{v_i,v_{i+1},\ldots,v_{i+k-1}\}$ векторлар шакллантирилади.

4-қадам: $x_i=f\left(W\cdot v_{i:i+k-1}+b\right)$ қийматлари генерация қилинади, $W\in\mathbb{R}^{d\times k}$. Филтр ойнаси тўлиқ ўтказилганда $m=\left[x_1,x_2,\ldots,x_{n-h+1}\right]$ ҳосил бўлади. l марта филтрлаш орқали $M=\{m_1;m_2;\cdots,m_l\}$ ҳосил қилинади, $M\in\mathbb{R}^{l\times(n-k+1)}$, $E=n-k+1$.

5-қадам: Ҳужжатдаги сўзларнинг тўлиқ боғлиқлигини қамраб олиш мақсадида M чиқиш векторлари LSTM қатламга узатилади. LSTM яширин қатлами H билан белгиланади. t-қадамдаги учта эшик яъни, кириш i_t, чиқиш o_t ва унутиш f_t қуйидагича янгиланади.

$$i_t=\sigma\left(W_i m_t+U_i h_{t-1}+b_i\right),$$
$$o_t=\sigma\left(W_o m_t+U_o h_{t-1}+b_o\right),$$
$$f_t=\sigma\left(W_f m_t+U_f h_{t-1}+b_f\right),$$
$$c_t=f_t\otimes c_{t-1}+i_t\otimes\tanh\left(W_c m_t+U_c h_{t-1}+b_c\right),$$

$$h_t = o_t \otimes \tanh(c_t)$$

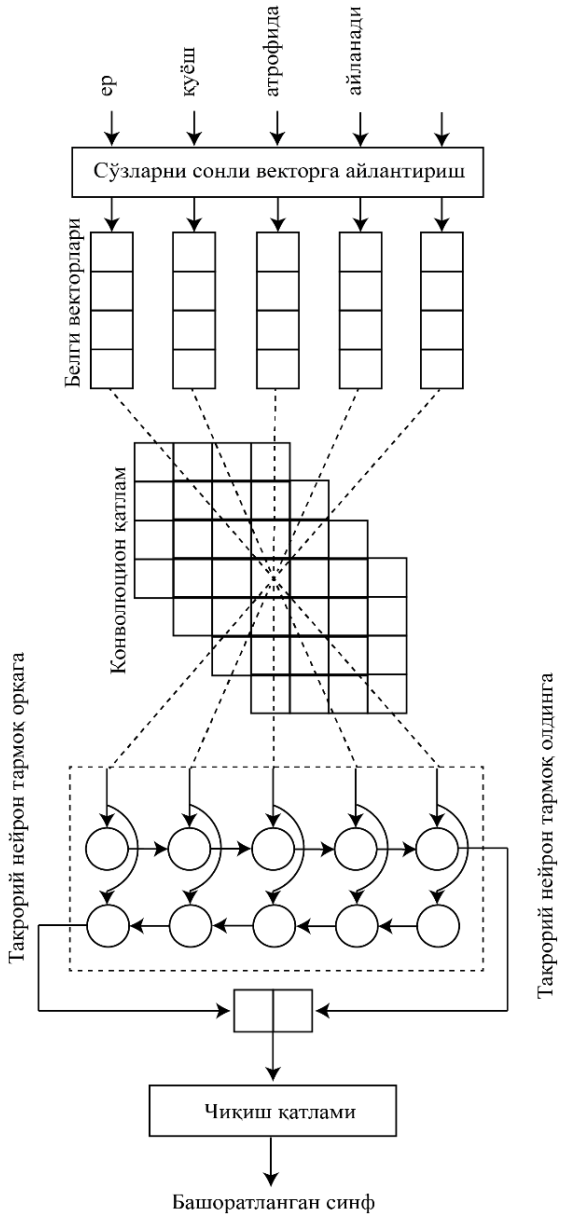
ҳисоблашлар икки йўналишда амалга оширилади. Бу ерда σ - сигмоид функция, $W \in R^{H\times E}$, $U \in R^{H\times H}$, $b \in R^{H\times 1}$ тармоқ параметрлари.

6-қадам: H ва H' конкатинацияланиб чиқиш қатламга киритилади.

7-қадам: *Softmax* функцияси орқали ҳужжат синфи аниқланади,

$$P_i(y) = \frac{\exp(y_i)}{\sum_{j=1}^{K} \exp(y_j)},\ i = 1,2,\ldots,K.$$

8-қадам: Тамом.

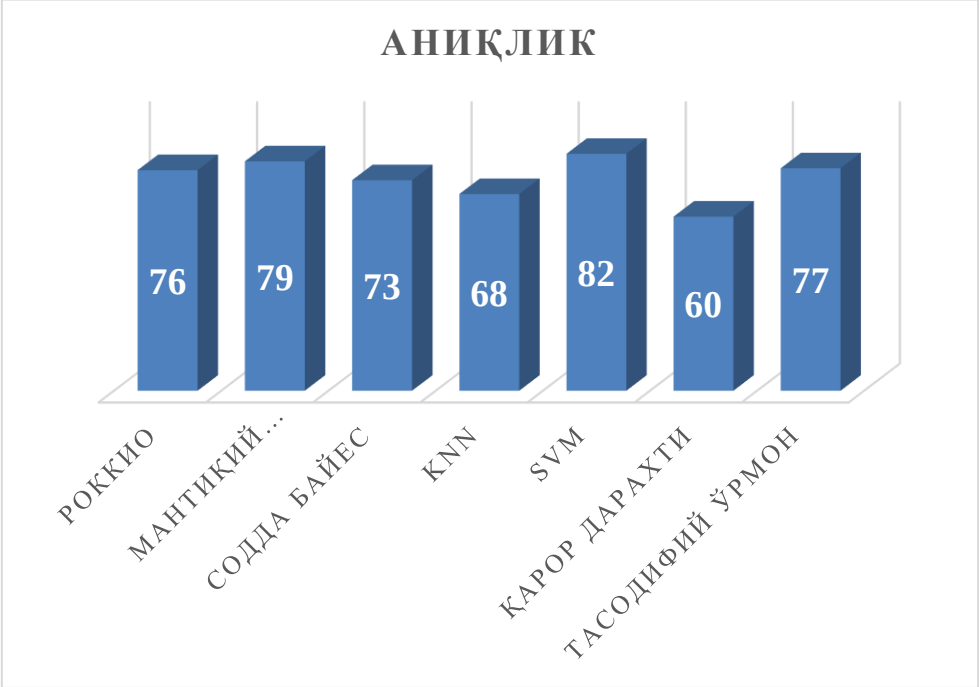


3-расм: Таснифлашнинг «CNN+RNN(BLSTM)» гибрид архитектураси

Диссертациянинг **«Матнли ҳужжатларни таснифлаш тажрибавий тадқиқот натижалари»** деб номланувчи тўртинчи боби тажрибавий тадқиқотлар натижаларини баён этиш ва амалий масалаларни ечишга бағишланган.

4.1-параграфда диссертация тадқиқоти объектлари тавсифи келтирилган бўлиб, унда асосий элементлар ажратиб тавсифланган. Мос равишда ҳар бир объектга қандай ёндашувларни қўлланиш усули баён этилган. Унга кўра амалий масалалар таркиби аниқланган ва предмет соҳа қамрови ажратилган. Танланган масалаларни ечиш учун турли алгоритм натижалари келтирилган.

Тажрибавий тадқиқотлар учун Ўзбекистон Миллий ахборот агентлиги давлат расмий ахборот манбаидан 10 та категорияга тегишли 1847 та энг охирги ўзбек тилидаги янгиликлар пости олинди. Ўтказилган ҳисоблаш тажрибаларида тадқиқот ишининг иккинчи ва учинчи бобларида тавсифланган усул ва алгоритмлар қўллаб кўрилди ва натижалар олинди. Анъанавий алгоритмлардан фойдаланган ҳолда олинган натижалар таҳлили қуйидаги кўринишда тасвирланган:



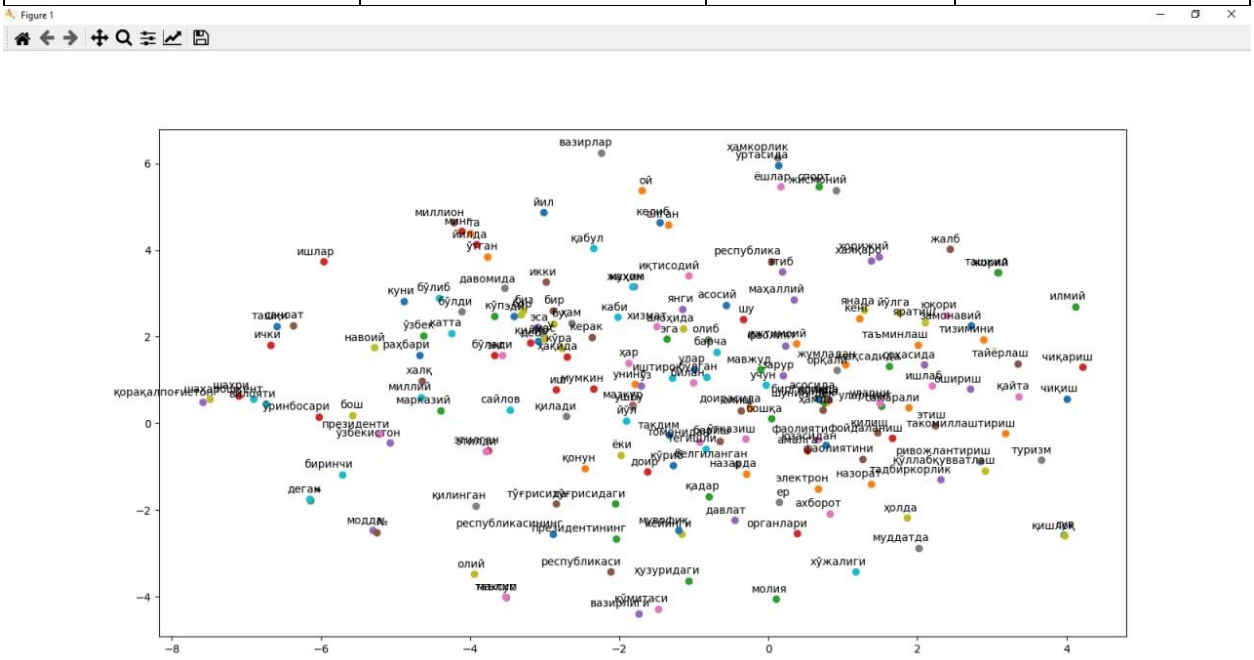
4-расм. Анъанавиалгоритмларида олинган натижалар

4.2-параграфда ишлаб чиқилган гибрид ёндашувга асосланган алгоритм натижалари акс эттирилган. Унга кўра тажрибавий тадқиқотлар учун Ўзбекистон Миллий ахборот агентлиги давлат расмий ахборот манбасидан 59299 та сўз учун вектор узунлиги 64 га тенг бўлган сўз вектори ҳосил қилинди. Қуйида сўзларнинг ўхшашлик жадвали келтирилган.

1-жадвал. CBOW моделига асосланган сўзлар ўхшашлиги

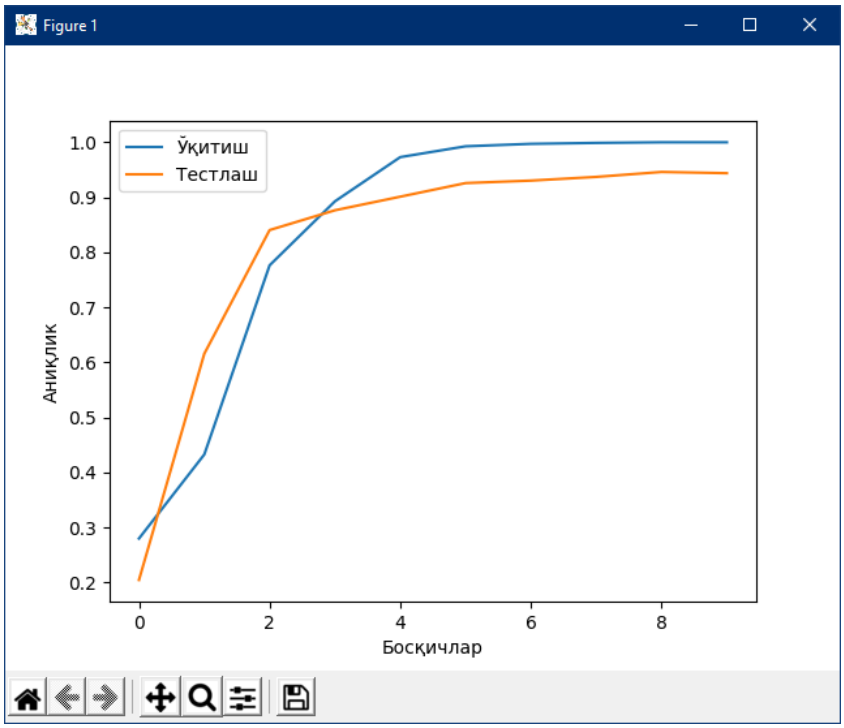
спорт	ахборот	сўм	талаба
тарбия	воситаларида	триллион	ўқиш
мусобақаларини	оммавий	миқдорида	кечки
қишки	воситалари	миллиард	олаётган
болалар	коммуникация	ажратди	таҳсил
турлари	технологиялари	маблағ	мактаб

атлетикага	вебсайтлари	доллар	етишмаслиги
ўсмирлар	дарча	миллион	амалиёт



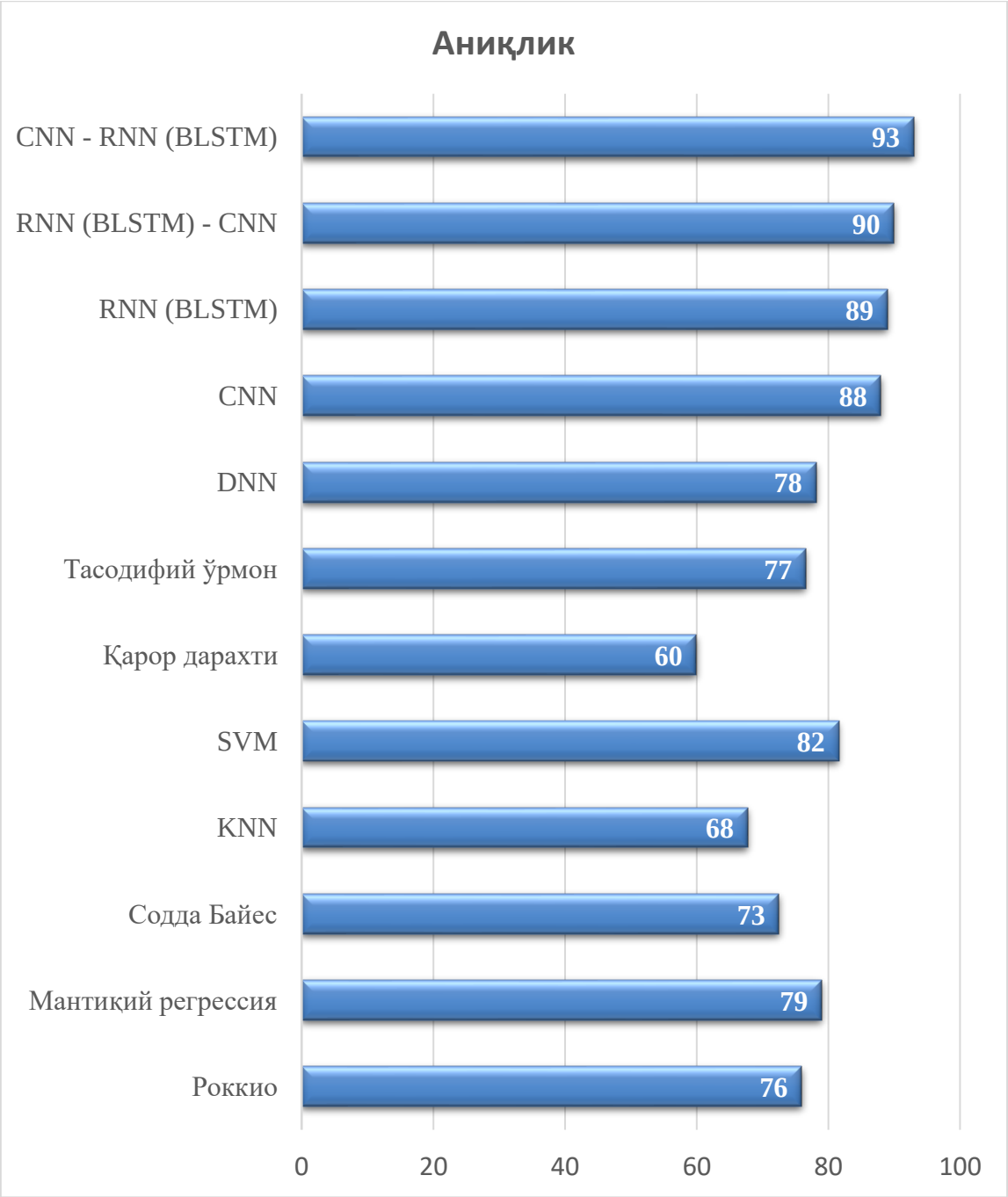
5-расм. «УзА» базаси энг кўп учрайдиган сўзларнинг икки ўлчамда ифодаланиши

Таклиф этилган CNN+RNN(BLSTM) (гибрид) алгоритми 93% аниқликда ишлаши тажрибавий тадқиқотларда аниқланди (6-расм).



6-расм. CNN+RNN(BLSTM) (гибрид) алгоритмида олинган натижалар

4.3-параграф ўтказилган тажрибавий тадқиқотлар таҳлилига бағишланган бўлиб, унинг асосида олинган натижалар ҳаққонийлиги ва самарадорлиги аниқланган.



7-расм. Таснифлаш алгоритмларида олинган натижалар

Маълум бўлган ва диссертацияда таклиф этилган усул ва алгоритмлар асосида матнли маълумотларга дастлабки ишлов бериш ва таснифлаш масалалари қараб ўтилган. Бу матнли ҳужжатларни таснифлаш га кетадиган вақтни 10-25%га ва таснифлаш аниқлигини 10%га оширилган.

Олинган натижалар мазкур соҳа экспертлари томонидан маъқулланган. Навбатдаги босқич тадқиқотларида таснифлаш аниқлигини таснифлашнинг чуқур ўқитишга асосланган гибрид ёндашувлари ҳамда белгилар фазосини қисқартириш орқали ошириш мақсадга мувофиқлиги белгиланди. Таклиф этилган усул ва алгоритмларга асосланган матнларни таснифлаш технологияси ёрдамида амалий тадқиқотларни режалаштириш мазкур соҳада кўпгина мутахассисларга қулайлик олиб келди ва харажатни камайтирди.

ХУЛОСА

Диссертацияда ўзбек тилидаги матнли ҳужжатларга дастлабки ишлов бериш, белгилар фазосини шакллантириш, белгилар фазосини қисқартириш ва таснифлаш алгоритмлари ишлаб чиқилди.

Тадқиқотни амалга оширишдан олинган асосий натижалар сифатида қуйидагиларни қайд этиш мумкин:

1. Матнли ҳужжатларга дастлабки ишлов бериш, таснифлаш ва таҳлиллаш билан боғлиқ бўлган тизимларни яратишда анъанавий ва чуқур ўқитишга асосланган усулларини ишлаб чиқишнинг назарий ва амалий жиҳатларининг замонавий ҳолатини ёритувчи илмий-техник адабиётлар тахлили амалга оширилди. Бу матнларни таснифлаш тизимларини ишлаб чиқишнинг концептуал тамойилларини, конструктив ёндашувларини, усул, модел ва алгоритмларини ишлаб чиқиш имконини беради.

2. Матнли маълумотларни таснифлаш масаласини ечишда асосий рол ўйнайдиган маълумотларга дастлабки ишлов беришнинг бир қатор ёндашувлари асосида таснифлаш масаласини ечиш самарадорлигини ошириш учун қўллаш таклиф этилган.

3. Матнларни таснифлашни самарали амалга оширишга хизмат қилувчи матн белгиларини қисқартириш механизми таклиф этилган.

4. Мантли ҳужжатларни таснифлаш масаласини ечишда белгилар фазосини шакллантиришнинг оддий санаш усули билан бир қаторда матн сўзларининг қўшниларини ҳисобга олган ҳолда самарадорликни ошириш муаллиф томонидан таклиф этилган ёндашуви асосида ишлаб чиқилиши мақсадга мувофиқлиги асосланган. Матнларни таснифлашда ҳозирда кенг қўллашга харакат қилинаётганларидан чуқур нейрон тармоқлари, такрорий нейрон тармоқ, конволюцион нейрон тармоқ ва бирикма усуллар каби нейрон тармоқларига асосланган матнли ҳужжатларни таснифлаш гибрид алгоритмлари таклиф этилган. Бу алгоритм ҳисоблаш тажрибаларида ўзининг самарадорлигини кўрсатди.

5. Таснифлаш натижалари ЎзА ахборот ресурсларида модел масала сифатида тажрибавий тадқиқот натижалари кўринишида келтирилган, олинган натижалар дастлабки ҳолатда ўртача 60-82% кўрсаткич берди. Чуқур ўқитишга асосланган алгоритмлар ўртача 78-90% аниқликни таъминлади. Гибрид алгоритм эса 93% дан юқори натижани таъминлади.

6. Ўзбек тилидаги матнли ҳужжатларни таснифлашга йўналтирилган дастурий воситалар Ўзбекистон Республикаси халқ таълими вазирлиги ҳузуридаги халқ таълими соҳасида ахборот коммуникация технологияларини ривожлантириш марказида, Ўзбекистон Республикаси Олий ва ўрта махсус таълим вазирлиги ҳузуридаги Таълим муассасаларида электрон таълимни жорий этиш марказида ва “Darakchi Inform Service” масъулияти чекланган жамиятида тадбиқ этилган, диссертация натижалари бўйича олинган далолатномалар олинган натижалар самарадорлигини тасдиқлайди.

РАЗОВЫЙ НАУЧНЫЙ СОВЕТ ПРИ НАУЧНОМ СОВЕТЕ DSc13/30.12.2019.T.07.01 ПО ПРИСУЖДЕНИЮ УЧЕНЫХ СТЕПЕНЕЙ ПРИ ТАШКЕНТСКОМ УНИВЕРСИТЕТЕ ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ

НАУЧНО-ИННОВАЦИОННЫЙ ЦЕНТР ИНФОРМАЦИОННО-КОММУНИКАЦИОННЫХ ТЕХНОЛОГИЙ ПРИ ТАШКЕНТСКОМ УНИВЕРСИТЕТЕ ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ

БОБОЕВ ЛОЧИНБЕК БОЙМУРОТОВИЧ

АЛГОРИТМЫ КЛАССИФИКАЦИИ ТЕКСТОВЫХ ДОКУМЕНТОВ

05.01.03 - Теоретические основы информатики

АВТОРЕФЕРАТ ДИССЕРТАЦИИ
ДОКТОРА ФИЛОСОФИИ (PhD) ПО ТЕХНИЧЕСКИМ НАУКАМ

Ташкент – 2021

Тема диссертации доктора философии (PhD) по техническим наукам зарегистрирована в Высшей аттестационной комиссии при Кабинете Министров Республики Узбекистан за B2019.4.PhD/T1419.

Диссертация выполнена в Научно-инноционном центре информационно-коммуникационных технологий при Ташкентском университете информационных технологий.

Научный руководитель:	**Бабомурадов Озод Жураевич** доктор технических наук
Официальные оппоненты:	**Мирзаев Номаз** доктор технических наук **Юлдашев Зафар Бахтиярович** доктора философии по техническим наукам
Ведущая организация:	**“UNICON.UZ” – Центр научно-технических и маркетинговых исследований**

Защита диссертации состоится «10» сентября 2021 г. в 14[00] часов на заседании Научного совета DSc.27.06.2017.T.07.01 при Ташкентском университете информационных технологий. (Адрес: 100202, г. Ташкент, ул. Амира Темура, 108. Тел.: (99871) 238-64-43; факс: (99871) 238-65-52; e-mail: tuit@tuit.uz).

С диссертацией можно ознакомиться в Информационно-ресурсном центре Ташкентского университета информационных технологий (регистрационный номер №219). (Адрес: 100202, г.Ташкент, ул. Амира Темура, 108. Тел.: (99871) 238-65-44).

Автореферат диссертации разослан «20» августа 2021 года.
(протокол рассылки № 1 от «23» июля 2021 г.).

Р.Х.Хамдамов
Председатель научного совета по присуждению учёных степеней, д.т.н., профессор

Ф.М.Нуралиев
Ученый секретарь научного совета по присуждению учёных степеней, д.т.н., доцент

М.А.Исмаилов
Председатель научного семинара при научном совете по присуждению ученых степеней д.т.н., профессор

ВВЕДЕНИЕ (аннотация диссертации доктора философии (PhD))

Актуальность и необходимость темы диссертации. В мире особое внимание уделяется разработке систем классификации текстовых документов. В этом направлении разработка методов и алгоритмов предварительной обработки, классификации и анализа текстовых документов является одной из актуальных задач. Кроме того, в развитых странах мира, таких как Китай, Россия, США, Англия, Германия, Индия, Франция и другие страны, большое внимание уделяется решению теоретических и практических задач в области предварительной обработки и классификации текстовых данных.

В мире ведутся исследовательские работы по усовершенствованию, разработке и созданию вычислительных алгоритмов для обработки текстовых данных, а также методов и алгоритмов классификации текстов. В этом отношении важнейшими задачами являются повышение качества и разработка эффективного инструмента классификации документов для интеллектуального анализа текстовых данных, разработка и совершенствование алгоритмов классификации и прогнозирования необходимых для интеллектуального анализа текстов, а также создание текстовых моделей и автоматизированных систем классификации.

В Республике особое внимание уделяется развитию программных средств предварительной обработки, интеллектуального анализа и классификации текстовых документов на узбекском языке. Реализация утвержденных в Дорожной карте задач по дальнейшему развитию Республики Узбекистан на 2017-2021 годы, внедрение информационно-коммуникационных технологий в экономике, социальной и управленческой сфере, совершенствование системы межведомственного документоборота, повышение качества предоставляемых государственных услуг, дальнейшее развитие системы “Электронное правительство”, где используется автоматический анализ электронных текстов, классификации и прогнозирования являются важными элементами информатизации общества. В этой связи, в качестве основных задач можно отметить исследования по развитию систем классификации и прогнозирования текстовых документов с учетом повышения их эффективности, обеспечения адресного рассмотрения электронных обращений на основе правильной их классификации, а также использования различных видов текстовой информации в эффективных системах анализа.

Настоящее диссертационное исследование в определенной степени служит для реализация задач, предусмотренных Законом Республики Узбекистан «Об электронном правительстве» принятым 18 ноября 2015 года, Указом Президента Республики Узбекистан № ПФ-4947 от 7 февраля 2017 года «О стратегии действий по дальнейшему развитию Республики Узбекистан», Постоновлениямием Президента от 17 июня 2021 года № ПП-4996 «О мерах по созданию условий для ускоренного внедрения технологий искусственного интеллекта», а также другими нормативно-правовыми актами, принятыми в данной сфере.

Соответствие исследования с приоритетными направлениями развития науки и технологий республики. Данное исследование выполнено в соответствии с приоритетным направлением развития науки и технологий IV. “Развитие информатизации и информационно-коммуникационных технологий”.

Степень изученности проблемы. Заслуживают внимания научные работы решения задач по разработке и усовершенствованию методов, алгоритмов и моделей обработки текстовых данных и классификации таких зарубежных учёных, как С. Агарвала, Т. Миколова, З. Янга и др.

В Узбекистане в развитие теоретических основ распознавание и интеллектуального анализа тектовых данных внесли свой вклад М.М.Камилов, Т.Ф.Бекмуратов, Ш.Х.Фазылов, С.С.Садыков, Н.С.Маматов, О.Ж. Бабомурадов и другие ученые.

На сегодняшний день повышение качества и эффективности классификации текстовых документов с помощью предварительной обработки данных, разработка информационных систем, ориентированных на управление и тхнологии обработки текстовых данных считаются наиболее перспективными направлениями исследований. Однако проблемы, возникающие при создании автоматизированных систем классификации и анализа текстовых документов на основе таких технологий все ещё не нашли своего окончательного решения. Кроме того, недостаточно изучена проблема разработки методов и алгоритмов, обеспечивающих надежную, быструю классификацию и анализ текстовых данных.

Связь диссертационного исследования с планами научно-исследовательских работ научно-исследовательского учреждения, где выполнена диссертация. Диссертационное исследование выполнено в рамках проектов плана научно-исследовательских работ Ташкентского университета информационных технологий имени Мухаммада ал-Хоразмий по следующим темам: № БВ-М-Ф4-09 “Интеллектуальные системы обработки сложноструктурированной информации, распознавания и прогнозирования ,основанные на классе логико-эвристических алгоритмов вычисления оценок в нечеткой среде” (2017-2020), № ПЗ-201906202 “Разработка системы анализа и классификации текстовых информаций узбекского языка” (2019-2021).

Целью исследования является разработка алгоритмов и программного комплекса классификации текстовых документов.

Научно-исследовательские цели:

анализ научных работ по обработке и классификации текстовых данных;

исследование методов и алгоритмов предварительной обработки текстовых данных;

разработка подхода к классификации текстовых документов на узбекском языке;

разработка алгоритма классификации текстовых документов на узбекском языке на основе глубокого обучения;

разработка программного комплекса классификации текстовых документов основанных на основе существующих и предложенных алгоритмов.

Объектом исследования является процесс классификации электронных текстовых документов.

Предметом исследования являются методы, алгоритмы и программное обеспечение для анализа и классификации текстовых данных.

Методы исследования. Для проведения теоретических исследований использованы методы системного анализа, имитационного моделирования, теории вероятностей, математической статистики, дискретной математики, обработки текста и методы классификации.

Научная новизна исследования заключается в следующем:

построены вероятностные и иерархически модели реализации и усовершенствования механизма автоматизированной классификации текстовых документов на узбекском языке;

усовершенствован модель CBOW(Continuous Bag of Words), основанная на последовательность предварительной обработки данных, формирующих признаковое пространство при реализации систем классификации текстовых данных на узбекском языке, содержашихся в информационных ресурсах;

разработана модификация алгоритма классификации текстовых документов на узбекском языке, основанная на глубокое обучение путем гипрпараметрической настройки;

разработан гибридный алгоритм CNN+RNN (BLSTM) классификации текстовых документов на узбекском языке, улучшающий результативность механизма правильного формирования и классификации текстовых элементов в текстовых документах.

Практические результаты исследования заключаются в следующем:

разработаны алгоритмы предварительной обработки и классификации текстовых данных;

разработан программный комплекс распознавания текстовых документов на узбекском языке «MT 1.0» на основе существующих и разработанных методов и алгоритмов.

Достоверность результатов исследования обосновывается корректностью применения математического аппарата при разработке алгоритмов обработки и классификации текстовых данных, а также положительными результатами экспериментальных исследований.

Научная и практическая значимость результатов исследования. Научная значимость результатов исследования объясняется вкладом разработанных алгоритмов в перспективное развитие теоретических основ обработки текстовых данных и классификации текстовых документов.

Практическая значимость результатов исследования заключается в том, что разработанный программный комплекс можно использовать при создании автоматизированных систем классификации текстовых документов.

Внедрение результатов исследования. На основе программного обеспечения, созданного на базе известных, а также предложенных в работе методов и алгоритмов решения задачи классификации текстовых документов:

программное средство классификации текстовых документов на узбекском языке, разработанное на основе CBOW модели, базирующейся на последовательность операций предварительной обработки и гибридного алгоритма классификации CNN+RNN(BLSTM) внедрен в деятельности в деятельности ООО «Darakchi Inform Service» (справка Министерства по развитию информационных технологий и коммуникаций №33-8/395 от 18 января 2021 г.). Разработанное программное средство производит классификацию новостей по рубрикам с точностью боле чем 90%, что позволят сокращать время классификации новостей на 50%.

программный комплекс автоматизированной классификации текстовых документов на узбекском языке, разработанный на основе вероятностных, иерархических и непараметрических моделей внедрен в деятельности Центра развития информационных и коммуникационных технологий в народном образовании при Министерстве народного образования Республики Узбекистан (справка Министерства по развитию информационных технологий и коммуникаций №33-8/395 от 18 января 2021 г.). Результаты внедрения позволили повысить эффективность классификации документов за счет предварительной обработки и нормализации текстовых документов на 20-25%.

программное средство, разработанное на основе алгоритма классификации текстовых документов на узбекском языке базирующееся на настройку гиперпараметров путем глубокого обучения внедрен в деятельности Центра внедрения электронного обучения в образовательных учреждениях Министерства высшего и среднего специального образования Республики Узбекистан (справка Министерства по развитию информационных технологий и коммуникаций №33-8/395 от 18 января 2021 г.). В результате обеспечивается высокая точность классификации за счет предварительной обработки и нормализации текстовых документов. Это позволяет улушчить эффективность классификации текстовых документов на узбекском языке на 15-20%.

Апробация результатов исследования. Теоретические и практические результаты диссертационного исследования доложены и обсуждены на 7 международных и 14 республиканских научных конференциях.

Опубликованность результатов исследования. По теме диссертации опубликовано 29 научная статья, из которых 6 статей опубликованы в научных изданиях, рекомендованных ВАК Республики Узбекистан, в том числе 1 зарубежных и 5 национальных журналах, получено 1 свидетельство об официальной регистрации программ для ЭВМ.

Структура и объем диссертации. Диссертация состоит из введения, четырёх глав, заключения, списка использованной литературы и приложений. Объем диссертации составляет 117 страниц.

ОСНОВНОЕ СОДЕРЖАНИЕ ДИССЕРТАЦИИ

Во введении обоснованы актуальность и востребованность темы диссертации, показаны соответсвие исследования приоритетным направлениям развития науки и технологий Республики Узбекистан. Определены цель и задачи, объект и предмет исследования, изложены научная новизна, практические результаты исследования, обоснована достоверность полученных результатов, раскрывается теоретическая и практическая значимость результатов исследования, приведены сведения о внедрении результатов исследования, об опубликованности результатов и структуре диссертации.

В первой главе диссертации **«Теоретические основы классификации текстов в системах обработки данных»** приведены теоретическая обоснованность классификации текстовых документов, результаты анализа текущего состояния проблемы, подходы к решению данной проблемы, а также достоинства и недостатки этих подходов. Кроме того, сформированы основные проблемы в области классификации текстовых данных. На основе проанализированных методов, моделей и алгоритмов предварительной обработки и классификации текстовых документов определены и уточнены направления развития исследований в этой области, поставлены цель и задачи исследований.

Параграф 1.1 посвящен выделению повторяющихся признаков, кластерному анализу, анализу отклонений и классификации, являющихся основными задачами интеллектуального анализа данных. Кроме того, в данном параграфе приведены особенности методов и средств интеллектуального анализа данных при обработке текстовых данных.

При анализе полученных результатов решения задач обучения и автоматической классификации при обработке данных, выделяются их особенности и целесообразности в каких случаях подходить их использования. Выявлено, что использование классических методов приводит к большим ошибкам при обработке больших массивов текстовых данных, а для применения подходов на основе глубокого обучения со сложной структурой требуется большой объем обучающей выборки для настройки параметров.

В параграфе 1.2 приведены результаты анализа мирового опыта иследований в области классификации текстов с целью определения основного направления исследований. Проанализированы основные направления исследований по классификации текстовых документов и полученные в этой области результаты. На основе этого анализа выявлены основные тенденции развития в области классификации текстов и определены основные аспекты, которые необходимо учитывать в ходе данного исследования. Анализ различных систем, разработанных на основе теоретических разработок, проведен в параграфе 1.3 данной диссертационной работы. В данном параграфе описана архитектура и структурные элементы системы классификации текстов, независимые от типа и состава естественного

языка. Определены достоинства и недостатки в этой области, основные факторы, которые влияют на эффективность системы и должны быть учтены при ее разработке.

На основе результатов аналитических исследований предметной области в параграфе 1.4 определена цель и сформулированы задачи данной диссертационой работы.

Вторая глава диссертации **«Процесс классификации текстовых данных»** состоит из четырех параграфов и посвящена предварительной обработке текстовых данных, подходам снижения размерности признаков текстов, традиционным моделям классификации текстовых документов и критериям оценки классификации.

В параграфе 2.1 рассматривается ряд подходов предварительной обработки данных, которые играют основную роль в решении проблемы классификации текстовых данных, и на их основе предлагаются подходы повышения эффективности решения задачи классификации текстовых данных. Для этого рассмотрены вопросы применения эффективного подхода на основе платформы выделения, преобразования, обработки частей и нормализации, модели TF-IDF, модели Word2Vec для реализации предварительной обработки.

В параграфе 2.2 исследуются подходы сокращения признаков в тексте, которые служат для повышения эффективности классификации, выработаны рекомендации по их использованию в различных ситуациях. Среди них подход к уменьшению размерности, основанный на традиционных методах, реализуется в пяти шагах в следующем виде:

1) После предварительной обработки реализуется выделение индексов терминов и очистка текста, в результате формируется n документов с m признаками;
2) Создание N-документа $\left(d \in \{d_1, d_2, \ldots, d_n\}\right)$, где вектор $a_{ij} = L_{ij} \times G_i$, L_{ij} - локальный вес термина i в документе j, а G_i - глобальный вес документа i;
3) Применение метода индивидуально ко всем условиям во всех документах;
4) Проекция введенного вектора документа в r-размерное пространство;
5) Отображение тестового множества в r - размерное пространство, используя ту же трансформацию.

Параграф 2.3 отражает результаты применения существующих алгоритмов классификации текстовых документов для классификации текста на узбекском языке и исследования алгоритмов классификации текста в различных ситуациях, направленных от простого к сложному. Приведены результаты применения таких алгоритмов, как алгоритм Роккио, алгоритм, основанный на известном ансамблемовом обучении (boosting и bagging), логическая регрессия, наивный Байес и k-ближайших соседей. Наряду с этим при классификации текстовых документов получены высокие показатели по

быстродействию и точности такими алгоритмами как SVM, дерево решений и случайные леса. Указанные алгоритмы в последующем применены при классификации текстовой информации на узбекском языке. В параграфе 2.4 исследованы критерии качества классификации, включающие в себе меры оценки классификации. Это обосновывает нецелесообразность применения одной модели в процессе классификации текстовых документов. Кроме того, в данном параграфе доказано, что на основе критериев оценки классификации целесообразно использование различных механизмов на разных этапах процесса классификации.

Третья глава диссертации **«Разработка гибридного алгоритма классификации текстовых документов»** посвящена построению модели последовательности текста, исследованию моделей классификации текстов на основе глубокого обучения, разработке гибридных алгоритмов классификации текстовых документов.

В параграфе 3.1 диссертации описан алгоритм, основанный на модели последовательности текстов.

Рассмотрение для j - входящего слова текста p-мерного размещения по всему корпусу дает $\bar{u}_j = (u_{j1}, u_{j2}, \ldots, u_{jp})$ и обеспечивает своеобразное размещение входящего текста $\bar{h} = (h_1 \ldots h_p)$. Скрытый слой дает следующий результат:

$$h_q = \sum_{i=1}^{m} \left[\sum_{j=1}^{d} u_{jq} x_{ij} \right] \qquad \forall q \in \{1 \ldots p\} \tag{1}$$

Хотя выражение не имеет значения для произведения, в большинстве случаев используется область определения m справа от выражения. Это соотношение также можно записать в векторной форме:

$$\bar{h} = \sum_{i=1}^{m} \sum_{j=1}^{d} \bar{u}_j x_{ij} \tag{2}$$

Для того, чтобы предположить птем размещения, что целевое слово $(h_1 \ldots h_p)$ будет одним из d-результата используется функция *softmax*. Веса выходящего слоя определяются матрицей $p \times d$ и представляются как $V = \left[v_{qj} \right]$. При значениях функции *Softmax* в единственном случае 1, а в остальных 0, $P(w \mid w_1 \ldots w_m)$-вероятность для y_j вычисляется следующим образом:

$$\hat{y}_j = P(y_j = 1 \mid w_1 \ldots w_m) = \frac{\exp\left(\sum_{q=1}^{p} h_q v_{qj} \right)}{\sum_{k=1}^{d} \exp\left(\sum_{q=1}^{p} h_q v_{qk} \right)} \tag{4}$$

В параграфе 3.2 рассмотены модель и архитектура многослойных, рекурентных (RNN) и сверточных нейронных сетей (CNN), экспериментально

обоснована эффективность классификации текстовых документов на узбекском языке на основе этих подходов.

Проведенные исследования показали, что методы глубокого обучения RNN и CNN дают результаты классификации с точностью 80% и 90% соответственно (рисунки 1 и 2).

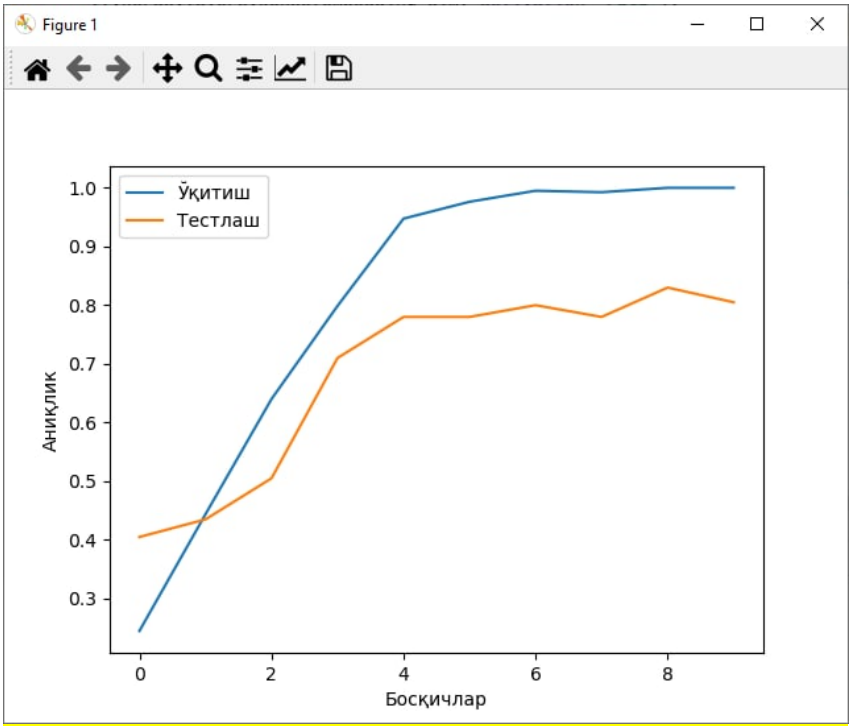


Рис. 1. Результаты классификации, полученные с помощью RNN

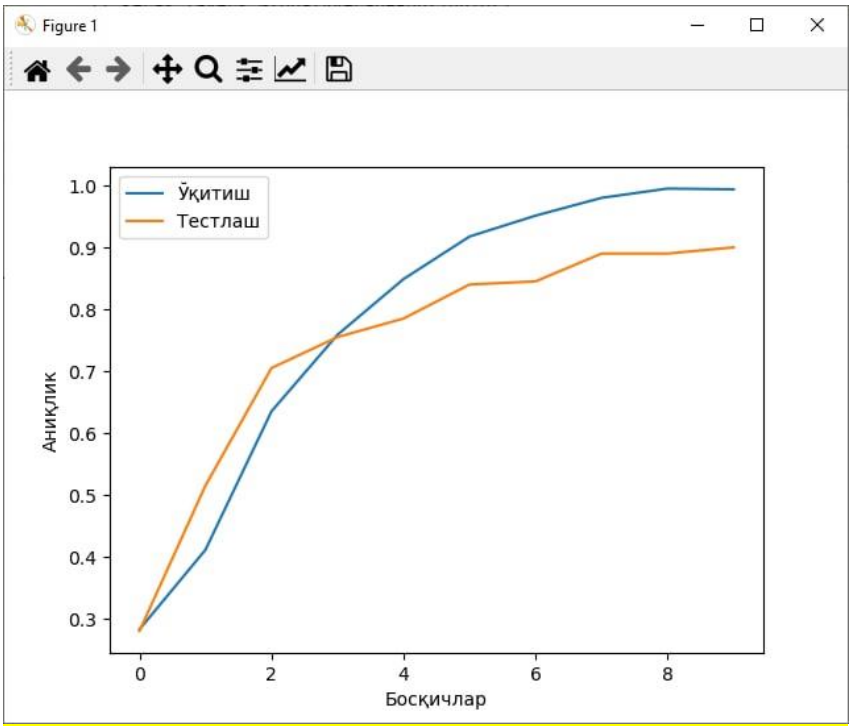


Рис. 2. Результаты классификации, полученные с помощью CNN

В параграфе 3.3 разработан гибридный алгоритм, основанный на глубокое обучение и обоснован метод применения его к текстовым документам на узбекском языке. Алгоритм реализован в 8 шагах:

Шаг 1: $D=\{w_1,w_2,\ldots,w_n\}$ вводной документ проходит стадию предварительной обработки.

Шаг 2: В документе каждое слово w_i преобразуется в вектор $v_i \in \mathbb{R}^d$.

Шаг 3: Для последовательности слов $w_{i:i+k-1}=\{w_i,w_{i+1},\ldots,w_{i+k-1}\}$ соответственно формируются векторы $v_{i:i+k-1}=\{v_i,v_{i+1},\ldots,v_{i+k-1}\}$.

Шаг 4: Генерируются значения $x_i=f\left(W\cdot v_{i:i+k-1}+b\right)$, $W\in\mathbb{R}^{d\times k}$. После полного проведение окна фильтра формируется $m=\left[x_1,x_2,\ldots,x_{n-h+1}\right]$. $M=\{m_1;m_2;\cdots,m_l\}$ формируется путем l разовой фильтрации, где $M\in\mathbb{R}^{l\times(n-k+1)}$, $E=n-k+1$.

Шаг 5: С целью полного охвата слов в документе выходные векторы M передаются на слой *LSTM*. Скрытый слой *LSTM* обозначим как H. Три элемента на шаге t, то есть вход i_t, выход o_t и забвение f_t обновляются следующим образом.

$$i_t=\sigma\left(W_i m_t+U_i h_{t-1}+b_i\right),$$
$$o_t=\sigma\left(W_o m_t+U_o h_{t-1}+b_o\right),$$
$$f_t=\sigma\left(W_f m_t+U_f h_{t-1}+b_f\right),$$
$$c_t=f_t\otimes c_{t-1}+i_t\otimes\tanh\left(W_c m_t+U_c h_{t-1}+b_c\right),$$
$$h_t=o_t\otimes\tanh\left(c_t\right)$$

расчеты реализуются в двух направлениях. Здесь σ - сигмоидная функция, $W\in R^{H\times E}$, $U\in R^{H\times H}$, $b\in R^{H\times 1}$ сетевые настройки.

Шаг 6: После концентрации H и H' вводятся в выходной слой.

Шаг 7: Класс документа определяется с помощью функции *Softmax,*

$$P_i(y)=\frac{\exp\left(y_i\right)}{\sum_{j=1}^{K}\exp\left(y_j\right)},\ i=1,2,\ldots,K.$$

Шаг 8: Конец.

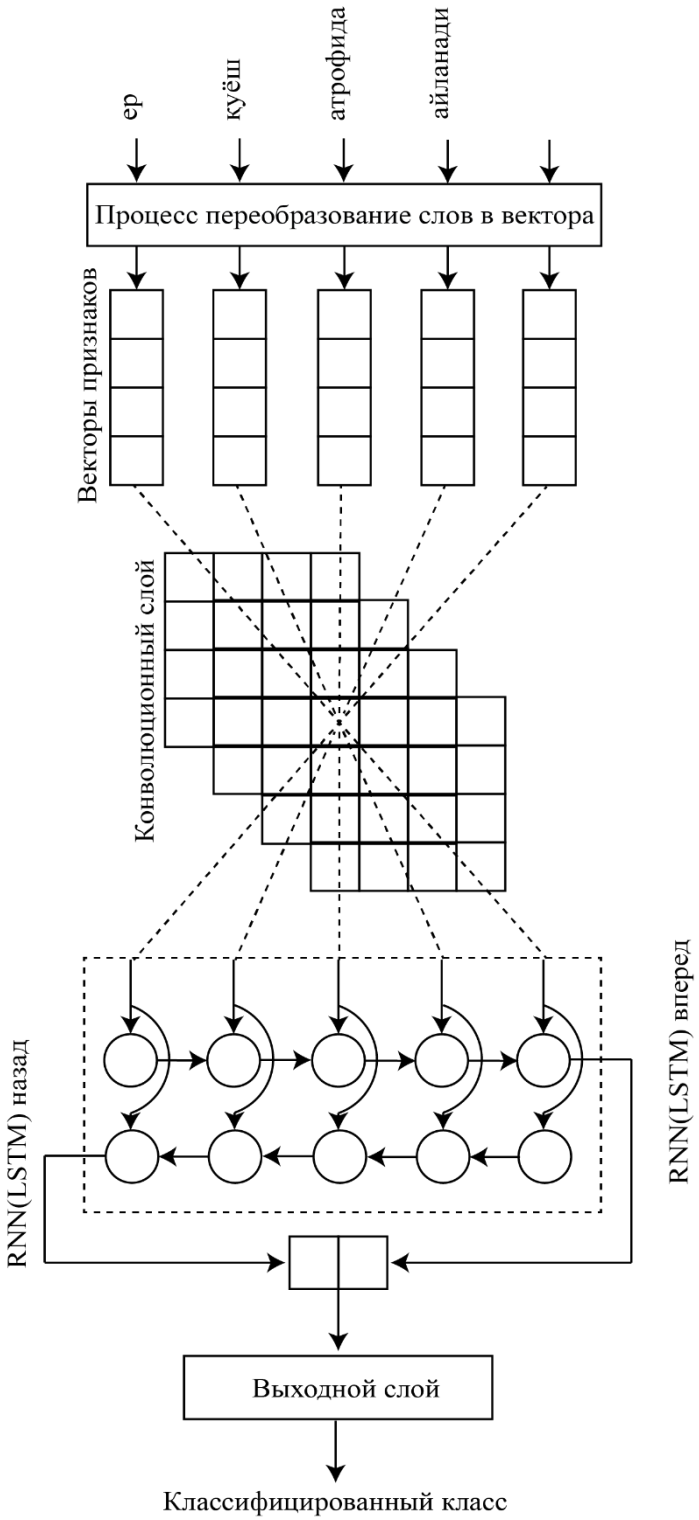


Рис. 3. Гибридная архитектура классификации CNN + RNN (BLSTM)

Четвертая глава диссертации **«Результаты экспериментальных исследований по классификации текстовых документов»** посвящена изложению результатов экспериментальных исследований и решению практических задач.

В параграфе 4.1 приведено описание объектов диссертационного исследования, при этом отдельно описаны основные элементы. Описан метод применения подходов к каждому объекту и определен соотвествующий состав практических задач, а также выделен охват предметной области. Приведены результаты различных алгоритмов решения этих задач.

Для проведения экспериментальных исследований из официального государственного источника Национального информационного агентства Узбекистана отобраны 1847 последних новостных постов на узбекском языке из 10 категорий. В проведенных вычислительных экспериментах использованы методы и алгоритмы, описанные во второй и третьей главах диссертационной работы. Результаты, полученные с использованием классических алгоритмов, проиллюстрированы на рисунке 4.

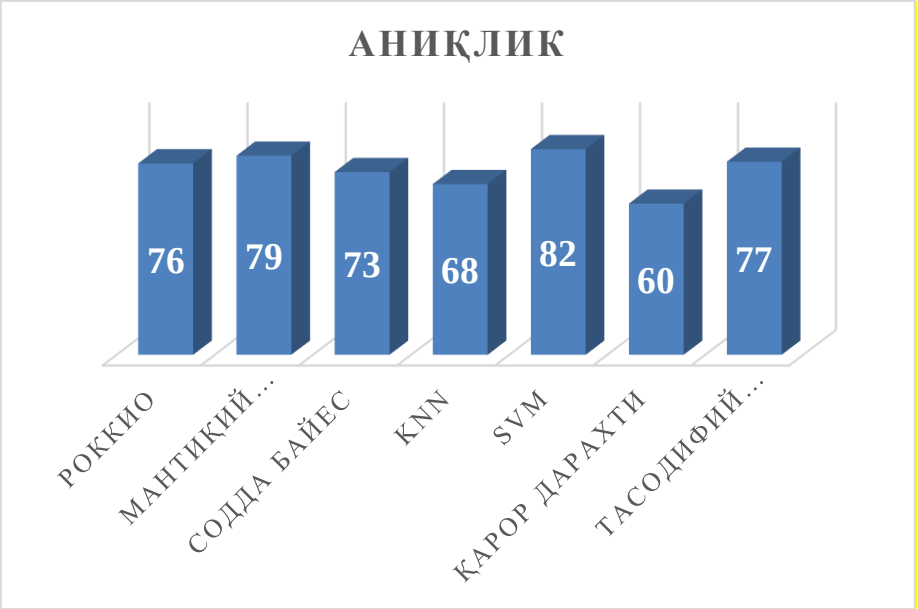


Рис. 4. Результаты классификации, полученые классическими алгоритмами распознавания

В параграфе 4.2 представлены результаты алгоритма, основанные на разработанном гибридном подходе. Для проведения экспериментальных исследований выбран в качестве объекта из официального государственного источника Национального информационного агентства Узбекистана отобраны 1847 последних новостных постов на узбекском языке на 10 категориях. В результате получился вектор длиной 64 слов для корпуса из 59299 слов. Ниже представлена таблица сходства слов.

Таблица 1. Сходство слов по модели CBOW

спорт	ахборот	сўм	талаба
тарбия	воситаларида	триллион	ўқиш
мусобақаларини	оммавий	миқдорида	кечки
қишки	воситалари	миллиард	олаётган
болалар	коммуникация	ажратди	таҳсил
турлари	технологиялари	маблағ	мактаб
атлетикага	вебсайтлари	доллар	етишмаслиги
ўсмирлар	дарча	миллион	амалиёт

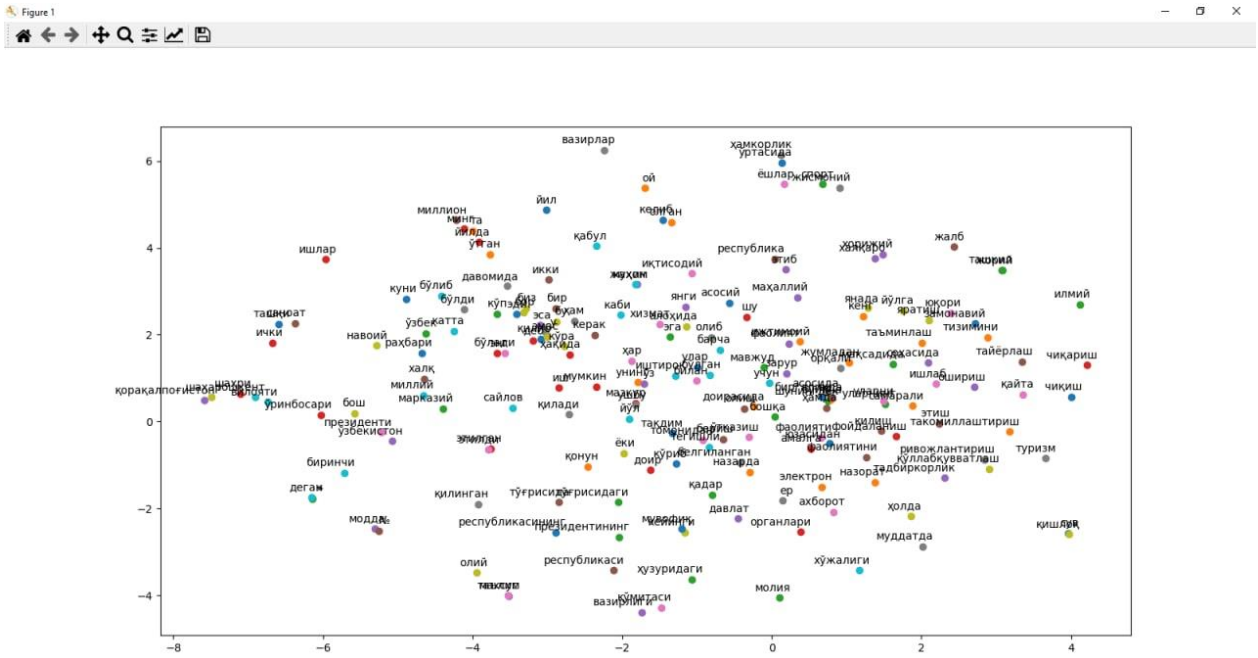


Рис. 5. Двумерное представление наиболее часто встречаюшихся слов в базе UzA

В экспериментальных исследованиях потдверждены эффективность предложенного (гибридного) алгоритма CNN+RNN(BLSTM), точность классификации которого составила 93% (рис.3).

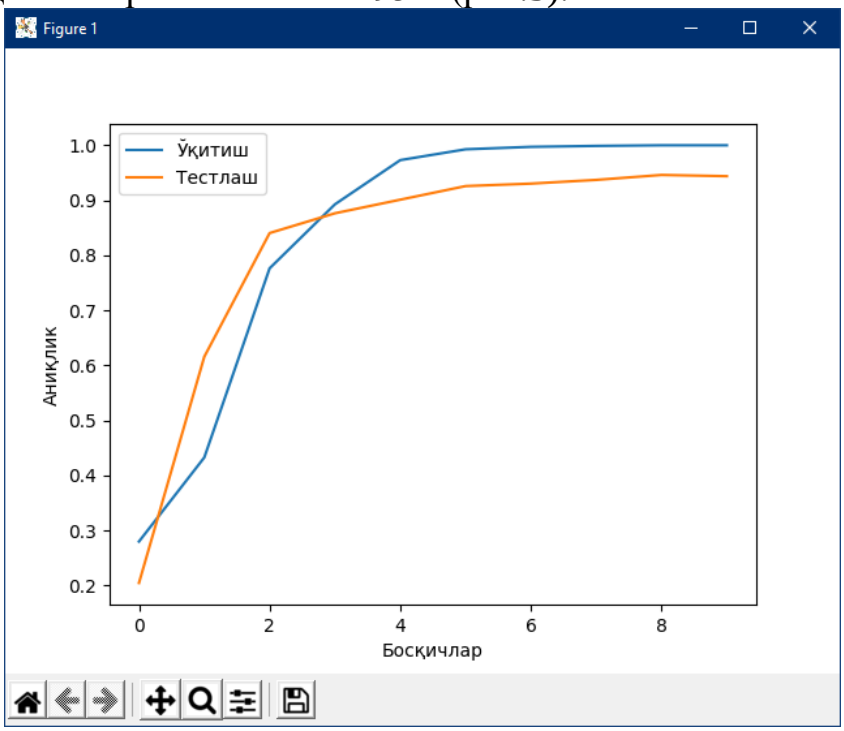


Рис. 6. Результаты классификации, полученые с помощью гибридного алгоритма CNN + RNN (BLSTM)

Параграф 4.3 посвящен анализу проведенных экспериментальных исследований, на основании которых определяется достоверность и эффективность полученных результатов.

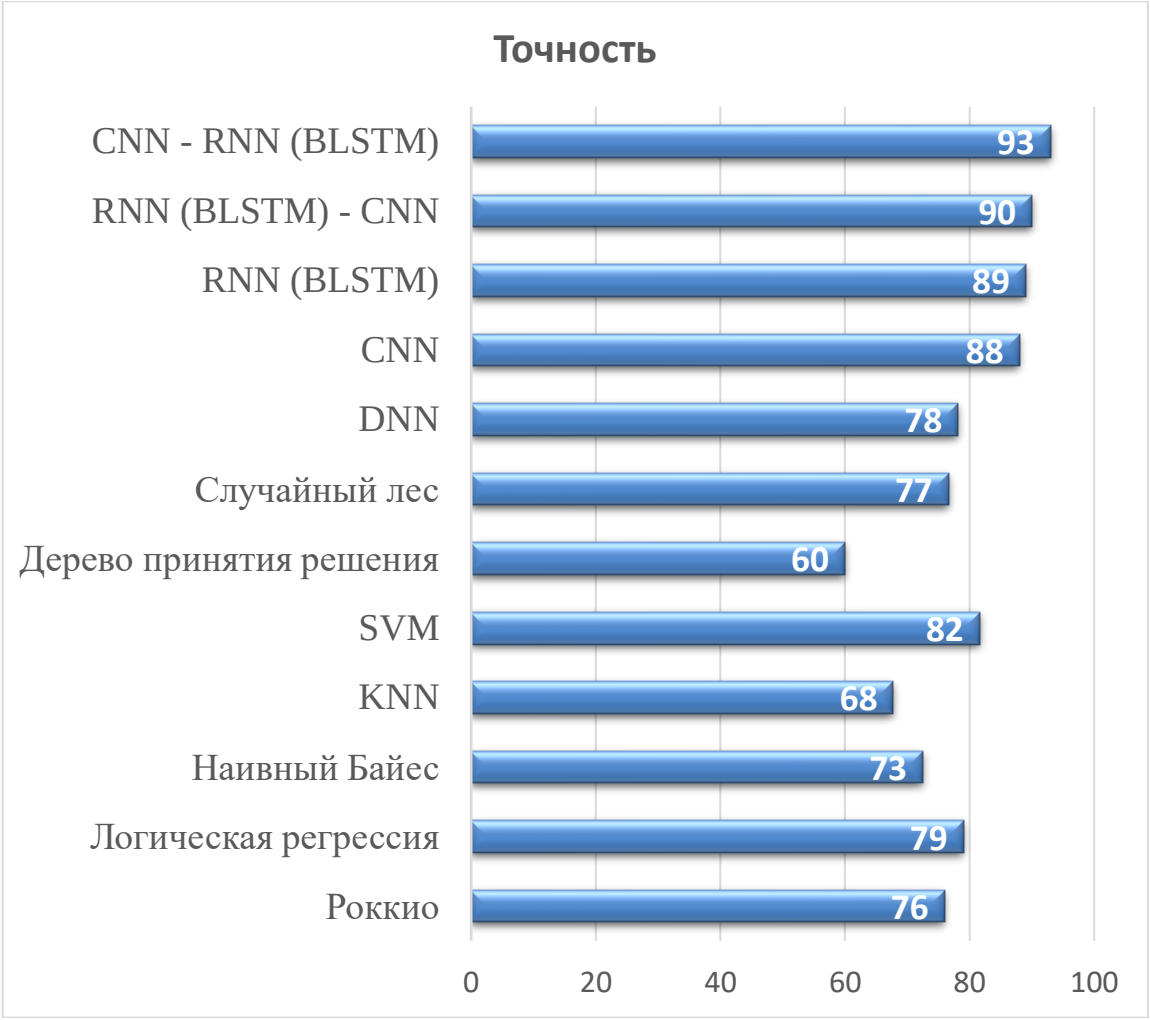


Рис. 7. Результаты классификации, полученные с помощью различных алгоритмов классификации

На основе известных и предложенных в диссертационной работе методов и алгоритмов решаются задачи предварительной обработки и классификации текстовых документов. Это позволило сократить время классификации текстовых документов в среднем на 10-25% и повысить точность классификации в среднем на 10%.

Полученные результаты были одобрены экспертами в этой области. В последующих исследованиях было признано целесообразным повышать уровень точности классификации за счет повышения эффективности применения других гибридных подходов классификации и уменьшение признакового пространства. Доказано, что планирование работ в области прикладных исследований посредством развития технологий классификации текстов на основе предложенных методов и алгоритмов повышает эффективность работы специалистов в этой области и позволяет снизить накладные расходы.

ЗАКЛЮЧЕНИЕ

В диссертации разработаны алгоритмы предварительной обработки, формирования признакового пространства, сокращения признакового пространства и классификации текстовых документов на узбекском языке.

Основные результаты исследования сводятся к следующим основным выводам:

1. Проведен анализ научно-технической литературы, освещающей современное состояние теоретических и практических аспектов развития традиционных и глубоких методов обучения при создании систем, связанные с предворительной обработкой, классификацией и анализом текстовых документов. Это позволило разработать концептуальные принципы, конструктивные подходы, методы, модели и алгоритмы систем классификации текстов.
2. Предложено эффективное решение задачи классификации текстов на основе некоторых подходов предварительной обработки данных, которые играют большую роль в решении задач классификации.
3. Предложен механизм сокращения признаков текста для эффективной реализации классификации текстов.
4. Обоснован эффективность подхода предложенный автором учитывающий соседных слов текста, чем простого считывания слов при формировании признакового пространство в задачах классификации текстовых документов. Предложен гибридный алгоритм классификации текстовых документов для повышения эффективности анализа текста. Этот алгоритм показал свою эффективность в вычислительных экспериментах.
5. Результаты классификации представлены в виде результатов экспериментальных исследований в виде модельной задачи на информационных ресурсах УзА, полученные результаты в первом случае дали в среднем 60-82%. Алгоритмы, основанные на глубоком обучении, показали в среднем 78-90% точности. Гибридный алгоритм дал результат выше 93%.
6. Программное обеспечение для классификации текстовых документов на узбекском языке был внедрён в информационных ресурсах Центра развития информационных и коммуникационных технологий при Министерстве народного образования Республики Узбекистан, Центра внедрения электронного оразования в образовательных учреждениях Министерства высшего и средне специального образования РУз, а также ООО «Darakchi Inform Service», получены акты, на основе результатов внедрения, которые подтверждают эффективность полученных результатов.

AD HOC SCIENTIFIC COUNCIL AT THE SCIENTIFIC COUNCIL AWARDING SCIENTIFIC DEGREES DSc13/30.12.2019.T.07.01 AT TASHKENT UNIVERSITY OF INFORMATION TECHNOLOGIES

SCIENTIFIC AND INNOVATION CENTER OF INFORMATION AND COMMUNICATION TECHNOLOGIES AT THE TASHKENT UNIVERSITY OF INFORMATION TECHOLOGIES

BOBOEV LOCHINBEK BOYMUROTOVICH

TEXT DOCUMENTS CLASSIFICATION ALGORITHMS

05.01.03 – Theoretical basis of computer science

DISSERTATION ABSTRACT
OF THE DOCTOR OF PHILOSOPHY (PhD) ON TECHNICAL SCIENCES

Tashkent-2021

The theme of doctor of philosophy (PhD) on technical sciences was registered at the Supreme Attestation Commission at the Cabinet of Ministers of the Republic of Uzbekistan under number B2019.4.PhD/T1419.

The dissertation has been prepared at Scientific and Innovation Center of Information and Communication Technologies at the Tashkent University of Information Techologies.

Scientific adviser:	**Babomuradov Ozod Jurayevich** Doctor of Technical Sciences
Official opponents:	**Mirzaev Nomaz** Doctor of Technical Sciences **Yuldashev Zafar Baxtiyarovich** Doctor of Philosophy (PhD) in technical sciences
Leading organization:	**"Unicon.uz" - The Scientific-engineering and marketing researches center**

The defense will take place "10" september 2021 at 14^{00} on the meeting of Scientific council No. DSc.27.06.2017.T.07.01 at Tashkent University of Information Technologies (Address: 100202, Tashkent city, Amir Temur street, 108. Tel.: (+99871) 238-64-43, fax: (+99871) 238-65-52, e-mail: tuit@tuit.uz).

The dissertation is available at the Information Resourse Centre of the Tashkent University of Information Technologies (is registered under No. 219). (Address: 100202, Tashkent city, Amir Temur street, 108. Tel.: (+99871) 238-64-43, fax: (+99871) 238-65-52).

Abstract of dissertation sent out on "20" august 2021 y.
(mailing report No. 1 on "23" july 2021 y.).

R.Kh.Khamdamov
Chairman of the scientific council awarding scientific degrees, Doctor of Technical Sciences, Professor

F.M.Nuraliev
Scientific secretary of scientific council awarding scientific degrees, Doctor of Technical Sciences, Docent

M.A.Ismailov
Chairman of the academic seminar under the scientific council awarding scientific degrees, Doctor of Technical Sciences, Professor

INTRODUCTION (abstract of PhD dissertation)

The aim of the research is to develop algorithms and software for classifying text documents.

Research objectives:

analyzing scientific work on the processing and classification of text data;

researching methods and algorithms of pre-processing text data;

developing an approach to the classification of text documents in the uzbek language;

developing a deep-learning based algorithm for the classification of text documents in the uzbek language;

developing a software for the classification of text documents on the basis of existing and proposed algorithms.

The object of research is the process of classification of electronic text documents.

The subject of the research is the methods and algorithms and software for the analysis and classification of text data.

Research methods. Theoretical research made use of system analysis, simulation modeling, probability theory, mathematical statistics, discrete mathematics, text processing, and classification methods.

The scientific novelty of the research is as follows:

probabilistic and hierarchical models has been built in the implementation and improvement of the mechanism of automated classification of text documents in the Uzbek language;

a sequence-based CBOW(Continuous Bag of Words) model of pre-processing of data has been improved, which serves to form the feature space in the implementation of the classification system of text documents in the Uzbek language available in information resources;

a modification of the algorithm for the classification of text documents in the Uzbek language has been developed, based on deep learning by adjusting the hyperparameters in the classification of text documents;

CNN+RNN(BLSTM) hybrid algorithm for the classification of text documents in the Uzbek language has been developed, which improves the efficiency of the mechanism for the correct formation and classification of text elements in text documents.

The practical results of the research are as follows:

algorithms for pre-processing and classification of text data have been developed;

based on existing and developed methods and algorithms, a software for identification of text documents in the Uzbek language «MT 1.0» was built.

The reliability of the research results is confirmed by the correct application of the mathematical apparatus for processing and classification of text data in developing algorithms, as well as the positive results of experimental research.

Scientific and practical significance of research results. The scientific significance of the research results is explained by the contribution of the developed

algorithms to the prospective development of the theoretical basis of pre-processing text data and text document classification.

Implementation of research results. On the basis of scientific results related to the solution of the problem of classification of text documents:
the sequence-based CBOW model software of pre-processing and classification hybrid algorithm CNN+RNN(BLSTM) in the classification of text documents in the Uzbek language was introduced for use in "Darakchi Inform Service" LLC (January 18, 2021 of the Ministry of Information Technologies and Communications) References No. 33-8/395). The resulting software works with more than 90% accuracy in classifying news into categories. This made it possible to reduce the time required to classify news by 50%.
The software package based on probabilistic, hierarchical and non-parametric models of automated classification of text documents in Uzbek language was introduced to the Center for Development of ICT in Public Education under the Ministry of Public Education of the Republic of Uzbekistan (Ministry of Information Technologies and Communications No. 33-8 / 395 of January 18, 2021) references). The results allowed reducing the time required to classify documents by 20-25% through initial processing and normalization of text documents;
software developed on the basis of the algorithm for classifying text documents in the Uzbek language based on deep learning by adjusting hyperparameters was introduced for use in the organization of distance and e-learning in the Center for implementation of e-learning in Educational Institutions under the Ministry of Higher and Secondary Specialized Education of the Republic of Uzbekistan. References No. 33-8 / 395 of January 18, 2021). As a result, it provides high-precision classification of documents through pre-processing and normalization of text documents. This allowed to increase the efficiency of the analysis of text documents in the Uzbek language by 15-20%.

Approbation of research results. The results of this research have been presented and discussed at 7 international and 14 national scientific conferences.

Publication of research results. A total of 29 scientific papers were published on the research, of which 6 articles were published in scientific journals recommended by the Higher Attestation Commission of the Republic of Uzbekistan, including 1 foreign and 5 national journals, and 1 computer software certificate.

The structure and scope of the dissertation. The dissertation consists of an introduction, four chapters, a conclusion, a list of references and appendices. The volume of the dissertation is 117 pages.

ЭЪЛОН ҚИЛИНГАН ИШЛАР РЎЙХАТИ
СПИСОК ОПУБЛИКОВАННЫХ РАБОТ
LIST OF PUBLISHED WORKS

I бўлим (I часть; I part)

1. Ниёзматова Н.А., Маматов Н.С., Отахонова Б.И., Бобоев Л.Б., Самижонов А.Н. Матнларни таснифлашда информатив белгилар мажмуасини аниқлаш усуллари // Мухаммад ал-Хоразмий авлодлари, № 4 (14), декабрь 2020. Б. 64-69. (05.00.00; №10)
2. Бабомурадов О.Ж., Маматов Н.С., Бобоев Л.Б., Отахонова Б.И. Қарор дарахти алгоритмидан фойдаланиб матнларни таснифлаш // Мухаммад ал-Хоразмий авлодлари, № 4 (10), декабрь 2019. Б. 20-23. (05.00.00; №10)
3. Babomuradov O.J., Mamatov N.S., Boboyev L.B., Otaxonova B.I. Text documents classification in uzbek language // International Journal of Recent Technology and Engineering (IJRTE), Volume-8, Issue-2S11, September 2019. pp. 3787-3789. (№41; SCImago; IF=0.158)
4. Babomuradov O.J., Boboyev L.B., Otaxonova B.I. A comparison of naive bayes models for text classification // Problems of computational and applied mathematics, № 1(19) 2019. pp. 39-43. (05.00.00; №23)
5. Бабомурадов О.Ж., Маматов Н.С., Бобоев Л.Б. Интеллектуал ахборот технологиялари асосида матнларни таснифлаш // Информатика ва энергетика муаммолари, № 5, 2018 й. Б. 35-39. (05.00.00; №5)
6. Бабомурадов О.Ж., Маматов Н.С., Бобоев Л.Б., Отахонова Б.И. Матнларга дастлабки ишлов бериш ва эҳтимолли баҳолаш моделлари асосида таснифлаш // Мухаммад ал-Хоразмий авлодлари, № 4 (6), май 2018. Б. 18-22. (05.00.00; №10)

II бўлим (II часть; II part)

7. Mamatov N.S, Samijonov. A.N, Boboyev L.B. The selection of informative features // International Journal of Computer Science Engineering and Information Technology Research 2249-7943 Vol. 8, Issue 4, Oct 2018, pp. 21-30.
8. Babomuradov O.J., Boboev L.B., Abduraxmanova N.N. Hybrid algorithm for classifying text documents // Олий таълим тизимида масофали таълимни жорий этишнинг техник-дастурий ва услубий таъминотини такомиллаштириш истиқболлари, Республика илмий-амалий конференсияси, Қарши, 28 май 2021, Б. 177-179.
9. Бабомурадов О.Ж., Бобоев Л.Б. Ўзбек тилидаги матнли ҳужжатларни таснифлашнинг мантиқий регрессия усули // Олий таълим тизимида масофали таълимни жорий этишнинг техник-дастурий ва услубий таъминотини такомиллаштириш истиқболлари, Республика илмий-амалий конференсияси, Қарши, 28 май 2021, Б. 110-112.
10. Бабомурадов О.Ж., Бобоев Л.Б. Матнли ҳужжатларни таснифлашнинг ансамбл усуллари // Олий таълим тизимида масофали таълимни жорий этишнинг техник-дастурий ва услубий таъминотини такомиллаштириш

истиқболлари, Республика илмий-амалий конференсияси, Қарши, 28 май 2021, Б. 107-110.

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

12. Бабомурадов О.Ж., Бобоев Л.Б. Анъанавий моделларга асосланган ўзбек тилидаги матнли ҳужжатларни таснифлаш алгоритмлари // Мирзо Улуғбек номидаги Ўзбекистон миллий университетининг Жиззах филиали, Фан, таълим ва ишлаб чиқариш интеграциясида рақамли иқтисодиёт истиқболлари, Республика миқёсидаги илмий-техник анжуман, Жиззах, 5–6 май 2021, Б. 203-210.

13. Бабомурадов О.Ж., Бобоев Л.Б. Матннинг CBOW модели // Ўзбекистон Давлат Жаҳон тиллари университети, Компьютер лингвистикаси: муаммо ва ечимлар мавзусидаги халқаро онлайн илмий-амалий конференция материаллари тўплами, Тошкент, 19 апрель 2021, Б. 80-83.

14. Бабомурадов О.Ж., Бобоев Л.Б. Матнли ҳужжатларни таснифлашда қўллаб-қувватловчи векторли машиналар (SVM) алгоритмидан фойдаланиш // Ўзбекистон Давлат Жаҳон тиллари университети, Компьютер лингвистикаси: муаммо ва ечимлар мавзусидаги халқаро онлайн илмий-амалий конференция материаллари тўплами, Тошкент, 19 апрель 2021, Б. 74-80.

15. Бабомурадов О.Ж., Бобоев Л.Б. Матнли маълумотларни таснифлаш учун қарор дарахти алгоритмини қўллаш // Муҳаммад ал-Хоразмий номидаги Тошкент ахборот технологиялари университети, Иқтисодиёт тармоқларининг инновацион ривожланишида ахборот-коммуникация технологияларининг аҳамияти, Республика илмий-техник анжумани, Тошкент, 4-5 март 2021, Б. 387-388.

16. Бабомурадов О.Ж., Бобоев Л.Б. Роккио алгоритми асосида ўзбек тилидаги матнли ҳужжатларни таснифлаш ёндашуви // Муҳаммад ал-Хоразмий номидаги Тошкент ахборот технологиялари университети, Иқтисодиёт тармоқларининг инновацион ривожланишида ахборот-коммуникация технологияларининг аҳамияти, Республика илмий-техник анжумани, Тошкент, 4-5 март 2021, Б. 385-387.

17. Бабомурадов О.Ж., Бобоев Л.Б. Таснифлашни баҳолаш ўлчовлари // Инновацион ёндашувлар илм-фан тараққиёти калити сифатида: ечимлар ва истиқболлар, Республика миқиёсидаги илмий-техник анжумани, Жиззах, 8-10 октябрь 2020, Б. 146-153.

18. Бабомурадов О.Ж., Бобоев Л.Б. Чуқур ўқитишга асосланган матнни таснифлаш моделлари // Инновацион ёндашувлар илм-фан тараққиёти калити сифатида: ечимлар ва истиқболлар, Республика миқиёсидаги илмий-техник анжумани, Жиззах, 8-10 октябрь 2020, Б. 153-158.

19. Бабомурадов О.Ж., Бобоев Л.Б., Дусанов Х.Т. Матнннинг кетма-кетлик модели // Инновацион ёндашувлар илм-фан тараққиёти калити

сифатида: ечимлар ва истиқболлар, Республика миқиёсидаги илмий-техник анжумани, Жиззах, 8-10 октябрь 2020, Б. 192-197.

20. Бабомурадов О.Ж., Маматов Н.С., Бобоев Л.Б. Интеллектуал ахборот технологиялари асосида матнларни таснифлаш // “Инновация-2018”, Халқаро илмий-амалий анжумани илмий мақолалар тўплами, Тошкент, 2018й. Б. 224-225.

21. Babomuradov O.J., Boboyev L.B., Khojiev S.A. Text documents classification in uzbek language // Topical issues of import substituting products based on the use of local raw materials in the Fergana valley, International Conference, Namangan, 27-28 October, 2018. P. 429-433.

22. Бабомурадов О.Ж., Маматов Н.С., Бобоев Л.Б. Интеллектуал ахборот технологиялари асосида матнларни таснифлаш // Математик моделлаштириш, алгоритмлаш ва дастурлашнинг долзарб муаммолари, Республика миқёсидаги илмий-амалий конференция. Тошкент, 17-18 сентябрь, 2018 й. Б. 357-360.

23. Бобоев Л.Б., Хасанов У.А. Маълумотлар таҳлилида информатив белгиларни танлаб олиш учун генетик алгоритм // Иқтисодиётнинг реал тармоқларини инновацион ривожланишида ахборот–коммуникация технологияларининг аҳамияти, Республика илмий-техник анжуманининг, Тошкент, 6-7 апрел 2017 й. Б. 165-167.

24. Нишанов А.Х., Бобоев Л.Б. Сотувчи саёҳати масаласида модификацияланган генетик алгоритм // Алгебра, амалий математика ва ахборот технологиялари масалалари, Республика илмий конференцияси, Наманган, 20-21 декабрь 2016 й. Б. 10-11.

25. Тургунов Б.А., Бобоев Л.Б. Review of network architectures, technologies for 5g systems // Современная наука: тенденции развития Материалы X Международной научно-практической конференции, Краснодар, 26 августа 2015 г. С. 214-217.

26. Тургунов Б.А., Жўраев Н.М., Бобоев Л.Б. Способ квантовой криптографии при обеспечении информационной безопастности на волс и проблемы в ее применении // Теория и практика актуальных исследований Материалы IX Международной научно-практической конференции, Краснодар, 24 июня 2015 г. С. 161-165.

27. Ruzibayev O.B., Khujaev O.K., Boboyev L.B. Comparative analyzing features selection methods for data mining tasks // WCIS – 2014 Eighth World Conference on Intelligent Systems for Industrial Automation, Tashkent, Uzbekistan, November 25-27, 2014. P. 332-338.

28. Бобоев Л.Б., Жабборов М.А. Интеллектуал таҳлил масалалари учун WEKA дастурий таъминоти ҳақида // Ахборот технологиялари ва телекоммуникация муаммолари Республика илмий-техник анжумани, Тошкент, 14-15 март 2013 йил. Б. 91-92.

29. Бабомурадов О.Ж., Рахимов Н.О., Бобоев Л.Б. Матнларни таснифлаш (MT 1.0) // Ўзбекистон Республикаси интеллектуал мулк агентлиги, электрон ҳисоблаш машиналари учун яратилган дастурнинг расмий рўйхатдан ўтказилганлиги тўғрисида гувоҳнома DGU 05474, 23.05.2018.

Автореферат “Мухаммад ал-Хоразмий авлодлари” илмий журнали таҳририятида таҳрирдан ўтказилди ҳамда ўзбек, рус ва инглиз тилларидаги матнларини мослиги текширилди.

Бичими 60x841/16. Рақамли босма усули. Times гарнитураси.
Шартли босма табоғи: 3,5. Адади 100. Буюртма № 90.

Гувохнома № 10-3719
“Tipograff” МЧЖ босмахонасида чоп этилган.
Босмахона манзили: 100011, Тошкент ш., Беруний кўчаси, 83-уй.