8D06103 – «Management information systems» білім беру бағдарламасы бойынша философия докторы (PhD) дәрежесін алу үшін ұсынылған «Онтологиялық тәсіл негізінде құқыққа қайшы мәтіндік ақпарат мониторингінің ақпараттық-талдамалық жүйесін әзірлеу» тақырыбындағы Ыбытаева Галия Сейткалиевнаның диссертациялық жұмысына

АҢДАТПА

Көптеген интернет-контентпен қанықкан заманауи ақпараттық кеңістік пайдаланушыға ақпарат алудың кең мүмкіндіктерін ұсынады. Алайда, қолда бар білім көлемінің өсуімен бірге жеке пайдаланушыларға да, жалпы қоғамға да зиян келтіруі мүмкін жалған, зиянды немесе құқыққа қайшы контентті анықтау мәселесі де артып келеді. Бұл зерттеу жұмысының маңыздылығы мен өзектілігі қылмыстық құрылымдар қызметінің күшеюіне және Интернет желісінде жалған ақпараттың таралуына негізделген.

Қазіргі уақытта әртүрлі елдердің құқық қорғау және мемлекеттік органдары қылмыс жасалғаннан кейін онымен күресуден гөрі қылмыс пен терроризмнің алдын алуға көбірек көңіл бөледі. Қылмыстың алдын алу парадигмасын ұстану үшін мәтіндік және дауыстық ақпаратты қоса алғанда, ақпараттың үлкен көлемін талдау, деректерді іздеу және мәтінді талдаудың озық технологияларын, сондай-ақ NLP құралдары мен тәсілдерін пайдалану қажет.

Жаңа ақпараттық технологиялардың әлеуетін кеңейтумен байланысты қазіргі қоғамның басты проблемаларының бірі – Интернетті деструктивті және қоғамға қарсы мақсаттары бар қылмыстың құралы ретінде пайдалану мүмкіндігі. Екінші жағынан, Facebook, Twitter, Instagram, YouTube сияқты компьютерлік-жанама коммуникациядағы (Computer-Mediated Communication (CMC)) құрылымданбаған ақпаратты талдаудың жаңа технологиялары мәтіндік деректерді алдын-ала өңдеуге және ықтимал қылмыстардың алдын алуға мүмкіндік береді. Алайда қылмысқа ниетті немесе дайындықты көрсетуі мүмкін барлық сайттардың мазмұнын қолмен табу және қадағалау мүмкін болмағандықтан, құқыққа қайшы Интернет ақпаратты анықтауды автоматтандыру қажет. Көптілді құқыққа қайшы контентті автоматты немесе автоматтандырылған іздеу және талдаудың мұндай жүйелері адамның криминалдық әрекетті жасауды жоспарлап отырғанын немесе жасағанын жоғары ықтималдықпен анықтау мүмкіндігіне ие болуы керек.

Жоғарыда айтылғандардың барлығы интернет-контенттегі ықтимал терроризмнің, экстремизмнің және басқа да құқыққа қайшы және зорлық-зомбылық әрекеттерінің цифрлық іздерін анықтау қажеттілігіне қатысты осы зерттеуге бағытталған **негізгі мәселені** анықтайды. Сонымен қатар, статистикалық әдістер мен машиналық оқыту әдістеріне негізделген тәсілдер оқытылған корпустың болмауына және әлсіз «лингвистикалық» маркерлерге ие құқыққа қайшы Интернеттің пәндік саласының (ПәС) анық еместігіне байланысты жақсы нәтиже бермейді. Демек, осы зерттеуде қарастырылатын

екінші мәселе – мәтіннің құқыққа қайшы контентке жататынын анықтау кезінде оның семантикалық (мағыналық) құраушысын ескеру қажеттілігі. Бұл мәселе қолданылатын машиналық оқыту әдістеріне семантикалық дифференциалданатын белгілерін қосудан тұратын онтологиялық тәсілді қолдану арқылы шешілуі керек. **Үшінші мәселе** қазақ тілі үшін «Құқыққа қайшы интернет-контент» пәндік саласында онтологияның болмауы, сондай-ақ орыс және ағылшын тілдері үшін мұндай онтологиялардың ашық қолжетімділікте болмауы болып табылады. Сонымен қатар, алдыңғысына қатысты мәселе мәтіндік корпустарға негізделген тиімді алгоритмдер мен автоматты онтологиялық генерациялау бағдарламалық құралдарының болмауы болып қала береді.

Жоғарыда айтылғандарға сүйене отырып, қазіргі уақытта көптілді құқыққа қайшы интернет-контентті автоматты түрде іздеу және талдау жүйесінің тиімді әдістері, модельдері мен бағдарламалық құралдарының қажеттілігі ерекше **өзекті** деген қорытындыға келуге болады.

Зерттеудің мақсаты Интернет желілерде қазақ және орыс тілдерінің құқыққа қайшы мәтіндерін автоматты сәйкестендіру жүйесінің ақпараттық моделін әзірлеу болып табылады.

Зерттеу міндеттері. Зерттеудің қойылған мақсаттарын іске асыру үшін мынадай мәселелер шешіледі:

1) Онтологиялық тәсіл негізінде құқыққа қайшы мәтіндік ақпарат мониторингінің заманауи әдістері мен модельдерін талдау.

2) Қазақ, орыс, украин және ағылшын тілдеріндегі Интернеттің криминалистік маңызды мәтіндерінің корпустарын әзірлеу.

3) Көптілді (қазақ, орыс және ағылшын тілдері) терминологиялық тезаурус құру.

4) «Құқыққа қайшы интернет-контент» онтологиясын құру.

5) Криминалистік маңызды мәтіндердің мамандандырылған корпустарын автоматты семантикалық белгілеу әдісі мен құралдарын әзірлеу.

6) Онтологиялық тәсіл негізінде құқыққа қайшы мәтіндік ақпарат мониторингінің ақпараттық-талдамалық жүйесін әзірлеу.

Ақпараттық-талдамалық жүйесі «Құқыққа қайшы интернет-контент» онтологиясын, мәтіндердің мамандандырылған корпустарын, криминалистік маңызды мәтіндердің мамандандырылған корпустарын автоматты семантикалық белгілеудің бағдарламалық құралын және әлеуметтік желілерде және басқа интернет көздерінде құқыққа қайшы контентті талдау мен мониторингінің интеграцияланған технологиясын қамтиды.

Әзірленген жүйені пайдалану қылмыстарды ашу және құқыққа қайшы әрекеттерді болдырмау ықтималдығын арттыру есебінен құқық қорғау және арнаулы мемлекеттік ұйымдар жұмысының тиімділігін арттыруға мүмкіндік береді. Бұл зерттеудің әлеуметтік әсері құқықтық және криминогендік жағдайды жақсарту және жалпы қоғамның өмір сүру сапасын жақсарту болып табылады.

Зерттеу нысаны құқыққа қайшы көптілді мәтіндік ақпаратты автоматты түрде іздеу және талдау жүйелері болып табылады.

Зерттеу пәні онтологиялық тәсілге негізделген құқыққа қайшы көптілді мәтіндік ақпаратты іздеуге және талдауға арналған модельдер мен әдістер, бағдарламалық құралдар болып табылады.

Зерттеу әдістері. Статистикалық ықтималдық әдістері, машиналық оқыту әдістері, корпустық лингвистика әдістері, интеллект теориялары, сондай-ақ табиғи тіл мәтіндерін семантикалық және грамматикалық талдау әдістері және сараптамалық бағалау әдістері.

Зерттеу жұмысының ғылыми жаңалығы.

– Интернеттің криминалистік маңызды көптілді мәтіндерінің корпустары және онтологиялық тәсілге негізделген криминалистік маңызды мәтіндердің корпустарын автоматты семантикалық белгілеу құралдары жасалды.

– Көптілді терминологиялық тезаурус, «Құқыққа қайшы интернет-контент» онтологиясы және ақпараттық модель құрылды.

– Онтологиялық тәсіл негізінде құқыққа қайшы мәтіндік ақпарат мониторингінің ақпараттық-талдамалық жүйесі әзірленді.

Жұмыстың ақпараттық қауіпсіздік және заңсыз контентпен күресу саласында маңызы зор, сонымен қатар мәтінді өңдеу және семантикалық талдау саласында одан әрі зерттеулер жүргізу үшін перспективалар ашылады. Жүйені іске асырудың экономикалық және индустриялық мүдделілігі алынған нәтижелерді мінез-құлықты талдау жөніндегі мамандардың, құқық қорғау органдарының және қауіпсіздік қызметтерінің ықтимал құқыққа қайшы қатерлерді бағалау рәсімдерін орындау немесе жақсарту кезінде жедел қолдануы үшін пайдалану мүмкіндігі болып табылады.

Алынған нәтижелердің **теориялық маңыздылығы** көптілді мәтіндік ақпаратты өңдеу мен талдаудың қолданыстағы модельдері мен әдістерін бейімдеу және әзірлеу болып табылады.

Алынған нәтижелердің **практикалық құндылығы** қорғауға шығарылған қағидалар негізінде ақпараттық- талдамалық жүйені әзірлеу болып табылады.

Қорғауға шығарылатын негізгі қағидалар.

1) Екі корпус және көптілді терминологиялық тезаурус құрылды:

– орыс, украин және ағылшын тілдеріндегі мәтіндерді (украин тілінде 3147 мәтін, орыс тілінде 5506 мәтін, ағылшын тілінде 300 мәтін) қамтитын көптілді корпус;

– орыс тіліндегі 3000 мәтінді және қазақ тіліндегі 3000 мәтінді, оның ішінде мағынасы жағынан тураланған қазақша-орысша сөйлемдерді құрайтын 2000 мәтінді қамтитын параллель қазақ-орыс корпусы;

– 600-ден астам негізгі сөздерді (330 зат есім, 107 сын есім және 170-ке жуық етістік) және 2500-ден астам негізгі сөз синонимдерін қамтитын тезаурус.

2) «Құқыққа қайшы интернет-контент» онтологиясы құрылды.

3) Криминалистік маңызды мәтіндердің мамандандырылған корпустарын автоматты семантикалық белгілеу әдісі мен құралдары әзірленді.

4) Онтологиялық тәсіл негізінде көптілді құқыққа қайшы интернет-контентті автоматты іздеу және талдау жүйесінің ақпараттық моделі мен бағдарламалық құралы әзірленді.

Сенімділік дәрежесі мен апробациялау нәтижелері. Диссертацияның негізгі нәтижелері халықаралық және шетелдік ғылыми конференцияларда, ғылыми семинарларда баяндалды және талқыланды:

1) Nina Khairova, Anastasiia Kolesnyk, Orken Mamyrbayev, Galiya Ybytayeva, Yuliia Lytvynenko. Automatic Multilingual Ontology Generation Based on Texts Focused on Criminal Topic // Proceedings of the 5th International Conference on Computational Linguistics and Intelligent Systems (COLINS 2021). – 2021. – Vol. 1. – P. 108-117.

2) Г.С. Ыбытаева, О.Ж. Мамырбаев, Н.Ф. Хайрова, Б.Ж. Жумажанов. Қазақ тіліндегі мәтіндерде коллокацияларды анықтаудың статистикалық әдістерін талдау // «Информатика және қолданбалы математика» VI Халықаралық ғылыми конференциясының материалдары. – Алматы, Қазақстан, 2021. – Б. 256-262.

3) Г.С. Ыбытаева, О.Ж. Мамырбаев, Н.Ф. Хайрова, Б.Ж. Жумажанов, К.Ж. Мухсина. Параллель корпусты әзірлеу мәселелері // «Информатика және қолданбалы математика» VII Халықаралық ғылыми конференциясының материалдары. – Алматы, Қазақстан, 2022. – Б. 175-182.

4) Ybytayeva, Galiya & Orken, Mamyrbayev & Khairova, Nina & Rizun, Nina & Berdali, Sanzharsultan & Kuralai, Mukhsina. (2023). Creating a Thesaurus "Crime-Related Web Content" Based on a Multilingual Corpus // Proceedings of the 7th International Conference on Computational Linguistics and Intelligent Systems (COLINS-2023). – 2023. – Vol. 3396. – P. 77-87.

5) Ybytayeva, G.; Khairova, N.; Mamyrbayev, O.; Mukhsina, K. and Zhumazhanov, B. Experimental Verification of Collocation Detection Methods // Proceedings of the 5th Workshop for Young Scientists in Computer Science and Software Engineering - CS&SE@SW, SciTePress. – 2023. – P. 13-18. DOI: 10.5220/0012008900003561.

6) Мамырбаев О.Ж., Хайрова Н.Ф., Ыбытаева Г.С**.** «Құқыққа қайшы интернет-контент» көптілді онтология қосымшасын құру // «Информатика және қолданбалы математика» VIII Халықаралық ғылыми-практикалық конференциясы. – Алматы, Қазақстан, 2023. – Б. 107-114.

Зерттеушінің жеке үлесі. Докторант диссертациялық жұмыстың міндеттерін өз бетінше орындап, шешті. Интернеттің криминалистік маңызды мәтіндерінің көптілді корпустарын жасады. «Құқыққа қайшы интернет-контент» онтологиясын құрды. Онтологиялық тәсіл негізінде құқыққа қайшы мәтіндік ақпарат мониторингінің ақпараттық-талдамалық жүйесін әзірледі. Әзірленген модельдер мен технологияларды эксперименттік бағалауды орындады.

Диссертация тақырыбының ғылыми-зерттеу жұмысының жоспарларымен байланысы. Диссертациялық жұмыс «Онтологиялық тәсіл негізінде көптілді құқыққа қайшы веб-контентті автоматты іздеу және талдау жүйесінің ақпараттық моделі және бағдарламалық құралы» – AP09259309 (2021-2023 жж.) ҚР ҒЖБМ гранттық зерттеулер жобасы аясында орындалды.

Диссертациялық зерттеудің негізгі нәтижелерінің жарияланымдары.

Диссертациялық жұмыс тақырыбы бойынша 2 авторлық куәлік алынды және 5 жұмыс жарияланды, оның ішінде 3 мақала ҚР ҒЖБМ ғылым және жоғары білім саласындағы сапаны қамтамасыз ету Комитеті ұсынған журналдарда жарияланды, 2 мақала Scopus және Web of Science базасымен индекстелген басылымдарда жарияланды, 1 монография жарыққа шықты:

1. N. Khairova, O. Mamyrbayev, N. Rizun, M. Razno and **G. Ybytayeva**, "A Parallel Corpus-Based Approach to the Crime Event Extraction for Low-Resource Languages," in IEEE Access, vol. 11, pp. 54093-54111, 2023, doi: 10.1109/ACCESS.2023.3281680. **(Scopus: Процентиль – 89, Q1; Web of science: IF – 0.89, Q2);**

2. Kartbayev, A., Mamyrbayev, O., Khairova, N., **Ybytayeva, G.,** Abilkaiyr, N., Mussayeva, D. Correction of Kazakh synthetic text using finite state automata (2021) Journal of Theoretical and Applied Information Technology, 99 (22), pp. 5559-5570. **(Scopus: Процентиль – 30, Q3**);

3. Картбаев, А., **Ыбытаева, Г.,** Мамырбаев, О., Мухсина, К., & Жумажанов, Б. (2022). Методы формального представления сущностей в криминальных новостях для автоматического построения онтологии преступлений. Известия НАН РК. Серия информатики, (3), 136-152. https://doi.org/10.32014_2518-1726_2022_343_3_136-152;

4. **Ыбытаева, Г**., Н.Ф. Хайрова, К.Ж. Мухсина, & Б.Ж. Жумажанов. (2022). Лингвистикалық онтологияны қолдану және қалыптастыру мәселелеріне шолу. ҚР ҰҒА жаңалықтары. Информатика сериясы, (1), 96-106. https://doi.org/10.32014/2022.2518-1726.121;

5. **Г.С. Ыбытаева,** О.Ж. Мамырбаев, Н.Ф. Хайрова, К.Ж. Мухсина, Б.Ж. Жұмажанов. Сарапшылар пікірлерінің келісім өлшемі ретінде Коэннің каппа коэффициентінің ерекшеліктері. ҚР ҰИА жаңалықтары. Ақпараттық технологиялар сериясы, № 3 (89), 2023, 139-151. https://doi.org/10.47533/2023.1606-146X.26;

6. Мамырбаев О., **Ыбытаева Г.,** Бердали С. Веб-приложение многоязычной базовой онтологии «Противоправный веб-контент». № 32055 от 26.01.2023 г.;

7. **Ыбытаева Г.,** Мамырбаев О. «Құқыққа қайшы интернет-контент» көптілді онтология қосымшасы. № 38766, 29.08.2023 ж.

8. O. Mamyrbayev, N. Khairova, W. Wójcik, **G. Ybytayeva**. Automatic identification of illegal texts in Internet. Almaty. Institute of Information and Computational Technologies. 2023. 151 p. (монография).

Диссертациялық жұмыстың құрылымы мен көлемі. Диссертациялық зерттеу жұмысы кіріспеден, 4 бөлімнен, қорытындыдан, 172 атаудан тұратын әдебиеттер тізімінен және 3 қосымшадан тұрады. Жұмыс 93 бетте берілген және 19 сурет, 13 кестеден тұрады.

Диссертациялық жұмыстың қысқаша сипаттамасы

Кіріспеде диссертациялық жұмыстың жалпы сипаттамасы берілген, өзектілігі анықталған, диссертацияның мақсаты мен міндеттері, зерттеу нысаны мен пәні тұжырымдалған, қорғауға ұсынылған қағидалар анықталған, орындалған зерттеулердің ғылыми жаңалығы, теориялық маңыздылығы, практикалық құндылығы негізделген, зерттеушінің жеке үлесі, сенімділік дәрежесі мен апробациялау нәтижелері, диссертация тақырыбының ғылыми-зерттеу жұмысының жоспарларымен байланысы, зерттеудің негізгі нәтижелерінің жарияланымдары және диссертациялық жұмыстың құрылымы мен көлемі келтірілген.

Бірінші тарауда көптілді құқыққа қайшы интернет-контентті автоматты түрде іздеу және талдау саласындағы бар мәселелерге аналитикалық шолу берілген. Семантикалық талдау және семантикалық корпусты белгілеу саласындағы негізгі мәселелер сипатталған. Сондай-ақ, лингвистикалық онтологияларды қолдану және пішімдеу мәселелеріне шолу жасалды. Онтологияны автоматты түрде генерациялауда қолданылатын негізгі ұғымдарды мәтіндерден шығарып алу әдістерінің күйі мен даму перспективалары қарастырылады. Құрылымдалмаған мәтіндерден data-driven және knowledge-driven оқиғаларды шығарып алу тәсілдерінің салыстырмалы талдауы берілген. Корпустарды талдау және семантикалық белгілеу үшін Text Mining әдістері мен құралдарын қолданудың бар мүмкіндіктеріне шолу жасалды.

Екінші тарауда криминалистік маңызды СМС ақпаратының құрылған мәтіндік корпустары және криминалдық лексикамен жасалған көптілді терминологиялық тезаурус кіретін құқыққа қайшы контенттің онтологиясын толтырудың негізгі көздері қарастырылады. Тарауда сонымен қатар құрылымдалмаған мәтіндерден фактілерді шығарып алудың қол жетімді моделіне негізделген тезаурусты автоматтандырылған толтыру және кеңейту үшін жасалған тәсілдің сипаттамасы бар. «Құқыққа қайшы интернет-контент» көптілді онтологиясын әзірлеу, оны визуализациялау және пайдалану шектеулері ұсынылған. Онтология құқыққа қайшы әрекеттер мен қылмысқа қатысты терминдердің синонимдік сөздігіне негізделген.

Үшінші тарауда криминалистік маңызды мәтіндердің мамандандырылған корпустарын автоматты семантикалық белгілеу әдісі мен құралдарын әзірлеу қарастырылады. «Құқыққа қайшы интернет-контент» онтология кластарының экземплярларын шығарып алу үшін лингвистикалық корпустарды пайдалану ұсынылады және корпус мәтіндерінде құқыққа қайшы контенттің лингвистикалық және лексикалық маркерлерін ерекшелеп көрсету әдісі енгізіледі. Фактілерді алудың логикалық-лингвистикалық моделіне негізделген мәтіндегі оқиғалар аргументтерінің рөлін анықтау моделі

келтірілген. Тарауда оқиғаларды триггер түрі бойынша анықтауға және оқиғаға қатысушыларды, оқиға атрибуттарын және қатысушылардың рөлдерін алу үшін онтологияны автоматты түрде генерациялау әдістері сипатталған.

Төртінші тарауда машиналық оқыту әдістері мен онтологиялық тәсілді қамтитын әлеуметтік желілердегі және басқа интернет көздеріндегі құқыққа қайшы контентті іздеу мен талдаудың интеграцияланған технологиясы қарастырылады. Бұл тарауда қазіргі Интернеттің криминалистік маңызды мәтіндерінің корпустарынан тұжырымдамаларды автоматты түрде шығарып алу нәтижелерінің дәлдігін бағалау үшін Коэннің каппа метрикасын қолдану көрсетілген. Әзірленген технологияның тиімділігінің эксперименттік дәлелі де сипатталған. Әзірленген технологияның тиімділігін оқиғаларды бақылаудың басқа тәсілдерімен салыстыру келтірілген.

Қорытындыда онтологиялық тәсіл негізінде құқыққа қайшы мәтіндік ақпарат мониторингінің ақпараттық-талдамалық жүйесін әзірлеу бойынша диссертациялық жұмыстың ғылыми және практикалық нәтижелері көрсетіледі.