АННОТАЦИЯ
диссертационной работы Ыбытаевой Галии Сейткалиевны на тему: «Разработка информационно-аналитической системы мониторинга противоправной текстовой информации на основе онтологического подхода», представленной на соискание степени доктора философии (PhD) по образовательной программе 8D06103 – «Management information systems»

Современное информационное пространство, насыщенное многочисленным интернет-контентом, предоставляет пользователю обширные возможности получения информации. Однако, вместе с ростом объема доступных знаний возрастает и проблема выявления недостоверного, вредоносного или противоправного контента, который может нанести ущерб как отдельным пользователям, так и обществу в целом. Важность и актуальность данной исследовательской работы обусловлены усилением деятельности преступных структур и распространением недостоверной информации в сети Интернет.

В настоящее время правоохранительные и правительственные органы различных многих стран склонны больше сосредотачиваться на предотвращении преступлений и терроризма до их совершения, чем на борьбе с преступлением после того, как оно было совершено. Для того, чтобы следовать данной парадигме предотвращения преступности необходимо анализировать огромный объем информации, включая текстовую и голосовую информацию, использовать передовые технологии интеллектуального анализа данных и текстового анализа, а также инструменты и подходы NLP.

Одной из главных проблем современного общества, связанных с расширением потенциала новых информационных технологий, является появившаяся возможность использования Интернета в качестве инструмента и средства преступления с деструктивными и антиобщественными целями. С другой стороны, появляющиеся новые технологии анализа неструктурированной информации, содержащейся в компьютерно-опосредованной коммуникации (computer-mediated communication (CMC)), подобных Facebook, Twitter, Instagram, YouTube, позволяют осуществлять превентивную обработку текстовых данных и предотвращать потенциальные преступления. Однако, в связи с невозможностью вручную найти и отследить содержимое всех сайтов, которые могут указывать на намерение или подготовку к преступлению, необходимо автоматизировать идентификацию противоправной информации Интернета. Подобные системы автоматического или автоматизированного поиска и анализа многоязычного противоправного контента должны с высокой долей вероятности определять планирует ли какой-либо человек совершить или уже совершил криминальное действие.

Все вышеизложенное определяет **основную проблему**, на решение которой направлено данное исследование, касающуюся необходимости выделения в интернет-контенте цифровых следов потенциального

терроризма, экстремизма и других противоправных и насильственных действий. При этом, подходы, базирующиеся на статистических методах и методах машинного обучения, не дают хороших результатов в связи с отсутствуем обученных корпусов и размытостью предметной области противоправного интернета, обладающей слабыми «лингвистическими» маркерами. Следовательно**, второй проблемой**, рассматриваемой в рамках данного исследования, является необходимость учета семантической (смысловой) составляющей текста при определении его принадлежности к противоправному контенту. Данная проблема должна быть решена благодаря использованию онтологического подхода, заключающегося в добавлении семантических дифференцирующих признаков к применяемым методам машинного обучения. **Третья проблема** заключается в отсутствии онтологии в предметной области «Противоправный интернет-контент» для казахского языка, а также отсутствия подобных онтологий в открытом доступе для русского и английского языков. Кроме того, проблемой, связанной с предыдущей, остается отсутствие эффективных алгоритмов и программных средств автоматической генерации онтологий на базе текстовых корпусов.

Исходя из вышеизложенного, можно прийти к выводу, что в настоящий момент необходимость в эффективных методах, моделях и программных инструментарий системы автоматического поиска и анализа многоязычного противоправного интернет-контента особенно **актуальна.**

Целью исследования является разработка информационной модели системы автоматической идентификации противоправных текстов казахского и русского языков в Интернет сетях.

Задачи исследования. Для реализации поставленных целей исследования решаются следующие вопросы:

1) Анализ современных методов и моделей мониторинга противоправной текстовой информации на основе онтологического подхода.

2) Разработка корпусов криминально окрашенных текстов Интернета казахского, русского, украинского и английского языков.

3) Создание многоязычного (казахского, русского и английского языков) терминологического тезауруса.

4) Создание онтологии «Противоправный интернет-контент».

5) Разработка метода и инструментария автоматической семантической разметки специализированных корпусов криминально окрашенных текстов.

6) Разработка информационно-аналитической системы мониторинга противоправной текстовой информации на основе онтологического подхода.

Информационно-аналитическая система включает онтологию «Противоправный интернет-контент», специализированные корпуса текстов, программный инструментарий автоматической семантической разметки специализированных корпусов криминально окрашенных текстов и интегрированная технологию анализа и мониторинга противоправного контента в социальных сетях и других интернет-источниках.

Использование разработанной системы позволяет повысить эффективность работы правоохранительных и специальных государственных организаций, за счет повышения вероятности раскрытия преступлений и предотвращения противоправных действий. Социальный эффект данного исследования заключается в улучшении правовой и криминогенной обстановки и улучшении качества жизни общества, в целом.

Объектом исследования являются системы автоматического поиска и анализа противоправной многоязычной текстовой информации.

Предметом исследования являются модели и методы, программные инструментарий для поиска и анализа противоправной многоязычной текстовой информации на базе онтологического подхода.

Методы исследования. Статистико-вероятностные методы, методы машинного обучения, методы корпусной лингвистики, теории интеллекта, а также методы семантического и грамматического анализа текстов естественного языка и методы экспертных оценок.

Научная новизна исследовательской работы.

– Разработаны корпуса криминально окрашенных многоязычных текстов Интернета и инструментарий автоматической семантической разметки корпусов криминально окрашенных текстов, базирующегося на онтологическом подходе.

– Созданы многоязычные терминологический тезаурус, онтология «Противоправный интернет-контент» и информационная модель.

– Разработана информационно-аналитическая система мониторинга противоправной текстовой информации на основе онтологического подхода.

Работа имеет важное значение в области информационной безопасности и борьбы с незаконным контентом, а также открывает перспективы для дальнейших исследований в области обработки текстов и семантического анализа. Экономическая и индустриальная заинтересованность от реализации системы заключается в возможности использования полученных результатов для оперативного применения специалистами по поведенческому анализу, правоохранительными органами и службами безопасности при выполнении или улучшения процедур оценки возможных противоправных угроз.

Теоретическая значимость полученных результатов заключается в адаптации существующих и разработке новых моделей и методов обработки и анализа многоязычной текстовой информации.

Практическая ценность полученных результатов заключается в разработке информационно-аналитической системы на основе положений, вынесенных на защиту.

Основное положение, выносимое на защиту.

1) Разработаны два корпуса и многоязычный терминологический тезаурус:

– многоязычный корпус, включающий тексты русского, украинского и английского языков, составляющий 3147 текстов на украинском языке, 5506 текстов на русском языке, 300 текстов на английском языке;

– параллельный казахско-русский корпус, составляющий 3000 текстов на русском языке и 3000 текстов на казахском языке, в том числе 2000 текстов, содержащих выровненные по смыслу казахско-русские предложения;

– тезаурус, содержащий более 600 основных слов (330 существительных, 107 прилагательных и около 170 глаголов) и более 2500 синонимов основных слов.

2) Создана онтология «Противоправный интернет-контент».

3) Разработаны методы и инструментарий автоматической семантической разметки специализированных корпусов криминально окрашенных текстов.

4) Разработаны информационная модель и программный инструментарий системы автоматического поиска и анализа многоязычного противоправного Интернет-контента на базе онтологического подхода.

Степень достоверности и апробация результатов. Основные результаты диссертации докладывались и обсуждались на международных и зарубежных научных конференциях, научных семинарах:

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.

Личный вклад исследователя. Докторант самостоятельно выполнил и решил задачи диссертационной работы. Разработал многоязычные корпуса криминально окрашенных текстов Интернета. Создал онтологию «Противоправный интернет-контент». Разработал информационно-аналитическую систему мониторинга противоправной текстовой информации на основе онтологического подхода. Выполнил экспериментальную оценку разработанных моделей и технологии.

Связь темы диссертации с планами научно-исследовательской работы. Диссертационная работа выполнялась в рамках проекта по грантовым исследованиям МНВО РК «Информационная модель и программный инструментарий системы автоматического поиска и анализа многоязычного противоправного веб-контента на базе онтологического подхода» – AP09259309 (2021-2023 гг.).

Публикация основных результатов диссертационного исследования.

По теме диссертационной работы было получено 2 авторских свидетельства и опубликовано 5 работ, из которых 3 статьи опубликованы в журналах, рекомендованных Комитетом по обеспечению качества в сфере науки и высшего образования МНВО РК, 2 статьи опубликованы в изданиях, индексирумых базой Scopus и Web of Science, 1 монография:

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).

Структура и объем диссертационной работы. Диссертационная исследовательская работа состоит из введения, 4 разделов, заключения, списка литературы из 172 наименований и 3 приложений. Работа изложена на 93 страницах и содержит 19 рисунков, 13 таблиц.

Краткое описание диссертационной работы

Во введении дана общая характеристика диссертационной работы, определена актуальность, сформулированы цель и задачи диссертации, объект и предмет исследования, определено основное положение, выносимое на защиту, обоснованы научная новизна, теоретическая значимость, практическая ценность выполненных исследований, приведены личный вклад исследователя, степень достоверности и апробация результатов, связь темы диссертации с планами научно-исследовательской работы, публикация основных результатов диссертационного исследования, структура и объем диссертационной работы.

В первой главе приведен аналитический обзор существующих проблем в области автоматического поиска и анализа многоязычного противоправного интернет-контента. Описаны основные проблемы в области семантического анализа и семантической разметки корпусов. Также осуществлен обзор проблем использования и форматирования лингвистических онтологий. Рассмотрены состояние и перспективы развития методов извлечения из текстов базовых концептов, используемых при автоматической генерации онтологии. Приведен сравнительный анализ data-driven и knowledge-driven подходов извлечения событий из неструктурированных текстов. Осуществлен обзор существующих возможностей использования методов и инструментария Text Mining для анализа и семантической разметки корпусов.

Во второй главе рассматриваются основные источники наполнения онтологии противоправного контента, к которым относятся разработанный многоязычный терминологический тезаурус криминальной лексикой и созданные текстовые корпуса криминально значащей информации СМС. Глава также включает описание подхода, разработанного для автоматизированного наполнения и расширения тезауруса, базирующегося на имеющейся модели извлечения фактов из неструктурированных текстов. Представлены разработка многоязычной онтологии «Противоправный интернет-контент», визуализация и ограничения ее использования. Онтология построена на базе синонимического словаря терминов, относящихся к противоправным действиям и преступности.

В третьей главе рассматривается разработка метода и инструментария автоматической семантической разметки специализированных корпусов криминально окрашенных текстов. Предлагается использование лингвистических корпусов для извлечения экземпляров классов онтологии «Противоправный интернет-контент» и вводится метод выделения лингвистических и лексических маркеров противоправного контента в текстах корпусов. Приведена модель определения роле аргументов событий в тексте, базирующая на логико-лингвистической модели извлечения фактов. Глава описывает методы автоматической генерации онтологии, позволяющем определять события по типу триггера, и извлекать участников события, атрибуты события и роли участников.

В четвертой главе рассматривается интегрированная технология поиска и анализа противоправного контента в социальных сетях и других интернет-источниках, включающая методы машинного обучения и онтологический подход. В данной главе показано использование метрики каппы Коэна для оценки точности результатов автоматического извлечения концептов из корпусов криминально значащих текстов современного Интернета. Также описывается экспериментальное доказательство эффективности разработанной технологии. Приведено сравнение эффективности разработанной технологии с другими подходами мониторинга событий.

В заключении отражаются научные и практические результаты диссертационной работы по разработке информационно-аналитической системы мониторинга противоправной текстовой информации на основе онтологического подхода.