АННОТАЦИЯ

диссертационной работы Карюкина Владислава Игоревича на тему «Исследование и разработка модуля интеллектуальной системы анализа и оценки социального настроения общества в медиапространстве Республики Казахстан», представленной на соискание степени доктора философии (PhD) по специальности 6D070300 – Информационные системы

Актуальность работы. В настоящее время развитие Интернет технологий способствовало значительному увеличению количества новостных сайтов и социальных сетей, описывающих различные события в мире. Публикация мнений, мыслей и идей о происходящих локальных и глобальных событиях в социальных сетях стала обычной практикой. Множество социальных сетей, Twitter, Facebook, YouTube и другие, остаются популярными и привлекают множество пользователей. А новые платформы, TikTok, Instagram, Pinterest и другие, набирают популярность в мире социальных медиа, подробно освещая огромное число происходящих в мире событий.

Так как число новостных тем и пользовательских мнений растет невероятно быстрыми темпами, появляется существенная необходимость отслеживать наиболее важные темы в различных сферах жизни (политика, экономика, гражданское общество, образование, здравоохранение, экология, культура и спорт и т.д.). Объем фактов и мнений о них, которыми делятся в социальных сетях, делает такое отслеживание практически невозможным без автоматизированных методов, увеличивших важность аналитических платформ. Наиболее существенной частью таких платформ является модуль анализа настроений. Хотя методы искусственного интеллекта и не способны полностью понять человеческие чувства, эмоции, культуру и менталитет, они позволяют определить тренд общественного мнения на те или иные события с помощью аналитических инструментов. Ручной анализ является очень долгим и ресурсоемким процессом, и он также оставляет неопределенности и неясности. Использование алгоритмов дает возможность быстрее получать оперативную аналитику и реализовывать разные гибридные подходы: словарный, с применением моделей машинного обучения и нейронных сетей.

В настоящее время существует целый ряд зарубежных аналитических платформ. Среди них особо выделяются Sproutsocial, Hubspot, Buzzsumo, Hootsuite, IQBuzz, Brandmention и Snaplytics. Эти платформы имеют схожие особенности, несмотря на их ориентацию на сферу бизнеса, что делает анализ социально-политических и экономических аспектов жизнедеятельности слабо представленным. Также данные платформы в основном работают с богатыми ресурсами языками, такими как английский, испанский, итальянский, французский и другими. Тексты же на русском и казахском языках имеют очень ограниченное представление. В связи с этим была разработана информационная система Opinion monitoring system (OMSystem), выполняющая мониторинг и анализ социального медиапространства Казахстана и уделяющая большое внимание различным актуальным темам, происходящим в стране. OMSystem анализирует ведущие казахстанские новостные порталы и социальные сети, такие как Facebook, Vkontakte, Instagram, Twitter и YouTube. Ключевым элементом OMSystem является модуль анализа социального настроения, использующий метод анализа данных с применением тональных словарей, моделей машинного обучения, нейронных сетей и маркетинговых показателей социального настроения.

В моделях машинного обучения для автоматического определения тональности была использована собранная база данных в размере 132 тыс. текстов из казахстанских новостных порталов и социальных сетей. Тексты прошли этапы предобработки, стемминга, извлечения признаков с помощью метрики *tf-idf* и метода встраивания слов FastText и балансировки классов для получения наилучших результатов классификации. На этапе классификации использовался ряд наиболее популярных алгоритмов машинного обучения (Support vector machine – SVM, Logistic regression – LR, Decision tree – DT, Random forest –

RF, Naïve Bayes – NB, k-nearest neighbors – k-NN, and XGBoost) и нейронных сетей (Deep neural networks – DNN, Convolutional neural networks – CNN и Recurrent neural networks – RNN) [36-40]. Результаты классификации представлены в виде сводных таблиц метрик оценки эффективности алгоритмов: правильности (accuracy), точности (precision), полноты (recall) и F-меры (F1-score), графиков кривых (Area under curve – Receiver operating characteristics – AUC–ROC) и матриц ошибок. Для анализа социального настроения общества разработаны модели с применением маркетинговых показателей в социальных сетях: уровня заинтересованности темой в обществе, активности обсуждения темы и социального настроения.

Эффективность разработанных моделей была оценена путем проведения эксперимента по теме вакцинации против Covid-19. Сводный анализ представил отношение общественности к кампании вакцинации, политике вакцинации, а также к действиям и методам правительства по борьбе с пандемией. Следующим этапом была разработка модуля electronic Social Mood (eSM), представляющего собой приложение, анализирующее данные, полученные с помощью платформы OMSystem.

Цель диссертационной работы – разработка метода оценки социального настроения общества в медиапространстве Республики Казахстан с использованием моделей машинного обучения, нейронных сетей и маркетинговых технологий.

Задачи исследования:

1. Проведение анализа архитектуры и функционала интеллектуальной системы OMSystem.
2. Разработка модуля анализа и оценки социального настроения общества в медиапространстве Республики Казахстан с использованием моделей машинного обучения, нейронных сетей и маркетинговых технологий системы OMSystem.
3. Оценка разработанного модуля на примере анализа темы вакцинации против Covid-19.
4. Разработка модуля electronic Social Mood (eSM), анализирующего данные, полученные с помощью системы OMSystem, и выполняющего оценку социального настроения общества.

Объект исследования: текстовые данные, публикации, новостные ресурсы, социальное медиапространство Республики Казахстан.

Методы исследования: Data mining, Web mining, Обработка естественных языков (Natural language processing – NLP), Анализ тональности (Sentiment analysis – SA), Машинное обучение, Нейронные сети, Маркетинговые технологии социальной аналитики.

Теоретическая значимость исследования: анализ архитектуры и разработка функционала интеллектуальной системы OMSystem, оценка эффективности модуля оценки социального настроения общества.

Практическая значимость исследования: анализ социального настроения общества с использованием разработанного модуля обработки и анализа данных интеллектуальной системы OMSystem.

Научная новизна проведенных исследований и полученных результатов:

1. Разработан метод анализа социального настроения, отличающийся использованием моделей машинного обучения и маркетинговых показателей заинтересованности пользователей темой, активности обсуждения темы и уровня социального настроения.
2. Разработана интегрированная модель обучения анализа социального настроения, включающая семь атрибутных моделей машинного обучения, а также четыре модели глубокого обучения.
3. Разработан тональный словарь для казахского языка, использующийся для интегрированной модели анализа социального настроения.

Положения, выносимые на защиту:

1. Разработанный метод анализа социального настроения общества с использованием моделей машинного обучения и маркетинговых показателей, позволяющий оценить различные социально-политические темы и реакцию пользователей на проводимые государственные кампании.
2. Разработанная интегрированная модель обучения анализа социального настроения, включающая семь атрибутных моделей машинного обучения, а также четыре модели глубокого обучения.
3. Экспериментальные результаты по теме вакцинации против Covid-19, продемонстрировавшие отношение общественности и деятельность правительства с помощью метода анализа социального настроения.

Объем и структура работы. Диссертационная работа состоит из 162 страниц и включает 69 рисунков и 30 таблиц. Содержание включает 6 разделов.

Во введении дано описание актуальности, цели, задач, объектов и методов исследования, теоретической и практической значимости, а также новизны диссертационной работы.

Первый раздел описывает основные аспекты информационно-аналитических систем мониторинга социальных сетей, подробно рассматривает зарубежные и отечественные платформы мониторинга и анализа социального медиапространства, выделяет их преимущества и недостатки. Приводится описание основных методов определения тональности текстов и социального настроения: словарный подход, методы машинного обучения, нейронных сетей и маркетинговые технологии социальной аналитики.

Во втором разделе подробно представлена разработанная аналитическая платформа OMSystem. Она специализируется на расширенном анализе и мониторинге социальных сетей и новостных порталов медиапространства Республики Казахстан. OMSystem включает русский и казахский тональные словари, модели машинного обучения и нейронных сетей, а также инструменты для моделирования и определения социального настроения и самочувствия общества. Также в данном разделе представлена разработка моделей бинарной и многоклассовой классификации текстов, что является важнейшей частью диссертационной работы. Результаты приведены в виде графиков, сводных таблиц и выводов. Представлена разработка моделей на основе методов управления маркетингом в социальных сетях, позволяющих определить показатели социального настроения общества по заданным тематикам.

В третьем разделе выполнен эксперимент, направленный на анализ социального настроения общества относительно вакцинации против Covid-19. Данная тема приобрела особую популярность в связи с быстрым распространением пандемии в мире. Она активно обсуждалась в новостных ресурсах и социальных сетях, были написаны тысячи комментариев под постами, посвященными данной тематике. Оценка пользовательских мнений выполнена с помощью тональных словарей, моделей машинного обучения и нейронных сетей и маркетинговых технологий.

В четвертом разделе представлен разработанный на фреймворке Django Python модуль electronic Social Mood (eSM), являющийся приложением, анализирующим данные, полученные с помощью платформы OMSystem. Данный модуль выполняет следующие основные функции: создание основных категорий тем анализа социального настроения общества, извлечение количественных данных по каждой из тем, подсчет уровней заинтересованности темой в обществе, активности обсуждения темы и социального настроения, визуальное представление полученных результатов в виде графиков и сводных таблиц.

В заключении обобщены теоретические и практические результаты данной диссертационной работы, приведены ее наиболее значимые аспекты в анализе настроения

общества с применением моделей машинного обучения и нейронных сетей и показателей социального настроения.

Личный вклад исследователя. Диссертантом был выполнен анализ существующих платформ мониторинга социального медиапространства, архитектуры и функционала аналитической платформы OMSystem; разработан модуль обработки и анализа данных с использованием моделей машинного обучения, нейронных сетей и маркетинговых технологий социальной аналитики; дополнительно разработан модуль electronic Social Mood (eSM), выполняющий оценку социального настроения общества.

Степень обоснованности и достоверности научных результатов. Результаты диссертации были представлены в 12 научных работах, из них 2 статьи и 1 глава в книге опубликованы в журналах и книжной серии, рецензируемых в базе Scopus, 4 статьи – в журналах, рекомендуемых Комитетом по обеспечению качества в сфере образования и науки Министерства образования и науки Республики Казахстан, и 2 статьи – в научных конференциях, рецензируемых в базе Scopus, и 3 статьи – в материалах международных конференций:

1. Karyukin, V., Mutanov, G., Mamykova, Z., Nassimova, G., Torekul, S., Sundetova, Z. & Negri, M. On the development of an information system for monitoring user opinion and its role for the public. *Journal of Big Data* 9, 110 (2022). **https://doi.org/10.1186/s40537-022-00660-w**.
2. G. Mutanov, V. Karyukin and Z. Mamykova, "Multi-class sentiment analysis of social media data with machine learning algorithms," *Computers, Materials & Continua*, vol. 69, no.1, pp. 913–930, 2021. **https://doi.org/10.32604/cmc.2021.017827**.
3. Mutanov, G., Mamykova, Z., Karyukin, V., Yessenzhanova, S. The Approach to Building a Context-Dependent Sentiment Dictionary. In: Mutanov, G., Serikbekuly, A. (eds) Digital Transformation in Sustainable Value Chains and Innovative Infrastructures. Studies in Systems, Decision and Control, vol 443, 2022. Springer, Cham. **https://doi.org/10.1007/978-3-031-07067-9_1**.
4. Мутанов Г.М., Мамыкова Ж.Д., Карюкин В.И., Жақсыкелді А.Ж. Разработка машинно-обучаемого алгоритма определения тональности пользовательского восприятия контента. Вестник КазНИТУ Серия Технические Науки, Казахстан, 135 (5), 2019.
5. Alimzhanova L.M. Karyukin V.I. A classification model based on decision-making processes. Вестник КазНИТУ Серия Технические Науки, Казахстан, 138 (2), 2020.
6. Рахимова Д.Р., Тұрарбек А.Т., Карюкин В.И., Карибаева А.С., Тұрғанбаева А.О. Қазақ тіліне арналған заманауи машиналық аударма технологияларына шолу. Вестник КазНИТУ Серия Технические Науки, Казахстан, 141 (5), 2020.
7. Karibayeva A., Karyukin V.I., Turganbayeva A., Turarbek A. The translation quality problems of machine translation systems for the Kazakh language. Journal of Mathematics, Mechanics and Computer Science, Kazakhstan, vol. 111, n. 3, 2021.
8. Vladislav Karyukin, Aidana Zhumabekova, and Sandugash Yessenzhanova. 2020. Machine Learning And Neural Network Methodologies of Analyzing Social Media. In Proceedings of the 6th International Conference on Engineering & MIS 2020 (ICEMIS'20). Association for Computing Machinery, New York, NY, USA, Article 9, 1–7. **https://doi.org/10.1145/3410352.3410739**.
9. Diana Rakhimova, Vladislav Karyukin, Aidana Karibayeva, Assem Turarbek, and Aliya Turganbayeva. 2021. The Development of the Light Post-editing Module for English-Kazakh Translation. In The 7th International Conference on Engineering & MIS 2021 (ICEMIS'21). Association for Computing Machinery, New York, NY, USA, Article 69, 1–5. **https://doi.org/10.1145/3492547.3492651**.
10. Карюкин В., Есенжанова С. Построение контекстно-зависимого тонального словаря. Международная научная конференция студентов и молодых ученых «ФАРАБИ ӘЛЕМІ», Алматы, Казахстан, 2020.

11. Карюкин В. Подход к построению приложения eSM. Международная научная конференция студентов и молодых ученых «ФАРАБИ ӘЛЕМІ», Алматы, Казахстан, 2020.

12. Карюкин В. Многоклассовая классификация с применением алгоритмов машинного обучения. Международная научная конференция студентов и молодых ученых «ФАРАБИ ӘЛЕМІ», Алматы, Казахстан, 2021.

Связь диссертации с научно-исследовательскими работами. Данное исследование выполнено в рамках проекта по коммерциализации результатов научной и (или) научно-технической деятельности “Информационная система мониторинга мнений OMSystem (Opinion monitoring system)” 0101-18-ГК. (Основная роль диссертанта заключалась в разработке модуля анализа и оценки социального настроения, моделей машинного обучения и нейронных сетей, проведения эксперимента по анализу общественного настроения по теме вакцинации против Covid-19 и разработке модуля electronic Social Mood).