INNOPOLIS UNIVERSITY

Автономная некоммерческая организация высшего образования «Университет Иннополис»

РЕЦЕНЗИЯ НА ВЫПУСКНУЮ КВАЛИФИКАЦИОННУЮ РАБОТУ (МАГИСТЕРСКУЮ ДИССЕРТАЦИЮ)

REVIEW ON MASTER GRADUATE THESIS

ФИО обучающегося	Оконича Озиома Ненубари
Student's full name	Okonicha Ozioma Nenubari
Тема выпускной квалификационной работы	Обработка естественного языка для аудита соответствия нормативным требованиям. Автоматизированная проверка соответствия соглашений об обработке данных на основе обработки естественного языка
Thesis title	NLP for Regulatory Compliance Audit. NLP-based Automated Compliance Checking of Data Processing Agreements against General Data Protection Regulation

Уровень образования	**высшее образование магистратура**
Level of Education	**Master's degree**

Наименование направления подготовки	**09.04.01 Информатика и вычислительная техника**
Program track	**09.04.01 Computer Science**

Направленность (профиль) образовательной программы	• Анализ данных и искусственный интеллект
Field of Study	• Data Analysis and Artificial Intelligence

Рецензия на выпускную квалификационную работу[1]

[1] Рецензент указывает в произвольной форме:

В дипломной работе Озиома исследует способы использования NLP моделей для автоматизации проверки соответствия документов политики конфиденциальности общему регламенту защиты данных (GDPR). В работе реализована система для оценки соответствия документов политики обработки персональных данных путём мультиклассовой классификации каждого принципа GDPR на уровне предложения и всего документа. Актуальность темы для потребностей отрасли полностью раскрыта во введении и обзоре литературы. Тема соответствует направлению обучения "Анализ данных и искусственный интеллект", так как включает в себя анализ текста с помощью методов NLP.

Дипломная работа хорошо структурирована. В обзоре литературы студент отмечает текущий прогресс в области автоматизированного анализа соответствия нормативным требованиям и обозначает проблемы и потенциальные преимущества использования методов NLP. Однако я рекомендую добавить более подробную информацию о конкретных моделях NLP, использованных в предыдущих работах.

Студент использует два набора юридических данных, OPP-115 и ACL Coling, для получения обучающих и тестовых данных. Для предсказания каждого из семи принципов GDPR используются такие модели, как SBERT, BERT, GPT-2 и GPT-3 embeddings + MLP, с анализом текста на уровне предложений и всего документа, хотя анализ документа ограничен размером контекстного окна модели. Результат работы системы включает в себя вывод о соответствии документа каждому из принципов GDPR со ссылками на текст. Качество системы оценивается с помощью стандартных метрик классификации с дисбалансом классов. Кроме того, анализ содержит классификацию соответствия на основе ChatGPT и примеры использования GPT-4o, однако не ясно, является ли использование SOTA LLMs более многообещающим, чем SBERT, BERT и GPT-2.

В результате работы студент формулирует следующие ответы на поставленные исследовательские вопросы (RQ):
RQ1 - модели NLP эффективны и применимы для анализа соответствия GDPR;
RQ2 - существующие модели страдают от вычислительной сложности и проблем с интерпретацией;
RQ3 - модели NLP способны автоматизировать анализ соответствия GDPR, идентифицируя проблемы соответствия в документе.

- соответствует ли тема ВКР направлению подготовки, области, объектам, видам и задачам профессиональной деятельности;
- является ли тема ВКР актуальной, соответствует ли современному состоянию и перспективам развития науки, техники и технологиям;
- насколько освещена данная тема ВКР в монографиях, статьях, научных докладах и т.д.;
- учитывает ли тема ВКР интересы и потребности индустрии;
- четко ли поставлены цели/задачи работы, насколько соответствует их уровень теме ВКР;
- какие методы, техники и методики использовались во время работы над темой ВКР;
- основывается ли тема ВКР на практической работе обучающегося;
- каков уровень сформированности компетенций показал обучающийся во время работы над темой ВКР;
- есть ли публикации, доклады обучающегося по теме ВКР;
- существуют ли перспективы дальнейших исследований по данной теме;
- было ли самостоятельным выполнение работы обучающимся, ответственно ли и организовано относился обучающийся к работе, своевременно ли выполнялись все этапы индивидуального плана работы над темой;
- является ли работа стилистически выдержанной, имеет ли смысловую законченность и оформлена ли в соответствии с требованиями;
- рекомендуется или нет работа обучающегося к защите на заседании ГЭК;
- какую оценку заслуживает работа.

Возможным направлением будущей работы является развитие использования современных LLM для классификации соответствия GDPR.
Дипломная работа рекомендована к защите на заседании Государственной аттестационной комиссии. Предлагаемая оценка диплома - A (отлично).

Review[2]

During thesis work, Ozioma investigates the ways of using NLP models for automation of compliance checking of privacy policy documents for general data protection regulation (GDPR). Thesis builds a pipeline for evaluation of policy documents compliance via multi-classification of each GDPR principle on sentence and entire policy levels. The relevance of the topic to the interests and needs of the industry is fully justified in the introduction and literature review. The topic corresponds to the Data Analysis and Artificial Intelligence track, as it includes text analysis with NLP methods.

The thesis is well-structured. In the literature review, the student observes current state-of-the-art on automated compliance analysis and outlines challenges: insufficient legislative datasets and potential use of NLP methods. However, I recommend to include more detailed information about specific NLP models used in prior works.

The student uses two policy datasets, OPP-115 and ACL Coling, to acquire training and testing data. To predict each of the seven GDPR principles, models like SBERT, BERT, GPT-2, and GPT-3 embeddings + MLP are used, with text analysis on sentence and entire policy levels, although the whole policy analysis is limited by model context size. The result of the pipeline includes a summary on compliance to each principle with references to the policy text. The pipeline quality is measured using standard classification metrics with focus on class imbalance. Additionally, analysis contains ChatGPT-based compliance classification and examples of GPT-4o use, but it is not clearly stated whether usage of SOTA LLMs is more prominent than SBERT, BERT and GPT-2.

As a result of the work, student formulates the following conclusions for the stated research questions (RQ):
RQ1 - NLP models are effective and applicable for compliance analysis;
RQ2 - current models suffer from computational complexity and interpretability issues;

---

[2] Internal Examiner freely states:

- Whether Thesis title corresponds to the program track, field, object, types and objectives of professional activity;
- Whether Thesis title is up-to-date and correlates with current state and prospects of science and technology development;
- The extent to which Thesis title is featured in monographs, articles, scientific reports;
- Whether Thesis title takes into consideration industry's concerns and needs;
- Whether goals/objectives of the work are clear, the extent to which their level corresponds to Thesis title;
- Which methods, techniques and approaches were used during the work on Thesis;
- Whether Thesis title is based on student's applied work;
- What level of competencies formed a student has demonstrated during the work on Thesis;
- Are there any papers or reports on Thesis title published;
- Are there prospects for any further research of the topic;
- Whether the student's working process was autonomous, did he/she do his/her work in conscientious and well-organized manner, did he/she do timely all the stages of individual work plan on the topic;
- Does Thesis stick to the style, does it have style completeness and complies with standard requirements;
- Whether student's Thesis is recommended for the Thesis Defense during State Attestation Committee hearing;
- What grade Thesis deserves.

RQ3 - NLP models can automate compliance analysis via proper identification of compliance problems.
A great direction of future work would be to extend the use of modern LLMs for GDPR compliance classification.
The thesis is recommended for the Thesis Defense during State Attestation Committee hearing.
The proposed grade for the thesis is A (excellent).

Letter Grade / Оценка (A-D)	A

Internal Examiner		Startseva Anna
Рецензент	Старцева подпись / signature	Старцева Анна Андреевна