МИНИСТЕРСТВО ОБРАЗОВАНИЯ РЕСПУБЛИКИ БЕЛАРУСЬ

БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

МЕХАНИКО-МАТЕМАТИЧЕСКИЙ ФАКУЛЬТЕТ

Кафедра дифференциальных уравнений и системного анализа

Аннотация к дипломной работе

ИНТЕЛЛЕКТУАЛЬНЫЙ АНАЛИЗ ВЕБ-ДАННЫХ

Ковалевская Варвара Сергеевна

Научный руководитель:
кандидат физ.-мат. наук,
доцент Л. Л. Голубева

2025

В дипломной работе 48 страниц, 10 иллюстраций, 13 источников, 1 приложение.

ИНТЕЛЛЕКТУАЛЬНЫЙ АНАЛИЗ ВЕБ-ДАННЫХ, ОБРАБОТКА ЕСТЕСТВЕННОГО ЯЗЫКА, BERT, ROBERTA

Объектом исследования дипломной работы является веб-контент с манипулятивными заголовками.

Целью дипломной работы является исследование методов интеллектуального анализа веб-данных и их применение для распознавания типов фейковых новостей.

Для достижения поставленной цели были использованы: язык программирования Python, библиотеки HuggingFace Transformers (источник датасетов и предобученных моделей), Scikit-Learn и PyTorch.

В дипломной работе получены следующие результаты:

1. Изучены и систематизированы подходы к интеллектуальному анализу веб-данных
2. Разработан и обучен классификатор на основе модели RoBERTa для определения типа текстового спойлера
3. Проведён лингвистический и семантический анализ цифрового контента платформ Reddit, Facebook и Twitter.

Дипломная работа является завершенной, поставленные задачи решены в полной мере, присутствует возможность дальнейшего развития исследований.

Дипломная работа выполнена автором самостоятельно.

Thesis project is presented in the form of an explanatory note of 48 pages, 10 figures, 13 references, 1 application.

WEB DATA MINING, NATURAL LANGUAGE PROCESSING, BERT, ROBERTA

The object of the study is web content with manipulative (clickbait) headlines.

The aim of the thesis is to investigate methods of web data mining and their application for recognizing types of fake news.

To achieve this goal, the following tools and libraries were used: Python programming language, HuggingFace Transformers (as a source of datasets and pre-trained models), Scikit-Learn, and PyTorch.

The thesis produced the following results:

1. Approaches to web data mining were studied and systemized
2. A classifier based on the RoBERTa model was developed and trained to determine the type of textual spoiler
3. Linguistic and semantic analysis of digital content from Reddit, Facebook, and Twitter platforms was conducted.

The thesis project is complete, all tasks have been successfully done, there is a possibility for further research and development.

The thesis project was done solely by the author.