

## АНОТАЦІЯ

Кваліфікаційна робота містить пояснювальну записку обсягом 78 сторінок, яка включає 18 рисунки, 10 таблиць і 4 додатки. Об'єктом дослідження є система автоматизованої модерації токсичного текстового контенту на основі методів обробки природної мови (Natural Language Processing, NLP). Предметом розробки є алгоритмічне та програмне забезпечення для детекції та інтерпретації неприйнятних висловлювань у текстах англійською мовою.

У межах роботи реалізовано програмну систему, що виконує: покомпонентний аналіз тексту; класифікацію речень за рівнем токсичності з використанням моделі unitary/toxic-bert; ідентифікацію токсичних лексем із застосуванням word-level підходу; визначення типів токсичності (insult, hate speech, threat тощо); візуалізацію результатів у графічному інтерфейсі; генерацію рекомендацій щодо нейтрального переформулювання.

У ході реалізації проведено критичний аналіз існуючих методів виявлення токсичності, встановлено їхні функціональні обмеження, визначено вимоги до системи. Розроблено модуль sentence-level класифікації з багатомітковим виходом, інтегровано word-level механізм підсвічування небажаних фрагментів, сформовано алгоритм візуалізації результатів із використанням тултіпів. Застосунок реалізовано у вигляді десктопного GUI-додатку на основі бібліотеки Tkinter, з підтримкою автономної роботи та розширюваною архітектурою.

Ключові слова: ОБРОБКА ПРИРОДНОЇ МОВИ, ВИЯВЛЕННЯ ТОКСИЧНОСТІ, BERT, ВІЗУАЛІЗАЦІЯ, TKINTER, МОДЕРАЦІЯ, UNITARY/TOXIC-BERT, ГРАФІЧНИЙ ІНТЕРФЕЙС, LLM, ПЕРЕФОРМУЛЮВАННЯ.

## SUMMARY

The qualification thesis comprises a 78-page explanatory report, including 18 figures, 10 tables, and 4 appendices. The object of study is an automated toxic text moderation system based on Natural Language Processing (NLP) techniques. The developed solution focuses on the algorithmic and software implementation of tools for detecting and interpreting toxic expressions in English-language texts.

The system supports: modular analysis of user input; sentence-level toxicity classification using the unitary/toxic-bert model; identification of toxic words via a word-level detection mechanism; categorization of toxicity types (INSULT, HATE SPEECH, THREAT, etc.); visualization of toxic spans within a graphical user interface; and automated generation of rephrasing suggestions for toxic sentences.

During development, existing toxicity detection methods were reviewed and their limitations analyzed. Functional and non-functional system requirements were formulated. A sentence-level classification component with multilabel output was implemented, along with a word-level highlighting module. A visualization algorithm with tooltip-based explanations was developed. The final product is a desktop GUI application based on the Tkinter library, featuring offline operation and a modular architecture suitable for further extension.

**KEYWORDS:** NATURAL LANGUAGE PROCESSING, TOXICITY DETECTION, BERT, VISUALIZATION, TKINTER, MODERATION, UNITARY/TOXIC-BERT, GUI, LLM, REPHRASING.