**УДК 004.85**

# ЭКСПЕРИМЕНТАЛЬНЫЕ ИССЛЕДОВАНИЯ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ В АНАЛИЗЕ ТОНАЛЬНОСТИ ТЕКСТОВ НА РУССКОМ ЯЗЫКЕ

**Быков Д.С.** (ИТМО)
**Научный руководитель – доктор технических наук, доцент Хитров Е.Г.**
(ИТМО)

**Введение.** Методы машинного обучения применяются в анализе текстов, например, для: определения языка текста, классификации текстов по жанрам, фильтрации спама, установления авторства текста, обнаружения «токсичных» комментариев, выявления кибер-буллинга и проч. В статье [1] авторы провели анализ современного состояния проблематики исследования. Решение задач, возникающих в области обработки естественного языка (NLP), считается весьма сложным с профессиональной и исследовательской точки зрения [1, 2].

**Основная часть.** Можно, с некоторым допущением, заметить, что для решения ряда из них уже предложены и апробированы работоспособные решения. Например, подготовить достаточно точную модель машинного обучения (ML) для определения языка текста, написанного на одном из европейских языков, можно при помощи следующего стека: сбор и предобработка данных для обучения модели, векторизация, обучение модели-классификатора, оценка качества классификации. При этом решается ряд задач: выбор моделей векторизации и классификации, настройка их гиперпараметров. Подобный стек может использоваться и для решения более «сложных» задач, например – выявления опасного явления, такого как буллинг в комментариях, тем более, в какой-либо узкой социальной прослойке. Формально, набор инструментов может остаться прежним. Однако, сбор и подготовка данных для обучения ML-модели станет нетривиальной практической задачей, а настройка гиперпараметров потребует отдельных исследований.

С точки зрения современной науки [3], элементы стека можно рассматривать как «черные ящики» и использовать их в качестве составных элементов экспериментального стенда, включающего в себя аппаратную часть и средства программной реализации вычислений (аналоги, например, [4 – 6]). Поскольку известны различные подходы к оценке качества ML-моделей (precision, recall, $F_1$-score), объединим их условно в класс performance. Множество гиперпараметров обозначим Lambda.

**Выводы.** В контексте предметной области, применительно к решаемой задаче, целесообразно выделить структурные элементы исследования:

1) Сбор и предобработка данных для обучения ML-модели.
2) Разработка методики экспериментального исследования «эффективности» (performance) стеков с различными «черными ящиками» и их параметрами (Lambda).
3) Апробация методики и получение результатов экспериментов с «черными ящиками» - функции performance (Lambda).
4) Подготовка и тестирование модуля распознавания тональности комментариев на русском языке.

**Список использованных источников**:

1. Khurana D., Koli A., Khatter K., Sukhdev S. Natural Language Processing: State of The Art, Current Trends and Challenges. // Multimedia Tools and Applications 82, 3713-3744 (2023).
2. Mikolov T. et al. Distributed Representations of Words and Phrases and their Compositionality // arXiv, 2013. http://arxiv.org/abs/1310.4546.
3. Латур Б. Наука в действии: следуя за учёными и инженерами внутри сообщества // пер.

с анг. К. Фёдоровой; научн. ред. С. Миляева. — СПб.: Издательство Европейского университета в Санкт-Петербурге, 2013. — 414 с.

4. Poon A., Sung J. Opening the black box of AI-Medicine // Journal of Gastroenterology and Hepatology 36 (2021) 581–584.

5. Ronald Y., Gabriele S. What's Inside the Black Box? AI Challenges for Lawyers and Researcher // Legal Information Management, 19 (2019), pp. 2–13.

6. Hassija V., Chamola V., Mahapatra A. et al. Interpreting Black-Box Models: A Review Explainable Artificial Intelligence // Cognitive Computation 16, 45–74 (2024).