УДК 004.855.5

СРАВНЕНИЕ ПОДХОДОВ ДЛЯ ПОВЫШЕНИЯ ТОЧНОСТИ АЛГОРИТМОВ КЛАССИФИКАЦИИ ЖАЛОБ НА ОСНОВЕ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ

Скворцов Д. А. (Университет ИТМО) **Леманов А. А.** (Университет ИТМО),
Ореховский И. А. (Университет ИТМО)

Научный руководитель – кандидат технических наук, доцент Фёдоров Д.А.
(Университет ИТМО)

Введение

Работники жилищно-коммунальной инспекции (ЖКИ) сталкиваются с проблемой обработки жалоб и запросов от граждан. Каждый запрос необходимо разбить по тематикам, чтобы отдать в подходящее подразделение. Чтобы оптимизировать этот процесс, предлагается разработать систему-классификатор на основе методов машинного обучения и обработки естественного языка.

Основная часть

Предлагаемая система состоит из двух основных компонентов: классифицирующей модели и модуля обработки и поиска данных. MLTC (Multi-label text classification) Модель работает с запросами, поступающими в виде векторов, полученных при помощи системы word2vec, и определяет их в одну из тематик, представленных в базе данных.

Для повышения точности системы используется несколько подходов.

1) Очистка данных, выбор классифицируемых тематик и сокращение признакового пространства для нормализации данных.

2) Выбор алгоритма классификации, показавшего лучший результат на этих данных, по сравнению с остальными. Для анализа рассматриваются следующие модели: метод Байеса, решающее дерево, метод опорных деревьев, метод ближайших соседей (KNN), модели-трансформеры (в частности, трансформер на основе предобученной модели BERT). Помимо этого, для сравнения берется подход без обучения - с использованием LLM и RAG-системы.

3) Смешение разных классификаторов в один путем перемножения векторов с вероятностями.

4) Сравнение по следующим метрикам: accuracy, accuracy-top-k (истинность одного из наиболее вероятных предсказаний), roc-auc (качество классификации, различимость классов в целом).

Также при помощи поиска данных, текст запроса дополняется полезной информацией, взятой из внешних источников. Модуль анализирует текст на упоминание вложений и законодательных актов, которые можно найти в базе данных. Так, более содержательный ввод позволяет избавиться от шума в сообщении и выделить основные детали для дальнейшей классификации.

В итоге, все алгоритмы сравниваются и ранжируются по успешности в своих задачах относительно метрик.

Выводы

Разработка системы классификатора тематик позволит значительно снизить нагрузку на операторов в сфере ЖКИ, автоматизируя анализ запросов. Используемые методы поиска информации и анализ существующих алгоритмов классификации повысит точность модели. Это откроет новые возможности для эффективной работы с базами текстовых запросов, анализа больших объемов информации и повышения качества обслуживания, в таких областях, как ЖКИ. Дальнейшие исследования могут быть направлены на расширение возможностей анализа запросов, исследование глубокого обучения моделей и методов предварительного анализа и очистки входных данных.

Список использованных источников:

1. Носков Дмитрий Владимирович Классификация текстов при помощи алгоритмов

машинного обучения // Вестник науки и образования. 2018. №4 (40). URL: https://cyberleninka.ru/article/n/klassifikatsiya-tekstov-pri-pomoschi-algoritmov-mashinnogo-obucheniya (дата обращения: 17.02.2025).

2. Воробьев Александр Викторович МЕТОД ВЫБОРА МОДЕЛИ МАШИННОГО ОБУЧЕНИЯ НА ОСНОВЕ УСТОЙЧИВОСТИ ПРЕДИКТОРОВ С ПРИМЕНЕНИЕМ ЗНАЧЕНИЯ ШЕПЛИ // Экономика. Информатика. 2021. №2. URL: https://cyberleninka.ru/article/n/metod-vybora-modeli-mashinnogo-obucheniya-na-osnove-ustoychivosti-prediktorov-s-primeneniem-znacheniya-shepli (дата обращения: 17.02.2025).

3. Misra, P. and Yadav, A. S. (2020). Improving the Classification Accuracy using Recursive Feature Elimination with Cross-Validation. International Journal on Emerging Technologies, 11(3): 659–665. URL: https://www.researchgate.net/publication/344181117_Improving_the_Classification_Accuracy_using_Recursive_Feature_Elimination_with_Cross-Validation (дата обращения: 17.02.2025).

4. Ikonomakis, M., Kotsiantis, S., and Tampakas, V. (2005). Text classification using machine learning techniques. WSEAS Trans. Comput. 4,966–974. URL: https://www.researchgate.net/publication/228084521_Text_Classification_Using_Machine_Learning_Techniques (дата обращения: 17.02.2025).

5. (2020). Learning to rank for multi-label text classification: Combining different sources of information. Natural Language Engineering, 27, 89 - 111. URL: https://doi.org/10.1017/S1351324920000029 (дата обращения: 17.02.2025).

6. González-Carvajal, S., & Garrido-Merchán, E. (2020). Comparing BERT against traditional machine learning text classification. ArXiv, abs/2005.13012. URL: https://doi.org/10.47852/bonviewJCCE3202838. (дата обращения: 17.02.2025).