УДК 004.8

Сравнительный анализ методов МО для классификации текста

Вансович А.И. (ИТМО)
Научный руководитель – кандидат технических наук, доцент Ключев А.О.
(ИТМО)

Введение. Обработка естественного языка (Natural Language Processing, NLP), представляет собой ряд задач, связанных с анализом и обработкой текста. В число таких задач входят машинный перевод, извлечение информации, анализ настроений, генерация текста, классификация текста и т. д. В свою очередь задачи классификации могут подразделяться также на отдельные типы: двоичная классификация, многоклассовая классификация, многозадачная классификация и т. п. Классификация текста имеет огромную практическую пользу в различных областях жизни: организация, фильтрация и анализ данных. В связи с развитием технологий растёт и количество информации, потребляемой человеком, бизнесом, что говорит об актуальности разработки системы, упрощающей процесс её анализа.

Основная часть. Алгоритмы машинного обучения давно используются для решения задач NLP, в частности для задач классификации.

Для данного типа задач используются как традиционные методы машинного обучения [1][2]:

1) Логистическая регрессия
2) k-NN классификация
3) Метод опорных векторов
4) Наивный Байесовский классификатор

А также находят широкое применение модели глубокого обучения [1][2]:

1) MLP
2) RNN
3) CNN

Это лишь небольшой список методов машинного обучения, с каждым годом растёт число архитектур моделей глубокого обучения, предназначенных для классификации текста, каждая из которых достигает большей точности в своей предметной области, к примеру, HDLTex [3], ULMFit [4].

В области анализа новостей различных компаний, представленных на бирже, классификация текста позволит автоматизировать процесс агрегирования, фильтрации данных для последующей обработки, к примеру, отбирать классифицированные тексты по какому-либо из признаков и передавать их на вход другой модели, чтобы получить краткое содержимое по интересующей теме.

Выводы. Проведен анализ методов машинного обучения для решения задачи классификации текста.

Список использованных источников:

1. Gasparetto, A.; Marcuzzo, M.; Zangari, A.; Albarelli, A. A Survey on Text Classification Algorithms: From Text to Predictions. // Information 2022, 13, 83.

2. Qian Li, Hao Peng, Jianxin Li, Congying Xia, Renyu Yang, Lichao Sun, Philip S. Yu, and Lifang He. 2022. A Survey on Text Classification: From Traditional to Deep Learning. // ACM Trans. Intell. Syst. Technol. 13, 2, Article 31 (April 2022), 41 pages.

3. Kamran Kowsari, Donald E. Brown, Mojtaba Heidarysafa, Kiana Jafari

Meimandi, Matthew S. Gerber, Laura E. Barnes. HDLTex: Hierarchical Deep Learning for Text Classification // 2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)

4. Jeremy Howard and Sebastian Ruder. 2018. Universal Language Model Fine-tuning for Text Classification. // In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 328–339, Melbourne, Australia. Association for Computational Linguistics.