УДК 004.62

ОБЗОР МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ ДЛЯ КЛАССИФИКАЦИИ ТЕКСТОВ

Полин Я.А. (Университет ИТМО, Санкт-Петербург)

Научный руководитель – к.т.н., доцент Ананченко И.В. (Университет ИТМО, Санкт-Петербург)

Введение. Постановка задачи. Задача классификации текстов одна из главных задач компьютерной лингвистики, так как к ней сводятся другие задачи: определение тематической принадлежности текстов, автора текста, эмоциональной окраски высказываний и др. В настоящее время классификация текстов все чаще реализуется с использованием методов машинного обучения. Актуальность исследования обусловлена растущим спросом на использование методов классификации. Методы классификации текстов лежат на стыке двух областей – информационного поиска и машинного обучения. На данный момент для задач классификации текстов существует множество различных методов и их вариаций. Решение задачи классификации состоит из четырех последовательных этапов:

- предварительная обработка документа и его индексация;
- уменьшение размерности пространства признаков;
- построение классификатора и его обучение с использованием методов машинного обучения;
- оценивание качества результатов классификации.

При выборе конкретного алгоритма классификации следует учитывать особенности каждого из них.

Цель работы. Исследовать наиболее популярные методы машинного обучения для задачи классификации текстов, сравнить возможности данных методов.

Промежуточные результаты. В процессе выполнения работы были рассмотрены формальная постановка задачи классификации текстов, общая схема классификации текстов, распространенные методы построения и обучения классификатора:

- NB (Naive Bayes);
- KNN (k Nearest Neighbors);
- SVM (Support Vector Machine);
- DT (Decision Trees)
- логистическая регрессия (logistic regression).

Основной результат. В рамках работы были рассмотрены основные методы машинного обучения для классификации текстов, а также был произведен анализ и сравнение по точности, времени работы, объему обучающей выборки и другим характеристикам.