А. В. Антонов

МОДЕЛИРОВАНИЕ НОВОСТНОГО КОНТЕНТА ПОСРЕДСТВОМ МАШИННОГО ОБУЧЕНИЯ (на материале портала BBC News)

Машинное обучение становится все более важным инструментом для анализа и обработки огромных объемов данных в различных областях. Анализ новостного контента находится среди практически значимых областей его применения. В данной статье будут рассмотрены некоторые методы и подходы к моделированию новостного контента на примере данных, собранных с портала *BBC News*.

Первый и самый главный этап – сбор и предобработка данных. Для анализа используются данные с портала *BBC News*, так как он предоставляет широкий спектр новостей по различным тематикам. Для сбора данных можно использовать веб-скрейпинг, *RSS*-ленты или *API*, предоставляемые самим порталом или сторонними сервисами (C. M. Bishop, 2006).

На этапе предобработки происходит удаление шума, токенизация текста, удаление стоп-слов, лемматизация и т.д. Ее цель – привести данные к формату, пригодному для дальнейшего анализа.

С помощью алгоритмов классификации можно определить тональность новости. Это полезно для оценки общего настроения в обществе или для отслеживания реакции на определенные события.

Другим подходом в моделировании является кластеризация новостных статей. Алгоритмы группируют новости по тематикам или ключевым словам, что помогает в создании тематических разделов на сайтах новостей или улучшение рекомендаций читателя на основе его интересов.

Также важным аспектом моделирования новостного контента является предсказание популярности новостей. С помощью методов регрессии или временных рядов можно предсказать, какие новости будут наиболее популярными у пользователей портала (T. Hastie, 2009).

Для оценки точности модели можно использовать различные метрики качества (*precision*, *recall*, *F1-score*) и провести визуализацию результатов (например, через *word clouds* или графики).

В заключение, машинное обучение предоставляет мощные инструменты для анализа и моделирования новостного контента. На примере данных с портала *BBC News* могут быть рассмотрены методы работы с новостными данными, подходы к их анализу и моделированию. Дальнейшие исследования в этой области могут помочь улучшить качество и релевантность новостного контента для пользователей.