**УДК 004.4'2**

# ИССЛЕДОВАНИЕ ВОЗМОЖНОСТЕЙ ПРИМЕНЕНИЯ НЕЙРОННЫХ СЕТЕЙ ДЛЯ ЛИНГВИСТИЧЕСКОГО АНАЛИЗА

**Поздняков М.В.** (федеральное государственное автономное образовательное учреждение высшего образования «Национальный исследовательский университет ИТМО»)
**Научный руководитель – кандидат технических наук, доцент Осипов Н.А.**
**(**федеральное государственное автономное образовательное учреждение высшего образования «Национальный исследовательский университет ИТМО»)

В докладе приводится описание системы кластеризатора отзывов покупателей к товарам интернет-магазина. Были сравнены самые распространённые метод машинного обучения для обработки текстовых данных, в том числе основанные на использовании нейронных сетей.

**Введение.** Анализ больших массивов текстовых данных является важным направлением машинного обучения. Отдельную роль в обработке текста играют нейронные сети. Они существенным образом повышают качество решения некоторых стандартных задач классификации текстов и последовательностей, снижают трудоёмкость при работе непосредственно с текстами, а также позволяют решать новые задачи (например, создавать чат-боты). В то же время нейронные сети нельзя считать полностью самостоятельным механизмом решения лингвистических проблем. В рамках данного доклада будут рассмотрены различные подходы к решению задачи кластеризации отзывов пользователей к товарам интернет-магазина.

**Основная часть.** Данная система была выбрана в качестве объекта исследования по следующим причинам. На данный момент существует большое количество решений для классификации отзывов. Например, подобные решения могут выделять среди отзывов положительные, нейтральные и отрицательные. Однако классификаторы ограничены в количестве возможных категорий отзывов. Тем не менее, для пользователя может быть полезно рассмотреть для каждого товара самые часто встречающиеся темы, поднимаемые в отзывах. Для этого необходимо выделять не заданные заранее категории, а кластеры, которые могут отличаться для каждого товара. Из существующих аналогов можно выделить классификаторы. Разнообразие библиотек для машинного обучения на языке Python позволяет за короткое время создать и обучить модель классификатора. Примером использования нейронных сетей для выделения наиболее популярных тем отзывов является решение от Яндекс.Маркета. Тем не менее, объединение всех тем в один отзыв может быть не всегда уместно. Пользователи отмечают абсурдность и противоречивость некоторых подобных пересказов. Если мнения пользователей о товаре разделились, то умный отзыв может похвалить товар и сразу же добавить, что его не стоит покупать.

Далее были рассмотрены наиболее часто используемые инструменты для автоматической обработки текстовых данных. Наивный байесовский классификатор представляет собой один из самых простых инструментов для анализа текста. Он основывается на формуле Байеса и предполагает, что какие-то слова чаще употребляются в одной категории текстов, а какие-то в другой. Главными достоинствами данного метода являются простота реализации и возможность дообучения модели. Тем не менее, данный метод не учитывает порядок слов, а также производит классификацию, а не кластеризацию. Метод TF-IDF предполагает назначение каждому упомянутому слову для каждого текста коэффициента, которых рассчитывается как произведение частоты слова в тексте и обратной частоты документа. В отличие от предыдущего метода, TF-IDF позволяет подготовить тексты к дальнейшей кластеризации. Тем не менее, данный метод всё ещё не учитывает порядок слов.

К сожалению, упомянутые выше методы при простоте реализации имеют недостатки, критические для рассматриваемой задачи. Модели, построенные на данных методах, чрезмерно упрощены и могут недостаточно точно различать тематику отзыва для определения принадлежности к кластеру. Применение нейронных сетей является важным шагом в сторону решения таких проблем. Наиболее часто применяемыми для анализа текста являются рекуррентные нейронные сети (RNN). В отличие от традиционных нейронных сетей, где подразумевается, что все входы и выходы независимы, рекуррентные нейросети учитывают предшествующие данные на каждом шаге. Разновидностью рекуррентных нейронных сетей являются LTSM. Их главными особенностями являются четыре слоя в повторяющемся модуле, а также такой ключевой компонент, как состояние ячейки (cell state). LSTM разработаны специально, чтобы избежать проблемы долговременной зависимости. Запоминание информации на долгие периоды времени – это их обычное поведение, а не что-то, чему они с трудом пытаются обучиться. Эти особенности выделяют LSTM среди других методов анализа текста и делают их крайне удобным инструментом для лингвистического анализа.

**Выводы.** В контексте лингвистического анализа нейронные сети являются наилучшим инструментом машинного обучения. Для разработки первого прототипа кластеризатора будет использоваться LTSM.

Поздняков М.В. (автор) Подпись

Осипов Н.А. (научный руководитель) Подпись