УДК 004.93

Использование нейронных сетей в задаче анализа тональности текста

С.В. Мирзаянова
(Университет ИТМО, г. Санкт-Петербург)
Научный руководитель – к. ф.-м. н., доцент, С.В. Рыбин
(Университет ИТМО, г. Санкт-Петербург)

Исследования выполнены за счет стартового финансирования университета ИТМО в рамках НИР № 618278 «Синтез эмоциональной речи на основе генеративных состязательных сетей».

Задача анализа тональности текста может быть интересна для применения в различных системах, таких как синтез речи или рекомендательные системы, а также при анализе удовлетворенности клиентов и пользователей какого-либо бизнес проекта.

В последнее время проведено немало исследований в области анализа тональности текстов с использованием пользовательских отзывов, а также коротких сообщений (твитов). На данный момент актуальна задача совершенствования предсказательных систем анализа тональности. Следует отметить, что данное исследование направлено на работу только с текстовыми данными, и не включает в себя обработку других модальностей.

С целью повышения точности определения тональности рассматривается необходимость разработки нейросетевых архитектур, позволяющих обрабатывать большие объемы данных с минимальной предварительной обработкой. Однако нейросетевые архитектуры требуют особого подхода к подбору их параметров. В связи с чем необходимо проведение многочисленных опытов для определения лучшего набора параметров нейросетевой архитектуры, что может позволить превзойти достигнутые пределы точности определения тональности текста.

Главной особенностью работы с данными при обучении нейросети является необходимость представления данных в виде вектора признаков[1]. При работе с текстовыми данными чаще всего выполняется векторизация слов, как единиц письменной речи. Для работы с векторами слов используется word embeddings [2,3]. Первые подходы к векторизации были неоптимальны, с точки зрения занимаемых объемов памяти и количества сохраняемой информации. Это такие алгоритмы как Bag-of-words и One-hot vectors. Для векторизации слов часто используются алгоритмы word2vec, что позволяет сохранять информацию о семантических связях и получать сжатые вектора.

При исследовании применимости нейросетевых архитектур сравниваются различные структуры рекуррентных нейросетей. Именно рекуррентные нейросети показывают лучший результат при работе с речевыми данными, как в случае аудио данных, так и в случае текстовых данных. Предлагается использование архитектуры LSTM с механизмом внимания для обработки длинных текстов.

Литература

1. Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng., Cristopher Potts // Learning Word Vectors for Sentiment Analysis. Association for Computational Linguistics. Portland, Oregon, 2011. P. 142 – 150.

2. Pengrei Liu, Shafiq Joty, Helen Meng // Fine-grained Opinion Mining with Recurrent Neural Networks and Word Embeddings. Association for Computational Linguistic. Lisbon, Portugal, 2015. P. 1433 – 1443.

3. Kaliyev A., Rybin S.V., Matveev Y. The Pausing Method Based on Brown Clustering and Word Embedding // Lecture Notes in Computer Science. –2017. –V. 10458. –P. 741–747