УДК 004.838.2

РЕАЛИЗАЦИЯ АЛГОРИТМА АНАЛИЗА ТОНАЛЬНОСТИ ТЕКСТА ЧАТОВ

Самигулин Т.Р. (Университет ИТМО)
Научный руководитель – профессор факультета инфокоммуникационных технологий доктор технических наук, доцент Басов О.О.
(Университет ИТМО)

Аннотация.

В работе демонстрируется способ анализа тональности текста с помощью нейронной сети LSTM. Описываются этапы подготовки данных и обучения классификатора.

Введение.

Анализ тональности текста чатов на сегодняшний день является актуальной задачей. Однако практически отсутствуют источники информации, описывающие полный цикл реализации алгоритма анализа тональности для решения прикладных задач. В имеющихся исследованиях либо отсутствует описание предобработки данных, что делает невозможным повторить эксперимент, либо исследование проводилось на синтетических данных и никак не тестировалось на реальных, что не позволяет точно оценить качество модели.

Основная часть.

В исследовании описывается процесс обучения нейронных сетей для классификации тональности текста. Результатом классификации является положительная или отрицательная тональность текста. Поскольку предполагается использовать нейросетевую модель для анализа тональности текста в чате Zoom-конференции, то её обучение проводилось на коротких сообщениях, для чего использовался датасет RuTweetCorp, содержащий ~220000 размеченных по тональности русскоязычных твитов.

На начальном этапе производилась предобработка данных: приведение к нижнему регистру, удаление пунктуации. Затем производились разбиение текстов на токены и их лемматизация – приведение слов в начальную форму, а также удаление стоп-слов, которые встречаются слишком часто и не несут для исследования полезной информации.

После очистки данных производилась векторизация –тексты преобразовывались в численные векторы для дальнейшей передачи их модели. Для векторизации использовался алгоритм word2vec, осуществляющий преобразование слов в вектора на основе их семантической близости.

После окончания этапа предобработки данные использовались для обучения нейронной сети LSTM, модификации рекуррентной нейронной сети, особенностью которой является способность обучаться долгосрочным зависимостям. Для тестирования модели использовалась собственная размеченная вручную выборка.

Выводы.

В проведенном исследовании оценена эффективность нейронной сети LSTM для задачи анализа тональности текста в чате Zoom-конференции. Результаты проведенного исследования могут быть использованы для создания собственного анализатора тональности текста.

Самигулин Т.Р. (автор) Подпись

Басов О.О. (научный руководитель) Подпись