МУЛЬТИАГЕНТНЫЕ ТЕХНОЛОГИИ И МАШИННОЕ ОБУЧЕНИЕ

А.Л. Антипин, Д.В. Блинова

АНАЛИЗ ЭМОЦИОНАЛЬНОЙ ОКРАСКИ СООБЩЕНИЙ СОЦИАЛЬНОЙ СЕТИ TWITTER

(Уфимский государственный авиационный технический университет)

Одной из проблем в задаче обработки естественного языка является определение эмоциональной окраски текста, то есть его тональности. Цель такой задачи состоит в автоматическом определении, является ли текст положительным, отрицательным, либо нейтральным по своему отношению к описанному в нем объекту. Трудность решения данной задачи состоит в том, что, как правило, нормы общения людей в социальных сетях заметно отличаются от каких-либо научных публикаций или литературных произведений. В сообщениях часто присутствуют грамматические ошибки, ошибки пунктуации, опечатки, неоднозначность, сарказм, сленг, а так же другие особенности, определяемые самим автором сообщений. Все эти факторы порой не может распознать и сам человек, не считая компьютер.

Целью данной работы является создание инструмента, позволяющего классифицировать с определенной точностью эмоциональную окраску сообщений, полученных из социальной сети Twitter для дальнейшего анализа полученных результатов.

Одним из путей решения данной задачи является использование нейронных сетей. Чаще всего для обработки последовательной информации, которой и является текст, используются рекуррентные нейронные сети. Самой распространенной архитектурой рекуррентной нейронной сети на данный момент является «долгосрочная краткосрочная память» (Long Short-term memory). В таких сетях внутренние нейроны «оборудованы» сложной системой, так называемых ворот (gates), а также концепцией клеточного состояния (cell state), которая и представляет собой некий вид долгосрочной памяти. Ворота же определяют, какая информация попадет в клеточное состояние, какая сотрется из него, и какая повлияет на результат, который выдаст нейронная сеть на данном шаге. Именно эти вариации рекуррентных нейронных сетей широко используется сейчас, например, для машинного перевода Google.

Для реализации такой LSTM-сети будем использовать наиболее популярную связку библиотек Theano и Keras. Theano - библиотека для символьных и тензорных вычислений в Python, работающая заметно быстрее аналога от компании Google, под названием TensorFlow за счет ее повсеместной оптимизации

вычислений. Keras - надстройка для Theano, упрощающая создание и обучение нейронных сетей, и дающая простой и удобный набор абстракций, методов и объектов. Для обучения данной нейронной сети будем использовать корпус твитов, подготовленный Юлией Рубцовой[1]. Данный корпус представляет собой два файла с именами «positive.csv» и «negative.csv», содержащие 114911 положительных и 111923 отрицательных записи. Перед началом обучения данные из корпуса нужно “очистить” от различно мусора, который не несет никакой смысловой нагрузки. А далее по средствам стемминга провести нормализацию слов. Стемминг - это процесс нахождения основы слова для заданного исходного слова. После этого данный корпус твитов можно использовать для обучения LSTM-рекуррентной нейронной сети. На данный момент не существует какого-то алгоритма для выбора топологии нейронной сети и большинство используемых нейронных сетей результат экспериментов с различными конфигурациями.

Для проверки способности LSTM-рекуррентной нейронной сети решать задачу определения эмоциональной окраски текста, создадим сеть с 2 слоями по 64 нейрона и слоем Dropout, отвечающим за переобучение и описанным в статье «Dropout: A Simple Way to Prevent Neural Networks from Overfitting»[2]. Библиотека Theano имеет возможность запуска обучения нейронной сети не только на центральном процессоре компьютера, но и на графическом процессоре с поддержкой ядер CUDA, что может обеспечить в некоторых случаях 20-ти кратный прирост скорости обучения.

Заключение

В процессе изучения проблемы определения эмоциональной окраски текста, которым являлись сообщения русскоязычного сегмента микроблогинговой социальной сети Twitter, была написана и протестирована LSTM-рекуррентная нейронная сеть. Обучение производилось на 200000+ сообщений, заранее распределенных по классам положительной или отрицательной эмоциональной окраски текста. В результате, данная нейронная сеть показала точность около 75% при определении тональности текста.

В дальнейшем, экспериментируя с различными топологиями сети и используя для обучения выборку большего объема, чем использовалась в эксперименте, можно повысить точность определения тональности.

Реализация на языке Python с использованием библиотек Theano и Keras позволит в дальнейшем встроить и использовать ее в структуре какого-либо веб-приложения или программного продукта.

Литература

1. Корпус коротких текстов на русском языке на основе постов Twitter: **http://study.mokoron.com/**

2. Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, Ruslan Salakhutdinov «Dropout: A Simple Way to Prevent Neural Networks from Overfitting», The Journal of Machine Learning Research Volume 15 Issue 1, 2014, 1929-1958