УДК 004

МЕТОДЫ ДЛЯ АВТОМАТИЗИРОВАННОГО ОПРЕДЕЛЕНИЯ ЭМОЦИОНАЛЬНОЙ ОКРАСКИ ТЕКСТА

Звонарев А.А. (Университет ИТМО, г. Санкт-Петербург)
Научный руководитель – **к.м.н., доцент Билый А.М.**
(Университет ИТМО, г. Санкт-Петербург**)**

Представленная работа посвящена сравнению эффективности различных методов анализа тональности текста. На корпусе русскоязычных твитов были протестированы три модели для решения проблемы бинарной классификации: логистическая регрессия (LR), классификатор XGBoost и сверточная нейронная сеть (CNN). Основываясь на полученных результатах CNN показала лучшие результаты, но при этом время, потребовавшееся на обучение LR, существенно меньше.

С каждым годом все больше общения, услуг, товаров переходит в сеть интернет и в основном вся информация предоставляется в виде текста. В связи с этим все более остро стоит задача определения эмоционального состояния человека без личного общения. Перспективы этого направления заключаются в том, что опираясь на текстовую информацию можно понять, в каком настроении находится собеседник, оценить успешность политических и экономических реформ, проверить, как человек реагирует на то или иное событие и решение. В настоящее время можно выделить несколько ключевых методологий для определения эмоциональности текста:

- Анализ с использованием заранее составленного словаря. Такие словари состоят из заранее подготовленных шаблонных слов, фраз и их сочетаний, где каждому элементу соответствует эмоциональная характеристика. Также для определения эмоциональной окраски используется корпусная лингвистика, которая позволяет улучшить точность оценки тональности. Оценка производится по совокупности найденных положительных и отрицательных паттернов. При явном выделении одного из них тексту или отрывку выставляется класс, набравший больше очков. Если явного перевеса нет – оценка выставляется как нейтральная. Основным недостатком является процедура составления словарей терминов с указанием веса фраз. Также эти словари необходимо подготавливать для конкретной области;
- анализ с использованием методов машинного обучения в последнее время получили наибольшее распространение, так как убирается человеческий фактор воздействия на оценку, ведь при словарном методе человек сам определяет вес того или иного слова, в свою очередь при использовании машинного обучения нейронные сети могут научиться самостоятельно выявлять закономерности в представленных текстах и достичь распознавания иронии и сарказма.

Предлагаемый метод основан на использовании методов машинного обучения. Разработанный прототип позволяет проводить оценку текста на “положительные” и “отрицательные”. Без явного выделения одного из типов относим к нейтральным. В качестве обучающей выборки был взят набор русскоязычных корпусов коротких текстов RuTweetCorp состоящий из 17,639,674 записей. Часть данных распределена на две группы: «заведомо положительные» (114,911 записей) и «заведомо отрицательные» (111,923 записей). Как и в любой задаче, связанной с использованием машинного обучения в первую очередь необходимо подготовить данные. В ходе работы были осуществлены следующие действия: весь текст приведен к одному регистру; удалены обращения к другим пользователям; удалены знаки пунктуации; слова с частицей «не» объединены; удалены ссылки; удалены стоп-слова.

В результате полученных действий был получен список всех уникальных используемых слов в данном наборе данных. Такой список все еще содержал в себе слишком много уникальных слов, что приводило к большому потреблению памяти и большим затратам на обучение модели, но не оказывало никакого положительного эффекта на качество распознавания. В связи с этим было принято решение о сокращении списка, а именно, все слова, встречающиеся реже трех раз были отброшены. Используя полученный словарь была подготовлена TF-IDF матрица. Это статистическая мера, используемая для оценки важности слова в контексте документа, являющегося частью коллекции документов или корпуса. Вес некоторого слова пропорционален частоте употребления этого слова в документе и обратно пропорционален частоте употребления слова во всех документах коллекции.

Так же следует помнить, что тексты имеют различную длину, но варьировать количество входных нейронов мы не можем, поэтому был проведен анализ длин текста на основании доступных данных твитов. Была выбрана размерность длины текста равным 23 словам, так как это значение покрывает 99.65% всех сформированных корпусов. В качестве архитектур выбор пал на XGBoost, логистическую регрессию и сверточную нейронную сеть. Данные модели отлично себя зарекомендовали в задаче анализа тональности текста.

Для оценки качества обучения были выбраны метрики точности и F1-score. В результате обучения XGBoost показал точность на тестовом сете в размере 72.8%, F1 – 71.3%. При этом обучение заняло почти 10 часов. Логистическая регрессия достигла точности 76.7% и F1 в размере 76.9%, при этом затратив на обучение всего лишь 45 секунд. Сверточная нейронная сеть продемонстрировала 79.5% при оценке доле корректных прогнозов и 78.1% на F1 мере и затратила на обучение 6 часов 11 минут. Анализируя полученные данные можно заметить, что сверточная нейронная сеть достигла наибольшего показателя F1 меры, тем не менее логистическая регрессия, показав чуть меньшую точность распознавания, обучилась за гораздо меньшее время. Следовательно, в зависимости от доступного времени и вычислительной мощностью логистическая регрессия может быть более предпочтительна в сравнении со сверточной нейронной сетью. Кроме того, было неожиданно, что классификатор XGBoost показал значительно более низкий результат, чем другие модели, в то время как он продемонстрировал высокую производительность на многих научных соревнованиях. Вероятно, на результате сказываются особенности русского языка и не оптимальный подбор гиперпараметров модели.