УДК 004.9

МЕТОД ОПТИМАЛЬНОЙ РАЗМЕТКИ ТЕКСТОВ ДЛЯ УЛУЧШЕНИЯ МЕТРИКИ КАЧЕСТВА КЛАССИФИКАЦИИ

Бабиков И.А. (Университет ИТМО)

Научный руководитель – к.т.н., рук-ль отдела научно-технологической инфраструктуры, Насонов Д. А.

(Университет ИТМО, Национальный центр когнитивных разработок)

Аннотация

Реализация метода улучшения качества исходной разметки текстовых данных, собранных из социальных сетей, с целью повышения качества классификатора на основе использования различных моделей кластеризации (метод k-средних, иерархическая кластеризация, модель гауссовой смеси). Новая разметка для отдельного кластера – это мода из распределения тегов (лейблов) в исходной разметке, попавших в кластер.

Введение. В настоящее время текстовые данные, собранные из социальных сетей (например, ВКонтакте) для обучения модели с учителем, могут иметь разметку; которая в большинстве случаев является неточной и привести к неудовлетворительным результатам при обучении модели с учителем (классификатор).

Для улучшения качества исходной разметки мы предлагаем метод, основанный на использовании моделей кластеризации, и проводим анализ его эффективности.

Таким образом, имея текстовые данные (предложения в строковом формате) и (частичную) разметку для них, мы хотим получить наиболее оптимальные теги (лейблы) для всех текстов (в частности для тех, на которых изначально их не было).

Основная часть.

Текстовые данные для загрузки в любую модель машинного обучения сначала необходимо преобразовать в текстовые эмбеддинги (представления слов в числовых векторах). Для получения эмбеддингов всех предложений (sentence embedding) используется модель BERT (Bidirectional Encoder Representations from Transformers), которая учитывает контекст слов в предложении. Их можно использовать на вход моделям кластеризации; в данной работе мы сравниваем три модели: метод k-средних, иерархическая кластеризация и модель гауссовой смеси.

Главная идея получения оптимальной разметки: численные представления текстов в одном кластере отражают одну тематику текстов, что позволит улучшить качество классификатора. Зная распределение тегов в одном кластере, мы можем присвоить всем попавшим в данный кластер текстам моду (самое часто встречающееся значение) среди тегов в нем.

На новой оптимальной разметке результаты классификатора, который также использует BERT, близки к современным результатам на текстовых данных с ручной разметкой.

Выводы. В ходе выполнения работы был разработан метод оптимальной разметки текстовых данных на основе кластеризации набора текстов с первоначальной (частичной) разметкой. Предложенная в работе система позволит получать более качественную разметку для текстов, собранных из социальных сетей, для дальнейшего построения модели машинного обучения с учителем (текстовый классификатор на основе модели BERT) с лучшим результатом по выбранной метрике.