УДК 004.048

МЕТОДЫ ВЫДЕЛЕНИЯ РЕКЛАМЫ И УТОЧНЕНИЯ ЕЁ КАТЕГОРИЙ НА ОСНОВЕ АНАЛИЗА НЕСТРУКТУРИРОВАННЫХ ДАННЫХ СОЦИАЛЬНЫХ СЕТЕЙ

Дощенко А.И. (Федеральное государственное автономное образовательное учреждение высшего образования «Национальный исследовательский университет ИТМО»)
Научный руководитель – младший научный сотрудник, аспирант Филатова А.А.
(Федеральное государственное автономное образовательное учреждение высшего образования «Национальный исследовательский университет ИТМО»)

Введение. Социальные сети – это платформа для публикации большого количества разнообразной информации, включая посты обычных пользователей, распространение рекламы или даже различного рода сообщений, затрагивающих политические и социально-значимые темы. Это означает, что социальные сети являются одним из основных способов общения и обмена информацией в окружающем нас мире. Многие компании и частные предприниматели используют их для продвижения своих товаров и услуг. Социальные сети с изображениями и возможностью отмечать геолокацию поста являются особенно популярными, так как они позволяют отслеживать распространение публикаций на карте, видеть и анализировать, где их больше и насколько они эффективны. То же самое можно применить и к рекламным публикациям. Выявление и детальный анализ рекламы по открытым данным социальной сети с изображениями и геолокациями, построение тематической карты предложений и анализ эффективности рекламы в разных частях города с последующим созданием методов рекомендации позволит потенциальным рекламодателям выбирать оптимальную геолокацию для размещения рекламы, а также поможет понять, каких тематических предложений не хватает в интересующем районе города.

Актуальность данной работы заключается в том, что на текущий момент область выделения рекламных публикаций по неразмеченным данным социальных сетей достаточно слабо покрыта исследованиями, а существующие решения покрывают выделение рекламы только в какой-то конкретной области, например реклама медицинских препаратов или незаконных веществ. Большинство работ также чаще всего в своих исследованиях задействуют социальные сети, которые у русскоговорящего населения не популярны. Несмотря на схожесть формата рассматриваемой в данной работе социальной сети и популярных сетей среди исследователей (возможность прикрепить изображение, хэштеги, чаще всего небольшой объём текста), недостатком других социальных сетей является практически полное отсутствие геолокаций, что не позволяет проанализировать размещение публикаций на карте города. В нашей же работе рассматриваются методы для выделения общей рекламы среди других данных, а также определение конкретных тематик рекламных предложений по выделенной рекламе.

Основная часть. Для решения задачи обнаружения распределённых в пространстве и времени рекламных публикаций, а также более узких тематик этих рекламных публикаций, были проанализированы существующие работы в области тематического моделирования, и разработаны методы выявления данного типа публикаций при помощи таких моделей, как BigARTM и Zero Shot learning, который сочетает в момент обучения контекстные эмбеддинги и мешок слов (Bag of Words). В качестве модальности для BigARTM учитывались такие мета-характеристики публикаций, как упоминания, хэштеги, ссылки, сама модель была обучена для выделения рекламных публикаций среди сырых данных социальных сетей.
Эффективное отделение такого рода публикаций является основополагающей частью данной работы, позволяющей в дальнейшем выделять более узкие тематики в рекламе. Для выделения таких тематик были также использованы модели тематического моделирования, а для определения необходимого количества тем была использована такая метрика моделей тематического моделирования, как интерпретируемость темы для человека – показатель

когерентности (coherence score). В дальнейших исследованиях выделенные тематические публикации будут проанализированы на эффективность (основываясь на метаданных публикаций, таких как количество лайков, комментариев, длина поста), а также будет составлена тематическая карта города (при помощи агрегации по количеству публикаций и по локациям).

Выводы. Методы, представленные в работе, позволили повысить эффективность выделения рекламных публикаций и рекламных категорий, основываясь на анализе неструктурированных данных социальных сетей города Санкт-Петербург. В дальнейших исследованиях разработанные методы будут использованы для определения эффективности рекламных публикаций (по метахарактеристикам постов), и разработки методов рекомендаций геолокаций для оптимального размещения рекламы в социальных сетях. Сам же пайплайн планируется применять в социальных сетях для более рационального размещения рекламы в городе, а также для анализа распределения тематик на карте с последующим применением результатов данного анализа в проектах, связанных с городской средой.

Дощенко А.И. (автор) Подпись

Филатова А.А. (научный руководитель) Подпись