0.1. *Желтова К.А.* Применение методов тематического моделирования для классификации пользователей социальных сетей

Современный темп развития социальных сетей делает их уникальным по своей полноте источником информации о людях и их интересах. Анализ социальных сетей предоставляет широкие возможности для решения исследовательских и прикладных задач, однако существенную проблему представляет наличие ложной информации и спама в профилях [1]. Одним из методов решения этой проблемы является дополнение описания профилей пользователей признаками, полученными с помощью анализа групп. В частности, применение методов тематического моделирования к описанию и контенту сообществ в социальных сетях позволяет выявить ключевые слова и латентные темы этих сообществ. Аддитивная регуляризация тематических моделей (АРТМ) — подход к построению тематических моделей, позволяющий комбинировать регуляризаторы, тем самым комбинируя тематические модели [2]. В данной работе АРТМ из библиотеки BigARTM была использована для анализа групп социальной сети «ВКонтакте» и последующей классификации пользователей по таким признакам как «пол» и «наличие детей». Исходная выборка состояла из 7000 профилей пользователей, 34884 профилей групп, а также информации о подписках пользователей на определенные группы. В результате работы тематической модели были получены 36 латентных тем сообществ. После этого профили пользователей были дополнены признаками вида «количество подписок на группы темы N» и «количество лайков в группах темы N». Для построения модели классификации использовалась библиотека градиентного бустинга CatBoost с метрикой качества $F1$-score. В результате качество на валидационной и тестовой выборках для задачи определения родительского статуса пользователя составило 0.923 и 0.902, а для задачи определения пола — 0.762 и 0.795 соответственно.

Список литературы

[1] Коршунов А. и др. Анализ социальных сетей: методы и приложения // Тр. Ин-та системного программирования РАН. 2014. Т. 26. № 1. С. 439–456.

[2] Воронцов К. и др. BigARTM: Библиотека с открытым кодом для тематического моделирования больших текстовых коллекций // Аналитика и управление данными в областях с интенсивным использованием данных. 2015. С. 28–36.