

УДК 004

НОРМАЛИЗАЦИЯ ТЕКСТОВ СОЦИАЛЬНЫХ СЕТЕЙ МЕТОДАМИ МАШИННОГО ОБУЧЕНИЯ

Щелкина О.Т. (ИТМО)

Научный руководитель – кандидат физико-математических наук Рыбин С.В.
(ИТМО)

Введение. Тексты социальных сетей содержат сленг, сокращения, ошибки, эмодзи и код-свитчинг [1], что затрудняет их обработку методами NLP. Существующие алгоритмы часто не адаптированы к неформальному стилю, снижая точность нормализации.

Цель данной работы — исследование проблемы нормализации текстов социальных сетей, существующих подходов решения, а также и разработка метода, основанного на дообучении модели RuT5-Small [2].

Основная часть. Для решения проблемы был проведен анализ текстов социальных сетей, выявлены их основные особенности и сложности нормализации. В качестве базового использовался корпус RuTweetCorp [3], содержащий реальные тексты пользователей. На его основе был подготовлен специализированный обучающий корпус с парами «исходный текст» — «нормализованный текст».

Выбранная модель RuT5-Small была дообучена на подготовленном корпусе. Обучение проводилось в течение 20 эпох с использованием графического процессора. В ходе экспериментов оценивалась точность нормализации, учитывались ошибки модели и возможные улучшения.

Результаты показали, что после дообучения модель достигла точности 80%, что подтверждает ее эффективность. Однако выявлены ограничения, связанные с небольшим размером модели и недостаточной представленностью редких конструкций в обучающих данных. Для повышения качества нормализации предлагается увеличить объем корпуса и рассмотреть использование более мощных языковых моделей.

Выводы. В работе проанализированы методы нормализации текстов социальных сетей, предложен и протестирован метод дообучения модели RuT5-Small. Полученные результаты подтверждают его эффективность, но дальнейшие исследования могут включать расширение корпуса и применение более сложных языковых моделей.

Список использованных источников:

1. Clark E., Araki K. Text normalization in social media: progress, problems and applications for a pre-processing system of casual English // *Procedia-Social and Behavioral Sciences* — Vol. 27. — 2011 — 2-11.
2. Рубцова Ю. Автоматическое построение и анализ корпуса коротких текстов (постов микроблогов) для задачи разработки и тренировки тонового классификатора // *Инженерия знаний и технологии семантического веба*. — 2012. — Т. 1. — С. 109-116.
3. RuT5-Small. — URL: <https://huggingface.co/AI4Bharat/ruT5-small>