Анализ данных с применением нейросетевых языковых моделей в исследовании мнений пользователей социальных сетей

Березин Сергей Андреевич, Антонова Яна Николаевна, Пирожкова Дарья Алексеевна, Молчанова Марина Геннадьевна

Студент

Новосибирский Государственный Университет, Новосибирск, Россия

E-mail: s.berezin@g.nsu.ru, y.antonova@g.nsu.r, d.pirozhkova@g.nsu.ru, mohai3737@gmail.com

Научный руководитель *— Медведев Алексей Николаевич*

Социальные сети занимают неотъемлемую часть жизни каждого из нас и позволяют формировать отношения между пользователями из разных социальных групп. Результат взаимодействий людей в виде историй, комментариев, дискуссий, мнений отображается в постах на таких площадках как Facebook, Twitter, YouTube, LinkedIn, Reddit.

Интернет-площадки все чаще становятся платформой для обсуждений различных явлений и событий. Одновременно с этим, построение моделей динамики мнений является актуальной и нетривиальной задачей на стыке математики, компьютерных наук и социологии. Социологи несколько десятилетий занимаются разработкой математических моделей динамики мнений в социальных сетях. Большие объемы данных (фото в Instagram, покупки на Amazon, посты в Twitter) трудно собирать и обрабатывать вручную. На основе этих данных появляются научные исследования взаимодействия пользователей под постами в комментариях под ними. Однако, математические модели работают с числами, а мнения обычно выражаются словесно.

Традиционно ученые решали эту проблему, используя различные формальные оценки: например, социологические опросы. В данной работе мы предлагаем новый подход к проблеме числового представления мнений, основанный на нейросетевых языковых моделях.

Для апробации предложенного подхода были выбраны сообщества сторонников основных кандидатов президентской гонки в США в 2020 году. Это политическое противостояние отражалось в ощутимой поляризации общества, и, следовательно, выражаемые мне-

ния должны являться хорошо отличимыми друг от друга [1]. Поэтому в качестве данных были взяты посты и комментарии с площадки Reddit за начало 2020 года из двух сообществ: “TheDonald” и “JoeBiden”.

Текстовые комментарии были переведены в многомерные числовые вектора с помощью трех NLP алгоритмов. В исследовании использовались и сравнивались алгоритмы векторизации текстов: Doc2Vec, Sentence-BERT и Universal Sentence Encoder (USE). После этого результаты векторизации каждым алгоритмом оценивались с помощью классификатора (логистическая регрессия).

Для интерпретации и визуализации результатов размерность полученных векторов снижалась до двух и трех измерений с помощью алгоритмов t-SNE и UMAP. В полученных визуализациях отчетливо наблюдались кластеры схожих мнений и разделимость подмножеств сторонников двух политиков.

В результате работы был получен инструмент, позволяющий определять полярность новых высказываний, изучать дисперсию мнений в сообществах и проводить кластерный анализ текстов.

Литература

1. Böttcher L., Gersbach H. The great divide: drivers of polarization in the US public //EPJ data science. – 2020. – Т. 9. – №. 1. – С. 1-13.