

Мясоедова Юлия Витальевна

магистрант, Белгородский государственный национальный

исследовательский университет,

Россия, г. Белгород

АНАЛИЗ РАСПРЕДЕЛЕНИЯ И ВЗАИМОСВЯЗЕЙ ЭМОЦИЙ НА ОСНОВЕ ДАТАСЕТА GOEMOTIONS

АННОТАЦИЯ

В статье рассматривается структура и особенности датасета GoEmotions, предназначенного для анализа эмоциональной окраски текстов на естественном языке. Проведен анализ распределения эмоциональных категорий, выявлен значительный дисбаланс классов, а также исследована частота встречаемости множественных эмоций в одном тексте. Особое внимание уделено изучению корреляций между наиболее распространенными эмоциональными состояниями. Полученные результаты позволяют глубже понять природу эмоциональных выражений в пользовательских комментариях и могут быть использованы при разработке моделей машинного обучения для задач многометочной классификации и анализа тональности.

ABSTRACT

The article examines the structure and features of the GoEmotions dataset, designed to analyze the emotional coloring of texts in natural language. The analysis of the distribution of emotional categories was carried out, a significant class imbalance was revealed, and the frequency of occurrence of multiple emotions in one text was investigated. Special attention is paid to the study of correlations between the most common emotional states. The results obtained allow for a deeper understanding of the nature of emotional expressions in user comments and can be used in the development of machine learning models for multi-metric classification and tonality analysis tasks.

Ключевые слова: эмоциональный анализ, GoEmotions, обработка естественного языка, анализ данных, классификация эмоций, корреляция эмоций.

Keywords: emotional analysis, GoEmotions, natural language processing, data analysis, emotion classification, emotion correlation.

В современную цифровую эпоху объем информации, публикуемой пользователями в социальных сетях и интернет-платформах, стремительно возрастает. Особую ценность представляет анализ не только содержания таких сообщений, но и их эмоциональной составляющей. Эмоции, выражаемые в текстах, играют ключевую роль в понимании психологического состояния автора, реакций аудитории, а также в разработке рекомендательных систем, систем модерации и даже в ранней диагностике психологических расстройств.

Датасет GoEmotions – это англоязычный корпус пользовательских комментариев с платформы Reddit, размеченный по эмоциональной шкале. Он был разработан исследователями из Google Research для обучения и оценки моделей распознавания эмоций в текстах на естественном языке. Данный датасет считается одним из наиболее полных и тщательно размеченных ресурсов в области эмоционального анализа: он включает 58 000 текстов, каждый из которых вручную размечен по одной или нескольким из 27 категорий эмоций, а также по дополнительной категории «нейтрально».

Эмоции в GoEmotions представлены разнообразными оттенками чувств, включая как базовые (радость, грусть, страх, гнев), так и более сложные и социально обусловленные (например, благодарность, смущение, разочарование, гордость, злорадство и др.). Это делает датасет особенно ценным для задач многоклассовой и многометочной классификации. Один текст может содержать несколько эмоций, что отражает реальную сложность и многогранность человеческих высказываний.

Комментарии в датасете были отобраны с учетом качества и нейтральности тематики – были исключены токсичные, оскорбительные и политически заряженные сообщения. Это делает GoEmotions пригодным не только для академических исследований, но и для промышленного применения, например, в разработке рекомендательных систем, чат-ботов и сервисов по мониторингу эмоционального состояния пользователей.

Вся разметка была проведена вручную с привлечением нескольких аннотаторов, после чего разногласия устранялись с использованием специальных процедур. Итоговая структура датасета включает текст комментария, список присвоенных эмоций в виде числовых индексов и вспомогательную информацию (например, принадлежность к обучающей, тестовой или валидационной выборке).

Для понимания структуры выборки необходимо изучить, как часто встречаются те или иные эмоции в исходном наборе данных. Это позволяет выявить наиболее доминирующие и редкие классы. На рисунке 1 представлен анализ распределения эмоций в датасете GoEmotions

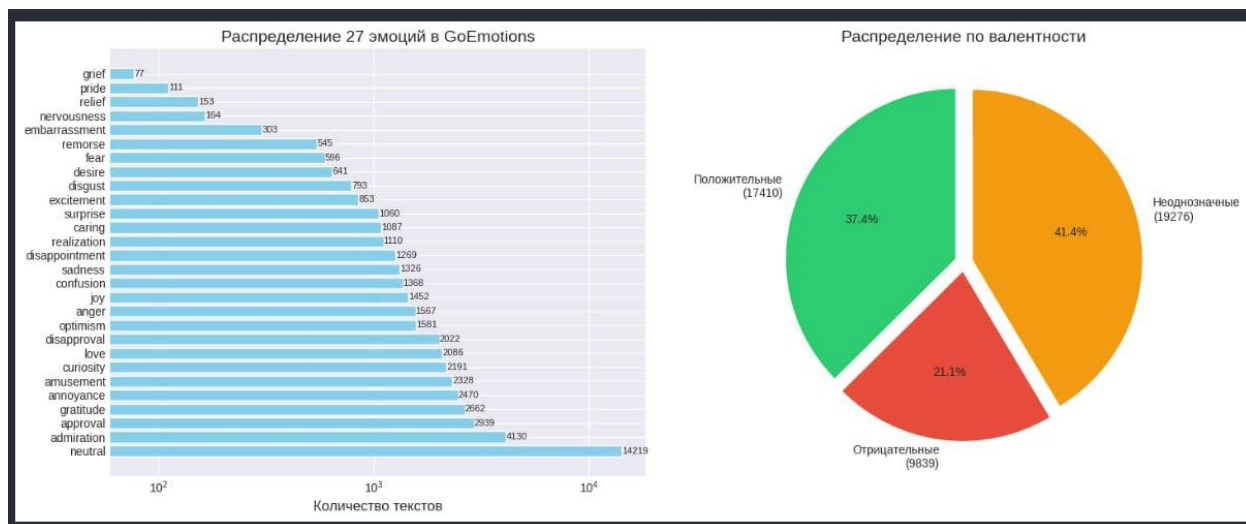


Рисунок 1. Анализ распределения эмоций

Анализ показал значительный дисбаланс классов. Наиболее часто встречающейся категорией является «нейтральная» (14 219 текстов). Среди выраженных эмоций лидируют «восхищение» (admiration) и «одобрение»

(approval), в то время как «горе» (grief) и «гордость» (pride) встречаются крайне редко. При группировке эмоций по валентности выяснилось, что 41.4% текстов имеют неоднозначную окраску, 37.4% – положительную и 21.1% – отрицательную.

Тексты в GoEmotions часто содержат более одной метки. Следующие визуализации демонстрируют, насколько часто встречаются «смешанные» чувства в одном комментарии. Распределение текстов по количеству эмоциональных меток и кумулятивная доля представлены на рисунке 2.

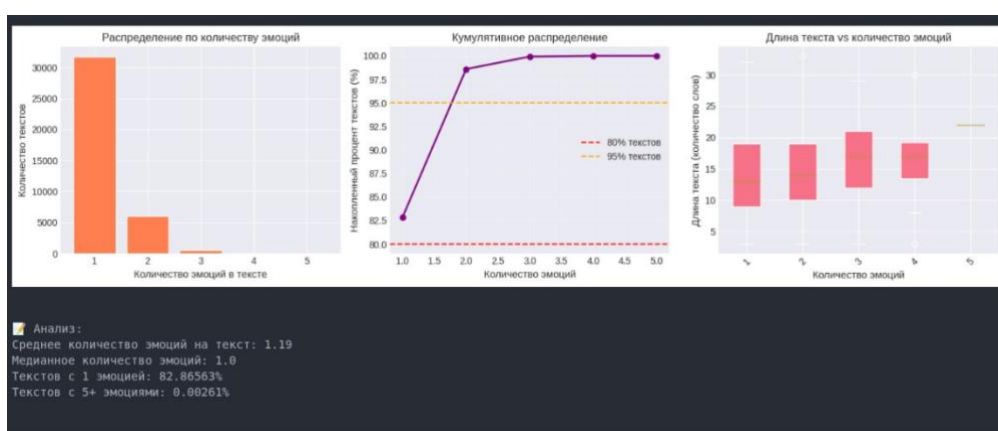


Рисунок 2. Количество эмоций на текст

На рисунке 3 изображена тепловая карта корреляций между 12 наиболее частыми эмоциями. Исследование связей между различными эмоциональными метками помогает понять, какие чувства часто проявляются одновременно. Тепловая карта наглядно демонстрирует силу этих взаимосвязей.



Рисунок 3. Матрица корреляции между эмоциями

Анализ корреляции между 12 наиболее частыми эмоциями показал отсутствие сильных линейных связей, за исключением небольших положительных корреляций между схожими состояниями таких как, «гнев» и «раздражение».

Проведённый анализ датасета GoEmotions показывает его высокую применимость для задач распознавания эмоций в текстах. Несмотря на значительный дисбаланс классов, ресурс отражает реальную многогранность человеческих высказываний, включая смешанные эмоциональные состояния. Отсутствие сильных линейных корреляций между большинством эмоций подтверждает сложность и неоднозначность эмоциональной окраски текстов.

Благодаря качественной ручной разметке, исключению токсичного и политически окрашенного контента, а также чёткому разделению на обучающую, валидационную и тестовую выборки, GoEmotions может служить надёжной основой для создания моделей эмоционального искусственного интеллекта в академических и прикладных областях.

Список литературы:

1. Data Science и интеллектуальный анализ данных : учебное пособие [Текст]/ Д.М. Назаров, С.В. Бегичева, Д. Б. Ковтун, А.Д. Назаров. — Москва : Ай Пи Ар Медиа, 2023. — 304 с.
2. Добдин С.Ю. Основы науки о данных с использованием языка Python (введение в Data Science) : учебное пособие [Текст]/ С.Ю. Добдин, А.В. Скрипаль. — Москва : Ай Пи Ар Медиа, 2024. — 210 с.
3. Татарникова Т.М. Интеллектуальный анализ данных : учебное пособие [Текст]/ Т.М. Татарникова. — Москва, Вологда : Инфра-Инженерия, 2024. — 172 с.