Техническое описание датасета "Toxic Russian Comments"

1. Общая информация

- **Название:** Toxic Russian Comments
- **Источник:** https://www.kaggle.com/datasets/alexandersemiletov/toxic-russian-comments/data
- **Описание:** Датасет содержит размеченные комментарии из русскоязычной социальной сети *ok.ru*. Использовался в соревновании на платформе *All Cups* для автоматической классификации комментариев по уровням токсичности.
  Классы:
  - нейтральные,
  - оскорбительные,
  - угрожающие,
  - содержащие описание или угрозы сексуального насилия.
- **Формат:** текстовый файл в формате *fastText* (каждая строка начинается с метки `__label__<CLASS>`, за которой следует текст комментария).
- **Язык:** русский
- **Лицензия:** CC BY-NC-SA 4.0 (некоммерческое использование, обязательное указание авторства, сохранение лицензии).
- **Ожидаемая частота обновления:** никогда (данные статичны).

---

2. Структура

- **Формат данных:** текстовый файл (`dataset.txt`) в формате *fastText*, где каждая строка = комментарий + метки.
- **Столбцы (fastText-формат):**
  - **Метки (`__label__<CLASS>`):** одна или несколько.
    - `__label__NORMAL` — нейтральный комментарий
    - `__label__INSULT` — оскорбление
    - `__label__THREAT` — угроза
    - `__label__OBSCENITY` — описание/угроза сексуального насилия
  - **Текст комментария:** полный текст на русском языке
- **Пример:**
- `__label__INSULT скотина! что сказать`
- `__label__NORMAL я сегодня проезжала по рабочей...`
- `__label__INSULT,__label__THREAT заколоть этого плешивого урода...`
- **Особенности:**
  - возможны множественные метки (мульти-лейбл классификация)
  - тексты содержат ошибки, сленг, эмоциональную лексику, локальные упоминания

---

3. Объем

- **Количество записей:** 248 290
- **Распределение по меткам:**

    - `__label__NORMAL` — 203 685 (82.0%)
    - `__label__INSULT` — 28 567 (11.5%)
    - `__label__INSULT,__label__THREAT` — 6 317 (2.5%)
    - `__label__THREAT` — 5 460 (2.2%)
    - `__label__OBSCENITY` — 2 245 (0.9%)
    - `__label__INSULT,__label__OBSCENITY` — 1 766 (0.7%)
    - `__label__INSULT,__label__OBSCENITY,__label__THREAT` — 176 (<0.1%)
    - `__label__OBSCENITY,__label__THREAT` — 74 (<0.1%)
- **Размер файла:** 39.27 МБ
- **Дисбаланс классов:** сильный (82% нейтральных комментариев). Требуются методы балансировки (oversampling, undersampling, class weights).

---

4. Качество разметки

- **Метод разметки:** создан для соревнования на All Cups. Подробности (число аннотаторов, критерии) отсутствуют.
- **Достоинства:**
    - чёткие 4 класса
    - удобный формат fastText
    - разметка в целом соответствует содержанию
- **Недостатки:**
    - нет описания процесса аннотации
    - субъективность в INSULT и OBSCENITY
    - мульти-лейблы могут требовать контекста
    - много ошибок, сленга, эмодзи
- **Рекомендации:**
    - выборочная валидация
    - кросс-валидация при обучении моделей

---

5. Применимость для задач классификации

- **Прямая:** классификация токсичности (multi-label). Подходит для задач с бинарной/трёхклассовой схемой.
- **Косвенная:**
    - выявление манипулятивного языка (например, "фейковые вакцины")
    - использование класса NORMAL для отделения нейтральных комментариев от токсичных
    - дообучение языковых моделей (ruBERT, XLM-RoBERTa)
- **Ограничения:**
    - дисбаланс классов (82% NORMAL)
    - нет контекста комментариев
    - требуется предобработка текста
- **Подходящие задачи:**
    - мульти-лейбл классификация
    - бинарная классификация (токсичные vs нетоксичные)
    - transfer learning
- **Технические аспекты:**
    - предобработка (очистка эмодзи, хэштегов, нормализация)

- токенизация/лемматизация (Natasha, pymorphy2)
- модели: ruBERT, XLM-RoBERTa, TF-IDF + SVM
- метрики: F1-score (micro/macro)

Техническое описание датасета "Jigsaw Toxic Comment Classification Challenge"

1. Общая информация

- **Название**: Jigsaw Toxic Comment Classification Challenge
- **Источник**: https://www.kaggle.com/datasets/julian3833/jigsaw-toxic-comment-classification-challenge/data
- **Описание**: Датасет создан для соревнования Jigsaw Toxic Comment Classification Challenge, целью которого является автоматическая классификация комментариев из Википедии по уровням токсичности. Комментарии размечены по шести категориям токсичности, включая нейтральные и оскорбительные классы. Датасет также использовался в других задачах, таких как Jigsaw Rate Severity of Toxic Comments.
- **Формат**: CSV-файл (train.csv).
- **Язык**: Английский.
- **Лицензия**: CC0: Public Domain (открытое использование без ограничений).
- **Размер файла**: 68.8 МБ (для train.csv).
- **Ожидаемая частота обновления**: Не указана (статичный датасет).

2. Структура

- **Формат данных**: CSV-файл с 8 столбцами, где каждый комментарий сопровождается уникальным идентификатором и метками для шести классов токсичности.
- **Столбцы**:
  - id: Уникальный идентификатор комментария (строка).
  - comment_text: Текст комментария на английском языке.
  - toxic: Бинарная метка (0 или 1), указывающая, является ли комментарий токсичным.
  - severe_toxic: Бинарная метка (0 или 1), указывающая на высокую степень токсичности.
  - obscene: Бинарная метка (0 или 1), указывающая на наличие непристойного содержания.
  - threat: Бинарная метка (0 или 1), указывающая на угрозы.
  - insult: Бинарная метка (0 или 1), указывающая на оскорбления.
  - identity_hate: Бинарная метка (0 или 1), указывающая на ненависть, основанную на идентичности (например, расизм, сексизм).
- **Пример данных**:
- `id,comment_text,toxic,severe_toxic,obscene,threat,insult,identity_hate`
- `0000997932d777bf,Explanation Why the edits made under my username Hardcore Metallica Fan were reverted? They weren't ...,0,0,0,0,0,0`
- `0002bcb3da6cb337,COCKSUCKER BEFORE YOU PISS AROUND ON MY WORK,1,1,1,0,1,0`

```
001810bf8c45bf5f,You are gay or antisemmitian? Archangel WHite Tiger
Meow! Greetingshhh! Uh, there are two ways, ...,1,0,1,0,1,1
```

- **Особенности**:
  - Поддерживает мульти-лейбл классификацию: один комментарий может иметь несколько меток (например, toxic=1, insult=1, obscene=1).
  - Тексты содержат орфографические ошибки, сленг, заглавные буквы для акцента (например, "COCKSUCKER") и контекст, связанный с Википедией (например, обсуждение правок).

3. Объем

- **Количество записей**: 159,571 комментариев (в train.csv).
- **Распределение меток** (на основе предоставленных данных):
  - toxic: 15,294 положительных (1) против 144,277 отрицательных (0) (~9.6% токсичных).
  - severe_toxic: 1,595 положительных (1) против 157,976 отрицательных (0) (~1.0% сильно токсичных).
  - obscene: 8,449 положительных (1) против 151,122 отрицательных (0) (~5.3% непристойных).
  - threat: 478 положительных (1) против 159,093 отрицательных (0) (~0.3% угроз).
  - insult: 7,877 положительных (1) против 151,694 отрицательных (0) (~4.9% оскорблений).
  - identity_hate: 1,405 положительных (1) против 158,166 отрицательных (0) (~0.9% ненависти по идентичности).
- **Дисбаланс классов**: Значительный перевес отрицательных меток (0) во всех категориях, особенно для threat (0.3%) и severe_toxic (1.0%). Это требует методов балансировки данных для обучения моделей.

4. Качество разметки

- **Метод разметки**: Датасет создан для соревнования Jigsaw, но детали аннотации (например, количество аннотаторов, критерии) не указаны в предоставленной информации. Вероятно, использовалась разметка экспертами или краудсорсинг.
- **Достоинства**:
  - Четкие бинарные метки для шести категорий, что упрощает задачу мульти-лейбл классификации.
  - Разнообразие комментариев (от нейтральных обсуждений правок до явных оскорблений) отражает реальные сценарии.
  - Примеры показывают логичную разметку (например, "COCKSUCKER..." помечен как toxic=1, severe_toxic=1, obscene=1, insult=1).
- **Недостатки**:
  - Отсутствие документации о процессе разметки снижает прозрачность.
  - Возможна субъективность в определении категорий, таких как insult или identity_hate, из-за культурных и контекстных различий.
  - Мульти-лейбл случаи требуют тщательной обработки, так как метки могут пересекаться (например, toxic и insult часто совпадают).
- **Рекомендации по валидации**: Провести выборочную проверку разметки или использовать кросс-валидацию для оценки согласованности. Дополнительно можно проверить корреляцию между метками (например, toxic и insult).

5. Применимость для задач классификации

- **Соответствие задаче**: Датасет идеально подходит для классификации комментариев как токсичных, оскорбительных или нейтральных, так как содержит разметку, соответствующую этим категориям:
  - **Нейтральные**: Комментарии, у которых все метки равны 0 (примерно 89.7% данных, если считать отсутствие метки toxic).
  - **Оскорбительные**: Комментарии с меткой insult=1 (~4.9%) или комбинации с другими метками (например, toxic=1, obscene=1).
  - **Токсичные**: Комментарии с метками toxic=1, severe_toxic=1, obscene=1, threat=1, или identity_hate=1 (~10.3% данных для toxic).
- **Типы классификации**:
  - **Мульти-лейбл классификация**: Прямая задача, где комментарий может иметь несколько меток (например, toxic=1, insult=1, obscene=1). Подходит для сложных моделей (BERT, RoBERTa).
  - **Бинарная классификация**: Упрощение до двух классов:
    - Нейтральные (все метки = 0).
    - Токсичные (хотя бы одна метка = 1, например, toxic=1 или insult=1).
  - **Трехклассовая классификация**: Разделение на:
    - Нейтральные (все метки = 0).
    - Оскорбительные (insult=1 или комбинации).
    - Токсичные (toxic=1, severe_toxic=1, obscene=1, threat=1, или identity_hate=1).
- **Преимущества для задачи**:
  - Большой объем данных (159,571 записей) обеспечивает достаточный материал для обучения моделей.
  - Разнообразие категорий токсичности (toxic, severe_toxic, insult, и т.д.) позволяет детализировать классификацию.
  - Английский язык упрощает использование с предобученными моделями (например, BERT, DistilBERT).
- **Ограничения**:
  - Значительный дисбаланс классов (например, threat=1 в 0.3% случаев) может привести к переобучению на нейтральных комментариях.
  - Контекст комментариев (обсуждения правок в Википедии) может быть специфичным, что требует осторожности при обобщении на другие платформы.
  - Орфографические ошибки, сленг и заглавные буквы для акцента требуют тщательной предобработки.