Отчет по проекту Comment Bot Detector (YouTube): ключевые признаки, влияющие на классификацию комментариев

22.10.2025

Подготовил:
Зеленов Никита

Цель отчета
Описать группы признаков, которые влияют на решение классификатора, и дать понятное объяснение, что именно является сигналом для модели.

Краткое резюме
- В текущей конфигурации основной классификатор опирается на текст комментария (семантика и лексика).
- В проекте также присутствуют модули для вычисления тональности (VADER) и анализа времени публикации; эти сигналы полезны как дополнительные признаки и для диагностических отчетов.
- Метаданные автора (например, возраст канала и частота комментариев) являются сильными поведенческими индикаторами и могут существенно повысить качество, если их добавить в модель.
- Для портфолио важно явно разделять: (а) признаки, которые реально использует модель сейчас, и (б) признаки, которые анализируются отдельно или могут быть добавлены в будущей версии.

Что считается "признаком" в этом проекте
Под "признаками" понимаются входные сигналы, на основании которых модель принимает решение о классе. В проекте можно выделить две категории:

- Входы, которые непосредственно подаются в модель (например, токены текста).
- Входы, которые рассчитываются/собираются и используются для анализа, проверки гипотез или как кандидаты для расширения модели (например, тональность, время, метаданные автора).

1) Текстовые признаки (основной сигнал)
Поскольку классификатор обучается на тексте комментария, ключевой вклад в решение дают контентные признаки. В трансформерных моделях (BERT и аналоги) эти признаки проявляются через токены, их контекст и устойчивые паттерны в тексте.

Наиболее влияющие текстовые паттерны (типовые для bot/spam)
- Шаблонность и повторяемость формулировок: одинаковые или очень похожие фразы в разных комментариях.
- Призывная лексика и спам-паттерны: подписки, переходы по ссылкам, "проверь мой канал", "напиши в телеграм" и т. п.

- Неестественные конструкции: много однотипных эмодзи/восклицаний, перегруженность ключевыми словами, странная пунктуация.
- Длина комментария и структура: очень короткие односложные похвалы или, наоборот, чрезмерно длинные рекламные тексты.
- Ссылки, упоминания, хэштеги: даже если URL удаляются на этапе очистки, следы таких конструкций и окружающий контекст часто остаются сигналом.

Почему эти признаки работают

Боты часто используют шаблоны (автогенерация или копипаст), поэтому распределение токенов и сочетания слов заметно отличаются от человеческих сообщений. Трансформер улавливает такие различия как на уровне отдельных токенов, так и на уровне контекста.

2) Тональность (sentiment) как дополнительный сигнал

В проекте рассчитывается тональность с помощью VADER (compound score) и дискретная категория тональности. Тональность обычно не является главным признаком сама по себе, но усиливает модель в сочетании с текстовыми паттернами.

- Сверхпозитивные однотипные комментарии ("best video", "amazing") часто коррелируют с накруткой.
- Резкие негативные выпады/провокации могут встречаться в бот-кампаниях и "рейдах".
- Sentiment удобно использовать для отчета: сравнить распределения тональности для разных групп комментариев.

3) Время публикации и поведенческие паттерны

Темпоральные признаки описывают, как именно распределены комментарии во времени. Они полезны для выявления "волн" активности и автоматизированных публикаций.

- Плотность: много комментариев за короткий интервал времени.
- Регулярность: комментарии с одинаковым шагом (признак автоматизации).
- Всплески: резкое увеличение активности по сравнению с базовым уровнем.
- Сдвиги по времени суток/дням: нетипичные для обычной аудитории периоды активности.

4) Метаданные автора (часто самые сильные сигналы)

Метаданные автора могут сильно улучшить качество детекции, потому что описывают не только текст, но и поведение пользователя. Типовые сильные признаки:

- Возраст канала/аккаунта: новые аккаунты при высокой активности часто подозрительны.

- Частота комментариев автора в собранном датасете: необычно высокая активность может быть признаком бота.
- Уникальность/повторяемость текста автора: один и тот же текст под разными видео.
- Комбинации: молодая учетная запись + высокая активность + шаблонные тексты.

Сводная таблица групп признаков

Группа	Примеры признаков	Статус в проекте	Ожидаемая важность
Текстовые (контент)	Слова/фразы, шаблонность, призывы к действию, длина, пунктуация, эмодзи	Да (основной сигнал: BERT по тексту)	Высокое
Тональность (sentiment)	VADER compound и категория (Very Negative ... Very Positive)	Сейчас используется для анализа/отчета; в базовой текстовой модели не обязателен	Среднее
Время публикации	Плотность комментариев, всплески активности, распределение по времени	Сейчас используется для анализа/отчета; в базовой текстовой модели не обязателен	Среднее
Метаданные автора	Возраст канала/аккаунта, частота комментариев автора, идентификатор автора	Собирается в пайплайне сбора; в базовой текстовой модели не обязателен	Высокое (если добавить в модель)
Поведенческие/графовые	Связи автор - видео - каналы, похожесть текста между авторами, кластеры	Не реализовано как отдельный модуль (можно расширять)	Высокое (для продвинутых версий)

Интерпретируемость

Для трансформерных моделей наиболее наглядный способ объяснения - подсветка вкладов токенов и фраз.

- Локальная интерпретация: LIME/SHAP по токенам или по словам (подсветка фраз, которые толкают в класс "bot").
- Диагностика ошибок: примеры FP/FN и анализ, какие паттерны приводят к ошибкам.
- Сравнение распределений: по длине текста, тональности и времени публикации для разных классов.

Ограничения и риски

- Если в предобработке удаляются ссылки/упоминания, модель может терять сильный сигнал "URL spam"; тогда важно компенсировать это контекстом или отдельной бинарной фичей.
- Данные из YouTube API ограничены квотами и могут содержать смещение (например, только выбранные видео).
- Разметка (label) - критический фактор: качество и согласованность разметки напрямую влияют на итоговую модель.
- Метрики должны считаться на отложенной выборке; для честного сравнения фиксируйте сплит, seed и параметры эксперимента.

Приложение: рекомендуемый список признаков для следующей версии

- Бинарная фича: содержит ли исходный текст URL/@/# (до очистки).
- Количество эмодзи, количество восклицательных/вопросительных знаков.
- Длина текста (символы, слова) и доля неалфавитных символов.
- Темпоральные агрегации по автору: комменты/час, комменты/день.
- Возраст аккаунта и активность автора как числовые признаки.