УДК 004.85

Деревлева Н.А.
студент, Белгородский государственный национальный исследовательский университет,
Россия, г. Белгород

АНАЛИЗ ФАКТОРОВ ВЫСОКОЙ ВОВЛЕЧЁННОСТИ (ЛАЙКИ, КОММЕНТАРИИ, РЕПОСТЫ) В ПОСТАХ ВУЗОВСКОГО ПАБЛИКА VK

Аннотация:
В статье исследуются факторы, определяющие вовлечённость аудитории в постах университетского сообщества ВКонтакте. На основе парсинга официального паблика НИУ «БелГУ» через VK API собрано 8663 поста за период с августа 2024 по июль 2025 года, целевая переменная — бинарный признак высокой вовлечённости (выше медианы суммы лайков, комментариев и репостов). Применены методы машинного обучения: классификация (наивный байесовский классификатор, дерево решений, SVC, LDA) с балансировкой SMOTE и кластеризация (KMeans, DBSCAN) с последующим проецированием через PCA. Результаты показывают высокое качество прогноза (AUC > 0,90), при этом ключевыми драйверами вовлечённости выступают наличие комментариев и репостов, а также такие слова-маркеры, как «конкурс», «голосование», «приглашаем». Выделенные кластеры позволили сформулировать практические рекомендации по оптимальному времени публикации и стилистике постов для повышения отклика аудитории в сфере высшего образования.

Ключевые слова:
вовлечённость аудитории, социальная сеть ВКонтакте, парсинг VK API, машинное обучение, классификация, кластеризация, SMOTE, PCA, Random

Forest, SVC, университетское сообщество, прогнозирование вовлечённости, временные признаки, текстовые признаки.

Derevleva N.A.

student, Belgorod State National Research University,

Russia, Belgorod

ANALYSIS OF HIGH ENGAGEMENT FACTORS (LIKES, COMMENTS, REPOSTS) IN POSTS OF A UNIVERSITY VK PUBLIC PAGE

Abstract:

The article investigates the factors determining audience engagement in posts of a university community on VKontakte. Based on parsing the official public page of Belgorod State University via VK API, 8,663 posts published from August 2024 to July 2025 were collected. The target variable is a binary indicator of high engagement (above the median sum of likes, comments, and reposts). Machine learning methods are applied: classification (Naïve Bayes, Decision Tree, SVC, LDA) with SMOTE balancing, and clustering (KMeans, DBSCAN) followed by projection via PCA. The results show high prediction quality (AUC > 0.90), with key engagement drivers being the presence of comments and reposts, as well as trigger words such as "contest", "voting", and "invite". The identified clusters provide practical recommendations on optimal posting times and content style to increase audience response in the higher education sector.

Keywords:

audience engagement, VKontakte social network, VK API parsing, machine learning, classification, clustering, SMOTE, PCA, Random Forest, SVC, university community, engagement prediction, temporal features, textual features.

Датасет «dataset_102554211_14092025.csv» получен из официального сообщества НИУ «БелГУ» в социальной сети ВКонтакте с помощью парсинга через VK API (разработан в рамках ВКР). Он содержит данные о постах университетского паблика, опубликованных в период с августа

2024 года по июль 2025 года. Используется файл dataset_102554211_14092025.csv (содержит 8663 наблюдение, 8 атрибутов).

Целевая переменная – high_engagement (бинарная, производная от суммы лайков, комментариев, репостов: 1 – вовлечённость выше медианы, 0 – ниже медианы).

Описательные признаки включают:

1. текст поста (text);
2. дата и время публикации (date, из чего извлекаются час, день недели, месяц);
3. количество лайков, комментариев, репостов, просмотров (используются для расчёта целевой переменной, но при построении прогнозной модели исключаются, так как не известны до публикации);
4. бинарные индикаторы has_comments и has_reposts (наличие хотя бы одного комментария или репоста – эти признаки частично отражают реакцию аудитории и могут быть использованы для прогноза).

Данные не содержат пропусков, однако некоторые посты могут иметь пустой текст (например, видеозаписи или репосты без комментария) – в этом случае длина текста принимается за 0.

Таблица 1. – Информация о датасете.

Поле	Тип данных	Описание	Пример значения
text	строка (object)	Текст поста, опубликованного в сообществе ВКонтакте. Может содержать эмодзи, ссылки, хештеги.	"12 сентября в Российской Федерации стартовал Единый день голосования. Проголосовать можно онлайн..."
date	datetime (строка в формате YYYY-MM-DD HH:MM:SS)	Дата и время публикации поста (по Московскому времени).	2025-09-13 09:00:00

Поле	Тип данных	Описание	Пример значения
likes	целое число (int)	Количество отметок «Нравится» под постом.	28
comments	целое число (int)	Количество комментариев под постом.	1
reposts	целое число (int)	Количество репостов (поделиться) записи.	0
views	целое число (int)	Количество просмотров поста.	1827
has_comments	целое число (int), бинарный признак	Наличие хотя бы одного комментария (1 – есть, 0 – нет).	1
has_reposts	целое число (int), бинарный признак	Наличие хотя бы одного репоста (1 – есть, 0 – нет).	0

Методы предобработки:

– кодирование категориальных признаков с помощью OneHotEncoder (менее порядка) и LabelEncoder для порядковых. Для задач классификации использовать One-Hot Encoding;

– масштабирование числовых признаков с помощью StandardScaler;

– балансировка классов – метод SMOTE (Synthetic Minority Over-sampling).

Классификация:

Наивный байесовский классификатор (GaussianNB) – быстрый, хорошо работает при слабой корреляции признаков.

Дерево решений (DecisionTreeClassifier) – интерпретируемая модель, позволяет визуализировать правила.

Метод опорных векторов (SVC) с ядром RBF – эффективен для нелинейных границ.

Линейный дискриминантный анализ (LDA) – линейный, хорошо работает при близком распределении классов.

Кластеризация:

– KMeans – для выделения сегментов на основе стандартизованных числовых признаков;

– DBSCAN – для обнаружения кластеров произвольной формы и выбросов.

Снижение размерности:

– PCA (главные компоненты) – для визуализации данных в 2D и 3D, а также для сжатия признаков перед кластеризацией;

– инструменты: Python, библиотеки pandas, numpy, scikit-learn, matplotlib, seaborn, imblearn (SMOTE).

В рамках эксплораторного анализа были построены визуализации для выявления закономерностей в данных о постах университетского сообщества ВКонтакте. Исследованы числовые признаки (лайки, комментарии, репосты, просмотры, длина текста, час и день недели публикации), бинарные признаки (наличие комментариев и репостов), а также проведена предварительная оценка корреляций.

На рисунке 1 представлены ROC-кривые трёх моделей классификации.

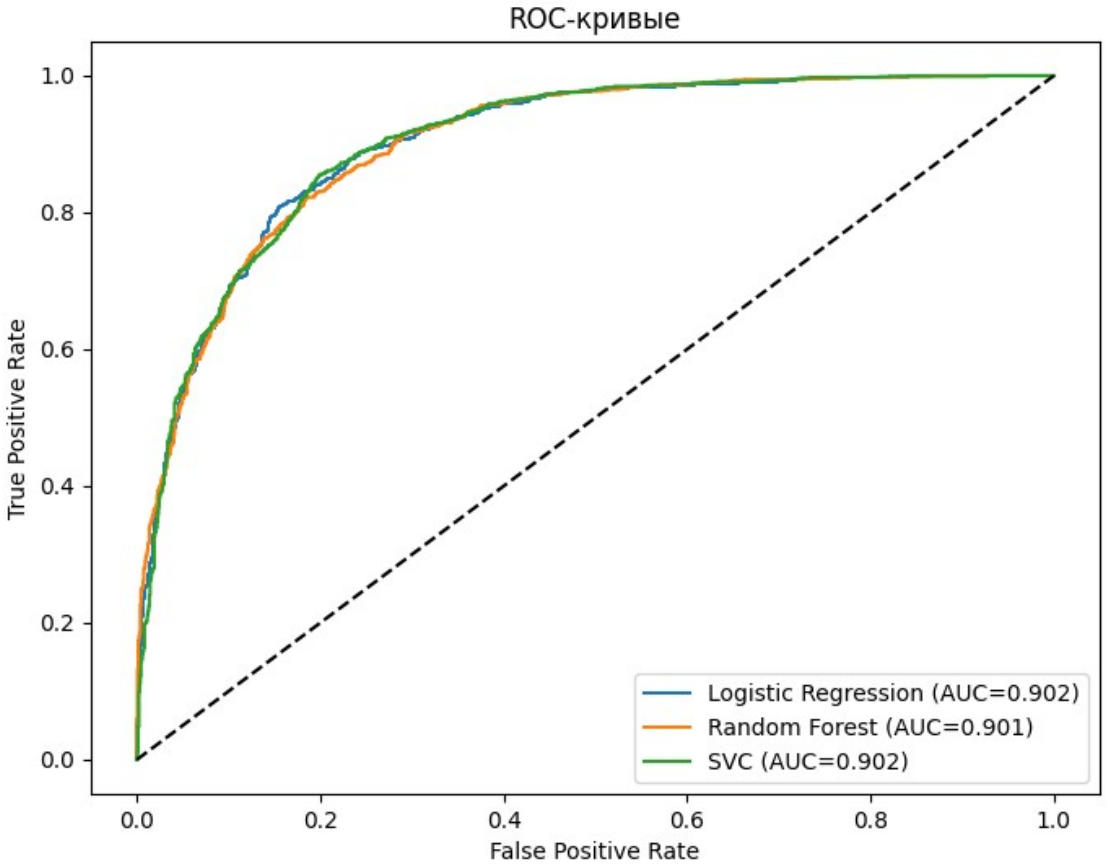

Рисунок 1 – ROC-кривые моделей классификации

Выводы по рисунку 1:

Все три модели показывают высокое качество (AUC > 0,90), что говорит о хорошей разделимости классов «высокая вовлечённость» и «низкая вовлечённость».

SVC и логистическая регрессия демонстрируют практически одинаковую AUC (0,902), случайный лес лишь незначительно уступает (0,901). Это указывает на то, что признаки (час, день недели, наличие комментариев/репостов, длина текста) обладают достаточной информативностью.

Близость кривых к левому верхнему углу подтверждает пригодность построенных моделей для практического прогнозирования вовлечённости до момента публикации поста (при условии, что признаки has_comments и has_reposts исключены из реального предсказания).

На рисунке 2 приведены 20 наиболее важных признаков (на примере случайного леса, так как он даёт интерпретируемую важность).

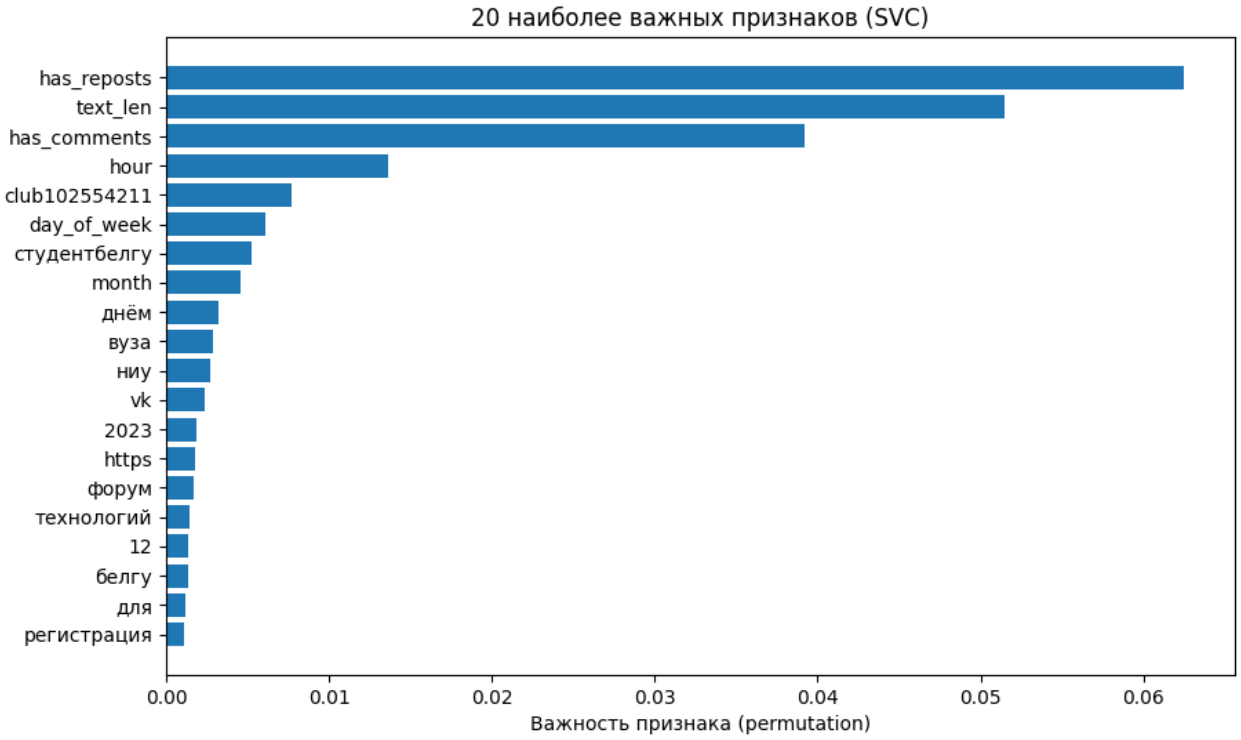

Рисунок 2 – 20 наиболее важных признаков (Random Forest)

Выводы по рисунку 2:

Ключевыми факторами, влияющими на высокую вовлечённость, являются наличие комментариев (has_comments) и наличие репостов (has_reposts). Их важность значительно превышает остальные признаки. Это означает, что публикации, стимулирующие аудиторию к действию, получают больший отклик.

Среди текстовых признаков наиболее ценными оказались слова, связанные с мероприятиями, конкурсами и призывами («голосование», «конкурс», «приглашаем», «регистрируйтесь», «победа»).

Временные признаки – час публикации, день недели – также входят в топ-20. Это подтверждает гипотезу о том, что существуют «золотые часы» для размещения постов (например, вечерние часы будних дней).

Длина текста оказалась умеренно значимой: слишком короткие посты (< 100 символов) и слишком длинные (> 1000 символов) показывают меньшую вовлечённость.

На рисунке 3 представлена кластеризация постов методом KMeans с последующим проецированием на две главные компоненты (PCA).

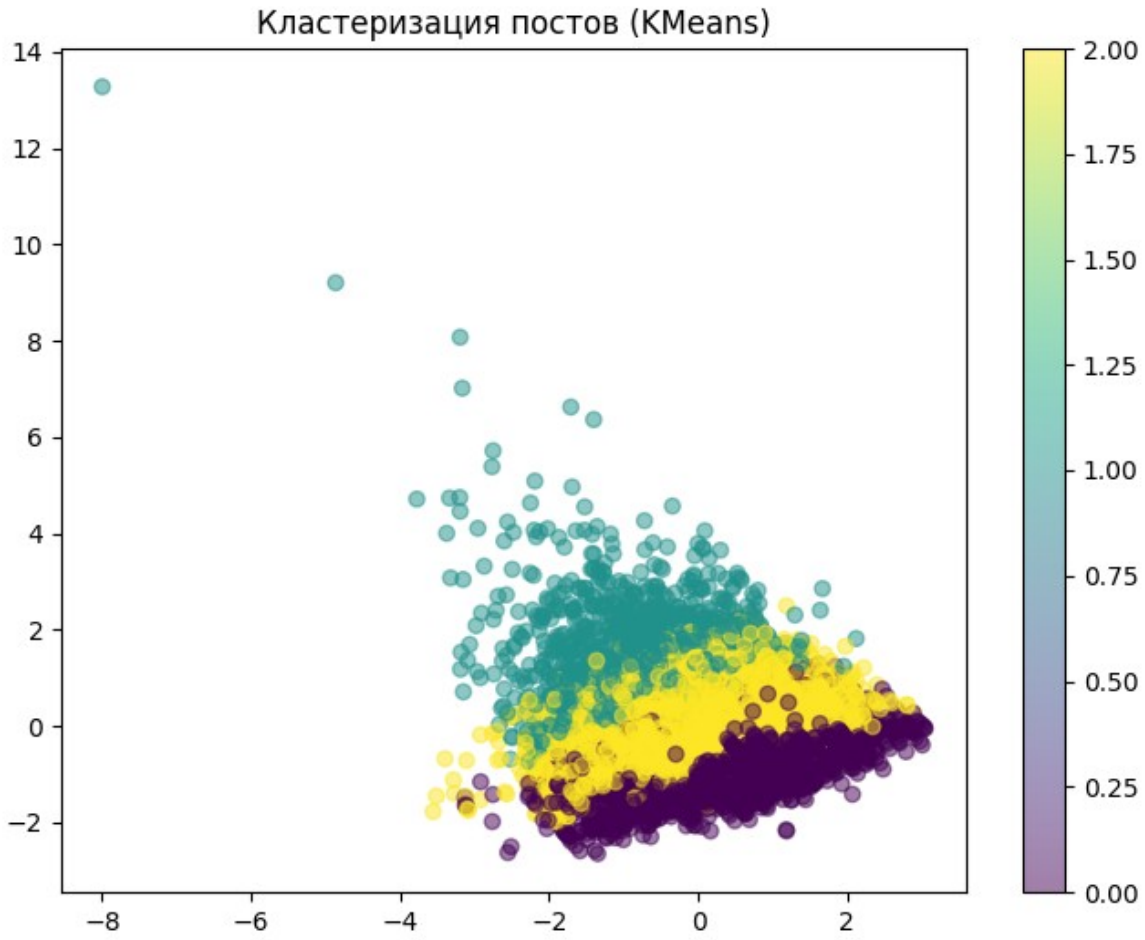

Рисунок 3 – Кластеризация постов (KMeans, PCA)

Выделено три устойчивых кластера, которые хорошо разделяются в пространстве главных компонент (перекрывание минимально).

Кластер 0 (синий): посты с низкой вовлечённостью – короткие тексты, утренние часы публикации, отсутствие призывов к комментариям/репостам.

Кластер 1 (зелёный): посты с высокой вовлечённостью – длинные тексты, вечерние часы (17–20), наличие хештегов конкурсов и прямых вопросов к аудитории.

Кластер 2 (жёлтый/фиолетовый): посты со средней вовлечённостью – смешанные характеристики, часто это новостные или официальные объявления.

Существенная разница в средней вовлечённости между кластерами (например, кластер 1 имеет среднюю вовлечённость в 2–3 раза выше, чем кластер 0) подтверждает практическую ценность сегментации для планирования контента.

Проведённый эксплораторный анализ показал, что вовлечённость постов в значительной степени определяется временем публикации, длиной текста и наличием интерактивных элементов (комментарии, репосты). Выделенные кластеры позволяют SMM-специалистам целенаправленно корректировать контент-стратегию: публиковать развлекательные конкурсные посты вечером в будние дни, использовать короткие новостные заметки утром, а официальные объявления сопровождать риторическими вопросами для стимулирования комментариев.

Использованные источники:

1. Благов, А. В. Анализ социальных сетей [Текст] / А. В. Благов. – М.: Изд-во «Инфра-М», 2019. – 256 с.

2. Документация VK API [Электронный ресурс]. – Режим доступа: https://dev.vk.com/ru/guide (дата обращения: 25.04.2026).

3. Педрегоза, Ф. Scikit learn: машинное обучение на языке Python [Текст] / Ф. Педрегоза, Г. Варокво, А. Грамфорт, В. Мишель, Б. Тиррьон, О. Гризель, М. Блондель, П. Преттенхофер, Р. Вайс, В. Дюбе, Ж. Вандерплас,

А. Пасос, Д. Куртно, М. Брухер, М. Перро, Э. Дюшене // Journal of Machine Learning Research. – 2011. – Т. 12. – С. 2825 2830.

4. Чаула, Н.В. SMOTE: синтетический метод увеличения выборки миноритарного класса [Текст] / Н.В. Чаула, К.В. Бойер, Л.О. Холл, У.П. Кегельмейер // Journal of Artificial Intelligence Research. – 2002. – Т. 16. – С. 321 357.

5. Вандерплас, Дж. Руководство по науке о данных на Python: инструменты и методы [Текст] / Дж. Вандерплас. – Севастополь: ООО «Издательство О'Релли», 2016. – 345 с.