УДК 004.912:004.85
DOI 10.17513/snt.40651

ВЫЯВЛЕНИЕ ТЕКСТОВЫХ ДРАЙВЕРОВ ВОВЛЕЧЕННОСТИ АУДИТОРИИ БРЕНДА В СОЦИАЛЬНЫХ МЕДИА НА ОСНОВЕ ГРАФО-РЕГУЛЯРИЗОВАННОГО ПОДХОДА

Родионов Д.Г. ORCID ID 0000-0002-1254-0464,
Поляков П.А. ORCID ID 0009-0008-6227-3625,
Конников Е.А. ORCID ID 0000-0002-4685-8569,
Старченкова О.Д. ORCID ID 0009-0009-1168-2362

Федеральное государственное автономное образовательное учреждение высшего образования «Санкт-Петербургский политехнический университет Петра Великого», Санкт-Петербург, Российская Федерация, e-mail: contact@polytech-invest.ru

В условиях экспоненциального роста объёмов текстовых данных в социальных сетях задача определения причинно-следственного влияния семантических конструкций на вовлечённость аудитории становится особенно актуальной. Целью работы является разработка комплексного метода анализа такого влияния, сочетающего метод частичных наименьших квадратов для снижения размерности данных и графовую регуляризацию для учёта семантической близости фраз. В рамках исследования предложена модель, которая на этапе обработки данных строит граф семантической близости на основе векторных представлений фраз и применяет комбинированную регуляризацию для стабилизации оценок. Метод был апробирован на корпусе постов бренда Nissan. Результаты показали, что предложенный подход существенно превосходит альтернативные методы по точности прогнозирования и позволяет выделить устойчивый набор фраз, оказывающих значимое влияние на вовлечённость. Выявлено, что метод корректно оценивает вклад отдельных выражений, устраняя смещения, характерные для наивных статистических подходов, и эффективно разделяет эффект контекста и эффект конкретных формулировок. В заключение отмечается, что разработанный метод обеспечивает интерпретируемый и устойчивый каузальный вывод, что делает его практическим инструментом для поддержки принятия решений в контент-маркетинге.

Ключевые слова: социальные медиа, вовлечённость пользователей, частичные наименьшие квадраты, графовая регуляризация, каузальный анализ текста

IDENTIFYING TEXTUAL DRIVERS OF BRAND AUDIENCE ENGAGEMENT IN SOCIAL MEDIA BASED ON A GRAPH-REGULARIZED APPROACH

Rodionov D.G. ORCID ID 0000-0002-1254-0464,
Polyakov P.A. ORCID ID 0009-0008-6227-3625,
Konnikov E.A. ORCID ID 0000-0002-4685-8569,
Starchenkova O.D. ORCID ID 0009-0009-1168-2362

Federal State Autonomous Educational Institution of Higher Education "Peter the Great St. Petersburg Polytechnic University", St. Petersburg, Russian Federation, e-mail: contact@polytech-invest.ru

Amid the explosive growth of textual content in social media, it becomes critically important to identify the causal impact of post characteristics on audience response rather than relying solely on correlations. Classical correlation tests poorly account for context and confounding factors, while deep neural networks often remain a “black box,” sacrificing interpretability. The authors propose a graph-regularized model that estimates the effect of semantic phrases on user engagement. First, textual features are compressed using partial least squares, after which semantic-graph regularization is introduced to smooth effect estimates across semantically similar phrases and reduce variance. Using a dataset from the Nissan brand, the method is shown to outperform alternatives in predictive accuracy and to produce a stable list of influential key phrases with confidence intervals. The approach provides interpretable estimates of the uplift of individual expressions, suppressing contextual “noise” and correcting inflated effects produced by naive methods, thereby revealing actionable engagement triggers to support marketing decisions. The graph is constructed based on cosine similarity of phrase embeddings; estimation is performed via cross-validation, and intervals are obtained via bootstrapping. Sensitivity analysis shows that the ranking of phrases remains stable under changes in the time period, increasing confidence in the findings and practical suitability for A/B testing.

Keywords: social media, user engagement, partial least squares, graph regularization, causal text analysis

Введение

Социальные медиа стали доминирующим каналом коммуникации между брендами и широкой аудиторией. Эффективность этой коммуникации измеряется вовлечённостью – метрикой, отражающей активную реакцию пользователей. Возникает сложная обратная задача, заключающаяся в определении текстовых компонент поста, которые вызывают рост вовлечённости. Актуаль-

ность вопроса обусловлена ограничениями существующих методов. Классические инструменты анализа текста развивались в парадигме тематического моделирования и классификации контента, тогда как выявление причинно-следственных эффектов текста на внешнюю реакцию изучено недостаточно. С одной стороны, попытки оценить влияние отдельных слов и фраз простыми статистическими тестами страдают конфаундингом – фраза может коррелировать с откликом не благодаря своей семантике, а вследствие присутствия в постах определённой популярной тематики. С другой стороны, в задачах прогнозирования популярности контента сегодня доминируют сложные нейросетевые модели, которые обеспечивают высокую точность ценой потери прозрачности. Существует потребность в новых методах, способных работать с высокоразмерными разреженными текстовыми данными, учитывать семантический контекст и при этом давать статистически корректные и интерпретируемые оценки влияния языковых конструкций на метрики вовлечённости.

Ряд недавних работ сфокусирован на том, какие характеристики текста связаны с реакцией аудитории. Например, тон и структура призывов к действию могут существенно влиять на отклик. Показано, что CSR-посты компаний, побуждающие аудиторию к участию в игровых акциях или программах, собирают больше лайков и репостов, тогда как избыточные призывы к обсуждению или одновременное использование нескольких разных призывов снижают вовлечённость [1]. Исследование Gkikas et al. выявило, что читаемость и объём текста, а также число хэштегов статистически значимо связаны с повышением пользовательской активности [2]. Легко читаемые и достаточно длинные описания (более 30 слов, >320 символов) с большим количеством меток показывают более высокий уровень вовлечённости аудитории. Анализ миллионов сообщений в X (бывший Twitter) также подтверждает решающую роль содержимого текста. Так, сравнительное исследование Toraman et al. (2022) показало, что семантика твита является основным драйвером вовлечённости, тогда как идентичность или популярность автора играет меньшую роль [3]. Помимо читаемости и семантики, важным фактором выступает эмоционально-смысловая окраска контента. Например, Saquete et al. (2022) применили анализ мнений и ассоциативных правил для выявления паттернов вирусного распространения сообщений и объяснения, почему одни посты становятся популярнее других [4]. Авторы показали, что определённые сочетания сентимента и тематики значительно повышают «виральность» контента. Существенное влияние оказывает и формулировка заголовков и текстов анонса. Даже при контроле темы и автора разные варианты заголовка заметно влияют на успех поста в Reddit [5]. В совокупности, предыдущие исследования подтверждают. Текстовые особенности публикаций оказывают измеримое влияние на вовлечённость аудитории.

Для количественного предсказания отклика широко используются модели машинного обучения – от регрессий до глубоких нейронных сетей [6]. Так, на данных конкурсов RecSys показано, что предобученные языковые модели способны довольно точно предсказывать метрики реакции по тексту поста [7]. Развиваются и гибридные подходы. Например, предложены графовые нейросети, учитывающие взаимосвязи пользователей при прогнозировании вовлечённости в X (бывший Twitter), а также быстрые сверточные модели, оптимизированные под соревнования рекомендаций контента [8, 9]. Вместе с тем в социальной информатике набирают популярность и каузальные подходы, нацеленные на выявление причинно-следственных связей. В период пандемии COVID-19 предпринимались попытки применить байесовские сети для отсеивания ложных корреляций и идентификации факторов, действительно влияющих на активность пользователей в соцсетях [10]. В сфере маркетинга появились и новые методы целенаправленно для текстовых данных. Lemaire et al. предложили фреймворк на основе эмбеддингов и инструментов причинного вывода, который изолирует вклад отдельных слов, контролируя фоновые переменные [11].

Структурные недостатки «черных ящиков» удаётся решить за счёт использования статистических моделей с латентными переменными. Перспективным инструментом зарекомендовал себя метод частичных наименьших квадратов (Partial Least Squares, PLS) [12]. В задачах с большим числом коррелированных признаков PLS позволяет одновременно выполнить снижение размерности и регрессионный анализ, максимально сохраняя связь «признаки – отклик». Он был успешно применён в различных областях – от хемометрии до анализа поведения пользователей соцсетей. В частности, Yang et al. (2021) интегрировали тематическое моделирование текстов с PLS-SEM для изучения экологических настроений в соцмедиа и подтвердили эффективность PLS-методов для выявления структурных

причинно-следственных связей в данных социального мониторинга [13]. В последние годы развиваются расширения классического PLS, сохраняющие его интерпретируемость на новых типах данных. Например, Vicente-Gonzalez et al. (2025) предложили бинарный PLS (BPLSR) для случаев категориального отклика, дополнив его наглядной визуализацией «триплот», что позволило интерпретировать взаимосвязи между наборами бинарных переменных по обе стороны модели [14]. Однако прямое применение линейных моделей (включая PLS) к текстовым признакам сталкивается с проблемой потери лексической структуры. Отдельные слова и фразы не независимы, а образуют группы синонимов и близких выражений. Стандартные методы регуляризации в таких случаях произвольно выбирают один из коррелированных признаков, обнуляя остальные, – что противоречит лингвистической интуиции и снижает устойчивость модели. Для решения этой проблемы в аналитике данных все шире применяется графовая регуляризация, вводящая априорные связи между признаками. Идея состоит в построении графа, где узлы – признаки (фразы), а ребра соединяют семантически сходные выражения [15]. Добавление в функционал регрессии штрафа за разрывы между соседями по графу сглаживает коэффициенты модели. Такой подход уже реализован, например, в задачах факторизации и кластеризации [16]. Регуляризация по графу позволяет учесть внутренние сходства в данных и за счет этого повысить устойчивость выделяемых факторов. В данном исследовании графовая регуляризация применяется впервые в сочетании с PLS для задач текстовой регрессии, что, по сути, встраивает знание о семантической близости слов в модель влияния контента [17].

Цель исследования – разработать и экспериментально верифицировать графо-регуляризованную PLS-модель, которая даёт статистически устойчивые и интерпретируемые оценки инкрементального прироста вовлечённости аудитории в публикацию, снижая смещения, вносимые контекстом, и тем самым обеспечивает прикладной инструмент для выявления формулировок-драйверов реакции аудитории и поддержки решений в контент-маркетинге.

Материалы и методы исследования

Рассматривается датасет из N социальных медиапостов одной тематики. Каждый пост имеет текст и числовой показатель вовлечённости – например, число комментариев. Обозначим через $X \in R^{N\times m}$ матрицу признаков, где $x_{ij} = 1$, если в j-м посте использована i-я фраза из словаря, и $x_{ij} = 0$ иначе. Вектор $y \in R^N$ содержит значение метрики вовлечённости для каждого поста (y_i – количество комментариев к j-му посту). Требуется построить модель $f: X \mapsto y$, которая позволяет предсказывать уровень отклика по содержимому поста и даёт интерпретируемые оценки вклада отдельных фраз, то есть выявляет фразы-драйверы вовлечённости и количественно оценивает их uplift – прирост отклика при присутствии фразы.

Метод частичных наименьших квадратов выполняет проекцию исходных признаков X в пространство латентных компонентов с одновременной оптимизацией их прогностической значимости для целевой переменной y. В данной работе используется PLS2 – вариант, позволяющий моделировать многомерный отклик. Алгоритм PLS итеративно извлекает набор скрытых компонентов $t_k = X^T w_k$ – линейных сочетаний исходных признаков, – которые максимизируют ковариацию с откликом:

$$t_k = \arg\max_{|w|=1} Cov\left(X^T w, y\right).$$

Компоненты вычисляются последовательно с ортогонализацией по предыдущим. Итоговая модель представляет собой регрессию y на d извлечённых компонент:

$$y \approx \beta_0 + \sum_{k=1}^{d} c_k t_k .$$

Вектор $c = (c_1,\ldots,c_d)$ определяется методом наименьших квадратов. Восстановление коэффициентов при исходных признаках происходит посредством разложения $X = TP^T + E$ (где $T = (t_1,\ldots,t_d)$, aP – матрица нагрузок), после чего оценка влияния i-го признака вычисляется как

$$b_i = \sum_{k=1}^{d} w_{ik} c_k .$$

Эти коэффициенты $b \in R^m$ представляют оценочные вклады каждой фразы в отклик. На малых выборках PLS обладает преимуществом перед обычной регрессией и даже RIDGE/LASSO, избегая проблемы мультиколлинеарности. За счёт ограничения пространства несколькими компонентами $d \ll m$ метод устойчиво оценивает влияния даже при $m \gg N$. Однако без дополнительной регуляризации PLS-модель будет выбирать из группы коррелированных фраз одну произвольную, присваивая другим нулевые коэффициенты. Для текстовых данных это означает, что синонимичные или схожие по смыслу выражения могут получить силь-

но различающиеся оценки b_i. Чтобы учесть априорные связи между текстовыми признаками, в модель вводится графовая регуляризация. Составляется ненаправленный граф семантической близости $G = (V,E)$, где вершины V соответствуют уникальным фразам. Ребро $(i,j) \in E$ проводится между двумя фразами, если они близки по смыслу. Для количественной оценки семантической близости используется косинусное сходство между векторными представлениями фраз. Если e_i – эмбеддинг фразы i, то вес ребра задаётся как w_j при превышении заданного порога. Полученный взвешенный граф отражает структуры синонимичных и тематически связанных выражений в корпусе. Регуляризация по графу вводится в функционал оптимизации модели в виде штрафа на разрыв значений коэффициентов соседних вершин. Для вектора регрессионных коэффициентов b добавляется пенальти вида формула (1):

$$\Omega(b) = \frac{\lambda}{2} \sum_{(i,j)\in E} w_{ij} \left(b_i - b_j\right)^2 = \lambda b^T L b. \quad (1)$$

В формуле (1) L – лапласиан графа G, $\lambda > 0$ – коэффициент регуляризации. Данный член штрафует ситуацию, когда две семантически близкие фразы имеют сильно различающиеся оценки влияния. Минимизация $b^T L b$ эквивалентна требованию гладкости распределения эффектов на графе. Модель по возможности будет присваивать схожие коэффициенты синонимичным выражениям. Это позволяет «заимствовать силу» между редкими и частотными синонимами. Даже если какая-то фраза встречается редко и её индивидуальный эффект статистически незначим, но у неё есть более частотные синонимичные соседи с уверенно положительным влиянием, регуляризация подтянет оценку редкой фразы вверх. Предлагаемый Graph-Regularized PLS сочетает описанные элементы. На первом этапе выполняется PLS. Вычисляются матрица компонент T и веса $W = (w_{ik})_{m\times d}$. Затем решается задача регрессии с графовой регуляризацией в пространстве исходных признаков. Эквивалентно можно рассматривать, что сразу получили оценки b_i из решения оптимизационной задачи (2):

$$\hat{b} = \arg\min_b \ \| y - X^T b \|_2^2 + \lambda b^T L b \ \ s.t. \ \ b \in Span(W), \quad (2)$$

где условие $b \in \mathrm{Span}(W)$ означает, что вектор коэффициентов лежит в пространстве, порождённом колонками W (то есть учитываются только компоненты, извлечённые PLS). Фактически, GR-PLS добавляет к PLS-регрессии квадратичный штраф на разности $b_i - b_j$ для связанных вершин графа. Задача является выпуклой и решается через систему нормальных уравнений с поправкой Лапласиана: $(XX^T + \lambda L)b = Xy$. В данном экспериментальном прототипе подбор оптимального λ осуществлялся по критерию минимизации ошибки предсказания на контрольной подвыборке. В качестве количественной меры влияния текстовой фразы на вовлечённость используется её uplift – относительный прирост целевой метрики при использовании данной фразы. Если $\hat{y}_j$ – предсказанное моделью значение отклика для поста j, то uplift фразы i определим как процентное изменение $\hat{y}_j$ при добавлении фразы i в текст (3):

$$\mathrm{U_i} = \frac{\hat{y}_{(x_{ij}=1)} - \hat{y}_{(x_{ij}=0)}}{\hat{y}_{(x_{ij}=0)}} \times 100\%, \quad (3)$$

где $\hat{y}(xij = 0)$ – прогноз модели для того же поста, но с обнулённым признаком i. В линейной модели это упрощается.

$$\hat{y} = \beta_0 + \sum_k b_k x_k,$$

поэтому $U_i \approx \frac{b_i}{y_{\text{base}}} \times 100\%$.

В GR-PLS расчёт uplift основан на регуляризованных коэффициентах $\hat{b}_i$. Таким образом, положительный U_i означает, что присутствие фразы i в тексте статистически увеличивает ожидаемую вовлечённость на U_i процентов против среднего уровня, отрицательный – снижает.

Результаты исследования и их обсуждение

Несмотря на то что эмпирическая проверка выполнена на данных одного бренда, такой дизайн апробации является методологически оправданным и достаточным для демонстрации работоспособности GR-PLS в прикладных условиях. Во-первых, фокус на одном бренде фиксирует аудиторию, тон коммуникации и стратегию публикаций, тем самым снижая межбрендовую гетерогенность и позволяет проверять именно целевую способность метода отделять вклад конкретных формулировок от эффекта контекста внутри однородного коммуни-

кационного потока. Во-вторых, выбранный корпус воспроизводит ключевые реальные сложности задачи высокой размерности и разреженности матрицы фраз, сильной коррелированности и синонимичности выражений, а также наличия контекстных смешивающих факторов, из-за чего требуется прохождение всей технологической цепочки метода от извлечения фраз и построения семантического графа до подбора регуляризации и получения интерпретируемых оценок эффектов с доверительными интервалами. В-третьих, устойчивость результатов дополнительно контролируется процедурно через валидацию на отложенных данных или кросс-валидацию, бутстрап-интервалы и анализ чувствительности по временным подвыборкам, что повышает доверие к выявленным драйверам и демонстрирует практическую применимость подхода для поддержки решений в контент-маркетинге при выборе формулировок и планировании A/B-проверок. Забегая вперёд, отметим, что экспериментальная апробация также проводилась на данных других брендов, где были получены аналогичные воспроизводимые результаты, согласующиеся с выводами, представленными ниже.

Для апробации метода использовались данные официальной страницы автоконцерна Nissan в социальной сети. Из выгрузки выбрана однородная выборка из 200 постов бренда, каждый с текстовым описанием и числом пользовательских комментариев, как основной метрикой вовлечённости. Тексты подвергнуты очистке. Словарь уникальных значимых фраз составил $m = 784$, матрица признаков X разреженная (плотность ~1.6%). Граф семантической близости фраз построен на основе дистрибутивных эмбеддингов. Использована модель Word2Vec по корпусу брендовых постов, для каждой фразы вычислен вектор как среднее слов, далее для каждой пары фраз с косинусным сходством >0.7 проведено ребро. Полученный граф имел 784 вершин и 4 690 рёбер. Параметры GR-PLS. Число латентных компонент $d = 10$, коэффициент регуляризации графа оптимизирован по отложенной выборке (лучший $\lambda \approx 0.1$).

Таблица 1 содержит качество предсказания вовлечённости для предлагаемого метода и базовой модели без учета графа. В качестве базы выбрана PLS-регрессия без регуляризации (с тем же $d = 10$). Видно, что включение графовой регуляризации существенно повышает объясняющую способность модели: R^2 увеличивается с ~0.17 до ~0.31 на тестовых данных. Также значительно снизилась среднеквадратичная ошибка (RMSE). Это свидетельствует, что учет семантических связей между фразами позволяет модели устойчивее выявлять истинные эффекты и лучше обобщать на новые наблюдения.

Главное преимущество GR-PLS – способность выявлять конкретные текстовые драйверы вовлечённости. Также стоит отметить, что графовая регуляризация явно сгладила оценки внутри семантических групп. Близкие по теме фразы получили сопоставимые веса. Например, кластер фраз, связанных с переходом на сайт, во всех случаях получил положительные коэффициенты $b_i \approx 0.25–0.3$. Это отличается от разрозненных результатов, которые показывала модель без графа. Аналогично, синонимичные призывы к участию в конкурсе сгруппировались и все получили очень высокий положительный вес.

Таблица 2 показывает топ-5 фраз с максимальным положительным uplift. Приведены также 95%-доверительные интервалы, полученные бутстрапированием выборки. Абсолютным лидером стала фраза «попасть в следующий пост» – т.е. призыв к пользователям участвовать в создании следующей публикации бренда. Согласно модели, наличие такого призыва повышает ожидаемое число комментариев на +367% по сравнению со средним уровнем (при $p < 0.01$). Этот результат отражает механику конкурсных активностей. Аудитория активно откликается, когда бренд обещает отметить или упомянуть лучших комментаторов в следующем посте. Высокий uplift (+136%) показали фразы «подробности на официальном сайте» и схожие обращения к переходу на сайт. На первый взгляд, это нетривиальный инсайт. Считается, что вставка внешней ссылки снижает вовлечённость в соцсети, отвлекая пользователя. Однако для автомобильного бренда обнаружено обратное. Аудитория, заинтересованная подробностями (спецификациями, ценами), напротив, более активно комментирует такие посты. Вероятно, детальный контент стимулирует обсуждение. Ещё одна группа – технические характеристики продукта. Фраза «система полного привода» стабильно даёт +65% комментариев. Это подтверждает, что целевую аудиторию Nissan сильно интересуют технические особенности (в данном случае – проходимость внедорожников), и посты с упором на эти свойства вызывают дополнительный отклик. Отметим, что для всех вышеперечисленных факторов PLS-оценки без графовой регуляризации были бы менее значимыми из-за мультиколлинеарности. Метод GR-PLS же «перенёс» значимость от родственных фраз, выдав более надёжные совокупные оценки эффектов.

Таблица 1

Качество моделей предсказания комментариев по тексту поста

Модель	R^2 на тестовой выборке	RMSE на тестовой выборке
PLS (10 компонент)	0,170	5,21
GR-PLS (10 компонент, графовая регуляризация)	0,310	4,47

Примечание: составлена авторами на основе полученных данных в ходе исследования.

Таблица 2

Топ-5 фраз-драйверов вовлечённости по оценке GR-PLS (данные Nissan)

Фраза (лемматизировано)	Uplift, %	95% ДИ, нижняя граница	95% ДИ, верхняя граница	Частота, n
«попасть в следующий пост»	+367,1	+80,6	+1108,0	17
«подробности на офиц. сайте»	+136,0	+39,7	+298,6	24
«прямая ссылка (http…)»	+110,6	−4,9	+366,6	10
«официальный дилерский центр»	+107,5	−3,4	+345,7	11
«система полного привода»	+65,2	+15,1	+148,3	23

Примечание: составлена авторами на основе полученных данных в ходе исследования.

Из таблицы видно, что помимо конкурсного механика («попасть в пост») и призывов к переходу на сайт, значимый эффект дает акцент на технических преимуществах продукта. Для маркетологов автомобильной отрасли это ценное указание. Подчёркивание конкретных характеристик (например, полноприводной системы) действительно стимулирует обсуждения больше, чем общие рекламные слоганы. Интересно, что некоторые фразы с высоким сырым эффектом при наивном анализе потеряли значимость после учета контекста. К примеру, выражение «благодарим за фото подписчика» в простой выборке ассоциировалось с резким ростом комментариев (+201% в t-тесте), но модель GR-PLS присвоила ей нулевой коэффициент. Причина выяснилась при изучении данных. Такие фразы встречались преимущественно в постах с пользовательскими фотографиями, которые сами по себе собирают много комментариев. Наивный анализ приписал весь эффект слову «благодарим», тогда как авторский метод корректно перераспределил эффект на фактор UGC-контента. Благодаря графовой регуляризации схожие благодарственные фразы тоже не были ошибочно отмечены как «магические» триггеры. Таким образом, GR-PLS успешно устраняет ложные драйверы, возникающие из-за спутанности признаков с темой поста. Полученные результаты подтверждают, что предложенная методика позволяет выявлять интерпретируемые текстовые детерминанты вовлечённости. В отличие от «чёрных ящиков», модель даёт маркетологам понятные рекомендации – на какие формулировки делать упор при подготовке контента, чтобы повысить отклик аудитории.

Заключение

В работе представлен новый подход GR-PLS – графо-регуляризованная регрессия на основе частичных наименьших квадратов – для анализа влияния семантических компонентов текста на вовлечённость в социальных медиа. Метод сочетает достоинства PLS, а именно устойчивость при $m \gg N$, выделение информативных латентных факторов с учётом семантических связей между фразами через графовую регуляризацию. Благодаря этому достигается более точное и интерпретируемое ранжирование текстовых триггеров вовлечённости. Схожие по смыслу фразы получают сглаженные коэффициенты, исключаются случайные всплески за счёт контекста.

На примере данных бренда Nissan показано, что GR-PLS существенно превосходит классические методы по качеству (R^2 повышается в ~1.8 раза) и выявляет нетривиальные инсайты. В частности, обнаружено, что конкурсные призывы, побуждение к изучению деталей на сайте, а также подчёркивание технических характеристик продукта являются сильными драйверами комментариев (+65–367% к среднему), тогда как вежливые благодарности или общие маркетинговые фразы сами по себе не увеличивают

активность аудитории. Эти выводы согласуются с интуицией и дают конкретные рекомендации для SMM-стратегии.

Список литературы

1. Chae M.-J. Driving Consumer Engagement through Diverse Calls to Action in Corporate Social Responsibility Messages on Social Media // Sustainability. 2021. Vol. 13. № 7. Art. 3812. DOI: 10.3390/su13073812.

2. Gkikas D.C., Tzafilkou K., Theodoridis P.K., Garmpis A., Gkikas M.C. How Do Text Characteristics Impact User Engagement in Social Media Posts? Modeling Content Readability, Length, and Hashtags Number in Facebook // International Journal of Information Management Data Insights. 2022. Vol. 2. № 1. Art. 100067. DOI: 10.1016/j.jjimei.2022.100067.

3. Toraman Ç., Şahinuç F., Yılmaz E.H., Akkaya I.B. Understanding Social Engagements: A Comparative Analysis of User and Text Features in Twitter // Social Network Analysis and Mining. 2022. Vol. 12. Art. 47. DOI: 10.1007/s13278-022-00872-1.

4. Saquete E., Zubcoff J.J., Gutiérrez Y., Martínez-Barco P., Fernández J. Why Are Some Social-Media Contents More Popular than Others? Opinion and Association Rules Mining Applied to Virality Patterns Discovery // Expert Systems with Applications. 2022. Vol. 197. Art. 116676. DOI: 10.1016/j.eswa.2022.116676.

5. Weissburg E., Kumar A., Dhillon P.S. Judging a Book by Its Cover: Predicting the Marginal Impact of Title on Reddit Post Popularity // Proceedings of the International AAAI Conference on Web and Social Media. 2022. Vol. 16. № 1. P. 1098–1108. DOI: 10.1609/icwsm.v16i1.19361.

6. Aldous K. K., An J., Jansen B. J. What Really Matters? Characterising and Predicting User Engagement of News Postings Using Multiple Platforms, Sentiments and Topics // Behaviour and Information Technology. 2023. Vol. 42. № 5. P. 545–568. DOI: 10.1080/0144929X.2022.2030798.

7. Volkovs M., Cheng Z., Ravaut M., Yang H., Shen K., Zhou J.P. Predicting Twitter Engagement with Deep Language Models // Proceedings of the Recommender Systems Challenge 2020 (RecSys Challenge ’20). New York: ACM, 2020. P. 38–43. DOI: 10.1145/3415959.3416000.

8. Arazzi M., Cotogni M., Nocera A., Virgili L. Predicting Tweet Engagement with Graph Neural Networks // Proceedings of the 2023 International Conference on Multimedia Retrieval (ICMR ’23). New York: ACM, 2023. P. 172–180. DOI: 10.1145/3591106.3592294.

9. Daniluk M., Dąbrowski J., Rychalska B., Gołuchowski K. Synerise at RecSys 2021: Twitter User Engagement Prediction with a Fast Neural Model // RecSysChallenge ’21: Proceedings of the Recommender Systems Challenge 2021. New York: ACM, 2021. P. 15–21. DOI: 10.1145/3487572.3487599.

10. Gencoglu O., Gruber M. Causal Modeling of Twitter Activity during COVID-19 // Computation. 2020. Vol. 8. № 4. Art. 85. DOI: 10.3390/computation8040085.

11. Lemaire A., Yin M., Netzer O. Words That Matter: Analyzing the Causal Effect of Words. SSRN. 2025. URL: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5205681 (дата обращения: 17.12.2025). DOI: 10.2139/ssrn.5205681.

12. Schuberth F., Zaza S., Henseler J. Partial Least Squares Is an Estimator for Structural Equation Models: A Comment on Evermann and Rönkkö (2021) // Communications of the Association for Information Systems. 2023. Vol. 52. P. 711–729. DOI: 10.17705/1CAIS.05232.

13. Yang C.-L., Huang C.-Y., Hsiao Y.-H. Using Social Media Mining and PLS-SEM to Examine the Causal Relationship between Public Environmental Concerns and Adaptation Strategies // International Journal of Environmental Research and Public Health. 2021. Vol. 18. № 10. Art. 5270. DOI: 10.3390/ijerph18105270.

14. Vicente-Gonzalez L., Frutos-Bernal E., Vicente-Villardon J.L. Partial Least Squares Regression for Binary Data and Its Biplot Representation // Mathematics. 2025. Vol. 13. № 3. Art. 458. DOI: 10.3390/math13030458.

15. Родионов Д.Г., Мугутдинов Р.М., Конников Е.А. Автоматизированный алгоритм системного анализа конкурентоспособности цифрового предприятия в рамках информационной среды // Экономические науки. 2021. № 200. С. 98-108. DOI: 10.14451/1.200.98. EDN: RRFFSY.

16. Liu Y., Wu J., Zhang J., Leung M.-F. Graph-Regularized Orthogonal Non-Negative Matrix Factorization with Itakura–Saito (IS) Divergence for Fault Detection // Mathematics. 2025. Vol. 13. № 15. Art. 2343. DOI: 10.3390/math13152343.

17. Барсков В.В., Белостоцкая А.А., Забелин Б.Ф., Конников Е.А. Актуальные вопросы производственного менеджмента в практической деятельности промышленного предприятия. Казань: Бук, 2017. 104 с. ISBN: 978-5-906954-04-6. EDN: YNVLWX.

Конфликт интересов: Авторы заявляют об отсутствии конфликта интересов.

Conflict of interest: The authors declare that there is no conflict of interest.

Финансирование: Работы выполнены в рамках реализации проекта «Разработка методологии формирования инструментальной базы анализа и моделирования пространственного социально-экономического развития систем в условиях цифровизации с опорой на внутренние резервы» (FSEG-2023-0008).

Financing: The work was carried out as part of the project “Development of a methodology for forming an instrumental base for analyzing and modeling the spatial socio-economic development of systems in the context of digitalization based on internal reserves” (FSEG-2023-0008).