

Прогнозирование высоковолатильных временных рядов социальных трендов и общественных интересов

Егор Валерьевич Задворнов

Московский физико-технический институт

Научный руководитель: А. С. Малков

2025

Цель исследования

Цель

- ▶ Разработать framework для:
 - ▶ кластеризации тем по семантической близости с минимальным шумом;
 - ▶ прогнозирования популярности кластеров на горизонте до 180 дней с минимизацией MAE и MSE.

Задачи

- ▶ Сравнить эффективность Prophet, ARIMA и Holt–Winters для разных паттернов.
- ▶ Оптимизировать число кластеров через метрику когерентности C_{cv} .

Уникальный датасет

- ▶ Объединены данные Twitter (топ-15 трендовых тем) и Google (топ-15 поисковых запросов) за 2019–2024 гг.
- ▶ Объём: 76 140 наблюдений. Примеры: «NFL», «Costco Travel», «Дрейк More Life».
- ▶ Многомерность медиаландшафта: социальные сети + поисковые системы.

Гибридный подход

- ▶ Комбинация тематического моделирования (NLP) и декомпозиции временных рядов.
- ▶ Применён для данных с высокой семантической неоднородностью и волатильностью.

Постановка задачи

$T = \{T^k\}_{k=1}^K$, $K = 27\,375$ — множество уникальных топиков.

На каждый день t_i задан набор популярных тем $\{T_{ij}\}_{j=1}^{M_i}$, $M_i = 30$.

Цель: построить модель, прогнозирующую

$y_m(t_i + h)$ — популярность кластера c_m в момент $t_i + h$

с минимизацией

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|, \quad \text{MSE} = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2.$$

Ключевая гипотеза: семантические кластеры учитывают уникальные жизненные циклы тем и повышают точность прогноза.

Визуализация данных

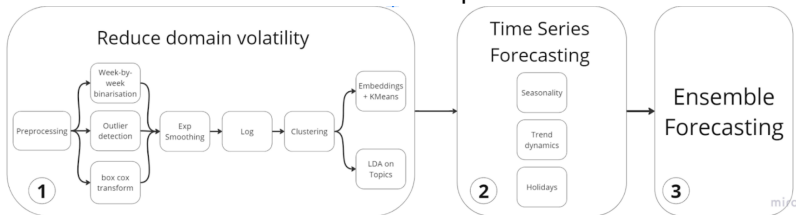
Time	Source	Topic
2024-03-03	Twitter	Rashford
2024-03-03	Twitter	#sundayvibes
...
2017-03-18	Google	Drake More Life
2017-03-18	Google	Drake More Life Download
2017-03-18	Google	Costco Travel

Таблица: Sample of data

Особенности данных:

- ▶ Цикличность: пики «NFL» в сезон (сентябрь–февраль).
- ▶ Аномалии: резкие всплески в кластерах.

Ключевые элементы решения:



Вычислительный эксперимент

1. Кластеризация T на n семантических кластеров

$c_m = \{T^k\}_{k=1}^{n_m}$, $m = 1, \dots, n$, где n_m – число топигов в кластере c_m .

$$n = \arg \min_{L \in \mathbb{N}} \frac{1}{L} \sum_{m=0, \dots, L-1} C_{cv}(c_m) \quad (1)$$

Алгоритмы:

- ▶ Embeddings (all-MiniLM-L12-v2, 384-меры) + K-Means:

$$\arg \min_{\{c_m\}} \sum_{m=1}^n \sum_{v_k \in c_m} \|v_k - \mu_m\|^2.$$

- ▶ LDA с регуляризацией α, β : $p(\theta, \phi, z, w | \alpha, \beta) =$

$$\prod_{k=1}^K p(\phi_k | \beta) \prod_{d=1}^D p(\theta_d | \alpha) \prod_{n=1}^{N_d} p(z_{dn} | \theta_d) p(w_{dn} | \phi_{z_{dn}}).$$

Вычислительный эксперимент

2. Для каждого кластера c_m строится временной ряд $\{t_i, y_m^i\}_{i=1}^N$, где y_m^i – число появлений топики из кластера c_m в день t_i .

Prophet

$$y(t) = g(t) + s(t) + h(t) + \epsilon_t \quad (2)$$

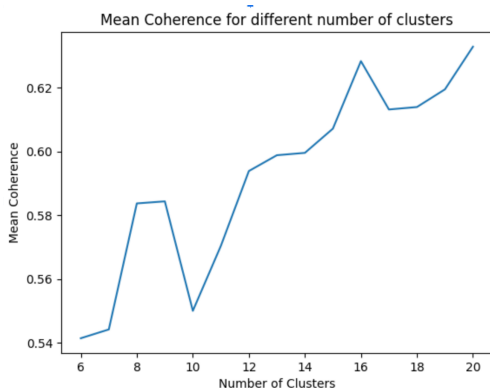
SARIMA

$$(1 - \phi_1 B)(1 - \Phi_1 B^m)(1 - B)^d(1 - B^m)^D y_t = (1 + \theta_1 B)(1 + \Theta_1 B^m) \epsilon_t \quad (3)$$

Метод Хольта-Винтерса (Exponential Smoothing)

$$\hat{y}_{t+h|t} = l_t + hb_t + s_{t+h-m(k+1)} \quad (4)$$

Результаты кластеризации



$$C_{cv} = \frac{1}{|S_{one}|} \sum_{(W_0, W_*) \in S_{one}} \tilde{m}_{\cos(nlr, 1)}(W_0, W_*).$$

Вывод: Оптимальное число кластеров $n = 16$ (максимальное $C_{cv} = 0.75$).

Результаты кластеризации

Метод	Средний C_{cv}	Время вычислений
Embeddings + K-Means	0.63	1 час
LDA	0.69	5 часов

Таблица: Сравнение методов кластеризации

Вывод: выбран K-Means как компромисс скорости и интерпретируемости.

Кластер	Тематика	C_{cv}
2	Мотивационные хештеги	0.75
0	NFL	0.63
...
5	Basketball	0.37

Таблица: Интерпретация ключевых кластеров

Выводы:

- ▶ «Мотивационные хештеги» — высокая семантическая согласованность.
- ▶ «Basketball» качество хуже бейзлайна - нуждается в улучшении фильтрации шума (намечено в Future Work).

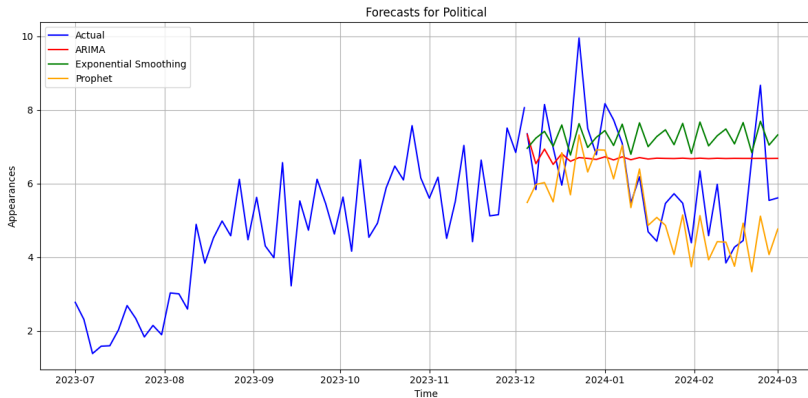
Результаты прогнозирования

Model	Average MAE	Average MSE
Prophet	0.251	0.101
ARIMA	0.454	0.322
Exponential Smoothing	0.422	0.293

Выводы:

- ▶ Обнаружены аномалии в данных (внезапные пики, квазипериодические паттерны), требующие алгоритмов обнаружения аномалий (намечено в Future Work).
- ▶ Prophet снижает MSE на 35% для сезонных кластеров (например, "American football") по сравнению с бейзлайном благодаря адаптивным рядам Фурье.
- ▶ ARIMA и Экспоненциальное сглаживание fails при наличии трендовых разрывов.

Результаты прогнозирования



Теория анализа высоковолатильных временных рядов

▶ Гибридный подход:

- ▶ Обоснована комбинация топик-моделирования и прогнозирования временных рядов для анализа высокодинамичных медиатрендов.

▶ Семантические кластеры:

- ▶ Доказано: разделение данных на кластеры с когерентностью до 0.75 учитывает уникальные жизненные циклы тем и улучшает результаты прогнозирования.

▶ Критика традиционных моделей:

- ▶ ARIMA и экспоненциальное сглаживание демонстрируют высокую ошибку ($MAE > 0.4$) на высоковолатильных временных рядах социальных трендов.
- ▶ Неспособность моделировать квазипериодические паттерны и шоки ограничивает их применимость.

Перспективные направления и заключение

- ▶ Разложение ошибки прогноза на вклад кластеризации и модели
- ▶ Адаптивная глубина тем по стабильности прогноза
- ▶ Регуляризация LDA по прогнозной ошибке (LDA-MSE)
- ▶ Весовая функция MSE для разлодок прогнозирования:

$$\text{MSE}_{\text{mod}} = \frac{1}{n} \sum w(t_i)(y_i - \hat{y}_i)^2, \quad w(t_i) = \begin{cases} 2, & \text{аномалия} \\ 1, & \text{иначе} \end{cases}$$

Заключение

- ▶ Framework: снижение MSE на 35% на новом датасете
- ▶ Патент на генерацию контента + грант ФСИ на Future Work