Прогнозирование высоковолатильных временных рядов социальных трендов и общественных интересов

Егор Валерьевич Задворнов

Московский физико-технический институт
Научный руководитель: А.С. Малков, к.ф.-м.н.

2025

Цели исследования

Основная цель

Прогнозирование популярности тем в медиа-ландшафте при высокой волатильности временных рядов.

- Разработать методы кластеризации топиков общественных интересов.
- Сравнить модели **Prophet**, **ARIMA** и **Exponential Smoothing** по метрикам MAE и MSE.

Научная постановка задачи и гипотеза

Проблема

Социальные тренды характеризуются высокой волатильностью, сменой интересов и редкими, но значимыми пиками.

Гипотеза

Кластеризация тематического пространства позволяет:

- стабилизировать поведение временных рядов,
- повысить интерпретируемость моделей,
- улучшить точность прогноза за счёт перехода от отдельных топиков к обобщённым интересам.

Постановка задачи

Пусть $T = \{T^k\}_{k=1}^K$ — множество уникальных тем, $\{t_i\}_{i=1}^N$ — временные метки, на каждой t_i есть $\{T_{ij}\}_{j=1}^{M_i}$. Цель — предсказать для каждого кластера c_m ряд $y_m^i = |\{j : T_{ij} \in c_m\}|$ в будущие моменты.

Оптимизация числа кластеров:

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

где C_{cv} — когерентность: $C_{cv} = \frac{1}{|S|} \sum_{(W_0, W_*) \in S} \tilde{m}_{\cos}(W_0, W_*)$.

Научная новизна

- Предложен гибридный подход **кластеризации** + **прогнозирования**, в котором оптимизация тематического пространства осуществляется с учётом прогностических целей.
- Критерий выбора кластеризации учитывает не только когерентность, но и устойчивость предсказаний.
- Обоснована применимость Prophet как базовой модели для волатильных рядов с учётом сезонности, трендов и выбросов.

Теоретическое обоснование Prophet

Prophet: $y(t) = g(t) + s(t) + h(t) + \epsilon_t$

- $g(t)$ — тренд (линейный/логистический), позволяет описывать рост/падение интереса.
- $s(t)$ — сезонность, выявляется с помощью ряда Фурье.
- $h(t)$ — эффекты праздников и редких событий.
- ϵ_t — шум, моделирует отклонения.

Подходит для задач с резкими пиками, нелинейностью и сменой режима, в отличие от ARIMA и HW.

Ограничения традиционных подходов

- **ARIMA** требует стационарности, плохо работает с внезапными изменениями.
- **Exponential Smoothing** предполагает устойчивость трендов, не справляется с нестабильными циклами.

Методы кластеризации

- **LDA**: классическая тематическая модель
- **Embedding + K-Means**: эмбеддинги Sentence-BERT

Метрики качества прогноза

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n} |Y_i - \hat{Y}_i|, \quad \mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2.$$

Данные

- Период: 2019-01-01 — 2024-03-03.
- Источники: Twitter и Google, топ-15 тем в день (76140 наблюдений).
- Целевая переменная: число упоминаний тем кластера.

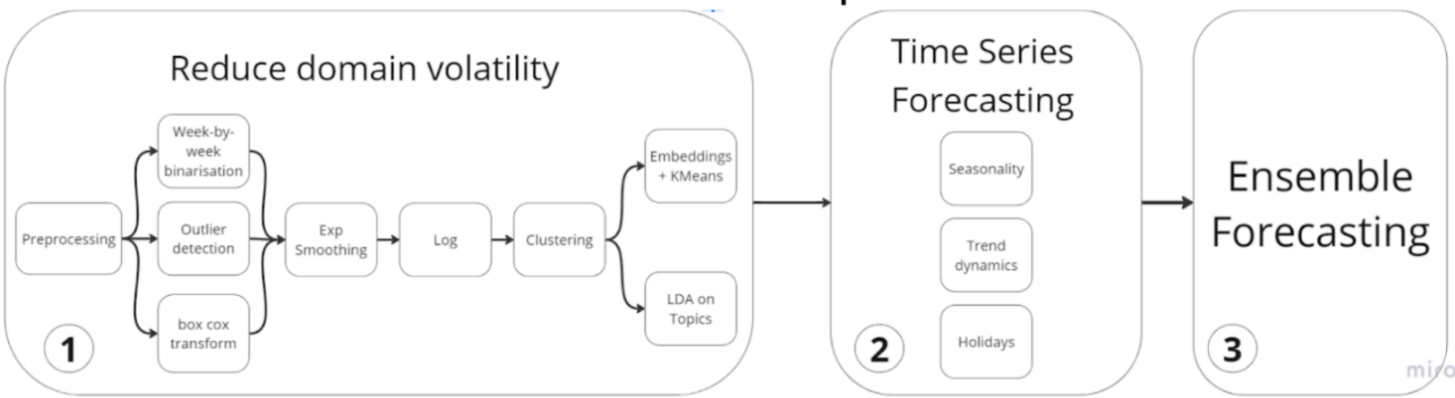
Визуализация данных

Time	Source	Topic
2024-03-03	Twitter	Rashford
2024-03-03	Twitter	#sundayvibes
2024-03-03	Twitter	Xavier Worthy
2024-03-03	Twitter	Foden
2024-03-03	Twitter	#UFCVegas87
...	...	...
2017-03-18	Google	Robert Osborne
2017-03-18	Google	Alejandra Campoverdi
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

Решение

Ключевые элементы решения:



Вычислительный эксперимент

1. Выполняется кластеризация множества T на n семантических кластеров $c_m = \{T^k\}_{k=1}^{n_m}$, $m = 1, \ldots, n$, где n_m – число топиков в кластере c_m.
Для определения оптимального числа кластеров n используется средняя мера когерентности C_{cv}, которая оценивает интерпретируемость кластеров человеком путем измерения семантической близости между словами внутри кластера:

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

$$C_{\text{cv}} = \frac{1}{|S_{\text{set}}^{\text{one}}|} \sum_{(W_0, W_*) \in S_{\text{set}}^{\text{one}}} \tilde{m}_{\cos(\text{nlr},1)}(W_0, W_*) \tag{2}$$

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 топиков из кластера c_m в день t_i.

Prophet

$$y(t) = g(t) + s(t) + h(t) + \epsilon_t \tag{3}$$

ARIMA

$$(1-\phi_1 B)(1-\Phi_1 B^m)(1-B)^d(1-B^m)^D y_t = (1+\theta_1 B)(1+\Theta_1 B^m)\epsilon_t \tag{4}$$

Метод Хольта-Винтерса (Exponential Smoothing)

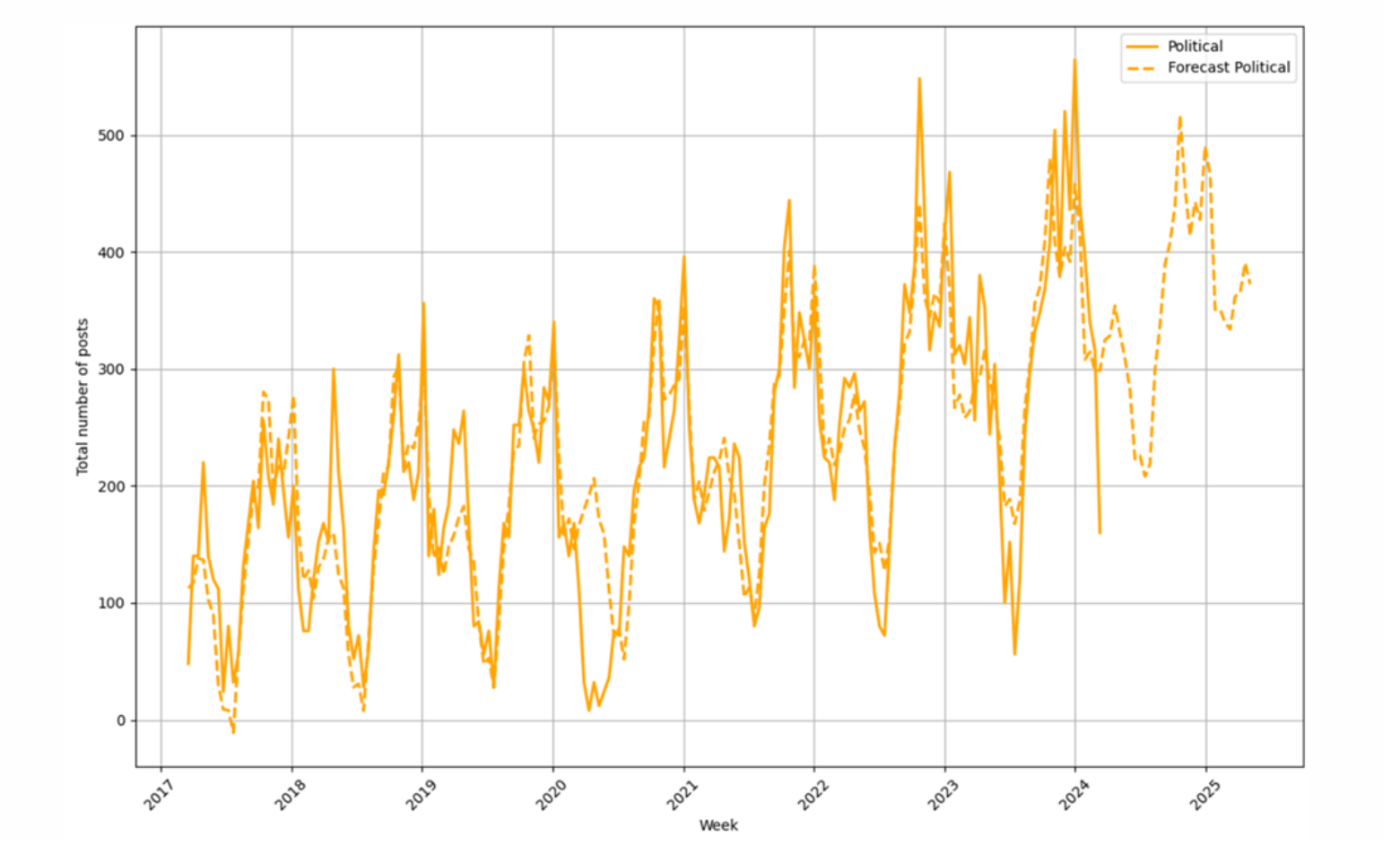
$$\hat{y}_{t+h|t} = l_t + hb_t + s_{t+h-m(k+1)} \tag{5}$$

Результаты кластеризации

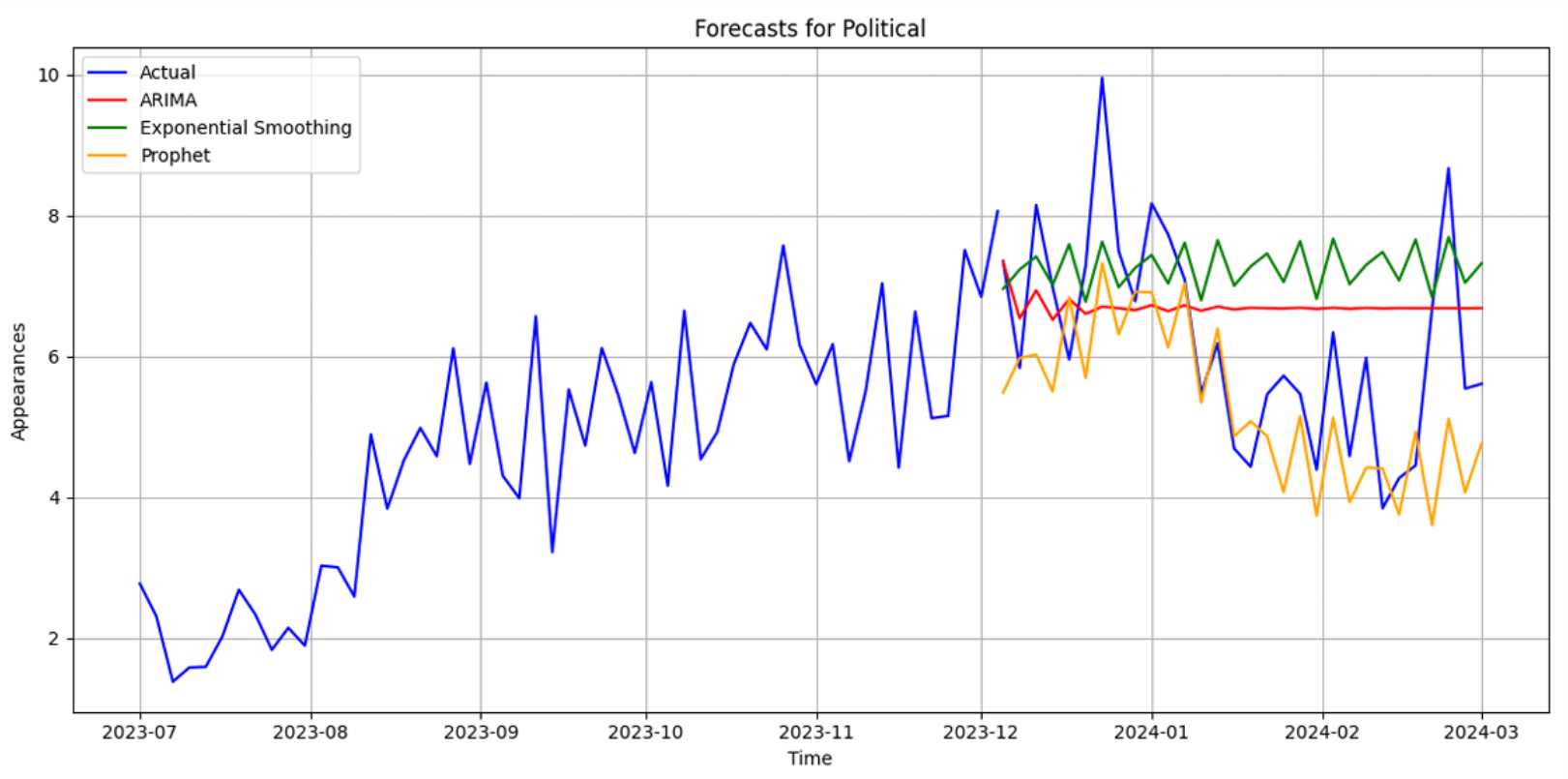
Модель	Средняя C_{cv}	Макс. C_{cv}
Embedding+K-Means	0.63	0.69
LDA	0.69	0.71

Результаты прогнозирования

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



Результаты прогнозирования



Заключение

- Гибридная схема «кластеризация + прогноз» повышает точность.
- Prophet лучше справляется с волатильностью.
- Необходима доработка методов детекции аномалий при резких пиках.

Будущие направления исследования

- **Регуляризованные тематические модели**, обучаемые с учётом предсказуемости.
- **Модели прогнозирования аномалий**, включая Prophet с custom loss или NeuralProphet.
- **Внешние переменные**. Модель может быть обобщена до мультивариативной с учётом дополнительных признаков, таких как тип медиа, регион или язык).