

Прогнозирование высоковолатильных временных рядов социальных трендов и общественных интересов

Егор Валерьевич Задворнов

Московский физико-технический институт

Курс: Автоматизация научных исследований
(практика, В. В. Стрижов)/Группа 128

Эксперт: А. С. Малков

Консультант: А. В. Мацейко

2024

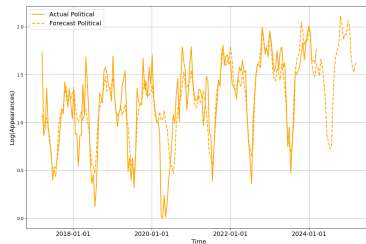
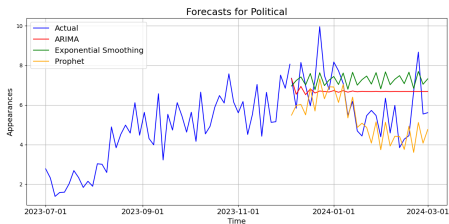
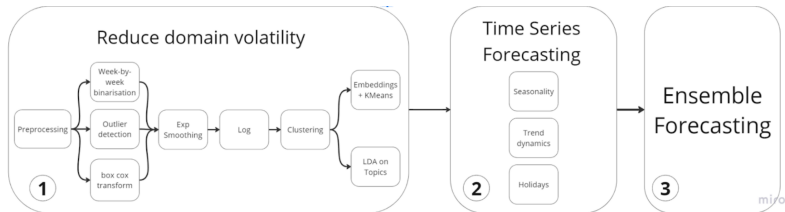
Цель исследования

Цели

Основная цель - прогнозирование временных рядов социальных трендов и общественных интересов, характеризующихся высокой волатильностью

- ▶ разработать методы кластеризации топиков общественных интересов
- ▶ сравнить качество моделей Prophet, Arima, Exp smoothig в задаче предсказания полученных кластеров по метрике семантического расстояния

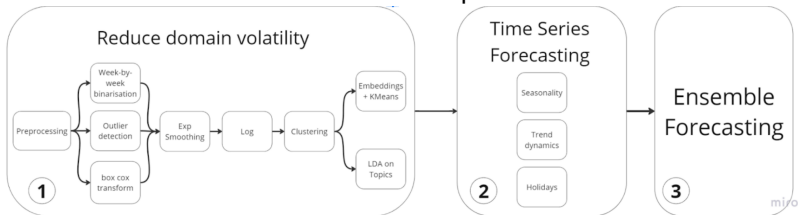
Доклад с одним слайдом



Постановка задачи

Пусть $T = \{T^k\}_{k=1}^K$ – множество уникальных топиков, где K – общее число различных топиков. Имеется временной ряд $\{t_i\}_{i=1}^N$, соответствующий последовательности дат, и на каждую дату t_i приходится набор $\{T_{ij}\}_{j=1}^{M_i}$ популярных топиков, где M_i – число популярных топиков в день t_i . Каждый $T_{ij} \in T$ представляет собой одно слово или фразу длиной до 4 слов. Цель данной работы – предсказать будущую популярность топиков. Для достижения этой цели предлагается следующий алгоритм:

Ключевые элементы решения:



Вычислительный эксперимент

1. Выполняется кластеризация множества T на n семантических кластеров $c_m = \{T^k\}_{k=1}^{n_m}$, $m = 1, \dots, n$, 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:

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

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

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

$$y(t) = g(t) + s(t) + h(t) + \epsilon_t \quad (3)$$

SARIMA

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

Метод Хольта-Винтерса (Exponential Smoothing)

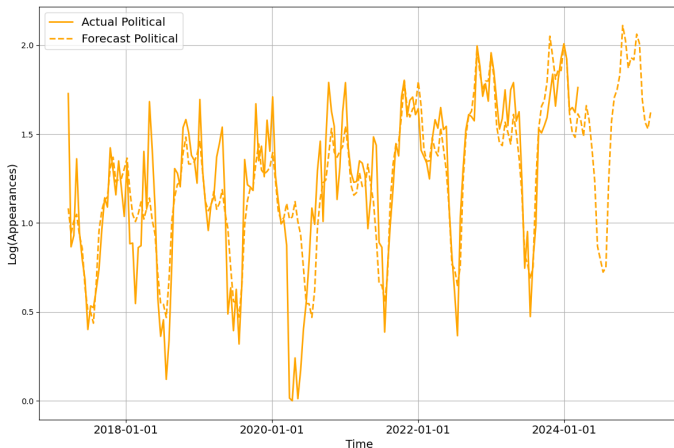
$$\hat{y}_{t+h|t} = l_t + hb_t + s_{t+h-m(k+1)} \quad (5)$$

Результаты кластеризации

Model	Mean Coherence	Coherence
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



Основные результаты

- ▶ Предложен гибридный подход, сочетающий методы прогнозирования временных рядов и тематического моделирования
- ▶ Разработан механизм оценки значимости трендов, повышающий точность прогнозирования
- ▶ Выявлены ограничения традиционных методов прогнозирования при наличии аномалий в данных
- ▶ Намечены пути дальнейшего развития, включая исследование алгоритмов обнаружения аномалий