

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

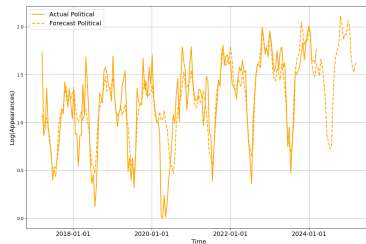
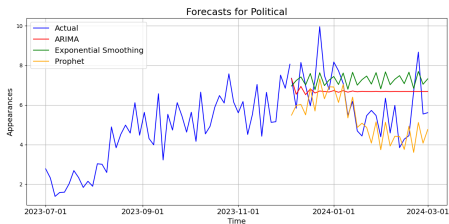
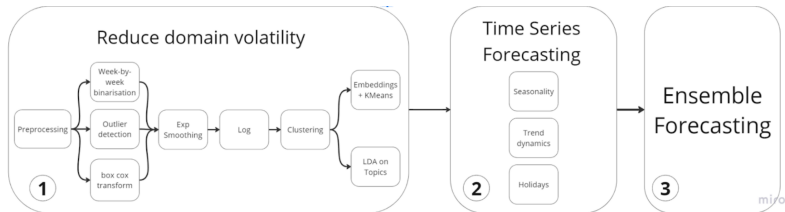
Цель исследования

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

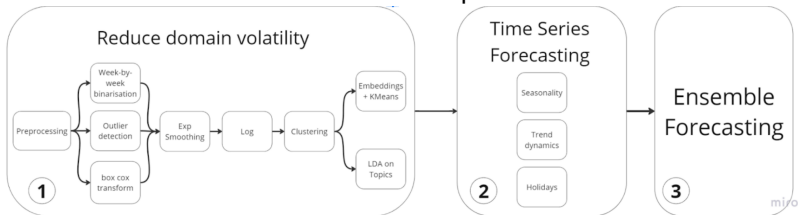
Доклад с одним слайдом



Постановка задачи

Задача заключается в разработке эффективного метода прогнозирования популярных тем и трендов в медиа-пространстве, учитывающего сложную временную динамику и тематическую структуру данных.

Ключевые элементы решения:



Вычислительный эксперимент

Результаты прогнозирования на реальных данных:

- ▶ Модель Пророка продемонстрировала высокую точность прогнозирования для кластеров, связанных с американским футболом и политикой
- ▶ Для некоторых кластеров, характеризующихся резкими пиками и изменениями трендов, традиционные методы прогнозирования показали ограниченную эффективность
- ▶ Средняя ошибка прогноза (MAPE) составила 28%, что характеризует хорошее качество прогнозов в рамках поставленной задачи

Результаты кластеризации

Model	Mean Coherence	Coherence
K-means	0.63	0.69
LDA	0.69	0.71

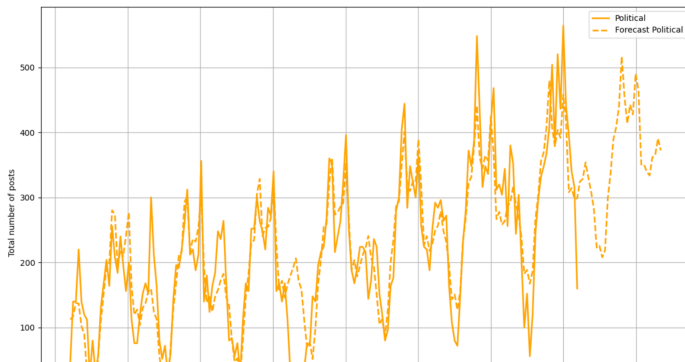
Таблица: Resulted metrics of each algorithm. The error calculated by “error of the mean” formula.

The LDA algorithm is better according to the Mean Coherence metric. Furthermore, it has more stable per-cluster Coherence than Embeddings + K-Means algorithm. Therefore, we have chosen to utilize LDA for further analysis. In the next section, we apply the forecast model to clusters-outcomes after LDA clustering.

Результаты прогнозирования

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

Таблица: Comparison of Average MAE and MSE for Prophet, ARIMA, and Exponential Smoothing models.



Анализ ошибки

Вычислим ошибку MAE каждого из прогнозов. В случае синтетического набора: $MAE_1 = 0.55$, $MAE_2 = 0.17$, $MAE_3 = 0.17$. На реальном наборе: $MAE_1 = 895.2$, $MAE_2 = 152.8$, $MAE_3 = 0.11$.

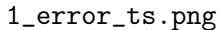
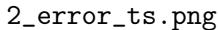
A placeholder for a plot titled '1_error_ts.png'.A placeholder for a plot titled '2_error_ts.png'.

Рис.: Средняя MAE прогноза попарных расстояний для каждого попарных расстояний для каждого

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

Представленный подход демонстрирует высокую эффективность в прогнозировании динамики медиа-ландшафта и может быть применен в различных областях, таких как анализ научных публикаций, прогнозирование спроса на продукты и мониторинг социальных тенденций.