
Trends prediction

Zadvornov Egor
MIPT
zadvornov.ev@phystech.edu

16 мая 2024 г.

Abstract

Analyzing and predicting trends in the media landscape is a complex task due to the volatility and instability of social trends and public interests. This study presents a novel approach that combines time series forecasting methods and topic modeling to tackle this challenge. The proposed framework leverages a multi-pronged clustering strategy, including embedding-based and topic modeling-based techniques, to identify thematic clusters within the media data. For each cluster, the study employs the state-of-the-art Prophet forecasting model to capture the unique characteristics and dynamics, enabling accurate predictions of future trends.

The results demonstrate the effectiveness of this hybrid approach, particularly in forecasting trends related to American football. However, the study also identifies limitations in predicting certain clusters that exhibit peculiarities, such as abrupt peaks and trend shifts. To address this, the study outlines future research directions, including the exploration of anomaly detection and forecasting algorithms specifically designed to handle complex time series patterns. By combining the strengths of time series forecasting and topic modeling, this work contributes to the advancement of trend prediction techniques in the dynamic and multifaceted media landscape, with potential applications in various domains beyond media, such as scientific publication trends and product demand forecasting.

1 Introduction

Analyzing and forecasting trends in the ever-evolving media landscape is a critical challenge with significant implications for various domains, including marketing, media production, public relations, and innovation research. The dynamic and volatile nature of social trends and public interests poses a formidable obstacle, as these phenomena often exhibit complex, non-linear, and rapidly changing patterns.

Our novel framework that integrates the time-series forecasting and topic modeling. Rather than relying on simplistic trend identification or manual curation, our framework employs a multi-pronged clustering strategy, including embedding-based and topic modeling-based techniques, to identify coherent thematic clusters within the media data. For each identified cluster, we then apply the state-of-the-art Prophet forecasting model, which has demonstrated exceptional capabilities in handling various time-series patterns, such as seasonality, trend changes, and outliers.

Currently, such models as ARIMA[link] or Exponential Smoothing[link] (e.g., Google Trends) are used to solve the task. These conventional approaches are applied directly to predict specific topics (trends) at the next point in time. The novelty of our method is the addition of a procedure for clustering topics and predicting the popularity of a whole class of trends. This approach seems to be more justified and useful, since it can be integrated into other algorithms, such as those for time series prediction on the predicted set of topics. One of the goals of this work is to check whether the algorithm we developed is superior to the methods described above. As demonstrated in Section “Forecasting”, the model is able to capture unique patterns in the data that simpler models, such as ARIMA or Exponential Smoothing, do not account for. Therefore, our framework demonstrates better metrics.

Furthermore, the framework can be extended to differentiate predictions based on social groups. This involves selecting a target social group (e.g., predicted cluster for sports), increasing the granularity of topics for that group (e.g., football, volleyball, Messi, etc.), and making predictions accordingly. This approach can provide more tailored and relevant insights for specific audience segments.

The potential impact of this work is vast, as the ability to reliably forecast media trends can have far-reaching implications across numerous industries and domains. From marketing and communication strategies to innovation research and product development, the insights gleaned from our framework can help organizations better anticipate and respond to the evolving interests and discussions of their target audiences.

Furthermore, the generalizability of our approach extends beyond the media domain, as the underlying principles can be applied to other complex and high-dimensional time-series data, such as trends in scientific publications, social movements, or consumer demand patterns. By pushing the boundaries of trend forecasting, this study aims to contribute to the advancement of time-series analysis and predictive modeling, with the ultimate goal of empowering decision-makers to navigate the dynamic and ever-changing landscapes of the modern world.

2 Постановка задачи

Пусть $T = \{T^k\}_{k=1}^K$ – множество уникальных топигов, где K – общее число различных топигов. 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 топигов, где M_i – число популярных топигов 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 топигов. Для достижения этой цели предлагается следующий алгоритм:

2.1 Кластеризация топигов

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 топигов 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:

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

Мера C_{cv} рассчитывается следующим образом:

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

где $S_{set}^{one} = \{(W_0, W_*) | W_0 = \{w_i\}, w_i \in W, W_* = W\}$ – сегментация, представляющая множество пар подмножеств слов, $\tilde{m}_{\cos(nlr, 1)}(W_0, W_*)$ – косинусное сходство между векторами контекста W_0 и W_* , оценивающее, насколько хорошо W_* "поддерживает" или "объясняет" W_0 .

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

2.2 Прогнозирование популярности топигов

Популярные модели обработки временных рядов, такие как Prophet, ARIMA и Exponential Smoothing, обучаются на тренировочных наборах данных $\{t_i, y_m^i\}_{train}$, $0 \leq i < N_{train}$, для каждого кластера c_m . Обученные модели затем применяются на тестовой выборке $\{t_i, y_m^i\}_{test}$, $N_{train} \leq i < N_{train} + N_{test}$, для предсказания числа появлений y_m^i топигов $\{T^k\}_{k=1}^{n_m}$ из кластера c_m в будущие моменты времени t_i .

2.2.1 Модель Prophet

Модель Prophet представляет собой аддитивную регрессионную модель, декомпозирующую временной ряд на следующие компоненты:

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

где $g(t)$ – компонента тренда, $s(t)$ – компонента сезонности, $h(t)$ – компонента, отражающая эффекты праздников, ϵ_t – случайная ошибка.

Компонента тренда $g(t)$ может быть представлена в виде логистического роста $g(t)_{log}$ или кусочно-линейного тренда с точками изменения $g(t)_{lin}$:

$$g(t)_{lin} = (k + \sum_{j=1}^S \delta_j)t + (m + \sum_{j=1}^S \gamma_j) \quad (\text{линейный тренд}) \quad (4)$$

$$g(t)_{log} = \frac{C(t)}{1 + \exp(-g(t)_{lin})} \quad (\text{нелинейный логистический рост}) \quad (5)$$

Сезонная компонента $s(t)$ моделируется с помощью ряда Фурье:

$$s(t) = \sum_{n=1}^N \left(a_n \cos\left(\frac{2\pi nt}{P}\right) + b_n \sin\left(\frac{2\pi nt}{P}\right) \right) \quad (6)$$

Эффекты праздников $h(t)$ моделируются как сумма индикаторных функций для каждого праздничного дня i :

$$h(t) = \sum_{i=1}^L \theta_i \mathbb{I}_{t \in D_i} \quad (7)$$

Оценивание и подбор оптимальных параметров модели $\hat{\theta}$ выполняется с использованием метода максимального правдоподобия:

$$\hat{\theta} = \arg \min_{\theta} -\log p(y|\theta) \left\{ = \frac{n}{2} \log(2\pi) + \frac{n}{2} \log(\sigma^2) + \frac{1}{2\sigma^2} \sum_t \left(y(t) - \sum_i f_i(t) \right)^2 \right\} \quad (8)$$

где θ – параметры модели, n – количество наблюдений, σ^2 – дисперсия ошибок, $y(t)$ – фактические значения, $f_i(t)$ – компоненты модели.

Прогнозирование осуществляется путем экстраполяции тренда, использования сезонной компоненты и эффектов праздников:

$$\hat{y}(t+h) = \hat{g}(t+h) + \hat{s}(t+h) + \hat{h}(t+h) \quad (9)$$

2.2.2 Модель SARIMA

Модель SARIMA (Seasonal Autoregressive Integrated Moving Average) является расширением ARIMA для учета сезонности во временных рядах. Модель $SARIMA(p, d, q)(P, D, Q)_m$ может быть представлена следующим образом:

$$(1 - \phi_1 B)(1 - \Phi_1 B^m)(1 - B)^d(1 - B^m)^D y_t = (1 + \theta_1 B)(1 + \Theta_1 B^m) \epsilon_t \quad (10)$$

где p – порядок несезонной авторегрессионной части, d – степень несезонного дифференцирования, q – порядок несезонной части скользящего среднего, P – порядок сезонной авторегрессионной части, D –

степень сезонного дифференцирования, Q – порядок сезонной части скользящего среднего, m – число периодов в сезонном цикле, ϕ_1 – параметр несезонной авторегрессионной части, Φ_1 – параметр сезонной авторегрессионной части, θ_1 – параметр несезонной части скользящего среднего, Θ_1 – параметр сезонной части скользящего среднего, B – оператор сдвига назад ($B^j y_t = y_{t-j}$), ϵ_t – белый шум.

Алгоритм модели SARIMA включает следующие шаги:

1. Идентификация модели: определение значений p, d, q, P, D, Q и m на основе анализа автокорреляционной (ACF) и частной автокорреляционной функций (PACF), а также тестов на стационарность.
2. Оценка параметров: оценка параметров $\phi_1, \Phi_1, \theta_1, \Theta_1$ и дисперсии шума σ^2 методом максимального правдоподобия или условным методом наименьших квадратов.
3. Диагностика модели: проверка адекватности модели путем анализа остатков на наличие автокорреляции, нормальности и гомоскедастичности.
4. Прогнозирование: получение точечных и интервальных прогнозов на основе оцененной модели:

$$\hat{y}_{t+h|t} = \phi_1 \hat{y}_{t+h-1|t} + \Phi_1 \hat{y}_{t+h-m|t} - \theta_1 \hat{\epsilon}_{t+h-1|t} - \Theta_1 \hat{\epsilon}_{t+h-m|t} \quad (11)$$

где $\hat{\epsilon}_{t+h|t} = 0$ для $h > 0$.

2.2.3 Метод Хольта-Винтерса (Exponential Smoothing)

Метод Хольта-Винтерса является расширением модели экспоненциального сглаживания для учета тренда и сезонности. Аддитивный метод Хольта-Винтерса представлен следующим образом:

$$\hat{y}_{t+h|t} = l_t + hb_t + s_{t+h-m(k+1)} \quad (12)$$

где $\hat{y}_{t+h|t}$ – прогноз на h шагов вперед, сделанный в момент t , l_t – уровень ряда, b_t – тренд, s_t – сезонная компонента, m – число периодов в сезонном цикле, $k = \lfloor (h-1)/m \rfloor$.

Алгоритм метода Хольта-Винтерса:

1. Инициализация начальных значений для уровня l_0 , тренда b_0 и сезонных компонент $s_0, s_{-1}, \dots, s_{-(m-1)}$.
2. На каждом шаге t обновляются уровень l_t , тренд b_t и сезонная компонента s_t по формулам:

$$l_t = \alpha(y_t - s_{t-m}) + (1 - \alpha)(l_{t-1} + b_{t-1}) \quad (13)$$

$$b_t = \beta(l_t - l_{t-1}) + (1 - \beta)b_{t-1} \quad (14)$$

$$s_t = \gamma(y_t - l_t) + (1 - \gamma)s_{t-m} \quad (15)$$

где α, β, γ – сглаживающие параметры.

3. Прогнозирование с использованием обновленных оценок уровня, тренда и сезонности:

$$\hat{y}_{t+h|t} = l_t + hb_t + s_{t+h-m(k+1)} \quad (16)$$

4. Оценка параметров α, β, γ путем минимизации суммы квадратов ошибок прогноза.

2.3 Критерии качества прогноза

Обозначим через $\mathbf{Y}_m \in R^{N_{test}}$ упорядоченный набор из фактических значений временного ряда кластера c_m в моменты t_i , $N_{train} \leq i < N_{train} + N_{test}$:

$$\mathbf{Y}_m = [y_m^{N_{train}}, \dots, y_m^{N_{train}+N_{test}-1}]^T. \quad (17)$$

Получим $\{\mathbf{Y}_m\}_{m=1}^n$

В качестве критериев качества прогноза временных рядов используются среднее абсолютное отклонение (MAE) и средний квадрат ошибки (MSE):

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |\mathbf{Y}_i - \hat{\mathbf{Y}}_i| \quad (18)$$

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n (\mathbf{Y}_i - \hat{\mathbf{Y}}_i)^2 \quad (19)$$

где $\{\hat{\mathbf{Y}}_m\}_{m=1}^n$ – прогнозные значения n – число временных рядов.

2.4 Задачи исследования

Основными задачами данной работы являются:

Выбор наилучшего алгоритма кластеризации топиков из двух вариантов: Embeddings + K-Means и LDA.

Сравнение популярных моделей для прогнозирования временных рядов, а именно ARIMA, Prophet и Exponential Smoothing, и проверка гипотезы о том, что модель Prophet дает наилучшие значения метрики MSE.

Сравнение метрик наилучшей модели для прогнозирования в случае предсказания числа вхождений y_i каждого отдельного топика T^k и в случае предсказания числа y_m^i внутри кластера c_m . Проверка гипотезы о том, что подход с кластеризацией топиков будет давать лучшие мет

3 Data construction

The dataset used in this study consists of publicly available posts on social media platforms, such as Twitter, and the most popular search queries through browsers, such as Google, over several years. The data was collected from media aggregation platforms and spans the period from January 1, 2019, to March 3, 2024.

The dataset includes the top 15 news topics from both Twitter and Google for each day during this time period, resulting in a total of 76,140 observations. The dataset has the following structure:

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 1: Sample of data

The dataset includes information on the timestamp, source (Twitter or Google), and the specific topic or search query. The target variable (Y) in this dataset is the Topic, representing the popular topics discussed on social media and search engines over time. The predictor variable (X) is the Time, which represents the date and serves as the input for forecasting the future trends in the media landscape. Prediction is feasible in this dataset due to the presence of cyclic/periodic patterns in the target variable (Y) over time (X). The topics discussed in the media often exhibit seasonal and temporal patterns, which can be leveraged to forecast future trends. To assess the generalizability of the proposed approach, the study also tested the model on two additional datasets: Twitter Trending Tweets [5]: This dataset contains information on the daily trending tweets on Twitter, including the topic and its significance. YouTube Trending Video Dataset [6]: This dataset

includes data on the daily trending YouTube videos, such as the video title, channel, and various engagement metrics. However, the results from these additional datasets were not as promising as the primary dataset, and the details are provided in the Appendix. The primary dataset used in this study offers a comprehensive representation of the media landscape, covering both social media and search engine trends. The combination of Twitter and Google data provides a well-rounded view of the evolving public interests and discussions, making it a suitable testbed for the proposed trend forecasting framework. We applied the model to this dataset, see this description in the next paragraphs.

4 Algorithms and Models

4.1 Pipeline of the research

The proposed approach aims to tackle the challenge of trend prediction in the media landscape. The scheme of the pipeline is presented at 4.1.

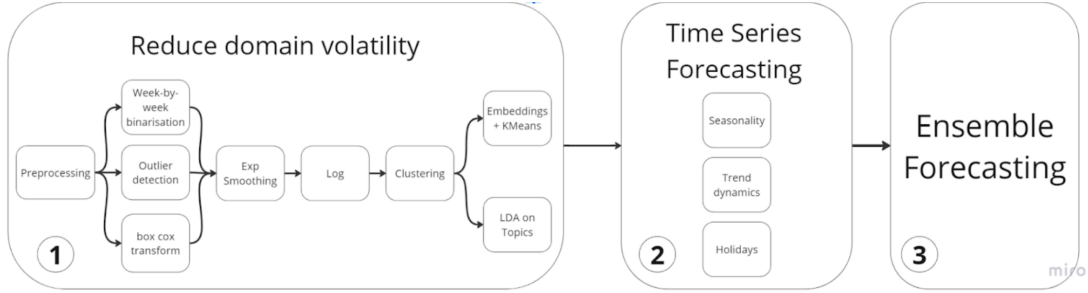


Рис. 1: Image of the pipeline structure.

The key steps of the pipeline are:

Preprocessing and Clustering: The raw media data (e.g., social media posts, search trends) is preprocessed to remove noise and extract relevant features. Reduce domain volatility through week-by-week binarisation. A multi-pronged clustering strategy is employed to identify coherent thematic clusters within the media data. This includes techniques such as: - Embedding-based clustering using sentence embeddings and K-Means - Topic modeling using Latent Dirichlet Allocation (LDA) on the topic representations

Significance Estimation: For each identified thematic cluster, the significance of the events or topics is estimated based on their position in the media landscape (e.g., ranking in trending topics). This significance estimation helps to prioritize the most relevant and impactful trends for the subsequent forecasting stage.

Time Series Forecasting: The time series of each thematic cluster is modeled using the state-of-the-art Prophet forecasting algorithm. The Prophet model is chosen for its ability to handle various time series patterns, such as seasonality, trend changes, and outliers. By applying the Prophet model to the individual thematic clusters, the framework can capture the unique characteristics and dynamics of each topic, leading to more accurate and nuanced predictions.

Ensemble Forecasting: The forecasts from the individual thematic clusters are combined to provide a comprehensive prediction of the future media landscape.

Anomaly Detection and Forecasting: The study identifies the need to explore existing algorithms specifically designed for anomaly detection and forecasting in time series data. These advanced techniques hold the potential to enhance the predictive capabilities of the proposed framework across a wider range of topic clusters, particularly those exhibiting complex or irregular patterns.

4.2 Clustering

4.2.1 Theory

We employed a multi-pronged approach to clustering the time series data, experimenting with several techniques to identify the most effective method. Here are the top 2 of them

- Embeddings + K-Means

- Topic Modeling on Topics using Latent Dirichlet Allocation (LDA)

Key steps of the first approach are Embedding Generation, which utilized the Sentence Transformers library [1] to generate contextual embeddings for each topic using the all-MiniLM-L12-v2 model [2], and then K-Means Clustering applied to embedded topics. Since both the embedding procedure and the k-means algorithm are widely known operations, we do not provide a detailed description of them in this work, leaving only a reference link.

The second algorithm works as follows. The preprocessing steps included lemmatization, stop word removal, and keeping words of length 2-3. Then the LDA model is created. [add description of how the LDA model works]. The aim of the section is to compare described approaches by clustering quality and choose one to use in the Forecasting section. The next criteria for selecting the best algorithm of these two, based on the coherence[add link!] metric, was elaborated: $\langle \text{rewrite as formula in latex with proper definitions} \rangle$ If $|\text{coherence embed} - \text{coherence LDA}| < 0.05$ AND $\text{coherence embed} > 0.5$ then: “We choose embedding” else: “We choose LDA” Thus, we allow a slight loss of the k-means algorithm to the LDA in metric value, since it is the more simple model and outperforms LDA.

The metric coherence was chosen since it ... $\langle \text{why it is a proper choice} \rangle$. The algorithm of coherence calculation is described at [add link!]. The main steps are $\langle \text{some summary of algorithm} \rangle$.

The next two subsections outline the main ideas of the compared algorithms and their clustering results.

4.2.2 Embeddings + K-Means

The Embeddings + K-Means approach was used to perform the initial clustering of the media topics. The key steps are as follows: Embedding Generation: We utilized the Sentence Transformers library [1] to generate contextual embeddings for each topic using the all-MiniLM-L12-v2 model [2]. These embeddings capture the semantic similarity between the topics. K-Means Clustering: The generated embeddings were then fed into the K-Means clustering algorithm to group the topics based on their semantic similarity. To determine the optimal number of clusters (K), we evaluated several metrics. The mean coherence score, which quantifies the interpretability of clusters to humans by measuring the semantic similarity among the top words within a cluster. The optimal number of clusters was found to be 16, as shown in the figure:

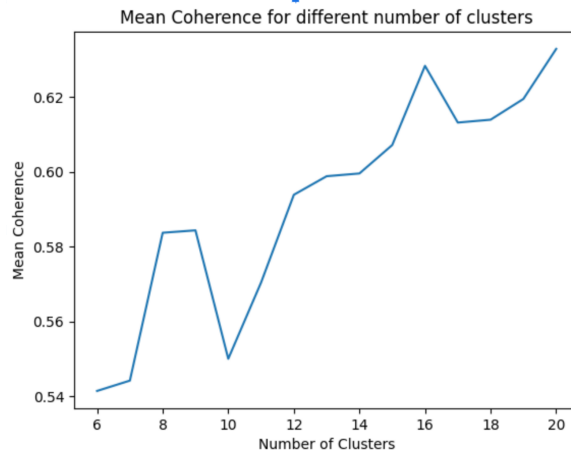


Рис. 2: Coherence score vs. number of clusters for the Embeddings + K-Means algorithm.

The resulting clustering with expert interpretation is presented at Table 2. The mean coherence score for the 16 clusters was 0.63.

The Embeddings + K-Means approach demonstrated good thematic coherence within the clusters, allowing for meaningful interpretation and subsequent forecasting. The clusters captured distinct themes, such as business, sports, politics, and entertainment, providing a solid foundation for the time series forecasting component of the study. Examples, references to the algorithms, clustering images and other details of the Embeddings + K-Means algorithms are given in the appendix.

Cluster #	Interpretation	Coherence
0	NFL	0.63
1	Football	0.43
2	Motivational Hashtags	0.75
3	Sports Personalities	0.58
4	Entertainment and Celebrities	0.58
5	Music	0.55
6	Politics and Holidays	0.73
7	UFC	0.74
8	Emotional Hashtags	0.70
9	Social events Hashtags	0.73
10	Political Figures and Events	0.59
11	Basketball	0.37
12	Celebrities and Personalities	0.65
13	Business	0.65
14	World Events and Countries	0.71
15	Entertainment	0.68
Mean	-	0.63

Таблица 2: Embed+Kmeans clustering result

4.2.3 Latent Dirichlet Allocation (LDA) on Topics

To address the limitations of the embedding-based approach, we explored topic modeling using Latent Dirichlet Allocation (LDA) [link!] directly on the topic representations. This allowed us to capture the latent thematic structures within the data. We implement the algorithm as follows. The preprocessing steps included lemmatization, stop word removal, and keeping words of length 2-3. Then the LDA model is created. [add description how the LDA model works]. The coherence score increased with a higher number of clusters, as shown in the Figure 3.

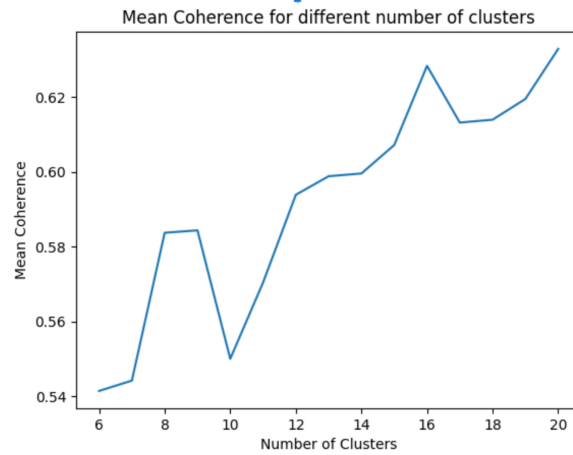


Рис. 3: Mean coherence score vs number of clusters

According to the plot there is no dependence of LDA on the number of clusters. Therefore, to provide a more accurate comparison, we set 16 clusters and calculated the Coherence metrics (Table 3).

4.2.4 Summary

According to criteria in the beginning of the Theory part, the algorithm “Embedding + K means” was chosen. The pivot table of metrics for the selected models are as follows:

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.

Cluster #	Interpretation	Coherence
0		0.69
1		0.71
2		0.68
3		0.67
4		0.70
5		0.71
6		0.67
7		0.71
8		0.69
9		0.68
10		0.68
11		0.72
12		0.71
13		0.71
14		0.69
15		0.67
Mean	-	0.69

Таблица 3: LDA clustering result

Model	Mean Coherence	Coherence
K-means	0.63 +- ?	0.69
LDA	0.693 +- 0.004	0.71

Таблица 4: Resulted metrics of each algorithm. The error calculated by “error of the mean” formula.

4.3 Forecasting

4.3.1 Theory

In this Section different architectures are applied to forecast popularity of topics. The input of the models is a week-by-week time series of clustered social media topics X , the prediction is a set of integer numbers N_i – the counts of posts, containing the cluster i in the next timestep (week).

Three models were investigated during the research. The first one is the Prophet, developed by researchers at Facebook’s Core Data Science team. It is a decomposable time series model that combines several fundamental components to capture the dynamics of a time series [4]. The key advantages of the Prophet model include its ability to handle various types of time series patterns, such as seasonality, trend changes, and outliers, as well as its scalability and ease of use. Basic steps of the algorithm are [add description from habr post]. For a more detailed description one can read [link! from habr post]. The second model called ARIMA works like [add description with links]. The last model we experimented with is Exponential Smoothing. The algorithm of its operation is [add description with links]. Since all the models were evaluated on the same dataset, the MAE and MSE metrics were chosen to compare their quality.

4.3.2 Prophet model

The forecasting component of the proposed framework represents a crucial step in predicting future trends in the media landscape. To address the challenge of accurately forecasting complex and volatile time series data, the study employed a state-of-the-art forecasting model, the Prophet algorithm [3]. After applying the Embedding algorithm, we conducted forecasts for each result. The following table presents the final metrics for the forecasts:

Clusters Count	Average MAE	Average MSE
5	0.22	0.08
7	0.23	0.08
9	0.25	0.10
16	0.28	0.13

Таблица 5: Resulted Prophet metrics.

For the American football cluster and Political cluster, the Prophet model was able to accurately capture the seasonality and life cycle of the topic, as shown in the following figures:

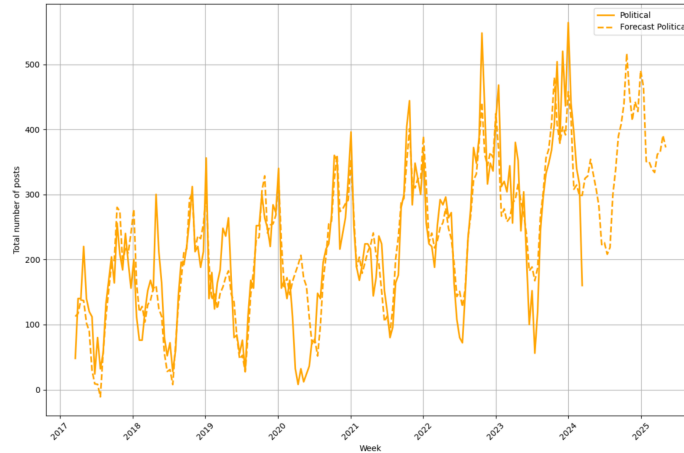


Рис. 4: Prophet forecast for American football cluster.

Список литературы

Taylor, S. J. and Letham, B. (2018). Forecasting at scale. The American Statistician, 72(1):37–45. doi:10.1080/00031305.2017.1380080. <https://peerj.com/preprints/3190/>

Reimers, N. and Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. arXiv preprint arXiv:1908.10084. <https://arxiv.org/abs/1908.10084>

Motrenko, A. and Strijov, V. (2014). Extracting fundamental periods to segment biomedical signals. <https://m1p.org/papers/MotrenkoStrijov2014RV2.pdf>

Das, A. (2020). Foundation of Variational Autoencoder (VAE). <https://ayandas.me/blog-tut/2020/01/01/variational-autoencoder.html>

Grabovoy, A. V. and Strijov, V. V. (2019). Quasi-periodic time series clustering for human activity recognition. <https://m1p.org/papers/Grabovoy2019QuasiPeriodicTimeSeries.pdf>

Uvarov, N. D., Kuznetsov, M. P., Malkova, A. S., Rudakov, K. V., and Strijov, V. V. (2018). Selection of superposition of models for railway freight forecasting. <https://m1p.org/papers/Uvarov2018SuperpositionForecastingeng.pdf>

Gardner, E. S. (1985). Exponential smoothing: The state of the art. Journal of Forecasting.

Bouchaud, J.-P., Bonart, J., Donier, J., and Gould, M. (2018). Trades, Quotes and Prices, Financial Markets Under the Microscope. Cambridge University Press. <https://www.cambridge.org/core/books/trades-quotes-and-prices/029A71078EE4C41C0D5D4574211AB1B5>

G^arleanu, N. and Pedersen, L. H. (2013). Dynamic Trading with Predictable Returns and Transaction Costs. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1364170 *Twitter Trending Tweets*. <https://www.kaggle.com/datasets/rsrishav/twitter-trending-tweets>

YouTube Trending Video Dataset. <https://www.kaggle.com/datasets/rsrishav/youtube-trending-video-dataset>

Sentence Transformers: all-MiniLM-L12-v2. <https://huggingface.co/sentence-transformers/all-MiniLM-L12-v2>