Forecasting highly volatile time series of social trends and public interests

Zadvornov Egor
MIPT
zadvornov.ev@phystech.edu

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.

Keywords A set of time series · Volatility · Social Trends · Prophet · Topic Modeling · ARIMA.

1 Introduction

The ability to accurately analyze and forecast trends in the ever-evolving media landscape is a critical challenge with far-reaching implications across various domains, including marketing, media production, public relations, and innovation research. This study presents a novel framework that combines state-of-the-art time series forecasting techniques and topic modeling to tackle this complex problem.

Key terms in this context are defined as follows:

- Media landscape refers to the dynamic and multifaceted ecosystem of digital media, including social media platforms, search engines, and news outlets.
- Trend forecasting involves the prediction of future patterns, themes, and discussions that are likely to emerge and gain prominence within the media landscape.
- Topic modeling is a technique used to identify and extract the latent thematic structure within a corpus of text-based data, such as social media posts or news articles.

The main idea of this work is to leverage a hybrid approach that integrates the strengths of time series forecasting and topic modeling to provide accurate and nuanced predictions of future media trends. By identifying coherent thematic clusters within the data and applying customized forecasting models to each cluster, the framework aims to capture the unique dynamics and patterns associated with different topics.

Recent developments in the field of time series analysis have demonstrated the potential for more sophisticated techniques to handle the complexity of media data. For instance, Shumway and Stoffer [1] investigated the time series analysis techniques, which could be adapted for predicting social trends. Bouchaud et al. [2] focused on modeling micro-level dynamics in financial markets, an approach that may be relevant for understanding the evolution of social trends. Additionally, the methods for quasi-periodic time series clustering discussed in Grabovoy and Strizhov [3] and the use of variational autoencoders [4] for learning lower-dimensional representations could be promising directions for the analysis of social trend data.

While existing approaches often rely on simplistic trend identification or manual curation, the current state of the field lacks a comprehensive and systematic framework for predicting trends in the media landscape [5]. Conventional models, such as ARIMA [6] or Exponential Smoothing [7], have been applied directly to forecast specific topics at the next point in time (e.g., Google Trends [8]). However, these methods may fail to account for the high-dimensional and volatile nature of the media data, as well as the need to differentiate predictions based on relevant social groups or audience segments.

To address these limitations, the proposed framework integrates a multi-pronged clustering strategy, including embedding-based [9] and topic modeling-based techniques [10], to identify coherent thematic clusters within the media data. For each identified cluster, the study employs the state-of-the-art Prophet forecasting model [11], which has demonstrated exceptional capabilities in handling various time-series patterns, such as seasonality, trend changes, and outliers.

This approach seems to be more justified and useful, since it can be integrated into other algorithms, such as those for conventional 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 Problem Statement

Let $T = \{T^k\}_{k=1}^{K}$ be the set of unique topics, where K is the total number of distinct topics. There is a time series $\{t_i\}_{i=1}^{N}$, corresponding to a sequence of dates, and for each date t_i there is a set $\{T_{ij}\}_{j=1}^{M_i}$ of popular topics, where M_i is the number of popular topics on day t_i. Each $T_{ij} \in T$ represents a single word or phrase of up to 4 words.

The goal of this work is to predict the future popularity of topics. To achieve this goal, the following algorithm is proposed:

2.1 Topic Clustering

1. The set T is clustered into n semantic clusters $c_m = \{T^k\}_{k=1}^{n_m}$, $m = 1, \ldots, n$, where n_m is the number of topics in cluster c_m. 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 [12] to generate contextual embeddings for each topic using the all-MiniLM-L12-v2 model [13], and then K-Means Clustering applied to embedded topics [9]. Since both algorithms are widely known operations, we do not provide a detailed description of them in this work, leaving only a reference links.

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 [10].

2.1.1 LDA

The Latent Dirichlet Allocation (LDA) model is a probabilistic topic model that allows discovering hidden (latent) topics in a collection of text documents. The LDA model is based on the assumption that each document is a mixture of several topics, and each word in the document belongs to one of these topics.

The LDA algorithm includes the following steps:

- For each topic $k \in 1, \ldots, K$:
  - Choose the word distribution $\phi_k \sim \text{Dirichlet}(\beta)$, where β is the hyperparameter that sets the prior Dirichlet distribution for the word distributions.
- For each document $d \in 1, \ldots, D$:
  - Choose the topic distribution $\theta_d \sim \text{Dirichlet}(\alpha)$, where α is the hyperparameter that sets the prior Dirichlet distribution for the topic distributions.
  - For each word w_{dn} in document d:
    - Choose a topic $z_{dn} \sim \text{Multinomial}(\theta_d)$
    - Choose a word $w_{dn} \sim \text{Multinomial}(\phi_{z_{dn}})$

The joint distribution of all the model variables (topic distributions θ, word distributions ϕ, topic assignments $\mathbf{z}$, and observed words $\mathbf{w}$) can be written as:

$$p(\theta, \phi, \mathbf{z}, \mathbf{w}|\alpha, \beta) = \prod_{k=1}^{K} p(\phi_k|\beta) \prod_{d=1}^{D} p(\theta_d|\alpha) \prod_{n=1}^{N_d} p(z_{dn}|\theta_d) p(w_{dn}|\phi_{z_{dn}}) \tag{1}$$

The inference task in LDA is to find the posterior distribution of the hidden variables (θ, ϕ and $\mathbf{z}$) given the observed words $\mathbf{w}$ and hyperparameters α and β. Since the exact inference in this model is computationally complex, approximate methods are used in practice.

2.1.2 Optimal Number of Clusters

To determine the optimal number of clusters n, the average coherence measure C_{cv} [14] is used, which evaluates the interpretability of clusters to humans by measuring the semantic similarity between the words within a cluster:

$$n = \arg\min_{L \in \mathbb{N}} \frac{1}{L} \sum_{m=0,\ldots,L-1} C_{\text{cv}}(c_m) \tag{2}$$

The C_{cv} measure is calculated as follows:

$$C_{\text{cv}} = \frac{1}{|S^{\text{one}}_{\text{set}}|} \sum_{(W_0, W_*) \in S^{\text{one}}_{\text{set}}} \tilde{m}_{\cos(\text{nlr},1)}(W_0, W_*) \tag{3}$$

where $S^{\text{one}}_{\text{set}} = \{(W_0, W_*)|W_0 = \{w_i\}, w_i \in W, W_* = W\}$ is the segmentation representing the set of pairs of subsets of words, $\tilde{m}_{\cos(\text{nlr},1)}(W_0, W_*)$ is the cosine similarity between the context vectors W_0 and W_*, evaluating how well W_* "explains" W_0.

2.1.3 Metric

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 metric, was elaborated:

$$\text{Model} = \begin{cases} \text{Embedding}, & \text{if } |C_{cv}^{Emb} - C_{cv}^{LDA}| < 0.1 \text{ and } C_{cv}^{Emb} > 0.5 \\ \text{LDA}, & \text{otherwise} \end{cases} \tag{4}$$

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.

2. For each cluster c_m a time series $\{t_i, y_m^i\}_{i=1}^N$ is constructed, where y_m^i is the number of occurrences of topics from cluster c_m on day t_i.

2.2 Topic Popularity Forecasting

Popular time series models, such as Prophet, ARIMA и Exponential Smoothing, are trained on the training sets $\{t_i, y_m^i\}_{train}$, $0 \le i < N_{train}$, for each cluster c_m. The trained models are then applied to the test set $\{t_i, y_m^i\}_{test}$, $N_{train} \le i < N_{train} + N_{test}$, to predict the number of occurrences y_m^i of topics $\{T^k\}_{k=1}^{n_m}$ from cluster c_m at future time points t_i.

2.2.1 Prophet Model

The Prophet model is an additive regression model that decomposes the time series into the following components:

$$y(t) = g(t) + s(t) + h(t) + \epsilon_t \tag{5}$$

where $g(t)$ is the trend component, $s(t)$ is the seasonality component, $h(t)$ is the holiday effects component, and ϵ_t is the random error.

The trend component $g(t)$ can be represented as a logistic growth $g(t)_{log}$ or a piecewise linear trend with change points $g(t)_{lin}$:

$$g(t)_{lin} = (k + \sum_{j=1}^{S} \delta_j)t + (m + \sum_{j=1}^{S} \gamma_j) \qquad \text{(linear trend)} \tag{6}$$

$$g(t)_{log} = \frac{C(t)}{1 + \exp\left(-g(t)_{lin}\right)} \qquad \text{(nonlinear logistic growth)} \tag{7}$$

The seasonal component $s(t)$ is modeled using a Fourier series:

$$s(t) = \sum_{n=1}^{N} \left(a_n \cos\left(\frac{2\pi nt}{P}\right) + b_n \sin\left(\frac{2\pi nt}{P}\right) \right) \tag{8}$$

The holiday effects $h(t)$ are modeled as a sum of indicator functions for each holiday i:

$$h(t) = \sum_{i=1}^{L} \theta_i \mathbb{1}_{t \in D_i} \tag{9}$$

The estimation and optimization of the model parameters $\hat{\theta}$ is performed using the maximum likelihood method:

$$\hat{\theta} = \arg\min_{\theta}(-\log p(y|\theta)) \left\{ = \frac{n}{2}\log(2\pi) + \frac{n}{2}\log(\sigma^2) + \frac{1}{2\sigma^2} \sum_t \left(y(t) - \sum_i f_i(t) \right)^2 \right\} \tag{10}$$

where $\theta$are the model parameters, n is the number of observations, σ^2 is the error variance, $y(t)$ are the actual values, and $f_i(t)$ are the model components.

Forecasting is done by extrapolating the trend, using the seasonal component, and incorporating the holiday effects:

$$\hat{y}(t+h) = \hat{g}(t+h) + \hat{s}(t+h) + \hat{h}(t+h) \tag{11}$$

2.2.2 SARIMA Model

The SARIMA (Seasonal Autoregressive Integrated Moving Average) model is an extension of ARIMA that accounts for seasonality in time series. The $SARIMA(p,d,q)(P,D,Q)_m$ model can be represented as:

$$(1-\phi_1 B)(1-\Phi_1 B^m)(1-B)^d(1-B^m)^D y_t = (1+\theta_1 B)(1+\Theta_1 B^m)\epsilon_t \tag{12}$$

where p is the order of the non-seasonal autoregressive part, d is the degree of non-seasonal differencing, q is the order of the non-seasonal moving average part, P is the order of the seasonal autoregressive part, D is the degree of seasonal differencing, Q is the order of the seasonal moving average part, m is the number of periods in the seasonal cycle, ϕ_1 is the parameter of the non-seasonal autoregressive part, Φ_1 is the parameter of the seasonal autoregressive part, θ_1 is the parameter of the non-seasonal moving average part, Θ_1 is the parameter of the seasonal moving average part, B is the backward shift operator ($B^j y_t = y_{t-j}$), and ϵ_t is white noise.

The SARIMA algorithm includes the following steps:

1.Model identification: Determining the values of p, d, q, P, D, Q and m based on the analysis of the autocorrelation (ACF) and partial autocorrelation (PACF) functions, as well as stationarity tests.

2. Parameter estimation: Estimating the parameters ϕ_1, Φ_1, θ_1, Θ_1 and the noise variance σ^2 using the maximum likelihood method or conditional least squares.

3. Model diagnostics: Checking the adequacy of the model by analyzing the residuals for autocorrelation, normality, and homoscedasticity.

4. Forecasting: Obtaining point and interval forecasts based on the estimated model:

$$\hat{y}_{t+h|t} = \phi_1 \hat{y}_{t+h-1|t} + \Phi_1 \hat{y}_{t+h-m|t} - \theta_1 \hat{\epsilon}_{t+h-1|t} - \Theta_1 \hat{\epsilon}_{t+h-m|t} \tag{13}$$

where $\hat{\epsilon}_{t+h|t} = 0$ for $h > 0$.

2.2.3 Holt-Winters (Exponential Smoothing) Method

The Holt-Winters method is an extension of the exponential smoothing model to handle trend and seasonality. The additive Holt-Winters method is represented as:

$$\hat{y}_{t+h|t} = l_t + hb_t + s_{t+h-m(k+1)} \tag{14}$$

where $\hat{y}_{t+h|t}$ is the forecast h steps ahead made at time t, l_t is the level of the series, b_t is trend, s_t is the seasonal component, m is the number of periods in the seasonal cycle, and $k = \lfloor (h-1)/m \rfloor$.

The Holt-Winters algorithm:

1. Initialization of initial values for level l_0, trend b_0 and seasonal components s_0, s_{-1}, …, $s_{-(m-1)}$.

2. At each step t, the level l_t, trend b_t and seasonal component s_t are updated according to the formulas:

$$l_t = \alpha(y_t - s_{t-m}) + (1-\alpha)(l_{t-1} + b_{t-1}) \tag{15}$$

$$b_t = \beta(l_t - l_{t-1}) + (1-\beta)b_{t-1} \tag{16}$$

$$s_t = \gamma(y_t - l_t) + (1-\gamma)s_{t-m} \tag{17}$$

where α, β, γ are smoothing parameters.

3. Forecasting using the updated level, trend, and seasonality:

$$\hat{y}_{t+h|t} = l_t + hb_t + s_{t+h-m(k+1)} \tag{18}$$

4. Estimate the smoothing parameters α, β, γ by minimizing the sum of squared forecast errors.

2.3 Forecast quality criteria

Let $\mathbf{Y}_m \in R^{N_{test}}$ be the ordered set of actual values of the time series for cluster c_m at moments t_i, $N_{train} \leq i < N_{train} + N_{test}$:

$$\mathbf{Y}_m = [y_m^{N_{train}}, \ldots, y_m^{N_{train}+N_{test}-1}]^T. \tag{19}$$

Obtain $\{\mathbf{Y}_m\}_{m=1}^n$

As performance metrics for time series forecasting, the Mean Absolute Error (MAE) and Mean Squared Error (MSE) are used:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|\mathbf{Y}_i - \hat{\mathbf{Y}}_i| \tag{20}$$

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{Y}_i - \hat{\mathbf{Y}}_i)^2 \tag{21}$$

where $\{\hat{\mathbf{Y}}_m\}_{m=1}^n$ are the forecasted values, and n is the number of time series.

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

Table 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 [15]: This dataset contains information on the

daily trending tweets on Twitter, including the topic and its significance. YouTube Trending Video Dataset [16]: 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 Experiment

4.1 Research Objectives

The main goal of this computational experiment is to verify the hypothesis that the proposed hybrid approach is superior to traditional time series models, such as ARIMA and Exponential Smoothing, in the task of forecasting trends in the media landscape.

To achieve this goal, the following is proposed:

- Develop methods for clustering topics of public interest
- Compare the quality of the Prophet, ARIMA, and Exponential Smoothing models in the task of predicting the obtained clusters using the semantic distance metric

We expect that using the Prophet model in conjunction with clustering will allow us to more accurately capture the specific dynamics of each cluster and, as a result, obtain more accurate predictions of topic popularity in the future.

4.2 Описание эксперимента

The scheme of the pipeline is presented at 4.2.

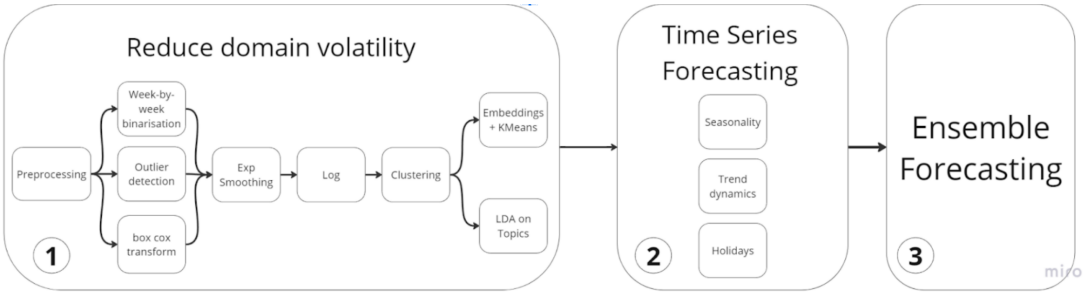

Figure 1: Image of the pipeline structure.

For the experiment, we use the dataset described in the "Problem Formulation" section. The data includes information about topics discussed on social media and search engines from January 1, 2019, to March 3, 2024.

In the first stage, we apply the preprocessing, clustering algorithms described in the "Topic Clustering" section: Embeddings + K-Means and LDA. Based on the clustering results, we choose the most suitable algorithm based on the coherence measure.

Next, for each cluster, we build time series of topic popularity and apply three forecasting models: Prophet, ARIMA, and Exponential Smoothing. We compare the quality of the forecasts obtained using these models using the Mean Absolute Error (MAE) and Mean Squared Error (MSE) metrics.

The code for this experiment is available at: `https://github.com/LoveMyWork/2024-Project-167/tree/master/src`

4.3 Experiment Results

4.3.1 Clustering Summary

The clustering results are presented in Table 2. According to the criteria described in the "Topic Clustering" section, the "Embedding + K means" algorithm was chosen, as it shows metrics slightly worse than LDA (by 0.06), but significantly outperforms it.

Model	Mean Coherence	Coherence
K-means	0.63	0.69
LDA	0.693	0.71

Table 2: Resulted metrics of each algorithm. The error calculated by “error of the mean” formula.

The next two subsections outline the hyperparetor optimisation of the compared algorithms and their clustering results.

4.3.2 Embeddings + K-Means

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:

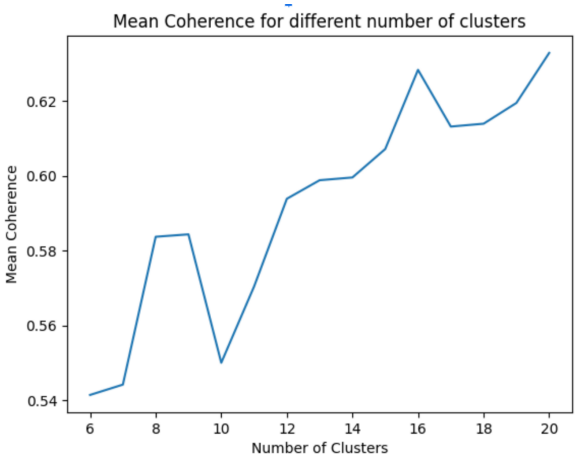

Figure 2: Coherence score vs. number of clusters for the Embeddings + K-Means algorithm.

The resulting clustering with expert interpretation is presented at Table 3. The mean coherence score for the 16 clusters was 0.63.

Cluster	Interpretation	C_{cv}
0	NFL	0.63
1	Football	0.43
2	Motivational Hashtags	0.75
3	Sports Personalities	0.58
4	Entertainment	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

Table 3: Embed+Kmeans clustering result

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.

4.3.3 Latent Dirichlet Allocation (LDA) on Topics

To address the limitations of the embedding-based approach, we explored topic modeling using Latent Dirichlet Allocation (LDA) [10] directly on the topic representations. This allowed us to capture the latent thematic structures within the data.

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).

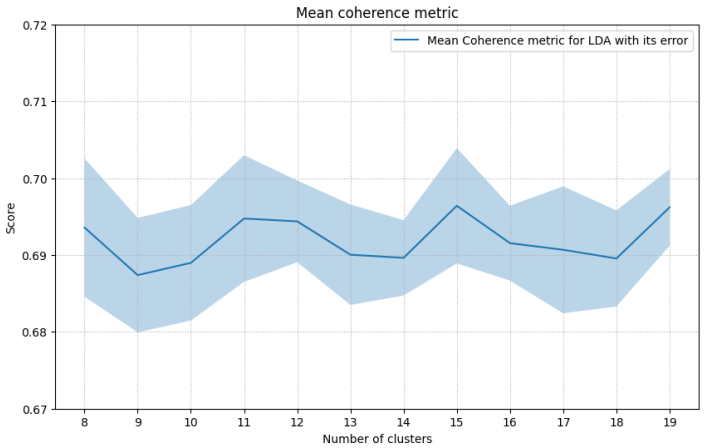

Figure 3: Mean coherence score vs number of clusters

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 4).

Cluster	C_{cv}
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

Table 4: LDA clustering result

In the next section, we apply the forecast model to clusters-outcomes after Embeddings + K-Means.

4.4 Forecasting

4.4.1 Prophet model

After applying the Embedding algorithm, we conducted forecasts for each result. The following table presents the final Prophet metrics for the forecasts for 30 days period 5:

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

Table 5: Resulted Prophet metrics.

Considering the complexity of the task, we find these results satisfactory.

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:

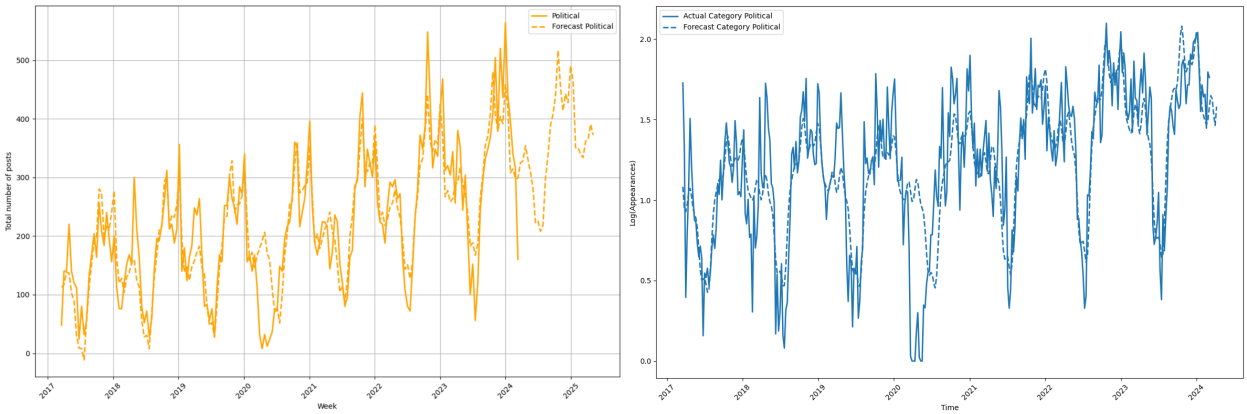

Figure 4: Prophet forecast for American football cluster.

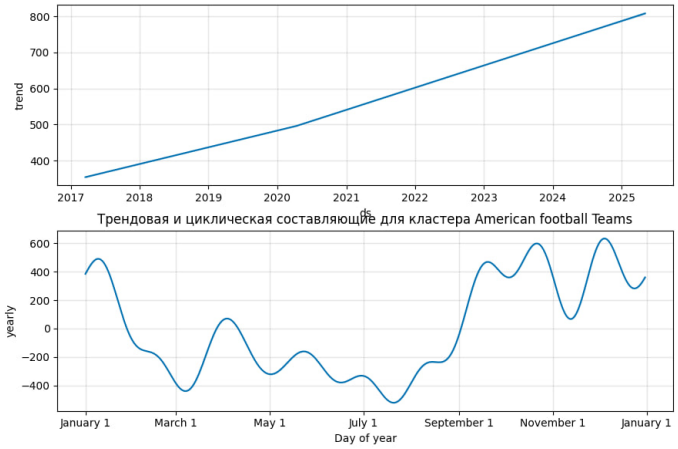

Figure 5: Prophet trend and cycle components for American football cluster.

Compared to the ARIMA and Exponential Smoothing models (Tables 6 and 7), the Prophet model shows better metrics, especially for the Political cluster, which demonstrates more complex temporal characteristics.

Cluster	Average MAE	Average MSE
Basketball players	0.413	0.240
Social events	0.390	0.226
Political	0.485	0.417
Business	0.497	0.379
Global News	0.445	0.325
American football Teams	0.615	0.532
Music	0.437	0.265
Motivational Day journaling	0.275	0.114
Football, Basketball, other Teams	0.527	0.403
Average	0.454	0.322

Table 6: Resulted ARIMA metrics.

Cluster	Average MAE	Average MSE
Basketball players	0.335	0.204
Social events	0.437	0.269
Political	0.514	0.482
Business	0.461	0.278
Global News	0.413	0.299
American football Teams	0.525	0.375
Music	0.459	0.255
Motivational Day journaling	0.058	0.008
Football, Basketball, other Teams	0.601	0.471
Average	0.422	0.293

Table 7: Resulted Exponential Smoothing metrics.

The ARIMA and exponential smoothing models demonstrated suboptimal forecasting accuracy for the "Political" cluster relative to our approach

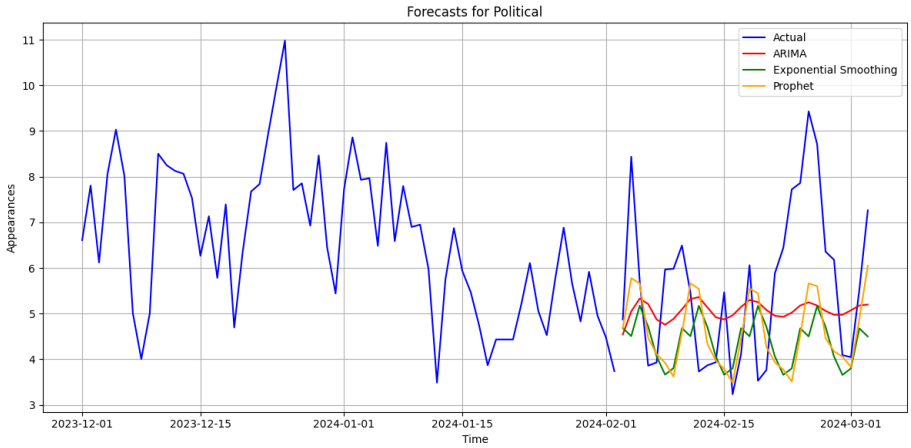

Figure 6: Forecast for 30 days.

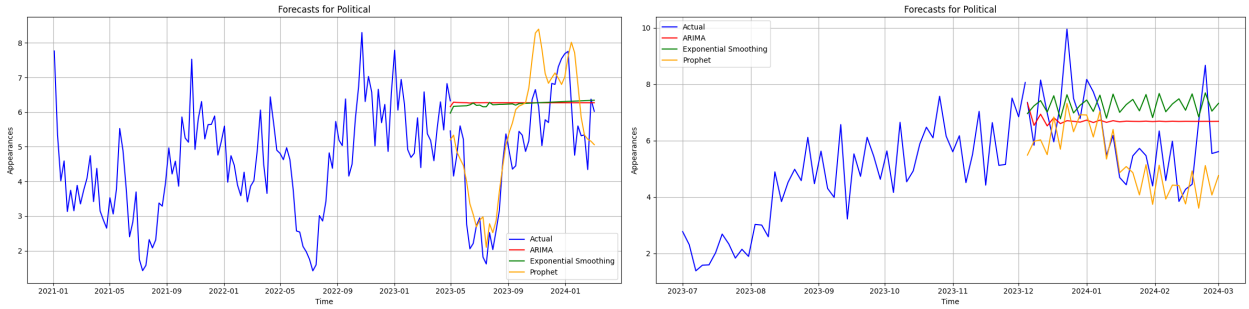

Figure 7: Forecast for 180 фтв 90 days.

This finding implies that the time series characteristics of the "Political" cluster, such as trend, seasonality, or other complex patterns , may not be adequately captured by the exponential smoothing or and ARIMA approaches for 180 or 90 or 30 days period.

5 Conclusion

By modeling the unique characteristics and life cycles of each thematic cluster, our framework is able to provide more nuanced and accurate predictions of future media trends. The model is able to capture unique patterns in the data that simpler models, such as ARIMA or Exponential Smoothing, do not account for.

In the appendix, we provide a detailed explanation of the calculation process and metrics for each cluster. You can find the code in the 'prophet_num.ipynb' notebook. We have also documented our experiments, successes, and failures in the 'research_Clustering_v3_LDA copy.ipynb' notebook.

However, the study also identified limitations in the forecasting performance for some clusters, which exhibited peculiarities such as abrupt peaks, trend shifts, and seasonal displacements. These anomalies in the time series presented challenges for traditional regression-based forecasting methods, including the Prophet model.

To address this issue, the study identified the need to explore existing algorithms specifically designed for anomaly detection and forecasting in time series data. The investigation and implementation of such algorithms are outlined in the "Future Research Directions" section, as they hold the potential to enhance the predictive capabilities of the proposed framework across a wider range of topic clusters.By leveraging the strengths of the Prophet model and addressing the limitations through the exploration of anomaly forecasting algorithms, this study lays the foundation for a robust and comprehensive framework for predicting trends in the dynamic and complex media landscape.

5.1 Future Research Directions

The key future research directions include:

- Refinement of Topic Modeling: Further investigation of topic modeling on longer text inputs, such as using a phrase-to-paragraph generation approach, to enhance the predictive capabilities of the proposed method.
- Anomaly Forecasting Algorithms: Implementation and evaluation of algorithms capable of accurately forecasting various anomalies in time series, such as abrupt peaks, trend shifts, and seasonal displacements.
- Multivariate Forecasting: Exploration of incorporating additional features, such as external events, sentiments, and demographic data, to improve the overall predictive performance.
- Differentiated Predictions for Social Groups: Implementation of the social group interest predictions, leveraging the unique relevance level metric based on the proximity of clusters on the dendrogram. This will provide more detailed and efficient prediction of future topics of conversation for specific audience segments.
- Cross-Domain Validation: Applying the proposed framework to different domains, such as scientific publication trends or product demand forecasting, to validate its generalizability and adaptability.

By addressing these future research directions, this study can contribute to the advancement of time series forecasting techniques, particularly in the context of complex, high-dimensional, and volatile media landscapes. The successful implementation of this approach can have significant practical implications for various industries, from marketing and media production to innovation research and development.

References

[1] R. H. Shumway, D. S. Stoffer, and D. S. Stoffer. Time series analysis and its applications, volume 3. Springer, 2000.

[2] J.-P. Bouchaud, J. Bonart, J. Donier, and M. Gould. Trades, Quotes and Prices, Financial Markets Under the Microscope. Cambridge University Press, 2018.

[3] A. V. Grabovoy and V. V. Strijov. Quasi-periodic time series clustering for human activity recognition. 2019.

[4] A. Das. Foundation of variational autoencoder (vae). 2020.

[5] Jérémie Rappaz, Dylan Bourgeois, and Karl Aberer. A dynamic embedding model of the media landscape. In The World Wide Web Conference, pages 1544–1554, 2019.

[6] Charles C Holt. Forecasting seasonals and trends by exponentially weighted moving averages. International journal of forecasting, 20(1):5–10, 2004.

[7] E. S. Gardner. Exponential smoothing: The state of the art. Journal of Forecasting, 1985.

[8] Hyunyoung Choi and Hal Varian. Predicting the present with google trends. Economic record, 88:2–9, 2012.

[9] Kristina P Sinaga and Miin-Shen Yang. Unsupervised k-means clustering algorithm. IEEE access, 8:80716–80727, 2020.

[10] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. Journal of machine Learning research, 3(Jan):993–1022, 2003.

[11] S. J. Taylor and B. Letham. Forecasting at scale. The American Statistician, 72(1):37–45, 2018.

[12] N. Reimers and I. Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. arXiv preprint arXiv:1908.10084, 2019.

[13] Sentence transformers: all-minilm-l12-v2 t.

[14] Michael Röder, Andreas Both, and Alexander Hinneburg. Exploring the space of topic coherence measures. In Proceedings of the eighth ACM international conference on Web search and data mining, pages 399–408, 2015.

[15] Twitter trending tweets.

[16] Youtube trending video dataset.