# PREDICT FUTURE SALE

**Zadvornov Egor**
MIPT
zadvornov.ev@phystech.edu

May 10, 2024

**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 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 [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.

**Task modeling.** We approach this task as a regression problem. For every item and shop pair, we need to predict its next month sales(a number).

**Construct train and test data.** In the Sales train dataset, it only provides the sale within one day, but we need to predict the sale of next month. So we sum the day's sale into month's sale group by item, shop, date(within a month). In the Sales train dataset, it only contains two columns(item id and shop id). Because we need to provide the sales of next month, we add a date column for it, which stand for the date information of next month.

# 3 Algorithms and Models

## 3.1 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 3.1.

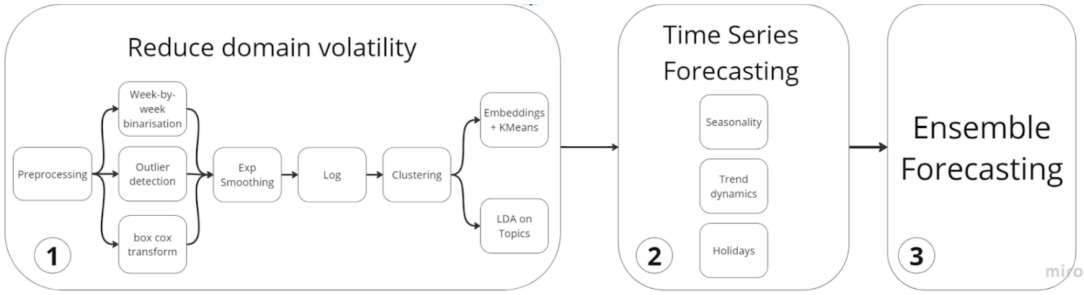


Figure 1: Image of the pipeline structure.

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.

## 3.2 Clustering

### 3.2.1 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: <rewrite as formula in latex with proper definitions> If | coherence embed - coherence LDA | < 0.05 AND 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 ... <why it is a proper choice>. The algorithm of coherence calculation is described at [add link!]. The main steps are <some summary of algorithm>.

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

### 3.2.2 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:

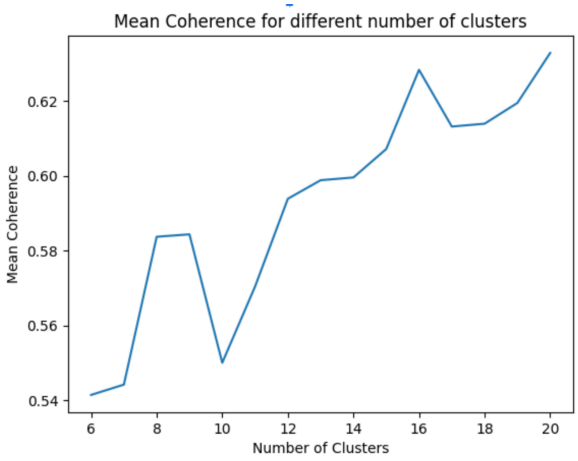


Figure 2: Coherence score vs. number of clusters for the Embeddings + K-Means algorithm.

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.

### 3.2.3 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.

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).

| 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 |

Table 2: Embed+Kmeans clustering result

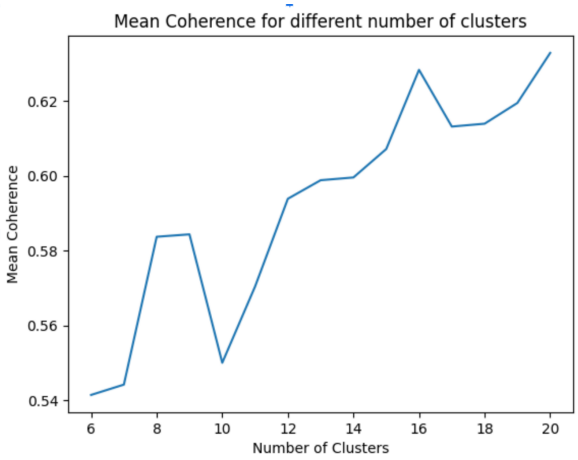


Figure 3: Mean coherence score vs number of clusters

#### 3.2.4 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.

### 3.3 Forecasting

#### 3.3.1 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

| 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 |

Table 3: LDA clustering result

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

Table 4: Resulted metrics of each algorithm. The error calculated by "error of the mean" formula.

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.

#### 3.3.2 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 |

Table 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:

## References

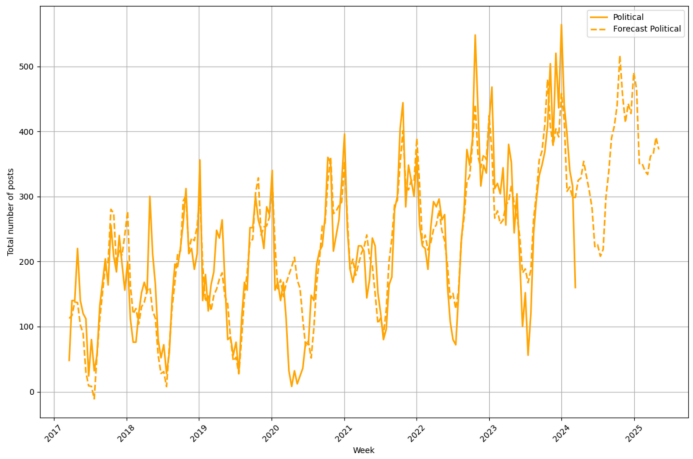


Figure 4: Prophet forecast for American football cluster.