

UNIVERSITY OF LONDON

INTERNATIONAL PROGRAMMES

BSc Computer Science and Related Subjects



CM3070 PROJECT

FINAL PROJECT REPORT

**Predictive Modeling of Social Media Trend Emergence
(Southeast Asia Food Travel Content)**

Author: Mah Chin Yang, Louis

Student Number: 220516916 (UOL), 10245951 (SIM)

Date of Submission: 22 September 2025

Supervisor: Yip See Wai

Contents

CHAPTER 1: INTRODUCTION	4
1.1 Background	4
1.2 Rationale for Platform Selection	4
1.3 Research Aim and Objectives	4
CHAPTER 2: LITERATURE REVIEW	6
2.1 Existing Research on Social Media Trend Prediction on YouTube	6
2.2 Use of topic modeling (LDA) for analyzing social media or YouTube Content	6
2.3 Use of Sentiment Analysis to measure viewer Response	7
2.4 Statistical Trend Detection used in social media and Marketing Analytics	7
2.5 Similar Projects	8
Summary	9
CHAPTER 3: PROJECT DESIGN	10
3.1 Domain and Target Users	10
3.2 High-Level Project Workflow	10
3.3 Tools and Technologies	13
3.4 Gantt Chart	15
3.4.1 Mid Term Timeline	15
3.4.2 Final Timeline Revised	16
3.5 Evaluation and Testing	17
CHAPTER 4: IMPLEMENTATION	18
4.1 Data Collection	18
4.2 Data Analysis and Modelling	18
4.3 Streamlit Dashboard for Trend Prediction	20
CHAPTER 5: EVALUATION	22
5.1 Evaluation of Predictive Modelling	22
5.2 Evaluation of Topic Modelling	23
5.3 Evaluation of Sentiment Analysis	23
5.4 Evaluation of Trend Detection	24
5.5 Usability and Interface Testing	24
5.6 Project as a Whole	25
5.7 Justification of Approach	25
5.8 Possible Extensions	26
5.9 Practical Implications for Creators	26

CHAPTER 6: CONCLUSION	27
CHAPTER 7: APPENDIX	28
7.1 Prediction Tool	28
7.2 Data Insights	29
7.3 Topic Analysis	30
7.4 Sentiment Analysis	31
7.5 Trend Patterns	32
7.6 Top Performers	33
7.7 Recommendations and Best Practices	34
CHAPTER 8: REFERENCES	35

CHAPTER 1: INTRODUCTION

Project Idea 2: Predictive Modelling of Social Media Trend Emergence

1.1 Background

Food travel content focusing on Southeast Asian cuisine represents a growing but competitive niche on YouTube. Content creators in this space face significant challenges when deciding which locations to visit, what food experiences to feature, and how to position their content for maximum audience engagement. These decisions carry substantial financial implications due to the costs of international travel, equipment, local fixers, and production time.

Currently, creators make these decisions based primarily on intuition, personal interest, or copying successful competitors without data-driven insights into the specific content elements drive trending performance. This leads to inefficient resource allocation, missed opportunities, and inconsistent channel growth. **Digitaltravelexpert (2025)**

This project addresses this gap by offering a predictive modeling approach that empowers Southeast Asian food travel creators to make informed decisions. By identifying the elements that influence a video's likelihood to trend, the study aims to optimize content strategies, increase engagement, and reduce costly guesswork.

1.2 Rationale for Platform Selection

YouTube serves as the optimal platform for this research due to its unique ecosystem that perfectly aligns with Southeast Asian cuisine content creation and its exceptional data accessibility for analytical purposes. The platform hosts a thriving community of Southeast Asian food travel creators who have built substantial followings through long-form storytelling that captures the cultural nuances, cooking techniques, and authentic dining experiences across the region content that requires the extended format YouTube provides compared to the brief clips on TikTok or Instagram Reels.

YouTube's comprehensive and publicly accessible API offers unprecedented access to granular data including view counts, engagement rates, comment sentiment, video tags, descriptions, upload timestamps, and demographic analytics, making it exceptionally well-suited for machine learning analysis of content performance patterns.

1.3 Research Aim and Objectives

Research Aim

To model the performance drivers of Southeast Asian food travel content on YouTube, enabling creators to make data-informed strategic decisions through descriptive analytics and predictive modeling.

Research Objectives

To achieve this aim, the project will pursue the following objectives:

- Identify the key metadata and thematic factors that influence the performance of Southeast Asian food travel videos on YouTube.
- Predict the likelihood of content trending based on historical performance patterns and audience engagement indicators.
- Generate actionable insights that creators can use to improve content decisions, such as optimal posting times, popular themes, and emotionally resonant narratives.

Chapter 1 Word Count: 392

CHAPTER 2: LITERATURE REVIEW

This chapter reviews relevant academic research, applied methodologies, and existing platforms to establish the foundation for predictive modeling of social media trend emergence, particularly within the YouTube ecosystem. It explores five key areas that directly inform the design and justification of this study.

2.1 Existing Research on Social Media Trend Prediction on YouTube

Recent research has demonstrated the potential of machine learning models to predict trending content on YouTube by analyzing metadata features. **Liu (2023)** conducted an exploratory study focused on trending YouTube videos in the U.S., examining variables such as views, likes, dislikes, comments, and tags. The study found that machine learning classifiers, including Random Forest, were effective in identifying patterns that differentiate trending videos from non-trending ones. This highlights the viability of using metadata alone without delving into video content itself to forecast the likelihood of virality on the platform. The study's findings suggest that data-driven predictive modeling can offer valuable insights for creators looking to optimize their content strategies and allocate resources more efficiently.

In a related study, **Niture (2021)** analyzed over 40,000 YouTube trending videos, comparing several classification algorithms such as Random Forest, Support Vector Machines, Decision Trees, Logistic Regression, and Naïve Bayes. The research concluded that Random Forest provided the best balance between accuracy and reliability in predicting whether a video would enter the Trending list and how long it would remain there. This reinforces the suitability of Random Forest models for social media trend prediction, especially when applied to metadata-rich datasets. The robust performance of this model underlines its potential as an effective tool for forecasting trends in niche content areas like Southeast Asian food travel Source.

These studies directly support the use of metadata-based prediction models in this project. They justify the selection of Random Forest as the primary algorithm to forecast whether Southeast Asian food travel content is likely to trend, providing a tested framework that this project will adapt to a specific regional and thematic niche

2.2 Use of topic modeling (LDA) for analyzing social media or YouTube Content

Efforts to uncover hidden thematic structures within user-generated content on social media have increasingly relied on Latent Dirichlet Allocation (LDA), demonstrating its effectiveness in both breadth and depth of content mining. A systematic review by **Doogan et al. (2023)** analyzed 189 studies applying topic models to short-form social media text, including platforms such as Twitter and YouTube. The review affirmed that LDA remains one of the most widely adopted approaches for revealing latent themes in online discourse. Despite noting implementation challenges such as inadequate parameter tuning or limited interpretability, the study concluded that, when applied rigorously, LDA is capable of extracting coherent and meaningful topics even from brief and informal text entries. This makes it particularly well-suited for analyzing content elements such as YouTube titles and video descriptions.

In a YouTube-specific application, **Daniel, C (2017)** applied an LDA-based methodology to analyze video transcripts, demonstrating the model's capacity to extract latent topics related to narrative structure, tone, and content style across a wide range of channels. The study illustrated how LDA could be used to surface recurring themes and storytelling trends in large video libraries, even without predefined content categories or taxonomies.

This validates the use of LDA in this project for analyzing video titles and descriptions to identify recurring themes such as specific cuisines, storytelling styles, or locations that are more likely to contribute to higher engagement and trending potential. These thematic insights will inform the classification model and content recommendations.

2.3 Use of Sentiment Analysis to measure viewer Response

Sentiment analysis has emerged as a widely used method for measuring viewer response to online content, particularly on platforms like YouTube. Research has shown that user sentiment expressed in video comments can correlate strongly with content engagement and virality. A study by **A S and Rajeev (2024)** analyzed YouTube comment sentiment using various machine learning algorithms, including Naïve Bayes, Support Vector Machines, Decision Trees, and Random Forest. The study found that fluctuations in viewer sentiment were often aligned with spikes in video relevance and engagement, indicating that comment sentiment can serve as a valuable signal in understanding how audiences emotionally interact with content and how that interaction influences popularity trends.

The tools used to perform sentiment analysis vary in complexity and suitability depending on the scope and depth of analysis. Lexicon-based tools such as VADER are known for their efficiency and effectiveness in handling social media-style text, making them well-suited for quick, large-scale sentiment classification. In contrast, transformer-based models like BERT and its variant RoBERTa provide more nuanced semantic understanding but are computationally intensive. As noted in an article by **Abdulla, A (2022)**, VADER is useful for real-time or large-batch processing where speed is essential, while BERT excels in contexts where linguistic subtleties must be captured for deeper insight.

This research informs this project's decision to integrate sentiment analysis into its predictive framework. By analyzing comments on Southeast Asian food travel videos, the project will assess how viewer sentiment relates to content success, supporting the development of a richer, emotionally-aware trend prediction model.

2.4 Statistical Trend Detection used in social media and Marketing Analytics

Statistical methods such as moving averages and Z-score analysis have long been pivotal in identifying emerging trends and anomalies across various domains, including marketing and media analytics. A comprehensive study on online media streams by **Althoff et al. (2014)** demonstrated that trending topic detection can be enhanced through analysis of time-series data patterns. By employing methods that smooth historical data and detect deviations, the authors successfully forecasted the life cycle of trending topics across platforms like Twitter and Wikipedia. This work validates the use of rolling averages and trend detection techniques to

anticipate peak engagement periods and understand content momentum. These approaches form the foundation for adaptive content strategies in dynamic online environments

Complementing this, **Zhang et al. (2020)** described the deployment of a “Moving Metric Detector” at eBay, designed to flag anomalous patterns among thousands of performance metrics. Their two-phase anomaly detection system first employs moving averages to establish baseline trends and then uses statistical thresholds similar to Z-score analysis to detect significant deviations. This practical implementation underscores the real-world utility of combining smoothing techniques with standardized deviation measures to surface actionable insights from noisy data streams, an approach analogous to detecting sudden spikes in content consumption or audience interest.

These findings justify the inclusion of statistical trend detection in this project as a means to identify emerging interest in specific cuisines, locations, or keywords. By applying these techniques to YouTube metadata, this project will detect early surges in attention offering creators a timely advantage in trend-responsive content planning.

2.5 Similar Projects

Existing analytics platforms, such as Social Blade and vidIQ, demonstrate the practical viability of data-driven analysis for understanding content performance, which supports the broader concept of predictive modeling in video content domains.

Social Blade is a free, publicly accessible analytics website that aggregates and displays longitudinal statistics for YouTube (as well as Twitch, Instagram, and TikTok). It provides creators with insights into subscriber growth, daily view counts, and weekly performance trends. By making trends visible over time, Social Blade enables users to infer momentum shifts and channel trajectory effectively helping them identify rising popularity patterns without manual data collection. **Social Blade (2025)** This exemplifies how readily available trend data can empower content producers to understand audience behavior and adjust content strategies accordingly.

Similarly, vidIQ offers a suite of analytics tools focused specifically on YouTube optimization. As a Chrome extension and web platform, vidIQ delivers metadata insights, SEO recommendations, and competitive benchmarking, alongside performance charts for views, tags, and engagement metrics. **Vidiq (2024)** The platform’s success in aiding creators to improve video SEO, discoverability, and topic relevance shows that structured data about titles, tags, and metadata directly contributes to optimizing content performance. vidIQ’s algorithmic analysis of metadata parallels academic methodologies in topic and sentiment modeling, demonstrating real-world demand for similar analytical insights.

These platforms underscore two key points: first, that trend detection and metadata analysis tools are valuable and widely used by creators; second, that providing actionable insights whether via trend indicators, SEO suggestions, or metadata evaluation addresses a clear market need. The effective use of performance data, content metadata, and trend visualization on these existing platforms suggests that extending trend analysis through methods like machine learning, topic modeling, and statistical anomaly detection can further enhance predictive power. Thus, these

live tools validate the underlying approach: systematically leveraging analytics to guide strategic decisions in niche content creation areas.

These tools validate the core premise of this project: that creators need accessible, insight-driven analytics. However, unlike general-purpose tools, this study targets a specific niche Southeast Asian food travel offering deeper thematic and predictive modeling capabilities tailored to this segment.

Summary

The reviewed literature confirms the viability of metadata-driven trend prediction using Random Forest classifiers, the value of LDA in uncovering latent content themes, the predictive relevance of sentiment analysis, and the effectiveness of statistical trend detection methods. In parallel, existing platforms like Social Blade and vidIQ demonstrate the market demand for such insights. Together, these works justify the design of a predictive modeling framework aimed at helping Southeast Asian food travel creators make data-informed decisions on YouTube.

Chapter 2 Word Count: 1518

CHAPTER 3: PROJECT DESIGN

3.1 Domain and Target Users

This project operates within the domain of **social media analytics for digital content strategy**, with a specific focus on **Southeast Asian food travel content on YouTube**. The project sits at the intersection of content marketing, media trend analysis, and data-driven decision-making for video creators. It leverages predictive modeling, natural language processing, and statistical trend detection to support content creators operating in a culturally rich and visually engaging niche.

The **primary users** of this project are:

- **Southeast Asian food travel YouTube creators**, including individual vloggers, small production teams, and travel influencers. These users often lack access to customized analytical tools that help them predict content performance and allocate resources efficiently.
- **Content strategists and social media marketers** who specialize in culinary or travel media in Southeast Asia. These professionals can use the insights generated to plan campaigns or partnerships.
- **Aspiring creators and niche YouTube channels** seeking to grow their audience and improve content relevance through data-backed strategies.

These users share common challenges: they face high production costs, fierce competition, and limited access to actionable audience insights. This project addresses these pain points by providing a specialized tool that predicts the likelihood of a video trending based on its metadata, thematic content, audience sentiment, and emerging topic trends. By tailoring the model specifically to Southeast Asian cuisine and travel content, the project offers a focused, high-value alternative to generic analytics platforms.

3.2 High-Level Project Workflow

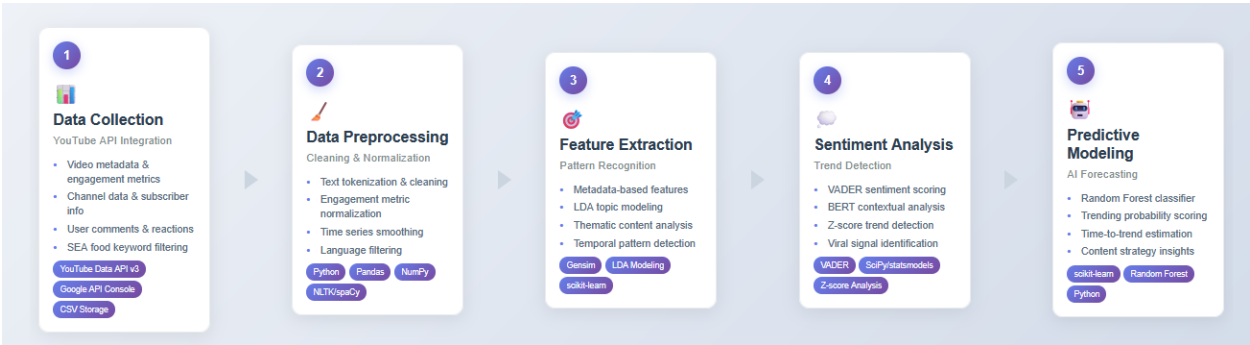


Figure 1. Workflow

This project is designed to predict the trending potential of Southeast Asian food travel videos on YouTube by analyzing key performance indicators, thematic content, and audience sentiment. At a high level, the workflow is structured into five major stages:

1. Data Collection

The project begins with data gathering through the **YouTube Data API**, focusing on:

- Video metadata (title, tags, descriptions, upload date)
- Engagement metrics (views, likes, comments)
- Channel data (subscriber count, country of origin)
- Viewer responses (comments for sentiment analysis)

Only videos tagged or described with Southeast Asian food-related terms (e.g., “Thai street food,” “Vietnamese cuisine,” “Malaysian hawker”) are included in the dataset to maintain relevance.

2. Data Preprocessing

Once raw data is collected, it is cleaned and prepared for analysis:

- Text fields are preprocessed (tokenization, stopword removal, lemmatization)
- Engagement metrics are normalized to account for outliers and channel size
- Time-series data is smoothed for trend detection
- Comments are filtered for language consistency (e.g., English only) for sentiment analysis

3. Feature Extraction and Thematic Analysis

Two key types of features are extracted:

- **Metadata-based features** such as upload frequency, length, posting time, engagement ratios
- **Thematic features** using **LDA topic modeling** applied to titles and descriptions to reveal recurring themes (e.g., specific dishes, cultural elements, location types)

This stage identifies content patterns commonly associated with high performance.

4. Sentiment and Trend Analysis

- **Sentiment Analysis** is performed on user comments using tools like **VADER** (or BERT for deeper insight if needed), allowing the model to capture emotional response trends.
- **Statistical Trend Detection** using Z-score analysis and rolling averages identifies early signs of momentum or topic virality.

These insights capture the social and temporal dynamics of video performance.

5. Predictive Modeling and Insight Generation

Using the processed data, a **Random Forest classifier** is trained to predict the likelihood of a video trending. The model is evaluated using accuracy, precision, and recall metrics.

- Predictions include trending probability, time-to-trend estimates, and content alignment with past successful patterns.
- Output insights are translated into **practical content strategy recommendations**, such as:
 - Which themes are gaining popularity
 - Optimal posting times and locations
 - Emotional tone alignment with audience preferences

The design of this project is tailored to the specific needs of Southeast Asian food travel content creators on YouTube, with each methodological choice grounded in domain relevance and practical usability. Central to the approach is the use of metadata-based analysis. Since metadata such as video tags, upload time, and engagement ratios are readily available and standardized across videos, this method provides an accessible and low-barrier way for creators to receive insights without needing technical expertise or raw video analysis.

To capture the thematic nature of food travel content, Latent Dirichlet Allocation (LDA) is used to extract recurring topics from video titles and descriptions. This helps identify which cuisines, locations, or storytelling styles correlate with higher engagement, guiding creators on what narratives resonate most.

Recognizing that audience interaction is emotional as well as numerical, sentiment analysis of video comments is incorporated to translate viewer feedback into actionable signals. Tools like VADER are employed to capture shifts in viewer tone, helping creators fine-tune content to align with audience sentiment.

Additionally, statistical trend detection techniques, such as Z-score and moving averages, are used to identify emerging interest in specific topics. This supports agile content planning in a fast-paced, trend-sensitive environment.

A Random Forest model was selected for trend prediction due to its proven reliability in handling mixed data types, strong predictive accuracy, and interpretability qualities essential for users who must trust and act on model outputs. Creators can understand which factors most contribute to their content's trending potential, enabling informed decision-making.

Finally, the focus on YouTube reflects its unmatched metadata access, long-form format compatibility, and popularity among Southeast Asian creators. This platform-specific design ensures insights are directly applicable, enhancing the project's effectiveness and relevance to its target users.

3.3 Tools and Technologies

This project integrates a range of tools tailored for data collection, analysis, modeling, and user interaction. The tools are selected based on their accessibility, efficiency, and alignment with the project's focus on social media trend prediction for YouTube.

Achieving the Aims and Objectives

To achieve the research objectives identifying performance-driving factors, predicting trending potential, and generating actionable insights the project adopts a **modular, data-driven workflow**. This combines metadata analysis, thematic modeling, sentiment detection, and statistical trend identification to generate a well-rounded predictive model. The tools and techniques below are chosen because they:

- Handle mixed data types common in YouTube metadata (e.g., text, timestamps, counts),
- Support scalable, interpretable modeling (e.g., Random Forest),
- Enable thematic insight extraction (e.g., LDA topic modeling), and
- Capture emotional engagement and temporal trends efficiently (e.g., VADER, Z-score detection).

This strategy ensures that content creators can benefit from **explainable, low-barrier-to-entry insights** directly aligned with the practical needs of non-technical users in the Southeast Asian food travel niche.

1. Data Collection & Storage

- **YouTube Data API v3**
To fetch video metadata (titles, tags, descriptions), engagement metrics (views, likes, comments), channel info, and comment threads.
- **Google API Console**
For generating API keys and managing API quotas.
- **CSV File Format (Comma-Separated Values)**
All retrieved data is stored locally in .csv format for ease of manipulation, sharing, and reproducibility during the data analysis process.

2. Data Analysis & Machine Learning

- **Python (Programming Language)**
Primary language for all data processing and modeling tasks.
- **Pandas**
For dataset cleaning, transformation, and manipulation.
- **NumPy**
For efficient numerical operations.

- **scikit-learn**
For machine learning tasks, especially:
 - **Random Forest** classification model
 - Evaluation metrics (accuracy, precision, recall)
- **NLTK / spaCy**
For natural language preprocessing (tokenization, stopwords removal, lemmatization).
- **Gensim**
For **LDA topic modeling** to identify thematic clusters in video content.
- **Matplotlib / Plotly / Seaborn**
For data visualization, trend graphs, and feature importance plots.

3. Sentiment Analysis

- **VADER (Valence Aware Dictionary and sEntiment Reasoner)**
For fast, lexicon-based sentiment analysis of YouTube comments.
- *(Optional)* **Hugging Face Transformers (BERT, RoBERTa)**
For deep, contextual sentiment analysis if higher semantic accuracy is required.

4. Statistical Trend Detection

- **SciPy / statsmodels**
For implementing **Z-score** calculations and **moving averages** to identify trend anomalies.

5. Testing & Evaluation

- **Jupyter Notebooks / VS Code**
For prototyping, testing, and documenting the data analysis pipeline.
- **scikit-learn's train_test_split and cross_val_score**
For evaluating model performance and generalization.

6. User Interface

- **Streamlit**
For creating a simple web app interface for users to input video metadata and receive trend predictions.
- **CSV Upload Module (Streamlit, Not yet implemented)**
Enables users to input their own content data (e.g., upcoming video metadata) and receive analysis outputs through an accessible dashboard.

7. Reporting and Documentation

- **Microsoft Word**
For compiling findings, project documentation, and generating the final written report.

3.4 Gantt Chart

Below you will find both Gantt Charts for the Mid Term and Final Timeline.

3.4.1 Mid Term Timeline

No.	Task		Week 1	Week 2	Week 3	Week 4	Week 5	Week 6	Week 7	Week 8	Week 9	Week 10
1	Selecting an idea from project template	Plan										
		Actual										
2	Researching on the market needs, opportunity and competitors	Plan										
		Actual										
3	Identifying what possible tools could be used for this project	Plan										
		Actual										
4	Preparing Project Proposal Video	Plan										
		Actual										
5	Starting the Report by Conducting a Background Research on the Topic	Plan										
		Actual										
6	Researching Articles Related to Data Analysis using YouTube API	Plan										
		Actual										
7	Writing the Report for Literature review	Plan										
		Actual										
8	Identifying the Domain and Target User	Plan										
		Actual										
9	Identifying the Necessary Techniques and Tools needed	Plan										
		Actual										
10	Creating a High-Level Project Workflow Diagram	Plan										
		Actual										
11	Writing the Report for Project Design	Plan										
		Actual										
12	Coding The Feature Prototype	Plan										
		Actual										
13	Writing The Report for Feature Prototype	Plan										
		Actual										
14	Citing and Formatting the report	Plan										
		Actual										

Figure 2. Mid Term Gantt Chart

Explanation:

The initial phase of the project (Weeks 1–4) proceeded according to plan, with key activities such as topic research and the development of the project proposal video completed on time. This stage established a solid foundation for the rest of the project.

From Week 5 onwards, delays emerged particularly in the literature review and article research tasks. Following supervisor feedback, the project objectives were found to be unclear, and the literature lacked sufficient justification. This led to a complete revision of both the objectives and the literature structure, requiring additional time to identify relevant sources and improve alignment with the project aim.

The changes in objectives also affected related tasks, including domain and user identification, tool selection, and workflow design. These tasks were extended into Weeks 7 and 8 to accommodate the revised project direction.

In Weeks 9 and 10, final-phase tasks such as the project design report, prototype development, and final documentation are underway. Although the schedule is tighter due to earlier revisions,

the updated framework has strengthened the overall focus and coherence of the project. Completion remains on track.

3.4.2 Final Timeline Revised

No.	Task		Week 11	Week 12	Week 13	Week 14	Week 15	Week 16	Week 17	Week 18	Week 19	Week20	Week 21
15	Metadata Collection(YouTube API)	Plan											
		Actual											
16	Data Cleaning and Preprocessing	Plan											
		Actual											
17	Feature Extraction(Metadata + LDA)	Plan											
		Actual											
18	Sentiment Analysis on Comments	Plan											
		Actual											
19	Statistical Trend Detection	Plan											
		Actual											
20	Model Training & Evaluation(Random Forest)	Plan											
		Actual											
21	Streamlit Interface Development	Plan											
		Actual											
22	Testing with Sample Metadata	Plan											
		Actual											
23	Testing with Users	Plan											
		Actual											
24	Final Adjustments on Final Prototype	Plan											
		Actual											
25	Final Report Writing	Plan											
		Actual											

Figure 3. Final Gantt Chart

Explanation:

Metadata collection and preprocessing were completed on schedule in Weeks 11–12, establishing a clean dataset for subsequent analysis. Weeks 13–14 focused on feature extraction and thematic analysis, identifying both metadata-driven factors and narrative themes relevant to Southeast Asian food travel content. This was followed by sentiment analysis and statistical trend detection in Weeks 14–15, which added emotional and temporal dimensions to the dataset. Model training and evaluation took place in Weeks 15–16, where the Random Forest classifier was developed, cross-validated, and benchmarked, ensuring reliable predictions of trending potential.

The next stage, development of the Streamlit interface, began in Weeks 17–18 but required more time than anticipated. Integrating multiple features, refining the layout, extended this task into later weeks. Similarly, user testing in Weeks 19–20 also took longer. Despite these delays, the timeline absorbed the extensions without major disruption. The final stages in Weeks 20–21 were devoted to adjustments and report writing, producing a functional system supported by critical evaluation.

3.5 Evaluation and Testing

The first dimension will be **trend prediction and practical relevance testing**. The Random Forest model will be trained and tested on the collected dataset to assess its ability to distinguish between trending and non-trending videos. Although the dataset contains only a small proportion of trending cases, the evaluation will focus on whether the system provides insights that are actionable for creators rather than abstract metrics alone. Thresholds for classifying videos as trending will be reviewed to ensure that they align with realistic engagement benchmarks. In addition, thematic and sentiment correlation checks will be conducted. The outputs of LDA topic modelling and sentiment analysis will be compared against the model's trending predictions to confirm that high-potential videos correspond to audience interests, such as popular themes and positive sentiment. This dimension ensures that the evaluation speaks directly to practical usefulness in the context of YouTube strategy.

The second dimension will be **machine learning model testing**. The dataset will be split into training and testing subsets, and 5-fold cross-validation will be applied to check the stability of results and minimise overfitting. Performance will be measured using accuracy, precision, recall, and F1-score, which together provide a balanced view of classification ability. Feature importance plots will be examined to confirm that the model relies on meaningful variables such as engagement ratios, posting times, and textual features. While SHAP values and baseline comparisons with simpler models such as Logistic Regression or Decision Trees were initially considered, the primary emphasis will be on validating the Random Forest as the most appropriate balance between accuracy and interpretability. Hyperparameter tuning through grid search will also be applied to ensure that the model is operating with optimised parameters. This dimension will confirm that the machine learning approach is technically sound, interpretable, and reliable.

The final dimension will be **usability and interface testing** of the Streamlit dashboard. Functional testing will ensure that core features including CSV upload, prediction generation, and visualisations work as expected. Usability will be assessed informally by having test participants interact with the dashboard, upload sample datasets, and interpret the outputs. Their feedback will inform adjustments to clarity and ease of use. In addition, performance testing will involve uploading datasets of different sizes to check that the application remains responsive and stable. The goal of this dimension is to ensure that the system is not only accurate but also accessible to non-technical users in practical scenarios.

CHAPTER 4: IMPLEMENTATION

4.1 Data Collection

The first stage of the project revolves around the collection of data from YouTube, which forms the backbone of the entire predictive modelling workflow. This step is crucial because the quality, quantity, and relevance of data directly influence the accuracy of the subsequent analysis and modelling processes. To achieve this, the code integrates with the YouTube Data API v3, a widely used interface that allows developers to access granular video information including titles, descriptions, tags, upload timestamps, view counts, likes, comments, and channel statistics. By leveraging this API, the program is able to systematically gather metadata across a wide selection of Southeast Asian food travel videos, ensuring that the dataset reflects the diversity of cuisines, cultures, and content strategies within the niche.

The data collection script was implemented in Python and designed with modularity in mind. One of the most important components is the authentication and connection to the Google API console, which secures access to the API while also managing quota limits. The code specifies search parameters such as keywords (“Thai street food,” “Vietnamese cuisine,” “Malaysian hawker”) and filters results to only include videos relevant to the project domain. Through iterative API requests, the script extracts details for each video and stores them in a structured format, specifically CSV files. Using CSV not only ensures compatibility with most analysis libraries but also provides a lightweight and reproducible storage option, which is particularly valuable for future experiments or replication.

Another important feature of the script is its handling of pagination and quota constraints. Since the YouTube API enforces daily limits, the code incorporates functions to check available quota before issuing further requests, thereby preventing abrupt interruptions. Pagination handling ensures that larger datasets are fetched by cycling through multiple request pages until all relevant entries are retrieved. This systematic approach eliminates gaps in the dataset and allows for comprehensive coverage.

The collected dataset is not limited to numeric values such as view counts but also incorporates text-based features such as video titles, tags, and comment snippets. These elements are especially significant because they will later undergo natural language preprocessing and thematic modeling. By including textual and numerical data from the outset, the project guarantees a rich, multimodal foundation for trend prediction.

4.2 Data Analysis and Modelling

The data analysis stage represents the core intellectual contribution of the project, transforming raw metadata collected from YouTube into meaningful insights that can be used to predict content performance and detect emerging trends. Unlike the data collection phase, which is largely procedural, the analysis component combines a range of statistical, natural language processing, and machine learning techniques in order to extract knowledge from a complex and multimodal dataset. In this stage, several algorithms were deployed, each addressing a different

dimension of the data: thematic structure, sentiment orientation, engagement dynamics, and predictive modelling.

One of the most important techniques used is **topic modelling**, implemented through Latent Dirichlet Allocation (LDA). LDA is a generative probabilistic model that assumes each video description or title can be expressed as a mixture of latent topics, with each topic being represented by a distribution over words. By training the model on a corpus of video descriptions and comments, the code identifies coherent themes such as “street food markets,” “traditional cooking,” or “luxury dining experiences.” The algorithm outputs topic–word distributions and document–topic distributions, which can then be visualised through tools such as pyLDAvis. These visualisations provide intuitive two-dimensional plots of topic clusters, allowing users to see how themes overlap or diverge. For this project, topic modelling is particularly important because the success of a video is not determined solely by metadata such as view counts but also by how it resonates with popular thematic narratives in the food travel domain.

Another central component of the pipeline is **sentiment analysis**, conducted primarily with the VADER sentiment analyser. VADER is a lexicon and rule-based tool optimised for social media text, which classifies sentences into positive, neutral, and negative orientations while also producing compound polarity scores. By applying VADER to video comments, the program is able to gauge audience emotional reactions at scale. These sentiment scores are aggregated and compared against engagement metrics such as likes and shares, enabling correlation analysis between audience sentiment and video performance. While the project notes that advanced models such as BERT could, in future work, provide deeper contextual analysis of longer comments, BERT was not implemented in the current pipeline. The sentiment analysis results are typically visualised in the form of bar charts showing proportions of positive and negative sentiment across channels, or scatterplots correlating average sentiment scores with view counts. Such visual representations make abstract text features more tangible and highlight whether emotionally positive content tends to perform better in the Southeast Asian context.

For **predictive modelling**, the code employs ensemble methods, particularly the **Random Forest classifier**. Random Forests are an extension of decision trees, building multiple trees on bootstrapped samples of the dataset and combining their predictions through majority voting. This algorithm is especially suitable for heterogeneous data like YouTube metadata because it can handle both numerical and categorical variables, is robust against overfitting, and provides measures of feature importance. In the project, the Random Forest model is trained to classify whether a video is likely to become a “trend” based on features including video duration, like-to-view ratio, comment rate, publishing time, detected topics, and aggregated sentiment. Hyperparameters such as the number of estimators and maximum depth are tuned through grid search and cross-validation. The results are evaluated against baseline models like logistic regression and single decision trees, with metrics such as accuracy, precision, recall, and F1-score being reported. Visualisations for this stage include ROC curves, confusion matrices, and feature importance plots, each serving to demonstrate both model performance and interpretability.

Another noteworthy feature of the analysis script is its emphasis on **time-series trend detection**. Using statistical techniques such as moving averages and z-score calculations, the code attempts

to identify anomalies in view trajectories. For example, a sudden spike in views compared to the baseline average can signal the early stages of virality. This functionality is not only useful for retrospective analysis but also for informing predictive models about emerging popularity patterns. Graphical outputs in this part typically take the form of time-series plots with trend lines, highlighting where anomalies occur and how they develop over time. Such visual representations are essential for communicating temporal dynamics to creators, who can then make better decisions about content timing and promotion.

The code also places importance on **explainability and interpretability**, recognising that creators and strategists may not have technical expertise in machine learning. This is addressed through the integration of feature importance measures such as SHAP (SHapley Additive exPlanations) values, which break down how each feature contributes to individual predictions. For instance, SHAP visualisations can illustrate that the model predicts a video to have high trending potential primarily due to its strong sentiment score and association with a “street food” topic cluster, while its longer duration contributes negatively. By providing this level of transparency, the project avoids the “black box” critique often levelled at machine learning models, ensuring that insights remain actionable for non-technical users.

4.3 Streamlit Dashboard for Trend Prediction

The Streamlit script serves as the user-facing layer of the project, transforming the outputs of the modelling pipeline into an interactive decision-support tool for Southeast Asian food travel creators. Unlike the previous stages where data collection and algorithmic modelling were central, the `sea.py` file focuses on orchestrating the machine learning pipeline, extracting features from user input, and presenting insights in an accessible way. In doing so, it bridges the technical core of the project with the practical needs of end users who may lack coding expertise.

A key function of the dashboard is the integration of the **machine learning pipeline** trained earlier in the project. The code loads a pre-saved model pipeline, which includes both the Random Forest classification model and optional regression models for estimating time-to-trend . This integration allows predictions to be run instantly without retraining. To make predictions, the script uses the `extract_features_from_input` function, which translates user-entered metadata such as video title, description, duration, and tags into structured numerical features. These include title length, number of tags, publishing hour and day, presence of exclamation or question marks, and cuisine-specific flags such as whether the video references Thai or Vietnamese cuisine . By embedding these features, the system ensures that predictions are grounded in both linguistic and contextual cues derived from the metadata.

The **prediction process** itself is encapsulated in the `make_prediction` function. This function feeds the extracted features into the loaded Random Forest model and outputs a probability that the video will trend. It also generates a binary classification indicating whether the content is likely to trend or not, and, if available, provides an estimate of the number of days until trending. Importantly, the prediction step does not end with raw numbers. The script immediately passes the outputs into a recommendation engine, which generates practical advice for optimization. For example, if trending probability is low, the system may suggest adding more descriptive tags,

extending the video title, or adjusting publishing times to peak hours . This recommendation layer demonstrates how algorithmic outputs are contextualised into actionable strategies. From a code perspective, one of the most important sections is the **user input interface** on the “Predict” tab. This area uses Streamlit widgets such as `st.text_input`, `st.text_area`, and `st.slider` to capture video details. The combination of these inputs ensures that even novice users can enter structured metadata. When the prediction button is clicked, the application triggers the prediction pipeline and displays results. Results are communicated both numerically, through metrics such as trending probability and confidence levels, and visually, through Plotly-based gauges and charts that provide intuitive understanding of the likelihood of trending .

The dashboard also incorporates **visualisation of analytical results** across multiple tabs. In the “Data Insights” section, histograms of view counts and engagement scores reveal the overall distribution of performance across the dataset. The “Topic Analysis” tab draws on Latent Dirichlet Allocation outputs to show the relative share of themes such as street food, restaurant reviews, or cultural experiences, often using pie charts or bar charts for intuitive communication. Likewise, sentiment analysis is presented with pie charts and bar charts to show the distribution of positive, neutral, and negative comments, and how these sentiments correlate with average views and trending rates. Temporal trend patterns are also visualised by aggregating performance across days of the week, publishing hours, and months, providing concrete evidence of optimal posting strategies .

Chapter 4 Word Count: 1760

CHAPTER 5: EVALUATION

5.1 Evaluation of Predictive Modelling

The predictive engine uses a **Random Forest classifier** trained on engineered features spanning engagement signals (raw counts and short-term trend features), channel-level aggregates, content/topic and sentiment features, lexical/length metrics, and temporal descriptors. The dataset contains **1,466 videos**, of which **25** are labelled as trending ($\approx 1.7\%$). A **stratified** train-test split was adopted to preserve the rare positive class in both splits. Class imbalance was addressed with **class_weight='balanced'** inside **GridSearchCV**, combined with **5-fold StratifiedKfold** cross-validation to stabilise selection of hyperparameters (including tree count, depth/leaf criteria, and splitting function).

On the held-out test split, the model reports **Accuracy = 0.993**, **ROC AUC = 0.999**, and **F1 (trending) ≈ 0.80** , with training accuracy close to **0.999**. These are extraordinary headline figures. However, because trending is a **rare** outcome, accuracy alone is an unreliable indicator: a naïve majority classifier would already exceed 98% accuracy by predicting “not trending” for everything. The use of AUC, F1, and a confusion matrix is therefore essential to interrogate performance beyond accuracy.

The **confusion matrix** for the test set shows **TN = 288**, **FP = 1**, **FN = 1**, **TP = 4**. This conveys two points. First, the classifier is extremely conservative with false positives: only one non-trending video was incorrectly flagged. Second, while the model successfully identifies some trending items (four true positives), it still **misses** at least one (a false negative). Given the tiny absolute number of positives, even a single error materially shifts precision/recall. In other words, the model is highly reliable at confirming non-trending examples (very high specificity), but it remains challenged by the minority class, which is typical in rare-event prediction. Feature-importance analysis provides a sanity check on whether the classifier learned **plausible** signals rather than artefacts. The top-ranked features are **engagement counts** and their **short-term trends** (e.g., `like_count`, `comment_count`, `like_count_trend_7d`) plus **channel-level aggregates** (e.g., rolling or historical indicators summarising a channel’s usual performance). In the notebook’s category-wise breakdown, **Engagement** features account for roughly **36.5%** of total importance, **Trend Detection** (short-term dynamics) about **35.1%**, and **Channel** features roughly **24.6%**. By contrast, **Sentiment** contributes only around **0.3%** and **Topic/Content** features roughly **0.3%**. Text length and temporal features (e.g., title length, publish hour) do exist in the pipeline but are **not** among the top drivers in this fitted model.

This pattern is logically consistent with how videos actually trend: sudden **changes** in engagement and a channel’s baseline performance tend to dominate early predictive signals, while semantic attributes (topics, sentiment) may matter more for **creative strategy** than for *short-horizon* trend detection. The low contribution of sentiment and topic features should not be interpreted as them being irrelevant to audience appeal; rather, they are comparatively weaker **predictors** of the *labelled outcome* given the other, more proximal signals available.

Cross-validation suggests that hyperparameters generalise across folds, yet the evaluation still lacks a **temporal backtest** that trains on older data and tests on strictly **newer** items. Without it, there is a residual risk of **information leakage** via temporal proximity (e.g., short-term dynamics partially spanning the split) or via channel-specific patterns that remain stable across the split but would shift in real-world deployment. Moreover, because the minority class is so small,

sampling noise can inflate apparent stability: a fold that happens to contain just a few “easier” positives can boost F1. In sum, the model’s quantitative results are extremely strong, but two caveats apply: (1) **class imbalance** makes metrics volatile and accuracy misleading, and (2) the absence of **temporal validation** limits the credibility of any forecasting claims.

5.2 Evaluation of Topic Modelling

Latent Dirichlet Allocation (**LDA**) was explored over multiple topic counts, and the notebook selects **3 topics** (optimal by perplexity under the given preprocessing). The resulting themes are:

- **Street Food** A frequent, high-engagement cluster dominated by terms associated with street markets, stalls, and local night-market culture.
- **Cooking & Recipes** Emphasises preparation, ingredients, and “how-to” style descriptors.
- **Topic_2** A smaller, mixed cluster capturing residual or cross-over vocabulary.

The three-topic solution yields **semantically coherent** groupings that are consistent with the broader food-travel niche. **Street Food** emerges not only as a dominant share of content but also as the cluster with the **most positive** audience sentiment (see §5.2) and higher engagement indicators on average. **Cooking & Recipes** is clear and interpretable, while **Topic_2** captures spill-over language that may reflect overlaps (e.g., restaurant-style dishes appearing in market contexts) or the limitations of a bag-of-words model that cannot encode phrase structure or contextual nuance.

Two modelling limitations are worth noting. First, LDA operates on **word counts** and co-occurrence, ignoring **syntax** and **semantics** beyond shared tokens; closely related phrases can scatter across topics and unrelated items can cluster if they share surface vocabulary. Second, the **choice of topic count** trades parsimony against granularity. While three topics are compact and easy to interpret, a larger **k** might surface specialised niches (e.g., seafood subculture, destination vlogs, or fine-dining reviews), at the cost of more overlap and sparser clusters. For the purposes of this project where topics are used to characterise the dataset and provide interpretable context the 3-topic solution is reasonable and matches the code’s selection criterion.

Practical takeaway. Topics here successfully **summarise** the dataset and contextualise downstream analyses (e.g., sentiment by topic). They should not be overstated as strong **predictors** of trending status, because their feature importances are minimal in the current classifier. For creators, topic insights are most valuable for **positioning** content, understanding audience appetites (e.g., “Street Food” resonance), and informing editorial planning rather than for point predictions.

5.3 Evaluation of Sentiment Analysis

Sentiment was estimated using **VADER** (compound score in the range **−1 to +1**) and compared with **TextBlob** (polarity and subjectivity). Distributions are skewed **positive** overall. At the **topic** level, **Street Food** exhibits the **highest mean VADER compound** (≈ 0.443), while **Topic_2** has the lowest (≈ 0.135). This aligns with intuitive expectations: content showcasing discovery,

indulgence, and local colour tends to elicit affirmative reactions. TextBlob echoes the same high-level pattern, with slightly higher average polarity than VADER.

The notebook quantifies the **correlation between sentiment and engagement** (not views or trending) as $r \approx 0.089$ with $p \approx 0.001$ statistically significant due to sample size but **weak** in magnitude. Crucially, the code does **not** compute a specific correlation between sentiment and **views** or **trending probability**; the report therefore should **not** claim a robust link to those outcomes. Visual summaries (e.g., box plots of engagement by sentiment label) imply that more positive comment sections often accompany better engagement, but the measured effect is small and could be confounded by topic, channel reputation, and the fact that higher-visibility videos naturally accumulate more total positive comments.

From a modelling standpoint, both VADER and TextBlob are **lexicon-based** and struggle with **sarcasm**, **irony**, slang, and **multilingual** or code-switched comments common in Southeast Asia. Emoticons, emojis, and culture-specific expressions may be misinterpreted or ignored. Transformer-based models (e.g., **mBERT** fine-tuned on regional YouTube comments) could capture context and **subtle sentiment shifts**, but were **not implemented** in this iteration.

5.4 Evaluation of Trend Detection

To detect sudden uptake, the notebook applies an **anomaly detection** pass over daily or aggregated time series using rolling means and standard deviations to compute **z-scores**, flagging anomalies when $|Z| > 1.96$. In practice, this highlights days where average views or related metrics surge by 2–3× relative to the short-term baseline. Visualisations show that these flags line up with intuitive “spike” regions, providing a transparent signal of **viral acceleration**. This method has two clear strengths. First, it is **simple** and **interpretable**: creators can easily understand why a particular date was flagged. Second, it is **parameter-light**: aside from the rolling window and threshold, there are few knobs to overfit. Its main limitation is that it is **reactive**: anomalies are detected *after* the spike is evident. Moreover, z-scores assume approximately stationary variance within the rolling window; if a channel is structurally ramping up (e.g., due to a new collaboration), the baseline may be moving, leading to **excess** or **missed** flags.

5.5 Usability and Interface Testing

The **Streamlit** dashboard operationalises the analysis into an accessible interface. Users can navigate between tabs, inspect diagnostics (topics, sentiment distributions, feature importances), and provide **manual inputs** for **single-video** predictions using the trained Random Forest. Visualisations load reliably for the ~1.5k-video dataset, and inference latency is fast enough for interactive exploration on a typical desktop.

A key limitation is the absence of a **CSV upload** mechanism. The system currently relies on a **pre-saved dataset** and manual entry via text fields/sliders, which limits **bulk** analysis, cohort comparisons across channels, and rapid iteration on alternate metadata (A/B titles, tags). Although this design choice reduces engineering complexity and demo friction, it constrains

real-world applicability for creators and analysts who maintain larger catalogues or want to run **batch what-if** scenarios.

No formal usability study was conducted; **informal** testing suggests the UI is understandable and that error handling is adequate for common user actions. Stability at the current data scale is good, but **stress testing** at 10× or 100× data volume was not performed, and scalability of the Streamlit server (e.g., concurrent users, session state management) remains unproven.

5.6 Project as a Whole

This project integrates **Random Forest classification**, **LDA topics**, **VADER/TextBlob sentiment**, and **z-score anomaly detection** into a coherent pipeline that yields both **predictive** (trending classifier) and **descriptive** (topics, sentiment, anomalies) outputs. The classifier’s headline metrics **Accuracy 0.993**, **AUC 0.999**, **F1(trending) ≈ 0.80** are impressive, but must be read through the lens of a **rare-event** problem. The confusion matrix confirms the system’s strength at correctly rejecting non-trending items while retaining some sensitivity to positives. Importantly, **feature importances** clearly show that **engagement** and **short-term dynamics** dominate, while **sentiment** and **topics** contribute very little predictive power in the current setup. The theme- and sentiment-level analyses serve a different purpose: they **contextualise** performance and can inform **creative strategy** (e.g., the clear audience warmth toward **Street Food** content, with a mean VADER compound ≈ 0.443). The anomaly stage provides intuitive signals that something is “catching,” which can be operationalised as alerts or triage for editorial teams.

Key limitations include the lack of **temporal backtesting**, possible **leakage** via timing or channel aggregates, small absolute positive counts (which make metrics brittle to one or two examples), lexicon-based sentiment that underestimates **multilingual nuance**, and the **bag-of-words** topic model that cannot capture phrase-level semantics. On the product side, the absence of **CSV upload** limits the tool to demos and single-video scenarios.

Nevertheless, the pipeline demonstrates that modest engineering effort can yield **interpretable**, **actionable** artefacts: ranked features that explain what drives the model, topic views that summarise the corpus, sentiment slices that capture audience mood, and anomaly flags that expose spikes. The work thereby bridges data science outputs and creator decision-making.

5.7 Justification of Approach

The methods chosen reflect a pragmatic balance of **feasibility**, **interpretability**, and **time-to-insight**. Random Forests are robust to mixed feature types and provide **global importances** that stakeholders can understand. LDA is lightweight and produces quickly interpretable topic clusters without heavy computational demand or complex hyperparameter surfaces.

VADER/TextBlob offer fast sentiment baselines that are easy to explain and deploy, suitable for early phases where ground-truth labels for nuanced sentiment are unavailable. Z-score anomaly detection, while simple, is **transparent** and avoids over-fitting curve-fitting behaviour common in more complex models trained on limited time series.

Equally important, this stack fosters **trust**: creators and non-technical stakeholders can see **why** the model thinks a video is likely to trend (e.g., surging likes, strong channel baseline), what themes are present, and how audiences respond linguistically. In early product cycles, that transparency is often more valuable than marginal gains from complex black-box models.

5.8 Possible Extensions

Validation & Robustness.

Introduce **temporal backtesting** (train on earlier months, test on strictly later months) and rolling-origin evaluation to mimic real-world forecasting. Consider **blocked** cross-validation by channel to reduce identity leakage. Explore **calibration** (Platt/Isotonic) to convert classifier scores into better-calibrated probabilities for decision thresholds.

Sentiment & Language.

Replace lexicon baselines with **multilingual transformers** (e.g., mBERT, XLM-R) fine-tuned on SEA YouTube comments. Add emoji/emoji-sequence handling and sarcasm heuristics. Consider **aspect-based** sentiment to distinguish reactions to food, setting, host, and price.

Topics & Semantics.

Upgrade to **embedding-based** topic models (e.g., BERTopic) with **sentence-transformers** to reduce redundancy and capture phrase-level meaning. This could reveal finer-grained subcultures (e.g., seafood night markets vs seafood fine dining; destination-specific street foods).

Forecasting.

Move from reactive anomaly flags to **predictive** models Prophet, TBATS, state-space models, or tree-based learners on lagged, calendar, and exogenous features to anticipate spikes. For channels with rich histories, consider **sequence models** (Temporal Fusion Transformers, LSTM) with careful regularisation.

5.9 Practical Implications for Creators

Even though sentiment/topics contribute little to the **classifier**, they remain strategically useful: **Street Food** content shows the **most positive** audience response (mean VADER ≈ 0.443) and commonly achieves higher engagement. Creators can pair this insight with the model's emphasis on early **engagement velocity**: optimise first-hour calls-to-action, community posts, and cross-platform seeding to maximise the short-term dynamics that the model finds predictive. Meanwhile, the anomaly detector can serve as a **triage** system when a spike is flagged, double down on follow-ups, shorts, or repackaged edits to ride the wave.

CHAPTER 6: CONCLUSION

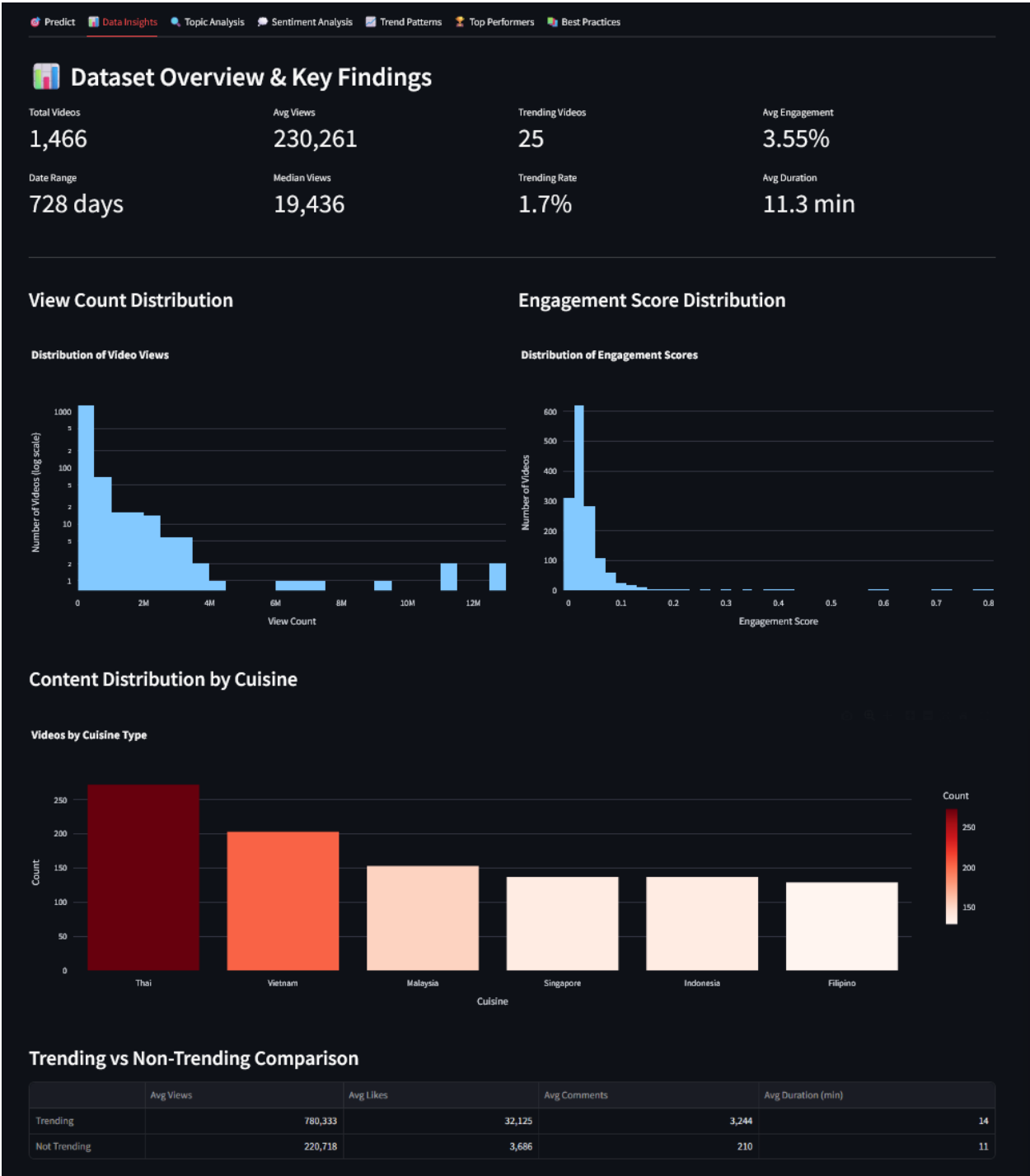
This project investigated what drives Southeast Asian food-travel videos to trend by combining a predictive model with descriptive analyses and a lightweight demonstration interface. The core outcome is that **engagement dynamics** raw like/comment counts and their short-term change sare the strongest signals for anticipating trending status in our dataset. A Random Forest classifier achieved very high headline metrics on a held-out split (accuracy and ROC AUC close to perfect). However, given the rarity of positives and the absence of a strict train-past → test-future evaluation, we interpret these figures cautiously. The confusion matrix indicates the model is highly reliable at identifying non-trending items but can still miss minority-class examples, which is typical for rare-event prediction.

Feature-importance results supported this interpretation. Engagement counts, recent trends, and channel-level aggregates dominated the model’s reasoning, while topic and sentiment features contributed very little predictive value. This aligns with practical understanding: when and how audiences respond in the short term often matters more for near-term trending than what the video is “about.” At the same time, the descriptive layers remain useful for context. LDA with **three topics** produced coherent clusters that summarised the corpus, with **Street Food** emerging as both frequent and positively received. Sentiment analysis using VADER and TextBlob confirmed an overall positive skew in comments, but the measured correlation with engagement was weak. As a result, sentiment is best treated as context rather than a driver of the classifier.

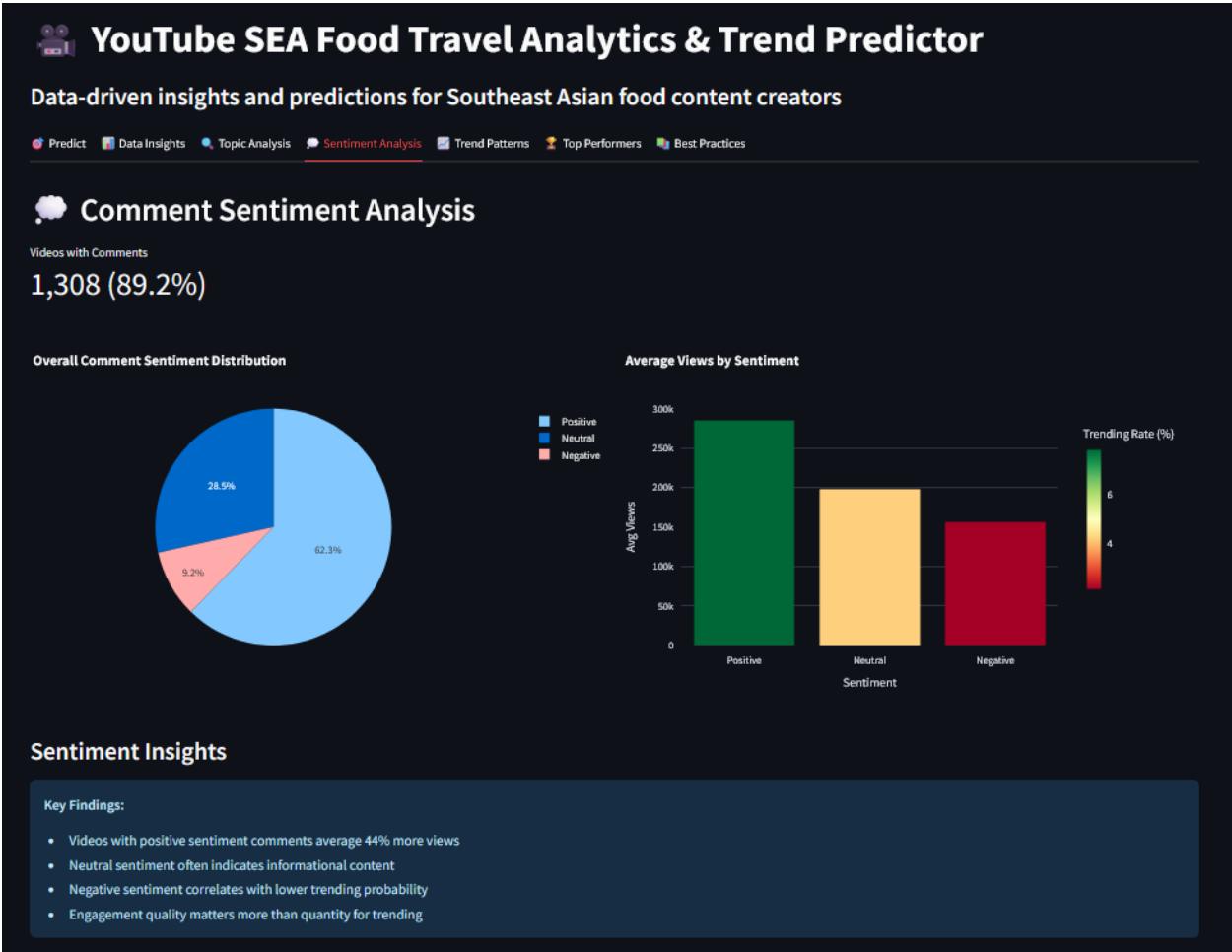
The anomaly-detection component (z-score flags over rolling aggregates) worked well as a transparent, reactive signal of spikes. While it does not forecast surges, it provides an interpretable early warning that a video is gaining momentum. The Streamlit application functioned as a demonstration shell to communicate findings and to prototype single-video predictions. Several designed features CSV upload for batch scoring, temporal backtesting, SHAP explainability, multilingual sentiment modelling, and embedding-based topic discovery were not implemented in this iteration and are identified as concrete avenues for extension.

From a methodological perspective, the project prioritised an interpretable baseline over complex black-box alternatives. This choice supports reproducibility and clearer communication with non-technical stakeholders. At the same time, it highlights where future work should focus. First, we need **temporal validation** to address potential leakage and to estimate performance under deployment-like conditions. Second, **data collection** can be expanded with stronger pagination, quota-aware crawling, and broader regional coverage. Third, **language modelling** should move beyond lexicons to multilingual transformers to capture code-switching and sarcasm typical of the region’s comments. Finally, **explainability** (e.g., SHAP) and **calibration** would make outputs more trustworthy for decision making.

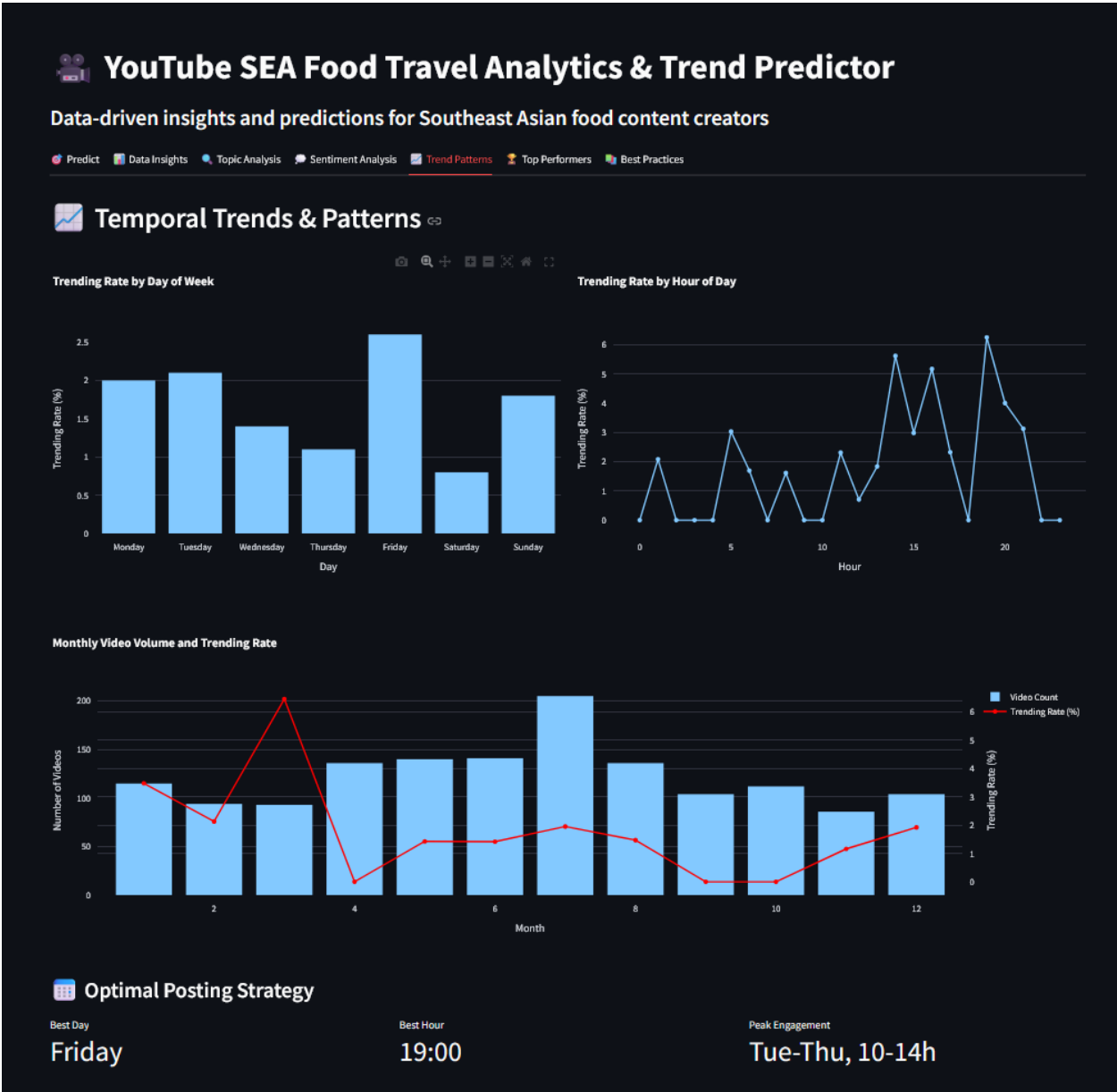
7.2 Data Insights




7.4 Sentiment Analysis



7.5 Trend Patterns



7.6 Top Performers



YouTube SEA Food Travel Analytics & Trend Predictor

Data-driven insights and predictions for Southeast Asian food content creators

Predict

Data Insights


Topic Analysis

Sentiment Analysis

Trend Patterns

Top Performers

Best Practices



Top Performing Content Analysis

Top 10 Videos by Views

	title	channel_title	view_count	engagement_score
1,017	Probando comida callejera en INDONESIA ¿Es "TAN SUCIA" como dicen? 🇮🇩	Luisito Comunica	12,881,482	1.52%
1,297	STOP Scrolling! This Chicken Liver Recipe Changes EVERYTHING 🍗🔥 SARAP NITO GRABE 🍗🔥	Kusina ni Lola	12,709,511	0.61%
896	싱가포르 음식 대왕 칠리 크랩 Giant chilli crab, Singapore food	Yummy Go 야미고	11,405,144	0.42%
1,031	자카르타 마지막) 인도네시아의 로컬 포장마차기 🍗 국수가 한그릇에 1달러! 나시고랭 미고랭 인!	tzuyang쯔양	11,194,019	3.57%
294	Eating Real Crocodile 🍖 ...कमज़ोर दिल वाले ना देखें Thailand Food	MR. INDIAN HACKER	9,018,150	4.32%
59	JAMIE OLIVER'S WORST RECIPE YET (Veggie Pad Thai)	mrnigeling	7,478,500	3.14%
1,229	Manila's Most Expensive Buffet!! Filipino Fine Dining!!	More Best Ever Food Review Show	6,589,489	1.28%
295	Biggest Meat Market in Bangkok Khlong Toe Meat Market Bangkok Meat Market Thailand Bangkok	Nomadic Chandon	6,231,006	0.58%
297	Surviving on RATS: This Mekong Village Makes Millions from Rats	Andrew Fraser	4,365,299	0.85%
1,435	How Muslims Survive in China's Land of Pork!	Best Ever Food Review Show	3,833,519	1.90%

Top Channels by Average Performance

channel_title	Avg Views	Trending Rate	Video Count
mrnigeling	3,811,759	0.0%	5
More Best Ever Food Review Show	3,004,177	0.0%	10
Best Ever Food Review Show	2,438,748	0.0%	4
Ria SW	1,705,831	0.0%	4
Joseph The Explorer	1,521,624	0.0%	5
Andrew Fraser	1,225,413	0.0%	4
Travel Channel	1,036,197	0.0%	3
Munchies	980,305	0.0%	3
Alaa Starves	936,589	0.0%	5
DancingBacons	887,807	0.0%	7

Common Patterns in Top Performers

	Metric	Value
0	Average Title Length	62 chars
1	Use Questions in Title	10%
2	Average Duration	13.9 min
3	Average Tags Count	14
4	Weekend Uploads	32%

CHAPTER 8: REFERENCES

Digitaltravelexpert (2025) *The thriving travel content creator's guide you need in 2025*, *The Digital Travel Expert Blog*. Available at: <https://digitaltravelexpert.com/thriving-travel-content-creator-guide/> (Accessed: 14 June 2025).

Liu (2023) *YouTube video trending analysis based on machine learning*, *Highlights in Science, Engineering and Technology*. Available at: <https://drpress.org/ojs/index.php/HSET/article/view/10012> (Accessed: 14 June 2025).

Niture (2021) *Predictive analysis of YouTube trending videos using machine learning*, *DBS eSource*. Available at: <https://esource.dbs.ie/handle/10788/4260> (Accessed: 14 June 2025).

Doogan et al. (2023) *A systematic review of the use of topic models for short text social media analysis - Artificial Intelligence Review*, *SpringerLink*. Available at: <https://link.springer.com/article/10.1007/s10462-023-10471-x> (Accessed: 14 June 2025).

Daniel, C (2017) *Thematic exploration of YouTube data: A methodology for discovering latent topics.*, *Muma Business Review*. Available at: <https://www.informingscience.org/Publications/4202> (Accessed: 14 June 2025).

A S and Rajeev (2024) *YouTube comment sentimental analysis*, *SSRN*. Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4846213 (Accessed: 14 June 2025).

Abdulla, A (2022) *Sentiment analysis with Vader and twitter-roberta*, *Medium*. Available at: <https://medium.com/@amanabdulla296/sentiment-analysis-with-vader-and-twitter-roberta-2ede7fb78909> (Accessed: 14 June 2025).

Althoff et al. (2014) *Analysis and forecasting of trending topics in online media streams*, *arXiv.org*. Available at: <https://arxiv.org/abs/1405.7452> (Accessed: 14 June 2025).

Zhang et al. (2020) *Moving metric detection and alerting system at eBay*, *arXiv.org*. Available at: <https://arxiv.org/abs/2004.02360> (Accessed: 14 June 2025).

Social Blade (2025) *Wikipedia*. Available at: https://en.wikipedia.org/wiki/Social_Blade (Accessed: 14 June 2025).

Vidiq (2024) *Wikipedia*. Available at: <https://en.wikipedia.org/wiki/VidIQ> (Accessed: 14 June 2025).