INSTAGRAM REACH ANALYSIS: FOOD BLOGGERS

Abstract

Bloggers

This project focuses on analyzing and predicting Instagram reach for food bloggers using data science methodologies. The main objective is to identify key factors that influence engagement and reach, such as content strategy, follower interaction, hashtag usage, and collaborations**.**

Data Collection

Instagram data (posts, likes, comments, shares, saves, and audience demographics such as follower locations and interests) will be collected using automated web scraping techniques. Selenium with ChromeDriver will be employed to extract the required data efficiently.

Data Preprocessing and Feature Engineering

The data will undergo preprocessing to clean and prepare it for analysis. This will involve handling missing values, encoding categorical variables, normalizing numerical features, and managing outliers.

During feature engineering, domain-specific insights will be incorporated:

- **Content Strategy:** Features will be created based on the type of content (The balance between these content types will be examined to understand their impact on engagement.
- **Follower Engagement**: Features reflecting likes, shares, comments, saves, and other interactions will be constructed.
- **Hashtag Analysis:** Features related to hashtag usage, such as niche food-related hashtags (e.g., #HomeCooking, #PlantBased) and location-based hashtags (e.g., #HyderabadFoodies), will be created to evaluate their effect on reach.
- **Collaborations:** Features measuring the impact of collaborations with restaurants or food brands on follower growth and engagement will be engineered.

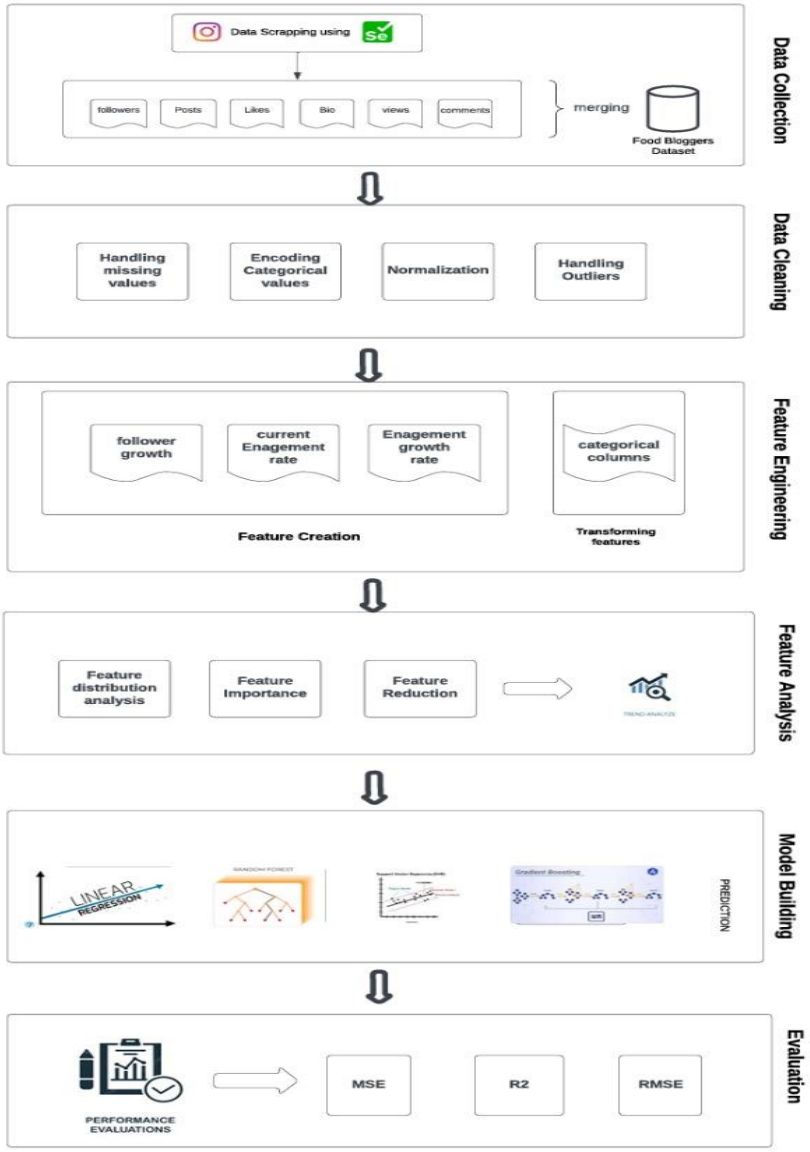
Model Training and Evaluation

After feature engineering, regression models will be trained to predict future reach. The models will be evaluated using metrics like R^2, MSE, and RMSE to ensure performance and reliability. Feature importance techniques will also be used to highlight the most influential factors contributing to engagement and reach.

Visualization and Deployment

The results will be visualized to present trends in engagement and projected reach, using visual tools like charts and graphs. These visualizations will help identify actionable insights for food bloggers to refine their content strategies. The trained models will be deployed to provide real-time projections of future reach, enabling food bloggers to optimize their Instagram presence and engagement strategies.

Methodology



Introduction

This project analyzes the reach and engagement of food bloggers on Instagram by examining metrics like followers, posts, comments, and sentiment of captions. The goal is to understand how these factors contribute to a food blogger's reach and audience engagement.

Importing Libraries

Key libraries used in the analysis:

Pandas: Data manipulation and analysis.

NumPy: Numerical operations and array manipulations.

Regular Expressions: Text processing for string handling.

Matplotlib & Seaborn: Data visualization tools.

Scikit-learn: Machine learning for modeling and preprocessing.

NLTK: Natural Language Processing tasks (tokenization, lemmatization).

VADER: Sentiment analysis for social media text.

WordCloud: Displays most common words in text data.

sample rows of the dataset

Initial Data Snapshot:

	username	posts_count	followers_count	following_count	Bio	post_url	likes	hashtags	location	post_date	comments	comments_count	overall_sentiment
0	vismaifoodies	1,323	1.4M	4	We help Foodies to get the Perfect Recipes.\nF...	https://www.instagram.com/vismaifoodies/p/...	2,944	#MAGGIApnaFoodBusiness, #AFB, #OnlineFoodChann...	Locations	2024-05-27T14:44:41.000Z	😍, Please don't associate with foods like Magg...	9	neutral
1	mysorefoodgram	129	2240	669	NaN	https://www.instagram.com/mysorefoodgram/p/CpR...	65	NaN	Locations	2023-03-02T03:57:48.000Z	NaN	0	neutral
2	mysorefoodgram	129	2240	669	NaN	https://www.instagram.com/mysorefoodgram/p/CpP...	105	NaN	Locations	2023-03-01T07:23:55.000Z	NaN	0	neutral
3	mysorefoodgram	129	2240	669	NaN	https://www.instagram.com/mysorefoodgram/p/CpR...	178	NaN	Locations	2023-03-02T03:58:37.000Z	NaN	0	neutral
4	mysorefoodgram	129	2240	669	NaN	https://www.instagram.com/mysorefoodgram/p/CpR...	101	NaN	Locations	2023-03-02T03:59:43.000Z	NaN	0	neutral
5	mysorefoodgram	129	2240	669	NaN	https://www.instagram.com/mysorefoodgram/p/CpP...	100	NaN	Locations	2023-05-30T17:23:23.000Z	Dosa point, Place?	2	interaction
6	mysorefoodgram	129	2240	669	NaN	https://www.instagram.com/mysorefoodgram/p/Cne...	78	NaN	Le Méridien Kochi	2023-01-16T12:07:33.000Z	NaN	0	neutral
7	mysorefoodgram	129	2240	669	NaN	https://www.instagram.com/mysorefoodgram/p/CpN...	106	NaN	Locations	2023-02-28T15:49:40.000Z	NaN	0	neutral
8	mysorefoodgram	129	2240	669	NaN	https://www.instagram.com/mysorefoodgram/p/C27...	181	NaN	Locations	2024-02-04T15:29:44.000Z	NaN	0	neutral
9	mysorefoodgram	129	2240	669	NaN	https://www.instagram.com/mysorefoodgram/p/Cqq...	130	NaN	Locations	2023-04-05T15:50:28.000Z	❤️	1	neutral
10	mysorefoodgram	129	2240	669	NaN	https://www.instagram.com/mysorefoodgram/p/Cqa...	154	NaN	Locations	2023-03-30T08:44:05.000Z	NaN	0	neutral
11	food_hunter_1	577	38.9K	0	Show best food	https://www.instagram.com/food_hunter_1/p/Cub...	378	NaN	Locations	2023-07-08T06:46:38.000Z	Bro meeku msg cheyyadaniki avvatle okasari cho...	1	neutral
12	food_hunter_1	577	38.9K	0	Show best food	https://www.instagram.com/food_hunter_1/p/C8A...	438	NaN	Locations	2024-06-09T16:45:15.000Z	Congratulations 👏, Heartily congratulations, C...	4	neutral
13	food_hunter_1	577	38.9K	0	Show best food	https://www.instagram.com/food_hunter_1/p/C_z...	665	NaN	Locations	2024-09-12T07:06:41.000Z	Congratulations 🎉 anna, Good... Keep Going 👌, C...	5	positive
14	foodonfarm	234	628K	1	All about food and nature \nTo order our produ...	https://www.instagram.com/foodonfarm/p/DAza5Ce...	15,590	NaN	Locations	2024-10-07T00:50:56.000Z	Namaste peddananna gaaru ❤️, Mi smile super th...	10	positive

Data Cleaning

Addressing missing data:

Missing "Bio" entries filled with "Unknown."

Missing "comments" replaced with "No Comments."

Missing "hashtags" replaced with "No Hashtags."
This ensures dataset completeness for accurate analysis.

Sentiment Analysis Setup

Using VADER to assess the emotional tone of posts and comments. This provides insights into how food bloggers interact with their audience and how their content is received.

Loading Data

The dataset containing Instagram profiles of food bloggers is loaded into a DataFrame for analysis. This provides a foundation for further analytical processes.

Data Cleaning

Unnecessary columns are removed to streamline the dataset and focus on relevant metrics.

Feature Engineering

Follower Count Conversion: A custom function converts follower counts from string (e.g., "k" for thousands, "m" for millions) to numeric values.

Cleaning & Converting Columns: Non-numeric entries (e.g., "Not Available") are replaced with NaN and converted to numeric types.

Handling Missing Values: Missing numeric values filled with median values.

Date Feature Extraction: Extracting year, month, day, and hour from 'post_date' for temporal analysis.

Text Preprocessing and Sentiment Analysis

Comment Preprocessing: Converts text to lowercase, removes punctuation, tokenizes, and lemmatizes for text standardization.

Sentiment Score Calculation: Using VADER, sentiment scores are calculated for each comment, ranging from -1 (negative) to +1 (positive).

Categorization of Sentiment: Comments are categorized as positive, negative, or neutral.

Post Type Classification: Posts are classified based on bio keywords into categories like "Recipe," "Food Post," or "Other."

Engagement Metrics

Likes and Comments Analysis: Aggregates total likes and comments per post to evaluate engagement.

Engagement Rate Calculation:

$$\text{Engagement Rate} = \frac{\text{Total Likes+Total Comments}}{Total\ Followers}\ X\ 100$$

Top Performers Analysis

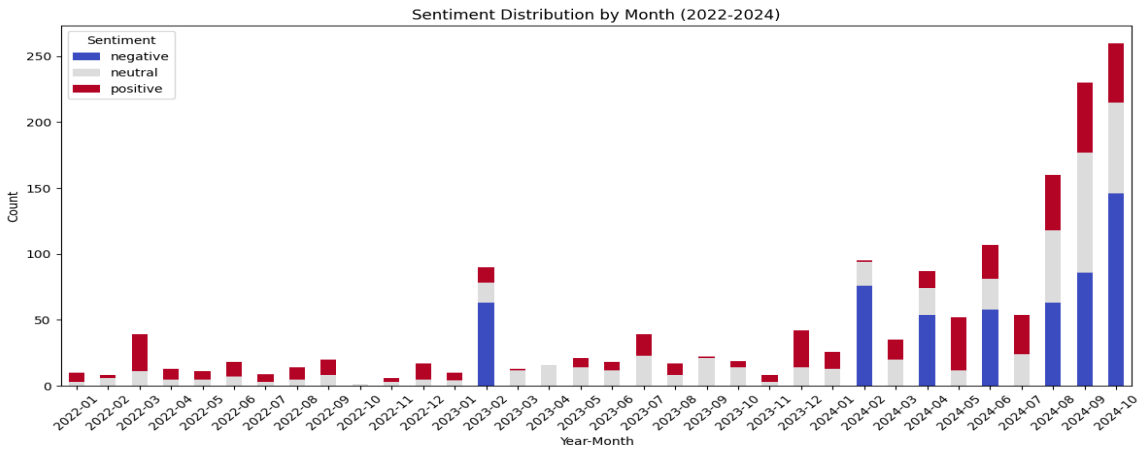
Top Bloggers by Engagement Rate: Highlights bloggers with the highest engagement rates.

Content Strategy Insights: Identifies content types (e.g., recipes) driving the most engagement.

Hashtag Analysis

Top Hashtags by Average Likes: Analyzes hashtags with the highest average likes.

Top Hashtags by Engagement Rate: Evaluates hashtags by their overall engagement (likes + comments).



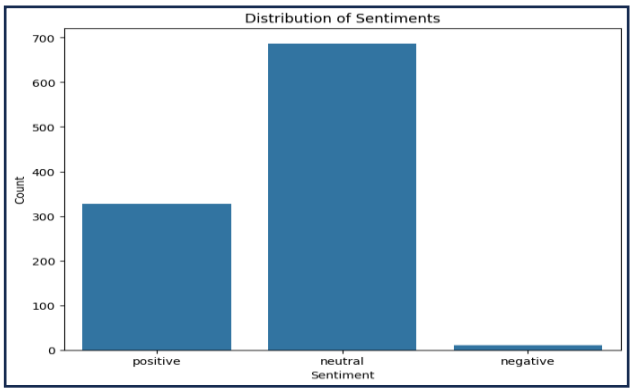
Engagement Metrics Over Time

A time series analysis of engagement metrics (likes, comments, engagement rate) by month helps identify trends over time.

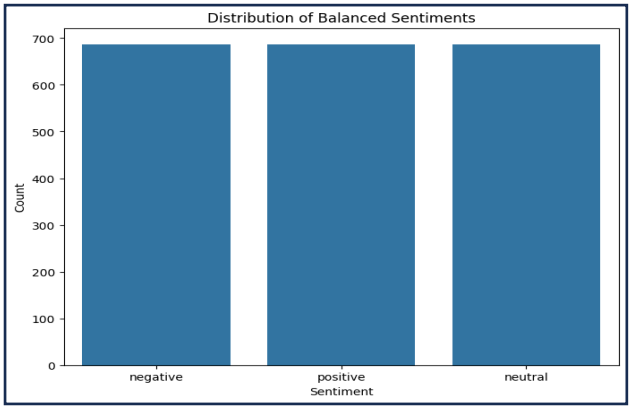
Sentiment Analysis

Sentiment distribution (positive, negative, neutral) is visualized to inform strategies for improving audience perception.

Before Resampling



After Resampling



Word Cloud Visualizations

Word clouds highlight the most frequent words in each sentiment category, providing insights into audience sentiment.

Negative

Neutral

Positive

Class Balancing Techniques

Sentiment classes are balanced through upsampling to ensure reliable predictions and avoid model bias.

A stacked bar chart shows how sentiment trends evolved over time, helping understand the impact of external factors on audience perception.

Dropping Unnecessary Features

Irrelevant columns are removed to streamline the dataset and focus on key features for analysis.

Feature Scaling and Encoding

Numerical features are scaled to ensure uniformity and prevent any one feature from disproportionately influencing the model.

Dimensionality Reduction with PCA

Principal Component Analysis (PCA) reduces the dimensionality of the dataset while retaining significant variance, improving model generalization.

Model Training and Evaluation

Various models (e.g., Linear Regression, Random Forest) are trained and evaluated using metrics like Mean Squared Error (MSE) and R-squared.

Hyperparameter Tuning with Grid Search

Grid search optimizes the Random Forest model, tuning hyperparameters to achieve the best performance.

Model Evaluation: Random Forest

Performance evaluated using metrics like Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-squared (R^2).

Linear Regression Performance

A baseline linear regression model is evaluated for comparison with more complex models like Random Forest.

Support Vector Machine (SVM)

SVM regression is explored to capture non-linear relationships in the data, compared with other models using evaluation metrics.

Gradient Boosting Model

Gradient Boosting is applied, and its performance is evaluated against Random Forest using similar metrics.

Model Comparison

Models are compared using RMSE values, and a box plot visually depicts performance differences.

Actual vs. Predicted Visualization

A scatter plot shows how closely model predictions match actual values, highlighting the performance of Random Forest and Gradient Boosting models.

Bar Chart of Evaluation Metrics

Bar charts visualize key metrics (MAE, RMSE, R^2) to compare model performance.

Conclusion

Random Forest and Gradient Boosting provided the most accurate predictions, with Gradient Boosting showing a slight edge. These models effectively estimate Instagram reach based on various features.