ESP IJAST
ESP International Journal of Advancements in Science & Technology
ISSN: 2583-9233/ Volume 4 Issue 2 April 2026
Paper Id: IJAST-V4I2P103/Page No: 18–21

Original Article

AI Enabled Adaptive Digital System

Jaspreet Sodhi[1], Dr.Anuradha Misra[2]
[1]*student, Department of computer science and engineering and technology, Amity University Luck now campus, India*
[2]*Assistant Professor, Department of computer science and engineering and technology, Amity University Luck now campus, India,*

Received Date: 06 March 2026 Revised Date: 18 March 2026 Accepted Date: 05 April 2026

Abstract: *Hash tags play a pivotal role in increasing the visibility, discoverability, and engagement of social media content, particularly in the tourism sector where users search for travel inspiration using thematic tags. This project presents a multi-label hash tag recommendation system for tourism related social media posts focused on Indian destinations. A key innovation is the use of weak supervision techniques to overcome the challenge of absent labeled data, a common limitation in real world applications. Two structured datasets containing tourism metadata such as site descriptions, geographical zones, ratings, and seasonal recommendations were used to generate synthetic social media style posts through template based natural language generation. A preprocessing pipeline including normalization, stop word removal, and lemmatization was applied, followed by TF-IDF factorization for feature extraction. A hybrid labeling approach combined rule based keyword matching with semantic similarity using cosine similarity between TF-IDF representations and prototype sentences. Manual filtering mechanism reduced noise and missioned hash tags. Finally, a logistic regression One-vs.-Rest multi-label model was trained and evaluated using micro, macro, and sample based F1 scores, demonstrating effective hash tag prediction.*

Keywords: *Machine learning, Natural Language Processing, Recommendation System, Tourism*

I. INTRODUCTION

With the rapid rise of digital content creation especially in the tourism sector, the demand for intelligent and adaptable software services has grown significantly. This project was envisioned to explore how modular, AI driven services can be effectively delivered within flexible software architecture. As part of this initiative, a multi label hash tag recommendation system was developed, specifically targeting tourism related social media posts. Designed as a smart backend service, the system analyzes natural language input and suggests contextually relevant hash tags to enhance content visibility and engagement. This aligns with the core philosophy of e delivering intelligent, customizable service that can be integrated seamlessly into broader digital platforms. To address the challenge of limited labeled data, a hybrid labeling strategy was implemented. This combined rule based heuristics with semantic similarity techniques, using cosine similarity to assign hash tags based on the content's closeness to predefined prototypes. Additionally, synthetic post data was generated from structured tourism metadata to simulate real world user generated content. Standard NLP preprocessing techniques were applied, followed by TF-IDF factorization for feature extraction. A logistic regression based multi label classification model was then trained to recommend suitable hash tags. By packaging this functionality as a modular service, the project showcased the application of machine learning in solving real world problems.

II. LITERATURE REVIEW

The growing demand for personalized travel experiences has led to significant advancements in recommender systems, especially through the integration of user generated content from social media platforms. Recent studies have explored innovative approaches to tailor recommendations based on individual preferences, emotions, and social interactions.

Coelho et al. [1] present a personalized travel recommendation system using Twitter data to identify user preferences through a machine learning model. The system classifies travel related tweets into categories like historical buildings, museums, parks, and restaurants, and also considers data from a user's friends and followers for better personalization. Evaluation against user surveys showed an accuracy of 68%, indicating moderate success. However, limitations include dataset quality and classification methods. The authors suggest improving techniques and expanding categories (e.g., entertainment, sports) to enhance performance, highlighting the potential of social media for travel recommendations.

South et al. [2] propose a hybrid travel recommendation system that combines emotion analysis from Integra comments with image recognition from user posts. The system uses image data, location details, and user interaction data, applying collaborative filtering to find similar users and content-based filtering to match destinations with user interests. It also features an interactive interface where users can upload images and captions to get real-time

recommendations. By combining emotional and visual insights, the system provides more personalized travel suggestions.

While both studies focus on personalized travel recommendations using social media, they differ in complexity and data modalities. While several existing studies focus on personalized travel recommendations leveraging social media data, such as tweets or Integra posts [1], [2], these systems predominantly target suggesting places or experiences rather than assisting content creators in generating relevant hash tags for enhanced social media engagement. This work addresses the challenge of limited labeled data by using a hybrid labeling strategy that combines rule-based methods and semantic similarity for automated multi label hash tag assignment. It provides a scalable solution tailored to Indian tourism content, helping improve content visibility and supporting automatic hash tag generation using weak supervision and natural language processing techniques.

III. MATERIALS AND METHODS

A. Data Collection and Understanding the Dataset

The foundation of this project was built on two comprehensive tourism datasets covering Indian tourist destinations, places, and cities. The first dataset provided complementary information at the city level, including aggregate travel ratings, detailed descriptions capturing cultural and historical insights, and the best time to visit.This narrative content was especially valuable for simulating natural, engaging social media posts.[3]

The second dataset contained rich metadata about tourist places, including geographical zones, state and city locations, types of places (like temples, natural parks, war memorials), historical or cultural significance, and visitor information such as entrance fees, Google ratings, and best time to visit [4].

Understanding the dataset's depth and variety was important because it helped in creating synthetic post texts that reflect real-world tourism content and support effective model training. The necessary libraries for data handling, NLP, and model building were imported, and the datasets (cities_df and places_df) were loaded using pd.read_csv(). From these datasets, fields such as city name, place name, type, significance, and "best time to visit" were used to support the post generation process.

cities_df.head()

	Id	City	Rating	About the city (long Description)	Best Time to visit
0	1	Gangtok	4.6	Incredibly alluring, pleasantly boisterous and...	Throughout the year
1	2	Udaipur	4.6	Udaipur, the "City of Lakes," stands as a jewe...	October to March
2	3	Gulmarg	4.4	Situated at an altitude of 2730 m above sea le...	October to June
3	4	Agra	4.9	Located on the banks of River Yamuna in Uttar ...	October to March
4	5	Andaman and Nicobar	4.6	Replete with turquoise blue water beaches and ...	October to Jun

Figure 1: Cities Dataset

places_df.head()

	Unnamed: 0	Zone	State	City	Name	Type	Establishment Year	time needed to visit in hrs	Google review rating	Entrance Fee in INR	Airport with 50km Radius	Weekly Off	Significance	DSLR Allowed	Number of google review in lakhs	Best Time to visit
0	0	Northern	Delhi	Delhi	India Gate	War Memorial	1921	0.5	4.6	0	Yes	NaN	Historical	Yes	2.60	Evening
1	1	Northern	Delhi	Delhi	Humayun's Tomb	Tomb	1572	2.0	4.5	30	Yes	NaN	Historical	Yes	0.40	Afternoon
2	2	Northern	Delhi	Delhi	Akshardham Temple	Temple	2005	5.0	4.6	60	Yes	NaN	Religious	No	0.40	Afternoon

Figure 2: Places Dataset

B. Synthetic Post Generation and Text Preprocessing

To build the dataset; tourism metadata was transformed into natural, social media like sentences to simulate real user-generated content. The functions generate_city_post (row) and generate_place_post (row) were created to construct meaningful posts using city and place details, and they were applied using .apply () on cities_df and places_df to generate a new post column. The posts from both datasets were then combined into one synthetic dataset to support effective NLP model training. After that, a preprocess(text) function was developed to clean the text by converting it to lowercase, removing special characters, removing stop words, performing lemmatization, and rejoining the words into a cleaned format. This preprocessing was applied to all generated posts and stored in a new cleaned column, improving the quality of input data for the machine learning model.

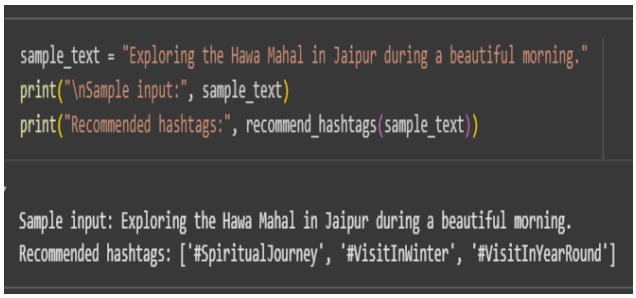
```
all_posts = pd.concat([places_df['post_cleaned'], cities_df['post_cleaned']], ignore_index=True)
```

Figure 3: Code Written To Do Concatenation Of Two Dataset

C. Initial Hash tag Labeling Using Rule Based Heuristics

A major challenge in building the hash tag prediction model was the absence of manually annotated labels. Manually tagging thousands of posts is time consuming and resource intensive, so a heuristic rule-based labeling approach was used to generate supervised labels automatically. This method relied on domain knowledge to assign hash tags based on keywords or metadata present in each post. For example, posts containing words such as "temple" or "religious" were assigned #Spiritual Journey, while terms like "mountain" or "trekking" resulted in #Adventure Seekers. To implement this, functions such as generate_place_hashtags (row) and generate_city_hashtags (row) were defined to examine metadata fields and keywords. These functions generated hash tag lists by combining attributes such as city

name, state name, and place type, and also analyzed the “Best Time to Visit” field to add tags like #Sunrise Spot or #Evening Views. The functions were applied to the datasets so each generated post received a synthetic set of hash tags using rule logic. Finally, the results were organized so that each cleaned post text was aligned with its corresponding rule



based hash tags for model training.

D. Hybrid Labeling Approach (Prototype Based Cosine Similarity and Rule-Based Method)

To improve rule-based labeling, cosine similarity between TF-IDF vectors of posts and predefined prototype texts for each hash tag was used. A dictionary hashtag_prototypes mapped hash tags to descriptive prototype sentences, which were normalized using the same preprocess () function. A TfidfVectorizer (vectorizer_for_labeling) was fitted on the combined set of cleaned posts and prototype descriptions. The posts and prototypes were transformed into TF-IDF vectors (post_vecs and proto_vecs), and cosine similarity was calculated between them. For each post, the top three most similar prototypes were selected and their hash tags were assigned. To reduce noise, manual filtering was applied by identifying frequently missioned hash tags and blacklisting them, improving label precision and overall relevance. Recognizing the strengths and weaknesses of both heuristic and similarity based labeling; the project combined these two methods into a hybrid approach. The rule based system provided precise and semantically accurate labels, while the cosine similarity method added flexibility and expanded coverage. By merging label sets and applying filtering, the project created a richer, more balanced training dataset, striking a balance between accuracy and generalizability.

Figure 4: Showing Assigned Hash Tags

E. Feature Extraction and Multi-Label Model Training for Hash tag Recommendation

The cleaned post texts were transformed into numerical features using TF-IDF (Term Frequency Inverse Document Frequency) factorization, which quantified the importance of words relative to their frequency within a document and across the corpus. TF-IDF highlighted significant terms in each post while down-weighting common words, making it suitable for classical machine learning algorithms. A TfidfVectorizer (max_features = 3000) was used, and .fit transform () was applied on the training cleaned texts to build the vocabulary. The .transform() method was then applied to the test texts (and new texts later) to obtain their TF-IDF representations, ensuring that only the training vocabulary was used and preventing data leakage. To handle posts containing multiple hash tags, a multi-label classification approach was implemented. MultiLabelBinarizer was used to convert hash tag lists into binary format, and the dataset was split into training and testing sets using train_test_split (). A Logistic Regression (max_iter=1000) model was trained using OneVsRestClassifier, enabling a separate classifier for each hash tag. For practical deployment, a recommend_hashtags (text, threshold=0.3) function was created to preprocess and victories user input and predict relevant hash tags. A confidence threshold was applied to filter low-probability predictions, along with fallback logic to return the top hash tags when no predictions met the threshold, ensuring consistent and reliable hash tag recommendations.

IV. RESULTS AND DISCUSSION

This project demonstrated how thoughtful integration of domain knowledge, heuristic labeling, and semantic similarity can compensate for the lack of manually labeled data. By building a multi label hash tag recommendation system trained on synthetically generated posts, it provided a practical solution to a common real world problem faced by

content creators in tourism. The project balanced technical rigor with resource constraints, delivering a scalable and effective model. The detailed iterative improvements and evaluation reinforced the importance of error analysis, data quality, and hybrid methodologies in machine learning projects.

V. REFERENCES

[1] J. Coelho, P. Nit, and P. Madera, "A personalized travel recommendation system using social media analysis," 2018 IEEE International Congress on Big Data (Big Data Congress), pp. 260–263, 2018.

[2] S. M. S, A. Krishna, and A. P. R. S, "Travel Destination Recommendation System based on Machine Learning by Analysing Integra Review Comments and Hash tags," 2024 5th IEEE Global Conference for Advancement in Technology (GCAT), pp. 1–10, 2024.

[3] Cities dataset from https://www.kaggle.com/datasets/kirtandwivedi02/most-traveled-cities-in-india

[4] Places dataset from https://www.kaggle.com/datasets/saketk511/travel-dataset-guide-to-indias-must-see-places

[5] A. H. Caldron, M. G. Pérez, F. J. Garcia Clemente, G. M. Pérez, "Design of a recommender system based on users' behaviour and collaborative location and tracking", Journal of Computational Science, vol. 12, pp. 83-94, Jan 2016.

[6] P. Martins, P. Madera, Personalizing Places of Interest Using Social Media Analysis, Milwaukee, WI, 2015.