REPUBLIC OF TÜRKİYE

ALTINBAŞ UNIVERSITY

Institute of Graduate Studies

Data Analytics Department

AUDIENCE BEHAVIOR AND ENGAGEMENT ON INSTAGRAM: A DATA-DRIVEN APPROACH

Oussama IGHIL GUITOUN

Master's Thesis

Supervisor

Assoc. Prof. Dr. Doğu Çağdaş ATİLLA

İstanbul, 2025

AUDIENCE BEHAVIOR AND ENGAGEMENT ON INSTAGRAM: A DATA-DRIVEN APPROACH

Oussama IGHIL GUITOUN

Data Analytics

Master's Thesis

ALTINBAŞ UNIVERSITY

2025

The thesis titled AUDIENCE BEHAVIOR AND ENGAGEMENT ON INSTAGRAM: A DATA- DRIVEN APPROACH prepared by Oussama IGHIL GUITOUN and submitted on 30/06/2025 has been **accepted unanimously** for the degree of Master of Science in Data Analysis.

Assoc. Prof. Dr. Doğu Çağdaş ATİLLA
Supervisor

Thesis Defense Committee Members:

Assoc. Prof. Dr. Doğu Çağdaş ATİLLA	Department of Electrical Electronic Engineering, Altınbaş University	
Asst. Prof. Dr. İncilay YILDIZ	Department of Management Information Systems, Altınbaş University	
Asst. Prof. Dr. Cahit KARAKUŞ	Department of Data Science and Analytics, İstanbul Esenyurt University	

I hereby declare that this thesis meets all format and submission requirements for a master's thesis.

I hereby declare that all information/data presented in this graduation project has been obtained in full accordance with academic rules and ethical conduct. I also declare all unoriginal materials and conclusions have been cited in the text and all references mentioned in the Reference List have been cited in the text, and vice versa as required by the abovementioned rules and conduct.

Oussama IGHIL GUITOUN

Signature

DEDICATION

I dedicate this thesis and work for all the members of my family that has supported my throughout the years.

Also I would like to thank myself yet again for putting the hours and the efforts to do this research and write this thesis.

And as usual I would like to thank the whole MAN CITY team for being the best team in the world.

ABSTRACT

AUDIENCE BEHAVIOR AND ENGAGEMENT ANALYSIS ON INSTAGRAM: A DATA-DRIVEN APPROACH

IGHIL GUITOUN, Oussama

M.Sc., Data Analytics, Altınbas University

Supervisor: Assoc. Prof. Dr. Doğu Çağdaş ATİLLA

Date: 06/2025

Pages: 79

This study presents a data-driven approach to understanding audience behavior and engagement on social media, with a focus on Instagram. By combining user survey responses with real-world post performance data collected via the Meta Graph API, the research explores how users interact with content and what drives engagement.

Content categories were identified using a fine-tuned BERT model, while survey data enabled demographic segmentation and behavior pattern mining. Post insights, such as likes, views, shares, and saves, were analyzed across different time slots and content types. The findings were compared with self-reported behaviors to uncover correlations and discrepan- cies. This integrated analysis provides valuable guidance for optimizing content strategies based on audience preferences and engagement trends.

Keywords: AI-driven Modelling, Social Media Engagement, Predictive Analytics, Meta Graph API, User Behavior, Content Optimization, Audience Segmentation.

TABLE OF CONTENTS

Pages

LIST OF TABLES

Pages

LIST OF FIGURES

Pages

ABBREVIATIONS

AI	:	Artificial Intelligence
BERT	:	Bidirectional Encoder Representations from Transformers
API	:	Application Programming Interface
SHAP	:	SHapley Additive exPlanations
LIME	:	Local Interpretable Model-agnostic Explanations

1. INTRODUCTION

1.1 BACKGROUND AND CONTEXT

Content creation has become an integral part of modern digital communication, with individuals either producing or consuming it daily across various platforms. Simultaneously, data has emerged as a critical resource in both personal and professional contexts, driving decisions

and strategies. The convergence of these two domains—content creation and data analysis holds immense potential. This thesis explores how the integration of organic and historical data from social media platforms can enhance content strategies, foster audience growth, and increase follower engagement through the application of Artificial Intelligence (AI). By leveraging AI, this study seeks to offer actionable insights and innovative methodologies for optimizing content creation and maximizing its impact.

1.2 SIGNIFICANCE OF THE STUDY

In recent years, content creation has evolved from a casual hobby into a viable source of income and a full-time profession for many individuals. This shift has led content creators to approach their work with increased seriousness and professionalism. However, despite its growing popularity, content creation presents significant challenges that are often overlooked by the average consumer, who may perceive it as easy or effortless. Success in this field is rarely due to mere luck; it requires strategic planning, data-driven analysis, and a deep understanding of the platform and audience.

Some of the key challenges faced by content creators include:

a. Generating Content Ideas: Many creators struggle with maintaining creativity, avoiding repetitive themes, and selecting ideas that drive engagement.

b. Identifying the Right Audience and Target: Successful content resonates with a specific audience, which requires creators to analyze and understand their followers' demographics and preferences.

c. Determining Optimal Posting Times: Even high-quality content can fail to perform

d. well if it is not shared at the right time when the target audience is most active.

e. Addressing Followers' Needs and Preferences: Creators must consistently meet their followers' expectations by producing content that aligns with their interests, or risk losing their audience.

f. Growing a Follower Base: Sustained growth is essential for long-term success, requiring strategies to attract new followers while retaining existing ones.

g. Selecting the Most Effective Content Type: Choosing the right content format (e.g., reels, stories, or longer videos) and production method (e.g., AI-generated or human-created) can significantly impact engagement and reach.

These challenges often lead to frustration and burnout among content creators, hindering their growth and potential. This study aims to address these issues by exploring applications of artificial intelligence (AI) in the content creation process. AI provides innovative solutions, such as audience analysis, content optimization, and predictive tools for improving engagement and ensuring consistent growth. By leveraging AI, content creators can overcome these obstacles and achieve greater success, transforming how content is produced and consumed.

1.3 RESEARCH MOTIVATION

As an active content creator for over a decade, working across multiple platforms and managing several digital accounts, I have personally faced many of the challenges mentioned earlier.

Despite the frustration of seeing videos I've worked hard on, receive low engagement, my passion for content creation keeps me motivated to continue improving and experimenting. At the same time, my studies in data analytics have deepened my understanding of how data- driven approaches can bring significant advantages to the field. The idea of merging these two areas of interest—content creation and data analytics—was a natural progression for me, especially as it allows me to address problems that I and many others in my field encounter. Understanding these challenges firsthand has given me a unique perspective on their impact and significance. The prospect of finding solutions through the application of artificial intelligence is not only exciting but also meaningful, as it has the potential to benefit a wide community of content creators. This dual passion for content creation and data analytics has driven me to explore how these fields can intersect to create practical, innovative tools for success.

1.4 OBJECTIVES OF THE STUDY

The primary objective of this project is to contribute to the scientific understanding of artificial intelligence (AI) applications in the field of content creation. By exploring how AI can be leveraged to address common challenges faced by content creators, this study aims to bridge the gap between data analytics and practical tools for content strategy optimization. The ultimate goal is to develop a user-friendly and accessible online platform designed specifically for content creators. This platform will provide actionable insights and predictive recommendations to simplify their workflow, enhance engagement, and foster growth. By doing so, the project seeks not only to advance academic knowledge but also to deliver a practical solution that content creators can integrate seamlessly into their daily lives.

1.5 LITERATURE SUPPORT

Content creation has become an integral part of the digital economy, but creators face numerous challenges in ensuring their work reaches and engages their target audience. A systematic review by Dolan et al. (2020) highlights the complexity of social media engagement, emphasizing that it is a multi-dimensional concept involving likes, comments, shares, and more nuanced actions such as audience loyalty and repeat interactions. The lack of standardized metrics

for measuring engagement adds to the difficulty, leaving creators unsure of how to optimize their strategies. Moreover, variations in engagement across platforms, demographics, and content types require constant adaptation, further complicating the process. Adding to these challenges, Floridi (2024) explores the transformative role of artificial intelligence (AI) in content creation. AI technologies have shifted the process from being solely human-driven to a collaborative effort between creators and AI tools. The paper discusses how tools such as automated editing software, personalized recommendations, and predictive analytics offer creators powerful ways to overcome these difficulties. However, the study also notes that the oversaturation of content on social media platforms makes it increasingly challenging for creators to stand out and capture audience attention. These studies underscore the need for innovative tools that address the core challenges faced by content creators, including audience understanding, engagement optimization, and idea generation. This thesis aims to contribute to this field by integrating AI into content creation strategies, offering actionable insights and practical solutions to enhance creators' success.

2. LITERATURE REVIEW

2.1 LITERATURE REVIEW TABLE

To better understand the current state of research and identify gaps in the existing literature, a review of relevant studies was conducted. The following table summarizes key papers related to the topic, outlining their objectives, methodologies, and major findings. This structured overview helps highlight how previous research has approached the subject and provides a foundation for the present study.

Table 2.1: Literature Review Table.

Paper Name	Authors & Date	Subject	Objective/Aim	Key Findings
Theoretical Foundations of Social Media Uses and Effects[22]	Patti M. Valkenburg, 2021	Media Effects Theories	To discuss communication and media effects theories applicable to social media use among adolescents.	• Introduces paradigms such as selectivity, transactionality, and conditionality to understand social media effects. • Emphasizes the integration of developmental and self-effects theories in social media research.
Paper Name	**Authors & Date**	**Subject**	**Objective/Aim**	**Key Findings**
Social Media Engagement Theory: Exploring the Influence of User Engagement on Social Media Usage[6]	Paul M. Di Gangi and Molly McCord, 2016	Social Media Engagement	To develop a theoretical model explaining the role of social media content in facilitating user engagement	• Proposes a typology of social media engagement behavior. • Utilizes uses and gratifications theory to model how organizations can stimulate positive engagement.

Table 2.1: Literature Review Table (Table Continued).

Theoretical Foundations of Customer Engagement in Social Media Implications for Businesses[23]	Malte Wattenberg, 2024	Customer Engagement	To examine prominent theories of customer engagement and develop recommendations foreffective corporate communication on social media	• Identifies factors influencing customer engagement based on selected theories. • Provides actionable recommendations for businesses to enhance social media strategies.
Paper Name	**Authors & Date**	**Subject**	**Objective/Aim**	**Key Findings**
Generative Artificial Intelligence, Content Creation, and Platforms[19]	Peukert, Christian; Wichmann, Tim, 2024	Economics of AI	To propose a theoretical model addressing the impact of generative AI technology on human-produced content.	• Analyzes how generative AI changes incentives for content production. • Explores the economic implications of AI-generated content on platforms.
Exploring the Impact of Artificial Intelligence on Content Creation: A Comprehensive Study[12]	Not specified, 2024	AI in Content Creation	To investigate how AI technologies influence creativity, productivity, and content quality across various industries	• AI has revolutionized digital content conception, creation, and consumption. • The integration of AI into content workflows enhances efficiency and introduces new creative possibilities.

Table 2.1: Literature Review Table (Table Continued).

Paper Name	Authors & Date	Subject	Objective/Aim	Key Findings
Harnessing Artificial Intelligence for Enhancing Digital Content Production[2]	Dr. Violet Barman, 2024	AI in Digital Media	To explore the intersection of AI and digital content production, focusing on automated content generation and personalization	• AI-driven tools can automate content creation, optimizing both quality and relevance. • Personalization through AI enhances user engagement by tailoring content to individual preferences.
AI-Generated Content (AIGC): A Survey[25]	Jiayang Wu, Wensheng Gan, Zefeng Chen, Shicheng Wan, Hong Lin, 2023	Artificial Intelligence	To provide an extensive overview of AI-generated content covering its definition, capabilities, and applications.	• Discusses the evolution and current state of AI-generated content technologies. • Explores various applications, including text, image, and multimedia content generation.
Paper Name	**Authors & Date**	**Subject**	**Objective/Aim**	**Key Findings**
Audience Segmentation in Social Media[11]	Alexander Lang, 2017	Social Media Analytics	To present an approach that detects various audience attributes, including author location, demographics, behavior, and interests, across multiple social media sources and languages	• Utilizes AI techniques to create author profiles for over 300 different analysis domains daily. • Demonstrates the effectiveness of AI in processing diverse and multilingual social media data for audience segmentation.

Table 2.1: Literature Review Table (Table Continued).

Paper Name	Authors & Date	Subject	Objective/Aim	Key Findings
Artificial Intelligence in Audience Segmentation: Enhancing Personalization in Digital Marketing[21]	Tran, H., Nguyen, D., 2021	AI in Marketing and Social Media Analytics	To explore how machine learning algorithms and natural language processing are used to segment audiences based on behavioral data from social media.	• Demonstrates the use of clustering algorithms like k-means and hierarchical clustering for audience segmentation. • Explains how sentiment analysis helps refine audience preferences.
Deep Learning for Social Media Audience Segmentation: A Case Study[27]	Zhang, Y., Li, F., 2020	Deep Learning and Social Media Analytics	To demonstrate how deep learning techniques like convolutional neural networks (CNNs) and recurrent neural networks (RNNs) areused to segment social media audiences dynamically.	• Introduces a pipeline for processing social media posts to identify audience clusters • Discusses challenges, such as handling unstructured data and ensuring scalability.

Table 2.1: Literature Review Table (Table Continued).

Paper Name	Authors & Date	Subject	Objective/Aim	Key Findings
Clustering Algorithms for Audience Segmentation in Social Media Advertising[1]	Ali, M., Chen, L., 2022	Social Media Marketing and Data Analytics	To analyze how clustering algorithms can improve targeting in social media advertising by grouping audiences with similar traits.	• Explores practical examples of clustering techniques such as DBSCAN and k-means. • Focuses on leveraging user interaction data (likes,shares, comments)for segmentation.
Natural Language Processing for Audience Profiling and Segmentation in Social Media[10]	Kumar, R., Sharma, P., 2021	NLP and Social Media Analytics	To discuss how NLP techniques can be used to analyze text data from social media for audience segmentation.	• Utilizes AI Demonstrates the use of topic modeling (LDA) and sentiment analysis to profile users. • Discusses case studies of segmenting audiences on platforms like Twitter and Instagram
Paper Name	**Authors & Date**	**Subject**	**Objective/Aim**	**Key Findings**
Predicting Social Media Popularity with Machine Learning[9]	J. Kim, 2021	Machine learning models for predicting social media popularity Analytics	To develop and compare machine learning models for predicting the popularity of Reddit posts based on textual content.	• Neural networks outperform other models, such as linear regression and random forest, in predicting post popularity. • Textual features and metadata are critical for accurate predictions.

Table 2.1: Literature Review Table (Table Continued).

Predicting Popularity Trend in Social Media Networks with Multi-Layer Temporal Graph Neural Network[18]	Not Specified, 2024	Graph neural networks for predicting social media content popularity	To propose a multi-layer temporal graph neural network framework to predict what content will become popular on social media platforms.	• Graph neural networks effectively capture temporal and structural patterns in social media interactions. • Demonstrates superior performance in identifying potential viral content
Paper Name	**Authors & Date**	**Subject**	**Objective/Aim**	**Key Findings**
Media and Artificial Intelligence: Current Perceptions and Future Outlook[8]	Kasneci, G., Jackson, P., Cao, L., et al, 2023	AI applications in media and content creation	To explore the integration of AI in the media industry, focusing on content generation, personalized recommendations, and audience analytics.	• AI models like GPT-3 enhance creative content generation and personalization but face challenges with coherence and biases.
The Impact of Predictive Analytics and AI on Digital Marketing Performance[1	Not specified, 2023]	Predictive analytics in digital marketing	To review how AI and predictive analytics improve marketing strategies, including email, social media, and audience segmentation.	• Businesses using AI report improved engagement, higher conversions, and better ROI through targeted marketing strategies.

Table 2.1: Literature Review Table (Table Continued).

Paper Name	Authors & Date	Subject	Objective/Aim	Key Findings
Impact of AI on Social Marketing and Its Usage in Social Media[14]	Not specified, 2023	AI in social marketing and social media	To explore how AI enables companies to predict consumer behavior and optimize social media content for engagement.	• AI enhances reach and engagement by analyzing emotions, reactions, and user behavior on social media platforms.
Leveraging AI for Social Media Trend Prediction[3]	Chen, R., Huang, L., 2022	AI in trend prediction for social media	To apply machine learning models to predict emerging Trends and user behavior on social platforms.	• Random forests and gradient boosting algorithms perform well in identifying patterns and predicting trends.

Table 2.1: Literature Review Table (Table Continued).

Paper Name	Authors & Date	Subject	Objective/Aim	Key Findings
Forwarding in Social Media: Forecasting Popularity of Public Opinion With Deep Learning[13]	Not specified, 2023	Utilizing deep learning algorithms to predict the popularity of public opinion in social media	To design a deep learning algorithm that predicts the trend of public opinion popularity (PPO) by analyzing the information entropy of forwarding circles.	• The proposed model accurately captures the time-series evolution trend of PPO on social media, outperforming traditional methods.
Mining Social Media Data for Sentiment Analysis and Trend Prediction[16]	Not specified, 2023	Application of deep learning in sentiment analysis and trend prediction using social media data	To propose a novel sentiment analysis model that incorporates contextual embeddings and deep learning architectures for enhanced sentiment classification and trend prediction.	• The model demonstrates significant advancements over traditional lexicon-based and machine learning approaches, providing more granular and accurate sentiment analysis.

Table 2.1: Literature Review Table (Table Continued).

Paper Name	Authors & Date	Subject	Objective/Aim	Key Findings
Predicting Social Media Popularity With Large Language Models[17]	Not specified, 2023	Utilization of large language models in predicting social media content popularity.	To integrate Low-Rank Adaptation (LoRA) with large language modelsfor interpreting and learningfrom semantic-rich sociamedia text,enhancing prediction accuracy.	• The approach facilitates a deeper understanding of social media content compared to conventional metadata-based methods, improving popularity prediction.
Sequential Prediction of Social Media Popularity with Deep Temporal Context Networks[24]	Bo Wu, WenHuang Cheng, Yongdong Zhang, Qiushi Huang, Jintao Li, Tao Mei, 2022	Sequential prediction of social media popularity using deep learning.	To propose Deep Temporal Context Networks (DTCN) that incorporate temporal context and attention mechanisms for sequential prediction of social media content popularity.	• DTCN outperforms
Paper Name	**Authors & Date**	**Subject**	**Objective/Aim**	**Key Findings**
PopALM: Popularity Aligned Language Models for Social Media Trendy Response Prediction[26]	Erxin Yu, Jing Li, Chunpu Xu, 2024	Aligning language models with popularity metrics for predicting trendy responses on social media.	To develop PopALM,a framework that aligns language models with popularity metrics using reinforcement learning, enhancing the prediction of popular responses to social media events.	• PopALM improves the performance of language models in generating responses that align with mainstream public reactions, utilizing curriculum learning in proximal policy optimization.

While prior studies have made significant progress in understanding social media engagement, several critical gaps remain:

a. Lack of Causal Insights – Most engagement models identify correlations but do not establish causal relationships between content features and user interactions.

b. Platform-Specific Biases – Studies rarely account for how social media algorithms influence engagement metrics, potentially skewing predictive models.

c. Limited Integration of User Preferences – Existing models rely on engagement metrics but do not incorporate direct user feedback from surveys.

d. Undefined Optimal Posting Times and Trending Topics – While some studies suggest best practices for posting, there is no AI-driven, data-backed framework to define optimal posting times and identify trending subjects.

This thesis aims to address these gaps by:

a. Helping understand the audience better through AI-driven segmentation and engagement analysis.

b. Categorizing the audience by integrating structured engagement data from Meta's

c. Graph API with qualitative user insights from surveys.

d. Defining good posting times and trending subjects by analyzing historical engagement trends and user preferences.

e. Increasing engagement by understanding the audience and developing predictive models that recommend optimized content strategies.

f. Applying both deep learning (BERT) and traditional models (Random Forest) to enhance engagement prediction accuracy.

g. Incorporating explainability tools (SHAP, LIME) to provide actionable insights for content creators and digital marketers.

2.2 CONCLUSION OF THE LITERATURE

This literature review highlights key findings in AI-driven audience segmentation and engagement prediction while identifying critical research gaps. By addressing these gaps, this study aims to enhance content strategy optimization using a data-driven approach. The next chapter details the research methodology employed to achieve these objectives.

3. METHODOLOGY

3.1 INTRODUCTION

This section outlines the methodology employed in this study, which integrates quantitative and qualitative approaches to analyze social media engagement and predict content performance. A mixed-methods approach was adopted to combine structured data from Meta's Graph API with insights from an online survey, ensuring a comprehensive understanding of user behavior and engagement trends. The data preprocessing steps, including cleaning, splitting, and merging, were carefully designed to prepare a robust dataset for analysis. The study leverages advanced machine learning models, such as BERT for text classification, and traditional models like Random Forest for engagement prediction, chosen for their proven effectiveness in handling structured and unstructured data. Tools such as Python, PyTorch, and the Meta Graph API were utilized to streamline data collection, preprocessing, and model implementation. Ethical considerations, including user privacy and data security, were prioritized throughout the research process. This methodology ensures a rigorous and reliable approach to achieving the study's objectives

3.2 RESEARCH DESIGN

This research adopts a mixed-methods design, combining quantitative data (e.g., engagement metrics from the Graph API) with qualitative and self-reported data (e.g., survey responses). This approach is supported by methodological frameworks presented in Creswell & Plano Clark (2017), Tashakkori & Teddlie (2010), and Johnson et al. (2007) [4, 20, 7]. This approach was chosen to address the multifaceted nature of social media engagement, where structured data (e.g., likes, comments, shares) provides measurable metrics, while survey data offers contextual insights into user preferences and behaviors.

a. Meta Graph API: Structured data, including post captions, engagement metrics (likes, comments, shares), and hashtag-related content, was collected to analyze engagement trends and develop predictive models. This data forms the foundation for understanding content performance on social media platforms.

b. Online Survey: A survey targeting social media users was conducted to gather demographic data (e.g., age, gender) and insights into user preferences (e.g., preferred content types, video lengths). The survey included both structured (e.g., Likert scale) and open-ended questions, providing a deeper understanding of user behavior and complementing the API data. By integrating these datasets, the study ensures a comprehensive analysis of social media engagement, enabling the development of actionable insights and predictive models tailored to audience preferences.

3.3 DATA COLLECTION

The data for this study was collected from two primary sources: an online survey and Meta's Graph API. These sources were selected to provide complementary datasets, combining structured engagement metrics from social media platforms with qualitative insights into user preferences and behaviors.

3.3.1 Online Survey

The online survey was conducted using Google Forms and consisted of 20 questions designed to gather information about users and their preferences when engaging with social media. The survey focused on two key categories of data:

i. User Information Data:

Table 3.1: User Information Data Table.

Data Name	**Data Type**	**Data Values**	**Description**
Age	Category	<18 — 18 - 24 — 25 - 34 — +35	Captures the user's age group to analyze engagement across different demographics.
Data Name	**Data Type**	**Data Values**	**Description**
Gender	Category	Male — Female	Records the gender of users to understand gender-based preferences.
City	String	*User Input*	Captures user-reported city information for geographical segmentation and analysis.
Status	Category	Working — Student — Nothing	Identifies the employment or education status of users to explore its influence on engagement.
Device Type	Category	Android — IOS	Determines the device type used to access social media for technical and preference analysis.

ii. User Preferences Data:

Table 3.2: User Preference Data Table.

Data Name	Data Type	Data Values	Description
Favorite App	Category	Instagram | Facebook TikTok	Identifies the social media platform users interact with most frequently.
Data Name	**Data Type**	**Data Values**	**Description**
Daily Time	Category	0-1h | 1-3h | 3-5h | +5h	Captures the time users spend on social media daily.
Favorite Content Type	Category	Comedy | Entertaining Educational | Lifestyle	Determines the type of content users find most engaging.
Engagement Content	Category	Images | Reels | Stories | Live	Identifies the formats users prefer to engage with.
Reactions	Category	Like | Share | Comment | Save	Tracks how users interact with content they like.
Engagement Motivation	Category	Title | Topic | Visuals | Humor | Education | Value	Explores reasons driving users to engage with specific content.
Influencers Engagement	Category	Never | Sometimes | Always	Measures user interaction with influencers and public figures.
Following Motivation	Category	Quality | Consistency Visuals | Shared Interests	Identifies reasons users decide to follow accounts.
Unfollowing Reasons	Category	Inconsistency | Ads Uninteresting Content	Captures why users disengage or unfollow accounts.
Preferred Video Length	Category	<30s | 30s-90s | +90s	Explores user preferences for video durations.

Table 3.2: User Preference Data Table (Table Continued).

Data Name	Data Type	Data Values	Description
Stops Watching	Category	Quality | Topic | Length	Examines reasons for users stopping content consumption.
Preferred Video Type	Category	AI | Visuals | Human	Identifies the types of videos users engage with most frequently.
Reaction to CTA	Category	Like | Comment | Share | Save	Captures user responses to Call-to-Action (CTA) prompts.
Preferred Time	Category	9am–12pm | 12pm–3pm | etc.	Analyzes the times users are most active on social media platforms.

3.3.2 Meta Graph API

The Meta Graph API is a tool developed by Meta to facilitate developers in accessing data from Meta's platforms conveniently. In this study, the API was employed for two purposes:

a. Public Account Data:

Data was collected from the writer's public account. The API output was stored in JSON format, containing the following details for each post:

i. Post Details: ID, Post Link, Caption, Media Type, and Posting Time.

ii. Engagement Insights: Likes, Comments, Shares, Saves, Initial Plays, and Accounts Reached.

b. Hashtag Search Results:

The API was also utilized to perform hashtag searches. This provided data on posts related to specific hashtags, including descriptions and their associated categories. The data was saved in JSON format for further analysis.

By leveraging the Meta Graph API, the study obtained structured and reliable data to analyze social media trends and engagement metrics.

The data collected from the online survey and Meta's Graph API serves distinct but complementary purposes in this study.

iii. User Information Data:

The demographic data (e.g., age, gender, status, and device type) obtained from the online survey is crucial for audience segmentation. By categorizing users into meaningful groups, this data helps to identify variations in social media preferences and behaviors across different demographics.

iv. User Preferences Data:

The preferences data, such as favorite content types, preferred video lengths, and reactions to content, provides insights into user engagement patterns. This data will inform the development of strategies to optimize content creation, align with user interests, and enhance engagement.

Meta Graph API Data:

The structured data retrieved via Meta's Graph API, including post captions, hashtags, and engagement metrics (likes, comments, shares, etc.), forms the foundation for predictive modeling. By analyzing trends in this data, the study aims to identify factors that contribute to content performance and develop algorithms to predict future engagement levels.

By integrating these datasets, the study ensures a comprehensive approach to understanding user behavior and content performance. The combined insights enable the development of predictive models and actionable strategies for social media content optimization.

3.3.2 Addressing Biases in Data Collection

Survey Bias:

The survey responses were gathered from individuals who voluntarily participated, which may introduce self-selection bias. This means that the data could overrepresent certain groups, such as highly engaged users, while underrepresenting those who engage less frequently. Additionally, repetitive patterns and similar answers may emerge due to the nature of voluntary responses.

However, this study is focused on a specific platform and account, rather than aiming for generalizable social media engagement trends. Engagement patterns vary significantly between different accounts and content categories, making it difficult to establish a universal formula. Therefore, if this study were to be repeated, the survey would need to be carefully redesigned to ensure a more balanced and representative dataset while maintaining a personalized approach that aligns with the unique nature of social media engagement.

That being said, this does not mean the findings cannot be generalized. A larger survey could be conducted and distributed across multiple accounts within the same country to increase the breadth of analysis. However, the primary focus remains on understanding what makes an individual's own content gain engagement and how insights can be personalized, rather than creating a broad, one-size-fits-all model.

Platform Specific Bias:
Social media platforms operate on proprietary algorithms that determine content visibility, engagement reach, and user interactions. While these algorithms are not publicly disclosed, it is well understood that multiple factors influence engagement, such as posting time, video type, content category, and trending topics.

Since the exact workings of these algorithms remain unknown, this study does not aim to define a universal formula for engagement but rather to establish practical insights based on real engagement data. By analyzing past engagement metrics alongside audience survey responses, this research seeks to provide a human-understandable interpretation of how these algorithms behave in real-world content interactions.

Ultimately, the goal is not to generalize a rigid engagement formula but to focus on what makes specific content successful on a given account. This personalized approach ensures that findings remain applicable to individual content creators while still offering insights that could be expanded in future studies.

3.4 SURVEY DATA INTEGRATION

3.4.1 Audience Segmentation

Audience segmentation was performed based on demographic and behavioral data, as supported by prior research in digital marketing and social media [11, 21, 27]. Audience segmentation refers to the process of dividing survey respondents into distinct groups based on shared demographic characteristics. This approach enables more personalized analysis of content preferences and behavior, making it possible to tailor strategies for specific user types. In this study, segmentation is based on variables collected in the first part of the survey: age group, gender, city, status (e.g., student, working, not working), and device type (Android or iOS).

These specific variables were selected because they are among the most influential factors in shaping digital content consumption. Age often affects preferred content formats and engagement styles; gender can reflect varying content interests and interaction patterns; city provides a geographical lens to capture potential regional or cultural differences in behavior. Status was included to account for lifestyle context, as students and professionals may differ in their content needs and usage times. Finally, device type may indirectly reflect user behavior or preferences tied to platform features or user experience, making it a useful technical variable for segmentation.

By grouping respondents based on these criteria, the study aims to extract patterns that reveal what type of content performs best within each segment. This segmentation will later support recommendation strategies by identifying, for instance, which content types should be posted for a specific audience at a specific time. It also allows for performance comparisons between content types within and across audience groups, offering deeper insight into what drives engagement on social media platforms.

3.4.2 Behavioral Insights and Modeling

Behavioral modeling was guided by prior studies examining how AI can enhance content strategy and audience interaction [2, 8]. To complement demographic segmentation, this study also analyzes behavioral data collected through the survey — specifically focusing on content-related preferences and interaction patterns. Key variables include Preferred Time, Daily Time, Favorite Content Type, Engagement Content, Preferred Video Length, Reaction, Follow Reason, Unfollow Reason, and Reaction to CTA. These features were chosen because they provide practical insights into how users consume, engage with, and respond to content on social media.

a. Preferred Time indicates when users are most active, helping to identify optimal posting windows.

b. Daily Time reflects the amount of time users spend on social media per day, which may correlate with engagement depth or content fatigue.

c. Favorite Content Type provides a direct look into what kind of content (educational, entertaining, lifestyle, etc.) resonates most with different segments.

d. Engagement Content shows which formats users prefer to engage with, such as stories, reels, or live videos.

e. Preferred Video Length reveals attention span and expectations for content pacing and duration.

f. Reaction captures the typical way users respond to content (like, comment, share, save), which can be compared to actual engagement metrics.

g. Follow Reason and Unfollow Reason highlight the main drivers for audience growth or loss — such as content quality, consistency, or topic interest.

h. Reaction to CTA (Call-to-Action) reflects how users respond when prompted to take an action like liking, saving, or following.

By examining these variables within each audience segment (age, gender, status, etc.), the study can uncover behavior patterns and preferences that are unique to different user groups.

These insights form the foundation for user profiling and can support predictive models and strategic content recommendations — such as which formats and times are best suited for a particular audience, or what factors most influence following or unfollowing behavior.

3.4.3 Prediction and Recommendation

The modeling framework draws inspiration from previous research on predicting engagement and social media popularity [9, 24, 26]. The combination of survey-based user profiles and performance metrics from the Graph API opens the door for predictive modeling and recommendation strategies. By linking user preferences (e.g., favorite content type, preferred format, time of engagement) with actual content performance (likes, shares, saves, reach), the system can begin to identify patterns and make predictions about future engagement.

For example, if a specific segment of users — such as students aged 18–24 — shows a strong preference for short educational reels and this content format also performs well in terms of saves and reach, the system can recommend that similar content be produced and scheduled during the users' preferred activity hours. This predictive capability can later be enhanced through machine learning models that incorporate both survey variables and post insights to forecast which content type, format, and timing would perform best for a particular target audience.

Ultimately, this stage of the study aims to move from observation to optimization — using data not just to describe user behavior, but to shape smarter, personalized content strategies that improve reach, engagement, and follower growth.

3.5 DATA PREPROCESSING

In this section, we focus on the methods used to preprocess the collected data. After the data was obtained and stored, it underwent a series of cleaning, transformation, and preparation steps to ensure its suitability for analysis and model training. These preprocessing steps were essential to eliminate inconsistencies, handle missing values, and structure the data for optimal use in machine learning models.

The results obtained from the Hashtag Search API were initially saved in a JSON file for use in model tuning and future predictions. Before proceeding with model development, the data underwent several preprocessing steps to ensure its quality and suitability for analysis:

i. Duplicate Removal:

Duplicate entries in the dataset were identified and removed to prevent biases, incorrect predictions, or unequal weight assignments during the model training phase.

ii. Character Filtering:

o Post descriptions containing non-English or non-Arabic characters were removed to maintain consistency and relevance to the study's focus.

iii. Column Separation:

The dataset was divided into two key columns:

Text Column: Contained the cleaned post descriptions.

Label Column: Represented the associated category or classification of the post, which was numerically encoded for machine learning.

iv. Dataset Splitting:

The preprocessed data was split into two subsets:

Validation Set: Used to evaluate the model's performance during training and fine-tuning.

Label Column: Represented the associated category or classification of the post, which was numerically encoded for machine learning.

After completing these steps, the dataset was fully prepared for model tuning and future predictions.

3.5.1 Graph API Posts

The data retrieved using the Graph API included two separate datasets: one containing post details and another containing post insights. These datasets were initially saved as JSON files. The following preprocessing steps were applied to ensure the data was organized, cleaned, and ready for analysis:

a. Merging *Datasets*: The two datasets (post details and post insights) were merged by matching each post with its relevant insights. This step ensured that the final dataset contained all necessary information, including post captions, media types, timestamps, and engagement metrics (likes, comments, shares, saves, etc.).

b. Columns *Cleaning*: Unnecessary columns from the merged dataset were removed to retain only relevant attributes for analysis.

c. Timestamp *Transformation*: Timestamps were transformed into a more readable format and split into two new columns:

i. Day: Representing the specific day of posting.

ii. Time: Representing the time of posting.

a. Column *Renaming*: Insight column names, originally in French, were translated into English for consistency and ease of understanding during analysis.

b. Post Caption *Extraction*: Captions from posts were extracted and saved in a separate file to facilitate category prediction using natural language processing (NLP) techniques.

c. Category Prediction:

i. NLP models were applied to the captions file to classify posts into predefined categories (e.g., educational, entertaining).

ii. The predicted categories were added as a new column, Category, in the final dataset.

After these preprocessing steps, the dataset was fully prepared, containing all necessary features for both theoretical analysis and technical model development.

3.5.2 Survey Data

The data collected from the online survey will be automatically saved in real time to a Google Sheets file. Once the data collection is complete, it will be downloaded to a local server for cleaning and preprocessing. The planned steps for processing the survey data are as follows:

i. Handling Missing Values:

a. The dataset will be reviewed row by row to check for missing or incomplete values.

b. If a missing value has minimal impact on the overall dataset, it will be replaced with an appropriate placeholder (e.g., the mean or mode).

c. If the missing values significantly affect the data quality, the corresponding row will be removed entirely.

ii. Data Splitting: The dataset will be divided into two categories for analysis:

User Information Data: Includes demographic attributes such as age, gender, and status, which will be used for audience segmentation.

User Preferences Data: Contains details about content preferences, reactions, and behaviors, which will be used to identify patterns and inform predictions.

iii. Data Integration:

Certain user preferences (e.g., preferred content types or reactions) will be merged with post insights from the Graph API dataset. This integration will facilitate future recommendations by combining user-specific behaviors with content performance metrics.

iv. Encoding Categorical Responses: Categorical survey responses will be numerically encoded for machine learning applications. For example:

Gender will be encoded as 0 for Male and 1 for Female.

Content preferences will be assigned numerical values (e.g., Comedy = 0, Educational = 1, Lifestyle = 2).

v. Readying the Dataset for Analysis:

After cleaning and encoding, the final dataset will be split into training and testing subsets to support predictive modeling and segmentation analysis.

3.5.3 Feature Selection

The selection of features was based on factors that influence engagement on social media, as identified in prior research and practical content strategies. The following features were chosen for the predictive models:

3.5.3.1 Text-Based features

a. Captions: The text accompanying a post often influences user engagement by providing context and encouraging interactions. Captions were processed using BERT to extract meaningful representations.

b. Hashtags: *These help* increase content visibility by making posts searchable. The number and type of hashtags were included as features.

3.5.3.2 Temporal features

a. Posting Time: The time of day and day of the week affect how many users see and interact with a post. These were included as separate variables to *analyze* their impact on engagement.

3.5.3.3 Engagement metrics (from previous posts)

a. Likes, Comments, Shares, Saves: The historical engagement of past posts was used to train the model, helping it learn patterns that contribute to high engagement.

3.5.3.4 Content type

a. Posts were categorized as images, videos, or carousels, as different formats tend to perform differently based on platform algorithms and user *behavior*.

These features were selected because they capture key engagement factors, balancing textual, *behavioral*, and temporal elements to improve model accuracy.

3.5.4 Interaction Terms

In addition to individual features, interaction terms were considered to *analyze* how combinations of factors influence engagement differently than standalone features. Interaction effects occur when two or more variables work together in a way that changes the engagement outcome.

For example, posting time and content type might not independently determine engagement, but their combination could reveal patterns—such as videos performing better at night while images get more engagement in the morning. Similarly, the relationship between hashtags and engagement metrics may vary depending on post type or audience preferences.

By incorporating interaction terms, the study aims to explore how different factors influence engagement when combined, rather than treating them as isolated elements. The final selection of interaction terms will be determined based on data analysis and model performance evaluation.

3.6 MACHINE LEARNING/DEEP LEARNING MODELS

This section outlines the machine learning and deep learning models employed in this study, detailing their functionality, training processes, and relevance to the research objectives. The models were selected based on their suitability for specific tasks, such as text classification, engagement prediction, and audience segmentation.

3.6.1 BERT for Text Classification

For the task of post classification based on caption text, this study adopts BERT (Bidirectional Encoder Representations from Transformers), a state-of-the-art transformer-based model developed by Google. BERT has demonstrated exceptional performance in a wide range of natural language processing tasks, particularly in text classification and sentiment analysis, due to its ability to understand contextual relationships between words in a sentence. Its multilingual capability also makes it suitable for this project, as the dataset includes content in both English and Arabic.

The pre-trained `bert-base-multilingual-cased` model was selected to reduce training time while benefiting from transfer learning. This decision is supported by the original BERT study by Devlin et al. (2019), which established BERT as a foundational model in NLP research [5].

3.6.1.1 Relevance to the study

Social media captions often consist of short, informal text, requiring a model that can understand nuanced language. BERT's ability to analyze bidirectional context makes it particularly well-suited for this task, aligning with the study's objective of providing actionable insights for content creators.

3.6.1.2 Why was BERT Chosen?

BERT was selected for this study based on the following considerations:

a. Pre-trained on Large Datasets: By leveraging a pre-trained model, the need for extensive training from scratch was minimized, significantly reducing computational resources and time.

b. Multilingual Support: BERT's support for multiple languages made it ideal for processing captions in both English and Arabic, ensuring inclusivity in the study's scope.

c. Proven Effectiveness: BERT has demonstrated superior performance in various natural language processing tasks, such as text classification, sentiment analysis, and question answering, making it a reliable choice for this study.

3.6.2 Traditional Machine Learning Models

The study employs a combination of deep learning and traditional machine learning models for engagement prediction. The choice of models was based on their effectiveness in handling the two main tasks:

a. Text Classification: Understanding and categorizing content from captions and hashtags.

b. Engagement Prediction: Forecasting engagement metrics such as likes, comments, shares, and saves.

3.6.2.1 Text classification: Why BERT?

For text classification, BERT was chosen due to its strong contextual understanding and ability to capture nuanced meanings in social media text. BERT processes words in relation to their surrounding context, making it well-suited for analyzing captions and hashtags, which often have multiple meanings.

Alternatives Considered:

a. RoBERTa: An optimized version of BERT that improves training efficiency and performance but requires more computational power.

b. GPT (Generative Pre-trained Transformer): While powerful, GPT models are designed primarily for text generation rather than classification, making them less suitable for this study's objectives.

3.6.2.2 Engagement prediction: Why random forest?

For engagement prediction, Random Forest was selected due to its robust performance on structured data, its ability to handle non-linearity, and interpretability. Random Forest works by constructing multiple decision trees and averaging their predictions, reducing overfitting while maintaining accuracy.

Alternatives Considered:

a. XGBoost (Extreme Gradient Boosting): Generally achieves higher accuracy than Random Forest but requires more computational resources.

b. Support Vector Machines (SVM): Effective for classification tasks but less efficient for large-scale engagement prediction tasks.

3.6.3 Future Directions

Additional NLP models, such as GPT or RoBERTa, may be explored to further analyze post captions and hashtags for trends and patterns. These models will complement the existing framework, providing deeper insights into content performance and user behavior.

4. IMPLEMENTATION

4.1 IMPLEMENTATION AND ANALYSIS

This section provides a detailed explanation of the practical aspects of this study. It covers the data collection methods and how the retrieved data was stored and preprocessed. Additionally, it discusses the models used, detailing their training, fine-tuning, and application for predictions and segmentation tasks. Finally, the programming languages, tools, and frameworks utilized throughout the implementation process are outlined.

4.1.1 Data Collection Using Graph API

The Meta Graph API was utilized to retrieve structured data from social media posts, including post captions, insights, and hashtags. The implementation process involved the following steps:

4.1.1.1 API setup

A developer account was created through the Meta Developer Platform, where the necessary permissions were granted to enable API searches. Once the permissions were in place, an access token was generated to authenticate API requests. Python was used to retrieve the page account name and page account ID, which were required for making API calls. These API requests provided access to Instagram account post information and allowed for hashtag searches, retrieving all relevant data for the study.

4.1.1.2 Graph API posts information and insights retrieval

The following figure illustrates the workflow for retrieving posts information and engagement insights using the Meta Graph API. The process begins with the API request to fetch post details, including post ID, caption, media type, link, and timestamp. Based on the media type, the workflow dynamically determines the relevant engagement metrics to retrieve, such as likes, comments, shares, impressions, reach, and plays. These insights are appended to the corresponding post data to form a unified dataset. Finally, all the collected data is saved in a structured JSON file for further preprocessing and analysis. This streamlined approach ensures comprehensive data collection for modeling and predictions in subsequent steps.

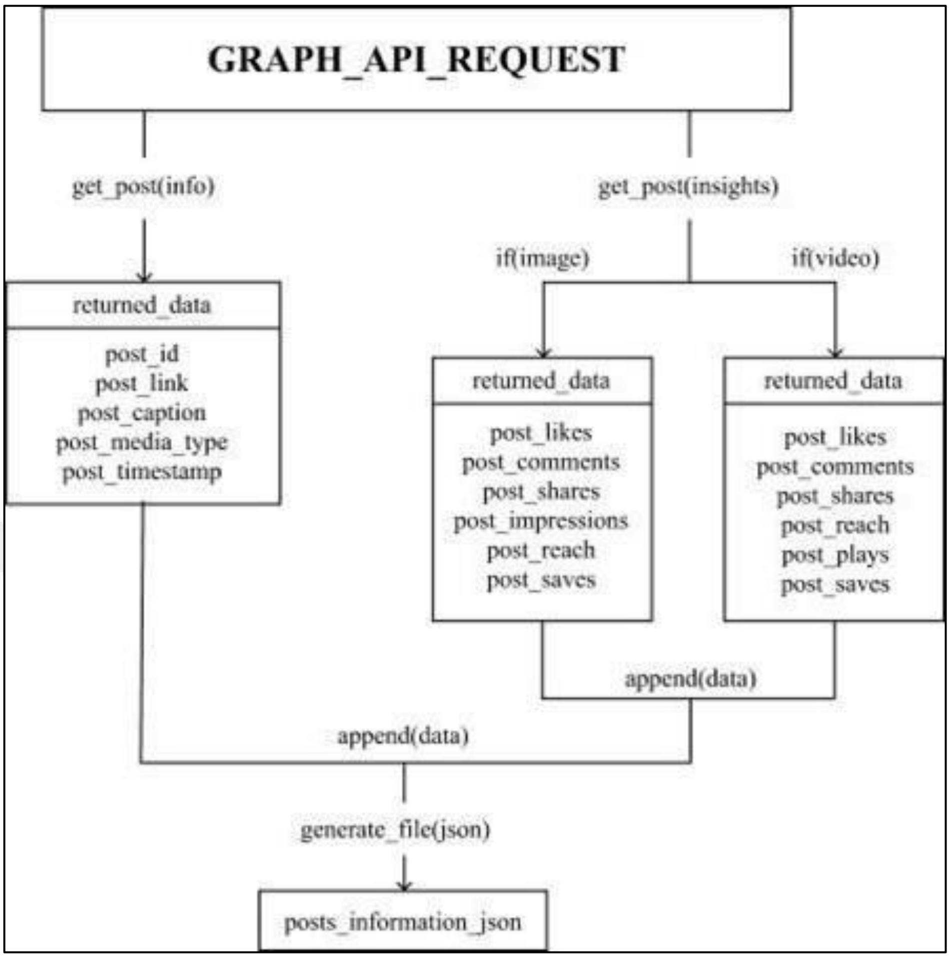

Figure 4.1: Workflow for Posts Information and Insights Retrieval.

4.1.2 Graph API Hashtag Search Graph API Hashtag Search

The following figure demonstrates the process of hashtag search and media retrieval using the Meta Graph API. The workflow begins with an API request to retrieve the hashtag ID using the hashtag name. This ID is then used to fetch both top and recent media posts associated with the hashtag. The data is stored in structured JSON files for further preprocessing and analysis.

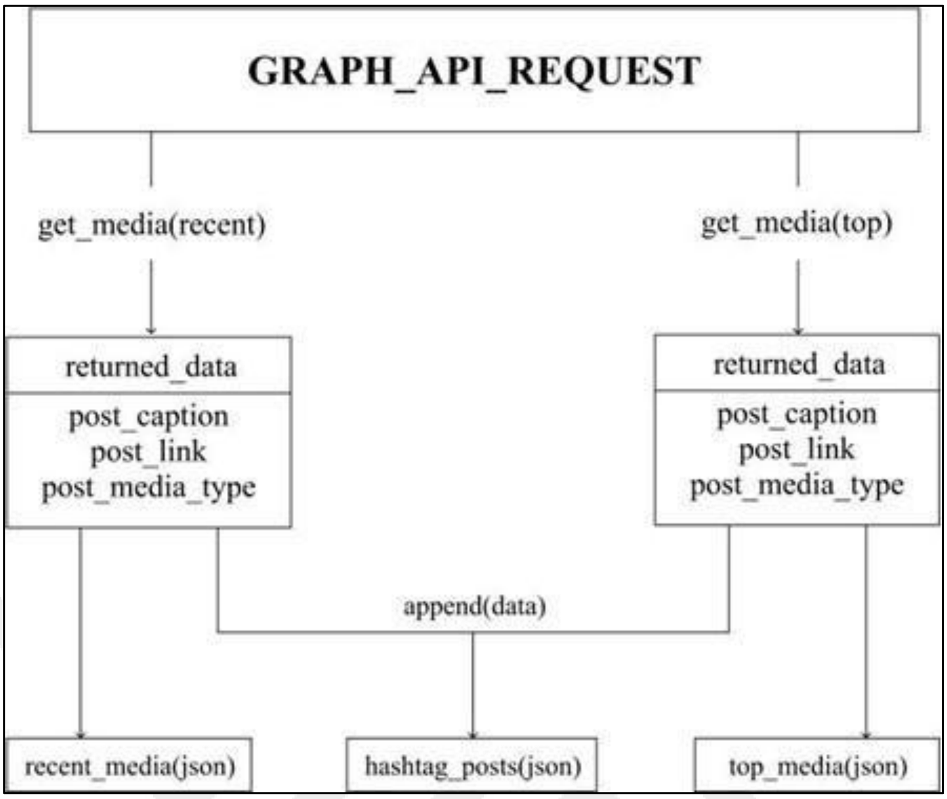

Figure 4.2: Workflow for Hashtag Search and Media Retrieval.

4.1.3 Survey Data Integration

The integration of survey data forms a critical step in the implementation process, providing the foundation for user segmentation, behavioral analysis, and eventual modeling. The survey responses included a mix of demographic information (such as age, gender, status, and device type) and behavioral preferences related to content types, video formats, engagement habits, and social media usage patterns. Before analysis, the data was cleaned and standardized to remove inconsistencies, non-English entries, and formatting noise such as special characters or mixed-language responses. Each survey dimension whether demographic or behavioral—was processed to allow for meaningful segmentation and comparison. The integration of this cleaned and structured dataset into the overall system enables the generation of tailored insights and supports future predictive applications. The following sections explain how each survey component was handled during implementation.

4.1.3.1 Audience segmentation

As part of the implementation phase, the cleaned and structured survey data was used to segment the audience based on key demographic attributes, including age, gender, status, and device type. This segmentation enabled the exploration of behavioral trends across distinct user groups. Multiple survey dimensions were analyzed, such as preferred content type, video length, engagement format, and social media usage habits. The goal was to determine how these preferences varied between segments and to prepare the data for later modeling and recommendation logic.

The segmentation was performed using Python's pandas library, with group-wise percentages calculated using the valuecounts(normalize=True) method. This allowed for the generation of normalized distributions for each behavioral variable within every demographic group. The results were visualized using seaborn heatmaps to reveal trends and differences across segments. Each segmentation variable was processed independently to ensure clarity and consistency in interpretation.

These visualizations and descriptive statistics are documented and discussed in detail in Chapter 5, where key insights and implications are derived from the observed patterns.

4.1.3.2 Multi-variable pattern mining

Beyond individual segmentation and descriptive analysis, this study incorporates a multi-variable pattern mining approach to uncover complex relationships between user demographics and content preferences. While segmentation provides valuable insights into isolated patterns (e.g., age vs. content type), real-world behavior is often influenced by multiple variables simultaneously.

This phase of the implementation aims to identify combined audience profiles and their corresponding behavior. For instance, discovering that users aged 25–34 who prefer educational content are also more likely to engage with videos between 30 and 90 seconds, or that males aged under 18 tend to engage more with comedic content during late-night hours.

To achieve this, combinations of key demographic variables (e.g., *age*, *gender*, *status*) are cross-analyzed with behavioral variables (e.g., *preferred video length*, *reaction type*, *preferred time to use social media*). The analysis is performed using Python, where filtered subsets are grouped by multiple attributes, and patterns are extracted based on frequency and engagement intensity.

This multi-variable approach not only supports more targeted content recommendations, but also contributes to the construction of predictive audience models in later stages of the research.

4.2 MODELS IMPLEMENTATION

This section provides a detailed explanation of the implementation processes for the models used in this research. It covers the training and fine-tuning phases, along with the steps taken to save and prepare the models for deployment.

4.2.1 Bert Model Tuning and Training

This section explains the implementation process of the BERT model used for text classification. The workflow is illustrated in Figure 3, which outlines the key steps involved in training and fine-tuning the model.

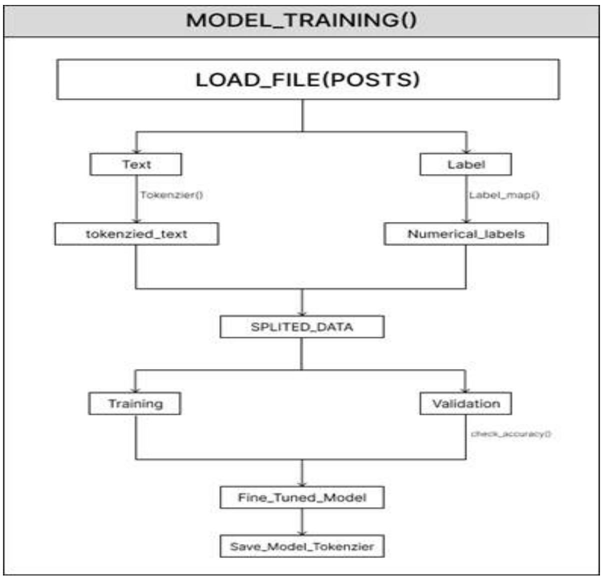

Figure 4.3: Workflow of Bert Model Training and Tuning.

Figure 4.3 illustrates the workflow for implementing the BERT model. The process starts with loading and preprocessing the dataset, including tokenizing the text and encoding labels. The data is then split into training and validation sets. The fine-tuning phase optimizes the pre-trained BERT model for the specific task of post categorization. Finally, the trained model and tokenizer are saved for inference and further analysis.

4.2.2 Bert Model Predictions

After the fine-tuned model and tokenizer were saved, they were utilized to classify Instagram post captions retrieved using the Graph API. The JSON file containing the captions was loaded, and the tokenizer was applied to tokenize the text into a format understandable by the model. A label map was then created, assigning a unique number to each predefined category, consistent with the training process. The tokenized captions, along with the label map, were fed into the model, which predicted the category for each caption based on the highest scoring logits. This process allowed for accurate classification of captions into categories such as Comedy, Sports, Educational, and Lifestyle.

4.3 TOOLS AND TECHNOLOGIES

The implementation of this study utilized a variety of tools and technologies to support data processing, model training, and website development. These included:

4.3.1 Python

Primary language for data preprocessing, model training, and integration with the Meta Graph API. Libraries used:

a. Transformers: For implementing and fine-tuning the BERT model.

b. PyTorch: For model training and optimization.

c. scikit-learn: For dataset splitting and preprocessing.

d. pandas: For handling and manipulating datasets.

4.3.2 Web Development

a. HTML & CSS: Used for structuring and styling the platform interface.

b. ReactJS: Used for building an interactive and dynamic user experience.

4.3.3 API

a. Meta Graph API: Used to retrieve Instagram post data, including captions, media types, timestamps, and engagement insights. The API also enabled hashtag searches to analyze trending content.

4.4 ETHICAL CONSIDERATION

The use of public data from social media platforms raises important ethical concerns, particularly regarding user privacy and data security. To address these issues, the following measures were implemented:

a. Anonymization: All data collected from the Meta Graph API and the online survey was anonymized to ensure that no personally identifiable information (PII) was stored or used in the analysis.

b. Data Security: Data was stored on secure servers with restricted access, and all API requests were encrypted to prevent unauthorized access.

c. Compliance with Platform Policies: The study adhered to Meta's API usage policies, ensuring that data was collected and used in compliance with platform guidelines.

d. Informed Consent: Survey participants were informed about the purpose of the study and how their data would be used. Participation was voluntary, and respondents could opt out at any time.

These measures were taken to ensure that the research was conducted ethically and responsibly, minimizing potential harm to users.

4.5 SUMMARY

This chapter detailed the implementation process of the study, covering data collection, preprocessing, model training, prediction, and platform development. Data was gathered from two primary sources: the Meta Graph API, which retrieved Instagram post captions, engagement insights (likes, comments, shares, reach), and hashtag-related content; and an online survey, which collected demographic data and user preferences related to social media content consumption. Once collected, the data underwent preprocessing, where duplicate and irrelevant information was removed, captions were tokenized, and labels were standardized for model training. Engagement insights were merged with post data to form a structured dataset, while survey responses were cleaned and categorized for integration with API data.

The BERT model was selected for content categorization and fine-tuned using a labeled dataset of social media captions. The training process was conducted using PyTorch optimizing hyperparameters to improve classification accuracy. After fine-tuning, the model was deployed for prediction, allowing new post captions to be automatically categorized into predefined groups such as Comedy, Sports, Educational, and Lifestyle. These predictions were then integrated into the platform for further analysis.

To present the results in an interactive and accessible format, the platform's frontend was developed using HTML, CSS, and ReactJS, ensuring a responsive user experience. The backend was structured to support the seamless integration of processed data and predictions, allowing content creators to access insights efficiently. The study relied on Python as the primary programming language, utilizing various libraries such as Transformers, PyTorch, scikit-learn, and pandas for data processing and model implementation. The Meta Graph API played a crucial role in obtaining Instagram-related data, while ReactJS was leveraged for frontend development.

Several challenges were encountered during implementation, including API rate limits, which imposed restrictions on data retrieval, requiring optimization in request handling. Additionally, data imbalance was observed in certain categories, affecting model performance, and was addressed during training. The computational demands of fine-tuning a deep learning model also required the use of GPU acceleration to improve efficiency. Despite these challenges, the implementation successfully established a structured workflow for content categorization, from data retrieval to model training and platform development. The next chapter will focus on evaluating the model’s performance and analyzing its effectiveness in classifying social media content.

5. RESULTS AND DISCUSSION

5.1 INTRODUCTION

This chapter presents the results of the implementation phase and discusses the key findings derived from the analysis of the survey and Graph API data. The focus is on interpreting the patterns and relationships uncovered through audience segmentation, behavioral modeling, and content preference analysis. Each result is accompanied by visualizations that illustrate how different user groups engage with content and how their preferences vary based on demographic and behavioral factors. The discussion contextualizes these findings within the broader objectives of the study, offering insights into how the extracted patterns can inform content strategy, personalization, and potential predictive modeling efforts. The chapter also reflects on the implications of these results and identifies notable trends, exceptions, or limitations observed during the analysis.

5.2 SURVEY DATA RESULTS

This section presents and discusses the results obtained from the survey data, supported by relevant graphs and descriptive statistics. The analysis highlights key insights and the conclusions that can be drawn from the observed patterns, helping to inform data-driven decisions.

5.2.1 Audience Segmentation

In this part, we present and discuss the results of the visualizations generated from the survey data, focusing specifically on audience segmentation. The analysis is based on three key demographic variables: age, gender, and status. These segments were used to explore variations in user behavior across several key survey features, including favorite content type, preferred video length, engagement preferences, and more. The goal is to understand how different audience groups interact with content and identify meaningful patterns that can support content strategy and personalization.

5.2.1.1 Favorite content type by age group

The heatmap in Figure 5.1 illustrates the distribution of favorite content types across different age groups, based on the survey responses. Users under 18 showed a strong preference for lifestyle content (52.6%), followed by educational (26.3%) and comedy (15.8%). The 18–24 age group displayed a more balanced distribution, with educational content leading at 47.5%, followed by lifestyle (28.1%) and comedy (13.5%). In the 25–34 age group, preference for educational content peaked at 64.5%, indicating a strong interest in informative media among young adults. Finally, the 35+ group showed a dual preference, split evenly between comedy and lifestyle content (50% each), with no recorded interest in educational or sports content.

These findings suggest that *younger audiences* lean toward lifestyle and entertainment, while *young adults* show a notable preference for educational media. This segmentation by age provides valuable insight for tailoring content strategies to meet the interests of each age group.

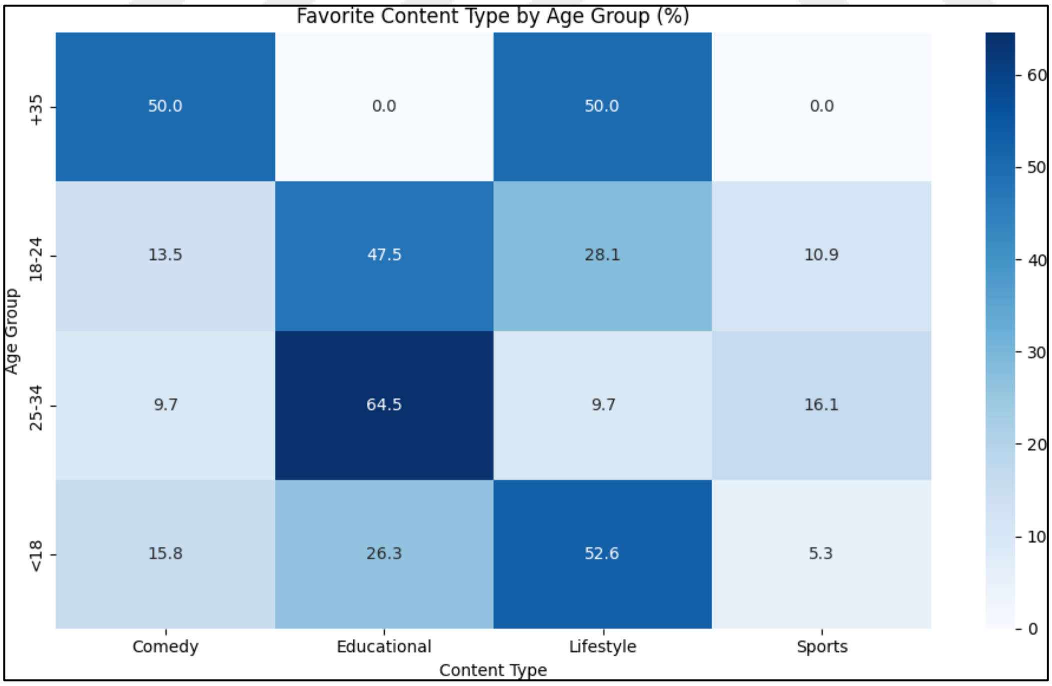

Figure 5.1: Favorite Content Type by Age Group (%).

5.2.1.2 Favorite content type by gender

Figure 5.2 displays how content preferences vary between male and female respondents. The data shows that both genders have a strong preference for educational content, with 51.4% of male respondents and 43.8% of female respondents identifying it as their favorite. Interestingly, female users displayed an equal preference for lifestyle content (43.8%), whereas males showed greater interest in sports (19.6%) and comedy (16.8%) compared to females (2.3% and 10.2%, respectively).

These results suggest that while educational content is widely favored across both genders, females tend to prefer lifestyle-oriented media, and males are more inclined toward sports and entertainment. These behavioral differences can inform more gender-aware content strategies for targeted engagement.

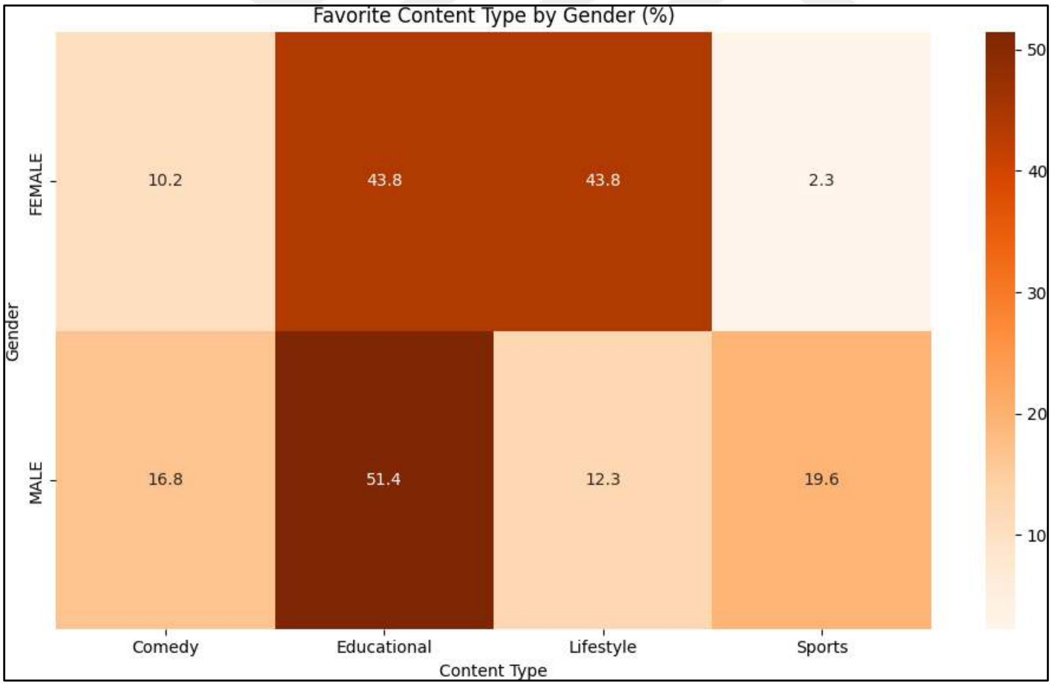

Figure 5.2: Favorite Content Type by Gender (%).

5.2.1.3 Favorite content type by status

Figure 5.3 illustrates the distribution of content preferences among users with different status categories: students, working individuals, and those currently not engaged in either. Across all groups, educational content stands out as the most favored, with preferences ranging from 47.1% (unemployed) to 50.0% (working individuals). Students also show a notable interest in lifestyle content (28.9%), while working individuals follow closely with 26.1%. The group marked as ”nothing” (i.e., neither working nor studying) demonstrated a more evenly distributed interest across all content types, with comedy, lifestyle, and sports each receiving 17.6%. This suggests a broader or more varied content appetite compared to the more structured interests of working and student users.

These patterns reinforce the value of considering status in audience segmentation, particularly in designing content strategies that align with user routines and informational needs.

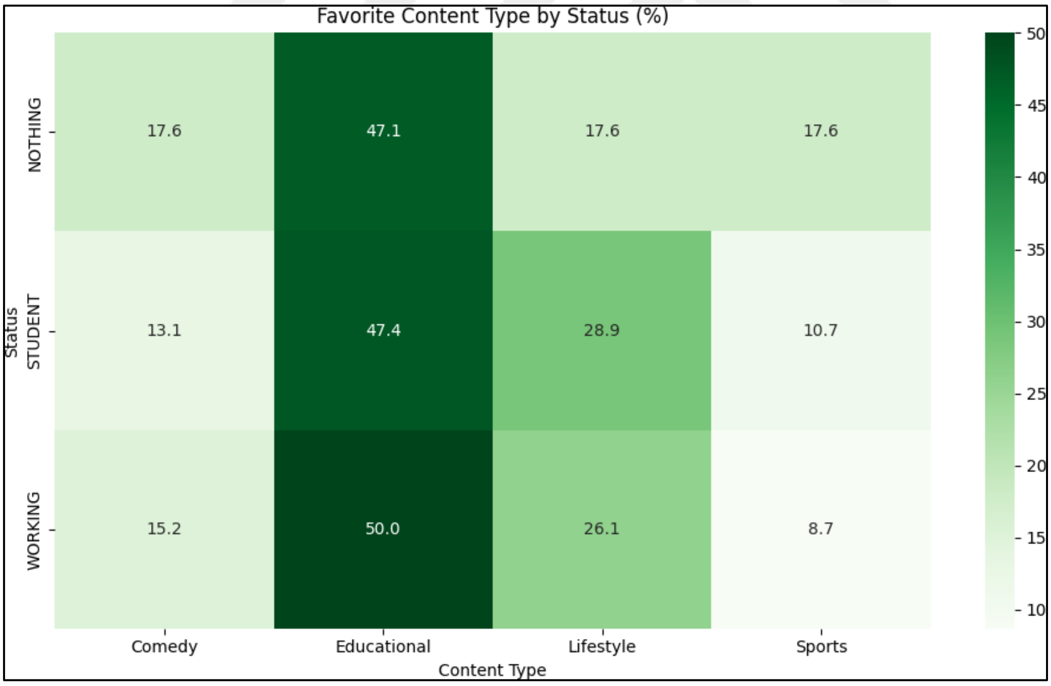

Figure 5.3: Favorite Content Type by Status (%).

5.2.1.4 Preferred video length by age group

Figure 5.4 displays the distribution of preferred video lengths among different age categories. Across all groups, medium-length videos (30–90 seconds) emerge as the most preferred format. The 25–34 age group shows the strongest inclination, with 75.0% favoring this length. Users aged 18–24 and those under 18 also demonstrate a majority preference for this format, at 58.0% and 57.9% respectively.

Short-form content (less than 30 seconds) is notably more popular among younger users. Under-18 users report a 26.3% preference for short videos, while the 18–24 segment follows closely with 27.5%. Meanwhile, longer videos (over 90 seconds) have limited appeal, especially among users aged 25–34 and 35+, where interest drops to 12.5% and 0.0%, respectively.

Interestingly, users aged 35 and above exhibit an even split between short and medium-length content (each at 50.0%), and show no preference for long videos. This suggests that while medium-length content remains universally favorable, tailoring short-form content to younger audiences could further boost engagement.

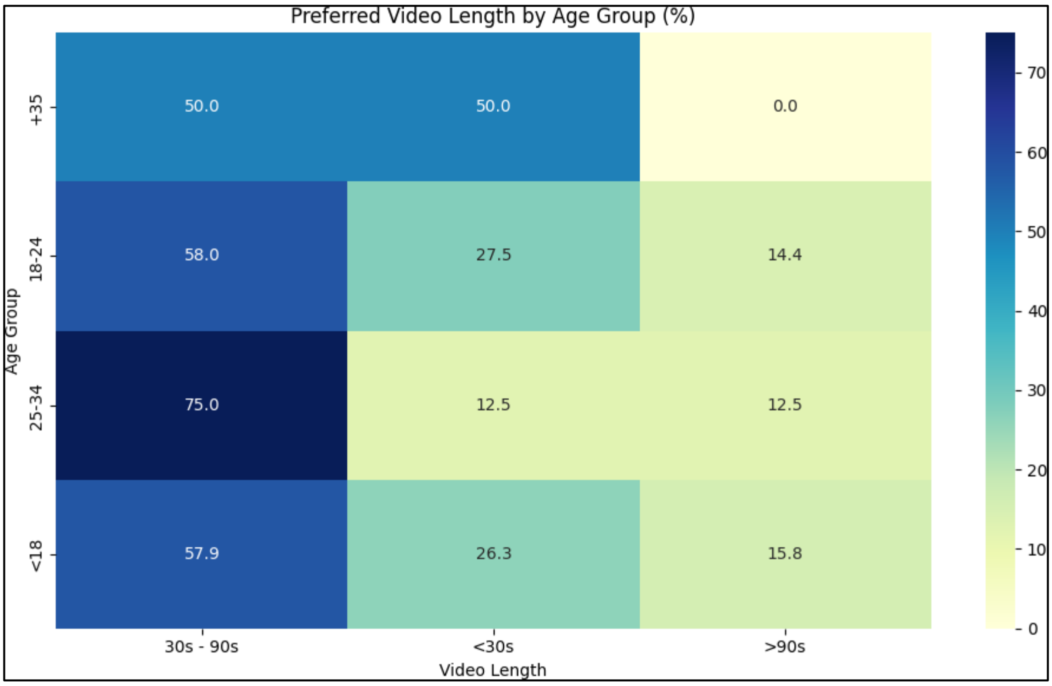

Figure 5.4: Preferred Video Length by Age Group (%).

5.2.1.5 Preferred video length by gender

Figure 5.5 presents the distribution of preferred video lengths across gender categories. The results show a strong overall preference for medium-length videos (30–90 seconds) among both male and female respondents. Female participants show a slightly stronger preference, with 60.8% favoring this duration, compared to 58.2% among males.

Short videos (under 30 seconds) are more favored by female users (30.7%) than their male counterparts (22.0%), suggesting that quick, easily consumable content may be more appealing to women. Conversely, male participants show a significantly higher interest in long-form videos (over 90 seconds), with 19.8% expressing this preference, compared to only 8.5% among females.

These insights highlight the importance of gender-based content formatting strategies, where shorter and medium-length videos are generally preferred, but longer content may resonate more effectively with male audiences.

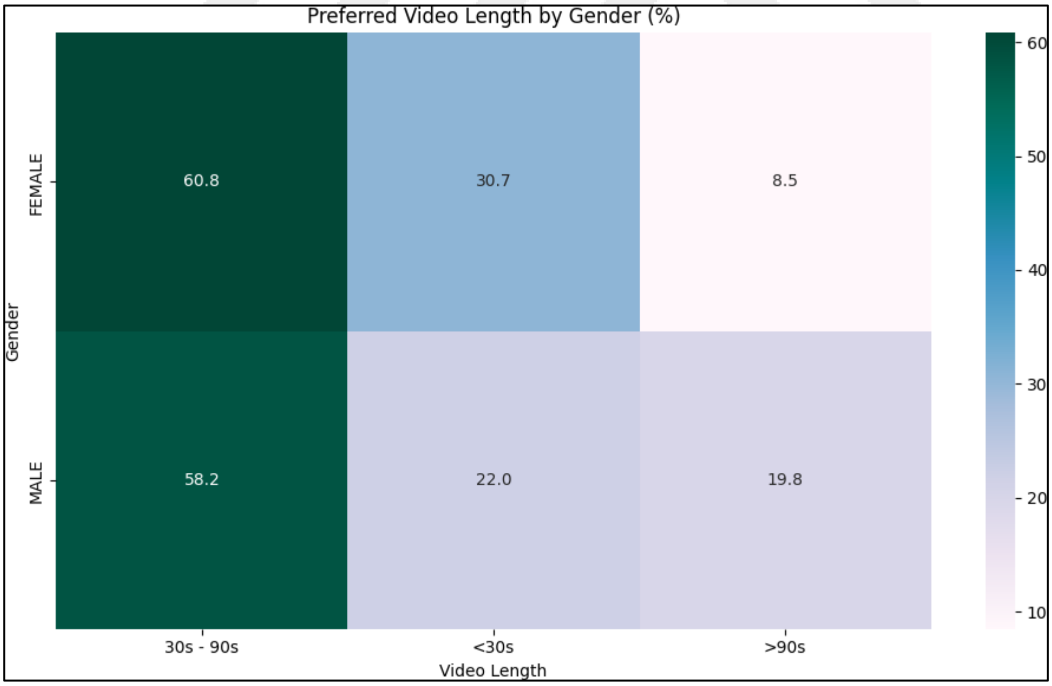

Figure 5.5: Preferred Video Length by Gender (%).

5.2.1.6 Preferred video length by status

Figure 5.6 shows the distribution of preferred video lengths among users with different status groups: working individuals, students, and those not currently studying or working.

Across all groups, there is a clear preference for medium-length videos (30–90 seconds), with percentages exceeding 59% for each category—61.1% for the unemployed, 60.9% for working users, and 59.4% for students. This highlights a universal leaning toward content that balances depth and brevity.

Students tend to show a stronger preference for short videos (27.3%) compared to working users (19.6%), possibly due to limited time or a higher exposure to fast-paced content. On the other hand, working individuals display the highest interest in longer videos (over 90 seconds), with 19.6% favoring extended content, compared to 13.3% among students and 16.7% among non-working participants.

These findings suggest that while medium-length videos are a safe and effective choice across all segments, content creators might consider varying length depending on user status to enhance engagement.

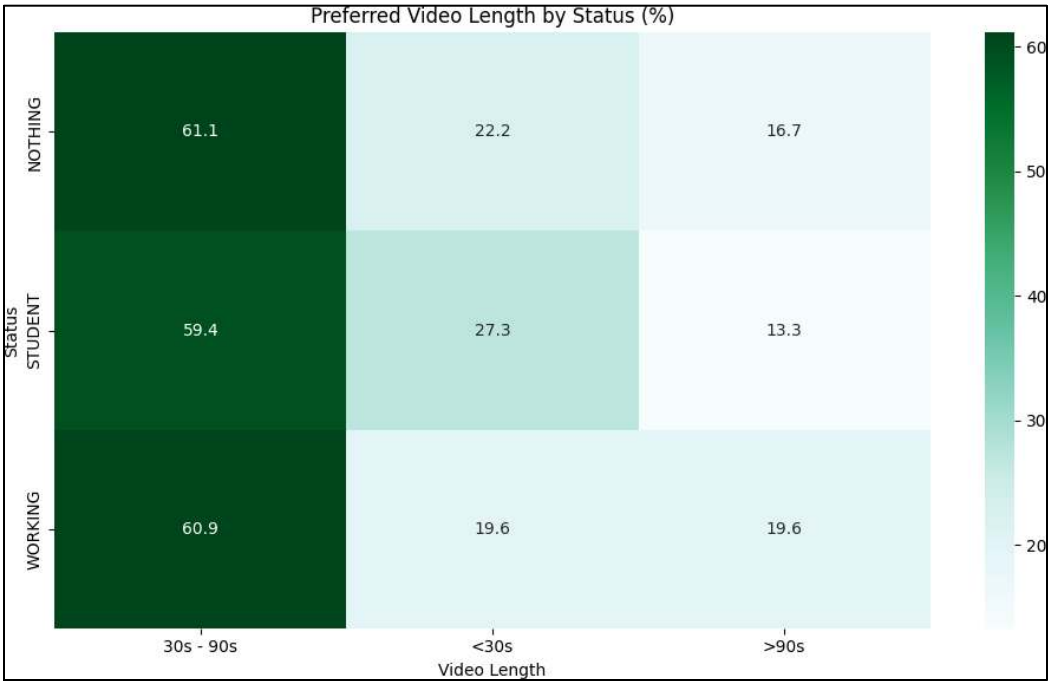

Figure 5.6: Preferred Video Length by Status (%).

5.2.1.7 Reaction type by age and gender

Figure 5.7 illustrates how users across different age groups and genders interact with content in terms of reaction types (Like, Share, Save, Comment). A consistent pattern emerges across most age groups, where likes dominate as the most common form of engagement.

For the under 18 group, both genders primarily respond with likes, with females showing a notably higher percentage (nearly 90%) compared to males (around 64%). Males in this group are slightly more likely to save content than share or comment.

Among users aged 18–24, the distribution becomes more balanced. While likes remain the most prevalent, there is a visible increase in content saves, especially among females (21% versus 16% for males). Commenting is minimal but slightly more common among males.
In the 25–34 category, likes drop slightly in prevalence, especially among males (60%), while shares and saves gain more weight. Females in this group display a stronger tendency to like and save content compared to males, while males are slightly more inclined to comment.

The 35+ group presents a more uniform engagement pattern, but only males were represented in this segment, with equal proportions of likes and shares, and no recorded saves or comments.

These results highlight that age and gender significantly influence the way users engage with content. Younger users tend to like and save content, while older users rely more on likes and shares. Gender-wise, females appear more likely to save content, whereas males distribute their interactions more evenly across reaction types.

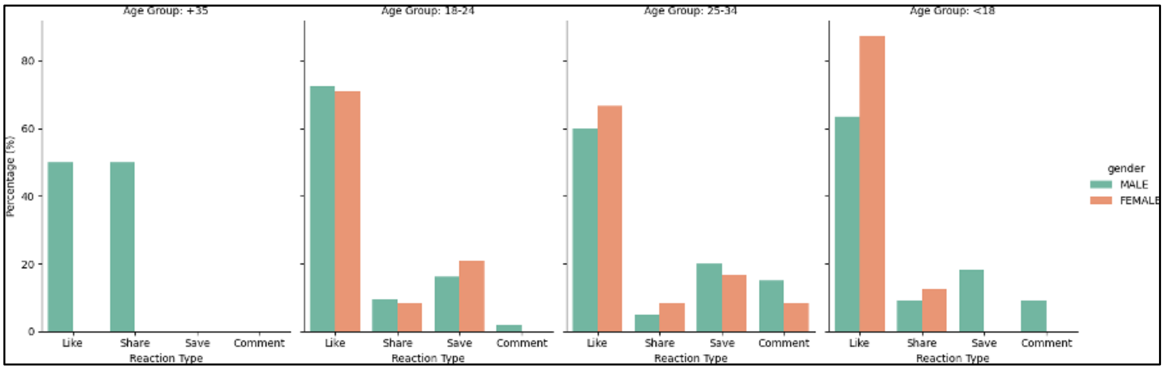

Figure 5.7: Distribution of Reaction Type by Age and Gender (%)^.

5.2.2 General Behavioral Insights

Beyond audience segmentation based on demographic variables, this subsection presents key behavioral patterns identified from specific survey questions. These patterns offer additional understanding of user motivations and preferences that influence engagement on social media. We focus here on responses related to following and unfollowing behavior, which provide actionable insights for content creators and digital marketers aiming to enhance retention and attract loyal followers.

5.2.2.1 Reasons for following accounts

Figure 5.8 presents the distribution of user responses regarding the factors that influence their decision to follow accounts. The results indicate that Content Quality is the leading reason, cited by 53% of respondents. This is closely followed by Shared Interests with 38%, suggesting that alignment in values or interests plays a significant role in follow behavior.

In contrast, Visuals and Consistency are less influential, with each accounting for only around 5% of responses. These findings suggest that content relevance and perceived value outweigh visual aesthetics or posting regularity when it comes to attracting followers.

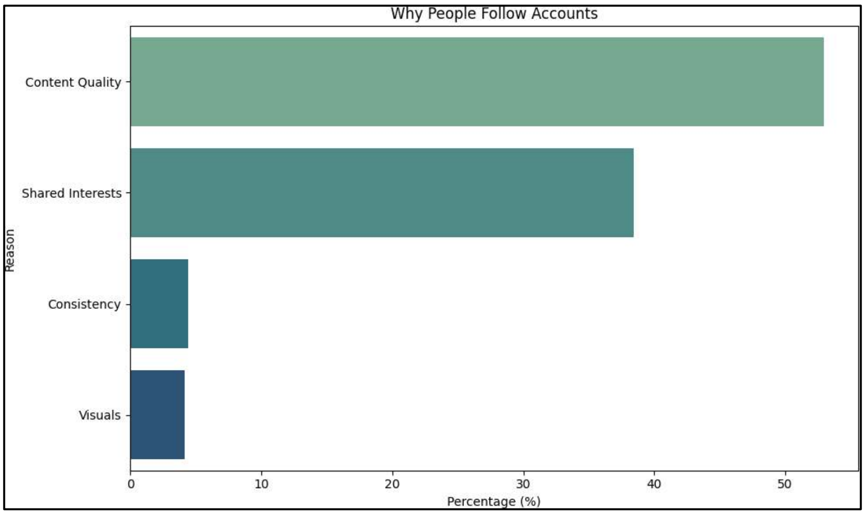

Figure 5.8: Why People Follow Accounts (%).

5.2.2.2 Reasons for unfollowing accounts

As shown in Figure 5.9, the most cited reason for unfollowing an account is Uninteresting Content, with a dominant 75% of responses. This clearly highlights the importance of maintaining engaging and relevant content to retain followers.

Ads come in second at 17%, indicating that overly promotional content can drive users away. Meanwhile, Inconsistent Posting is cited by only 8%, suggesting that while regularity matters, it's not as crucial as content value.

These insights underline the importance of sustained content quality and relevance in follower retention strategies.

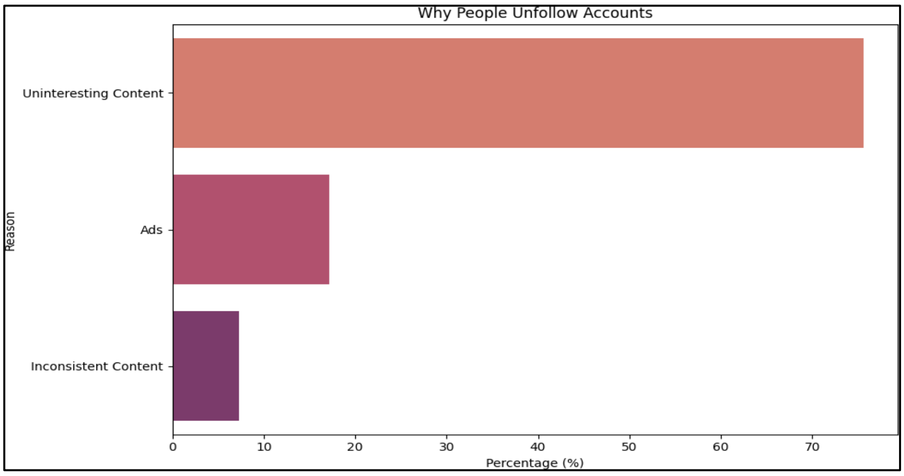

Figure 5.9: Why People Unfollow Accounts (%).

5.2.3 Multi-Variable Audience Pattern Mining

In this subsection, we explore key audience behavior patterns by analyzing combinations of multiple variables such as age, content type, video length, reaction type, and preferred usage time. These combined insights offer more personalized content strategy opportunities and reveal both strengths and limitations in user preferences and interactions.

5.2.3.1 Age + content type + preferred video length

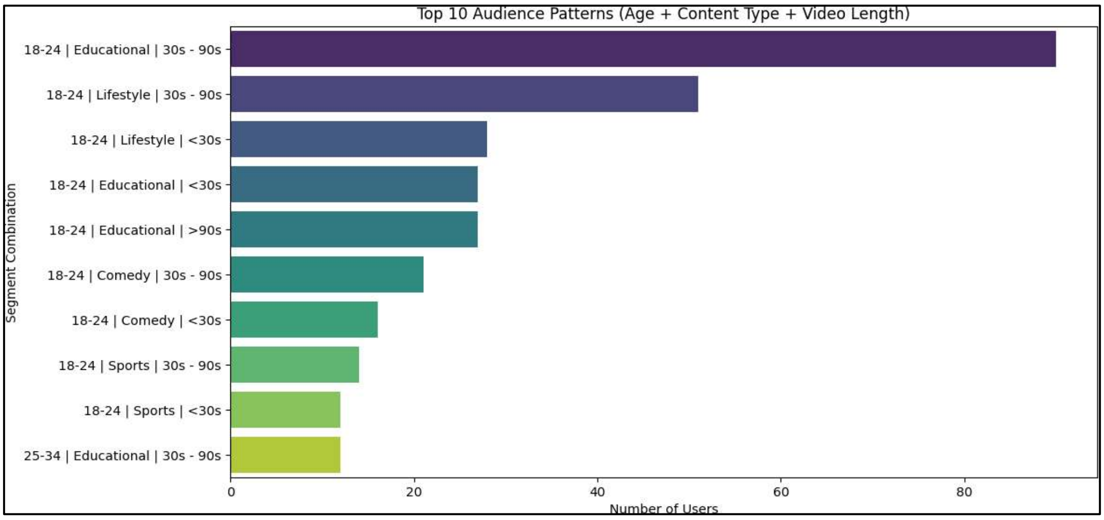

Figure 5.10: Top 10 Patterns: Age + Content Type + Video Length.

As shown in Figure 5.10, younger users aged 18–24 exhibit strong preferences for educational

and lifestyle content with mid-length videos (30s–90s) dominating the top combinations. This suggests that well-paced and informative content performs best for this age segment. Interestingly, although shorter and longer videos are also present, they appear far less frequently, implying reduced engagement outside the mid-length range.

A notable limitation is the absence of older age groups in the top patterns, particularly those above 35. This could be due to a smaller representation in the survey sample or a less defined preference among that demographic.

5.2.3.2 Age + content type + preferred time to use social media

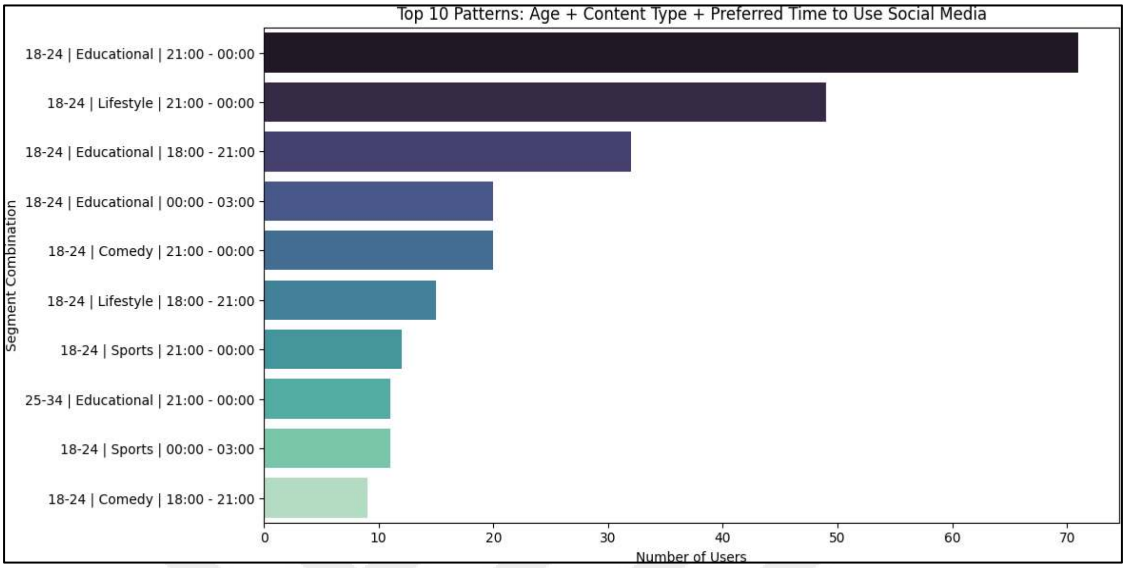

Figure 5.11: Top 10 Patterns: Age + Content Type + Preferred Time to Use Social Media.

Figure 5.11 reveals that the 18–24 age group, once again, dominates the engagement pattern, particularly with educational content consumed between 21:00 and 00:00. Late-night consumption appears to be the trend, especially for educational and lifestyle content. These findings align with common behavior among students or young adults who engage in passive content consumption late in the evening.

However, the results show a lack of clear engagement patterns in the morning or early afternoon slots, indicating an opportunity area for testing content at unconventional times. Additionally, no strong presence of the 35+ group or alternative content types like sports suggests that such patterns are either rare or underrepresented in the sample.

5.2.3.3 Age + preferred time to use social media + reactions

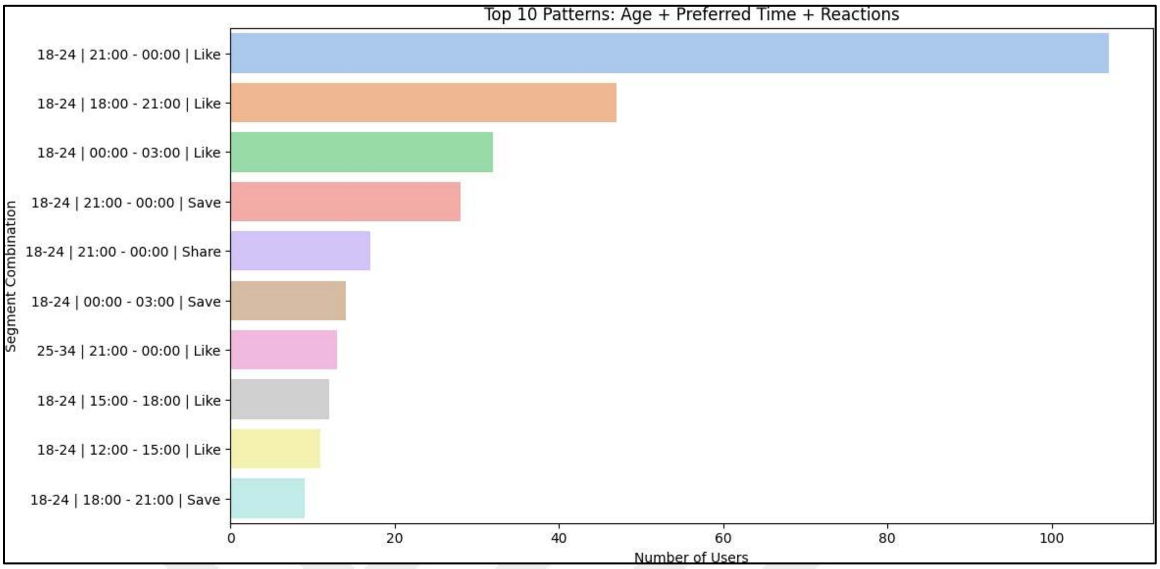

Figure 5.12: Top 10 Patterns: Age + Preferred Time + Reaction Type.

In Figure 5.12, the "Like" reaction dominates across all age and time segments, especially among users aged 18–24 during the late-night hours (21:00–00:00). The reaction distribution highlights a clear engagement gap, as "Save", "Share", and "Comment" are notably underrepresented, even though they are vital engagement signals for long-term content value.

This suggests that while users enjoy content enough to like it, they may not find it useful or engaging enough to save or share. This could indicate a need for deeper or more actionable content formats.

5.2.3.4 Content type + reaction + preferred time

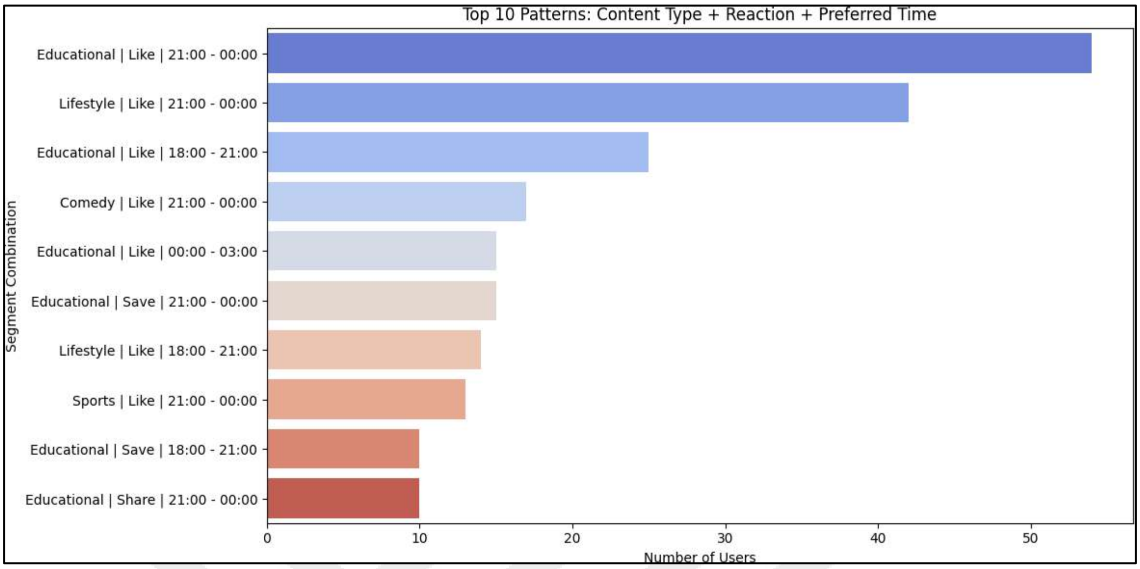

Figure 5.13: Top 10 Patterns: Content Type + Reaction Type + Preferred Time.

The final figure 5.13 confirms that educational content receives the highest number of "likes", particularly when posted during the 21:00–00:00 window. Lifestyle content follows closely. While this reinforces the popularity of these formats in late-night hours, it also uncovers another insight: sports and comedy content are much less present, especially in combination with reactions like "share" or "save".

Additionally, comment-based engagement is entirely absent from the top combinations, pointing to a weakness in generating conversational content. For content creators, this indicates an opportunity to experiment with calls-to-action or more interactive formats.

Across all four visualizations, a consistent pattern is observed: younger users (18–24) dominate most content interactions, with educational and lifestyle content performing strongest in evening time slots. However, engagement remains shallow—focused primarily on "likes"

with minimal depth in comments, shares, or saves. These insights suggest room for experimentation in format, timing, and calls to action to drive richer interactions.

5.2.4 BERT Model Performance

To assess the performance of the fine-tuned BERT model used for content classification, we evaluated it on a reserved portion of the dataset using standard classification metrics. The evaluation aims to measure how effectively the model predicts the correct content typeComedy, Sports, Educational, or Lifestyle—based on Instagram post captions.

5.2.4.1 Classification report and evaluation

Table 5.1 summarizes the classification performance of the model in terms of precision, recall, and F1-score. The overall accuracy reached 97.39%, indicating strong predictive power. The model shows consistently high performance across all four classes: Lifestyle, Comedy, Educational, and Sports.

Table 5.1: Updated Classification Report for BERT Content Classification.

Label	Precision	Recall	F1-score	Support
Lifestyle	0.95	0.95	0.95	22
Comedy	0.97	0.97	0.97	36
Educational	0.98	0.98	0.98	49
Sports	1.00	1.00	1.00	8
Accuracy			**0.97**	115
Macro avg	0.98	0.98	0.98	115
Weighted avg	0.97	0.97	0.97	115

As shown in Table 5.1, the model exhibits strong and balanced performance across all content categories. All four labels—Lifestyle, Comedy, Educational, and Sports—are well represented in the validation set, allowing for a fair evaluation of the model's effectiveness. Precision, recall, and F1-scores range from 95% to 100%, indicating a high degree of reliability. This consistent performance across categories suggests that the model is capable of generalizing well to diverse types of content, making it suitable for real-world applications in content classification.

5.2.4.2 Confusion matrix analysis

Figure 5.14 presents the confusion matrix for the BERT-based content classification model, showing the number of correctly and incorrectly predicted samples across the four content categories: Lifestyle, Comedy, Educational, and Sports.

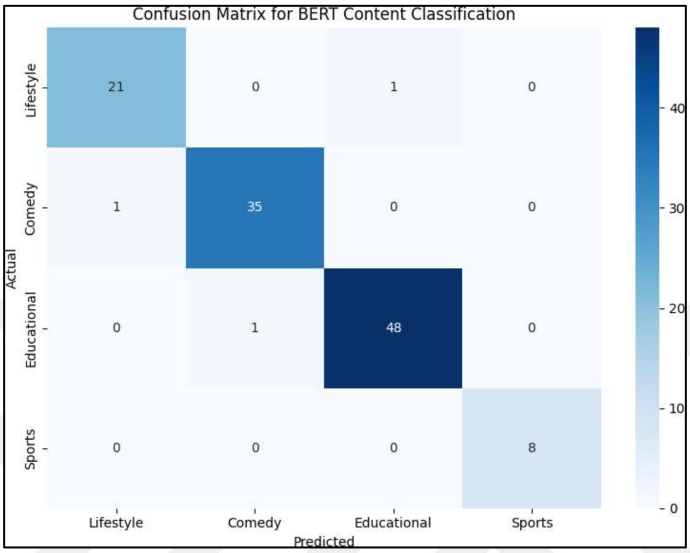

Figure 5.14: Confusion Matrix for BERT Content Classification.

The Confusion Matrix Reveals a Strong Performance in all Categories:

i. Lifestyle: 21 out of 22 samples were classified correctly, with only one misclassified as Educational.

ii. Comedy: 35 out of 36 were correctly identified, with one misclassified as Lifestyle.

iii. Educational: 48 out of 49 predictions were accurate, with one instance misclassified as Comedy.

iv. Sports: Achieved perfect accuracy with all 8 instances classified correctly.

This performance supports the high scores reported in the classification report, especially the perfect precision and recall for Sports. The few misclassifications occurred between semantically adjacent categories such as Educational and Lifestyle, which may reflect overlapping language or themes in post descriptions.

Overall, the model demonstrates strong discriminative capability, confirming its effectiveness in categorizing Instagram captions into content types based on text alone.

5.2.5 Graph API Data Analysis

To gain a deeper understanding of how content performance varies across different time slots and categories, we analyze the average engagement metrics (likes, shares, and views) for each predicted category and publishing time. This comparison is made possible by combining post insights from the Graph API with the predicted categories obtained from our BERT model. By doing so, we can identify patterns regarding when content of a specific type tends to perform better and which metric it excels at (e.g., views vs. likes).

5.2.5.1 Average likes by category and time

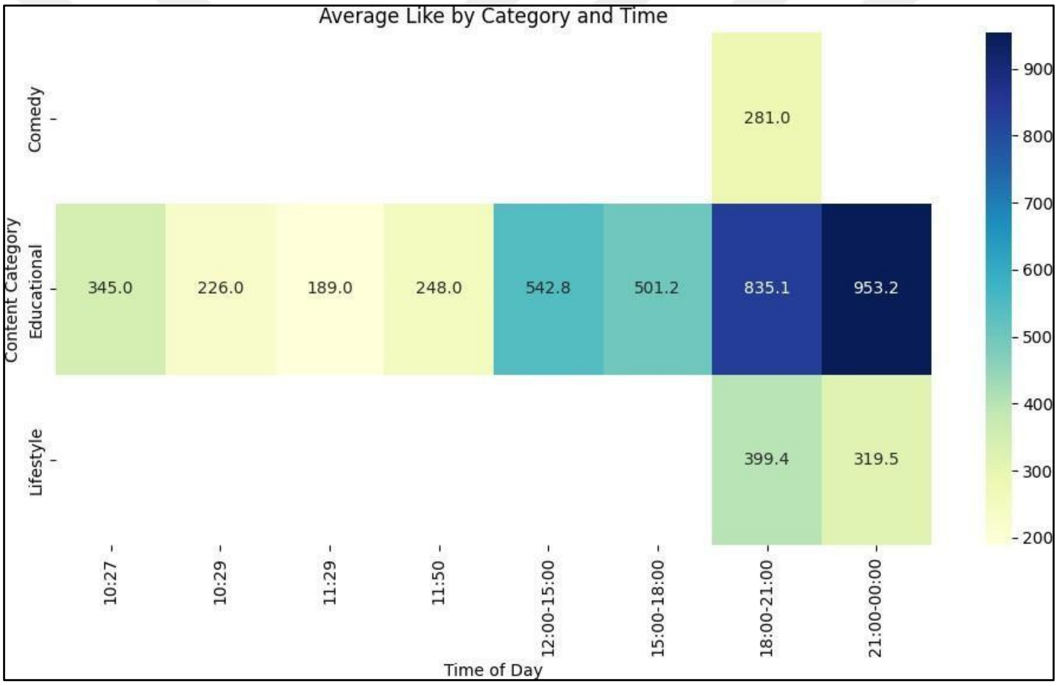

Figure 5.15: Average Likes by Content Category and Time.

Figure 5.15 illustrates the average number of likes received per post across time slots and categories. Educational content clearly dominates in terms of engagement, especially between 18:00–21:00 and 21:00–00:00, with averages exceeding 800 and 950 likes respectively. Lifestyle content shows a decent performance during 18:00–21:00 (around 399 likes), but other categories like Comedy are underrepresented in this dataset.

This suggests that audiences are most responsive to educational content in the late evening hours, aligning with common post-work hours when users are more likely to consume value-driven content.

5.2.5.2 Average shares by category and time

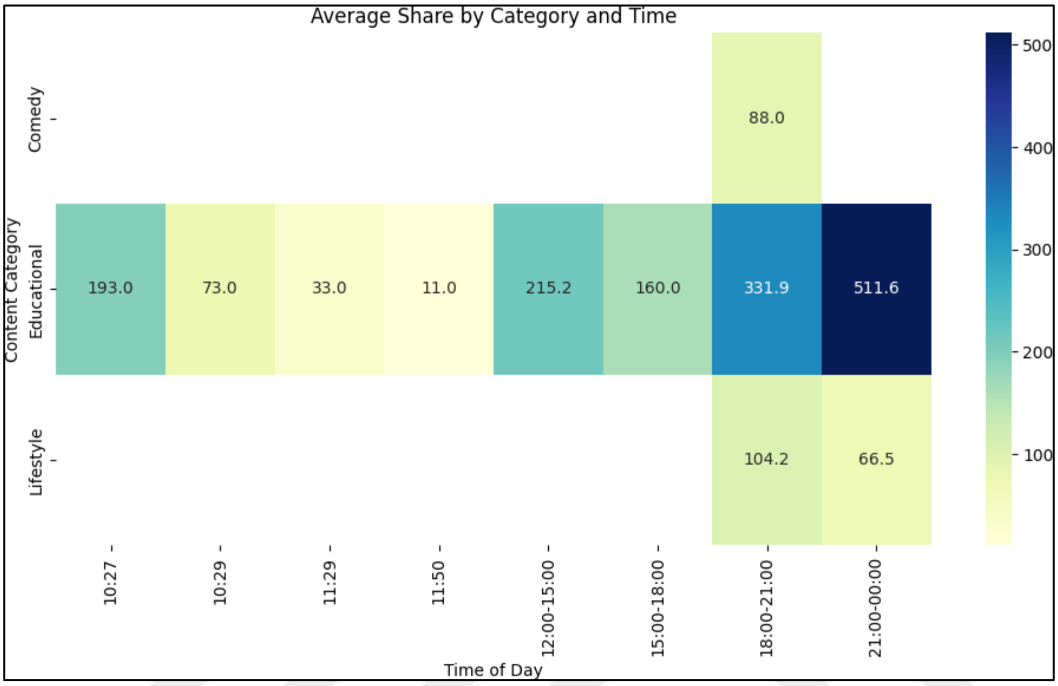

Figure 5.16: Average Shares by Content Category and Time.

As shown in Figure 5.16, Educational content again leads in shares, peaking during the 21:00–00:00 slot with an average of over 500 shares. The 18:00–21:00 slot also shows strong performance. Lifestyle content sees modest shares in comparison, and Comedy content appears only once with lower values.

These results reinforce the strength of Educational content not just in likes, but also in its ability to drive sharing behavior, especially at night when audiences may be more inclined to distribute informative material.

5.2.5.3 Average views by category and time

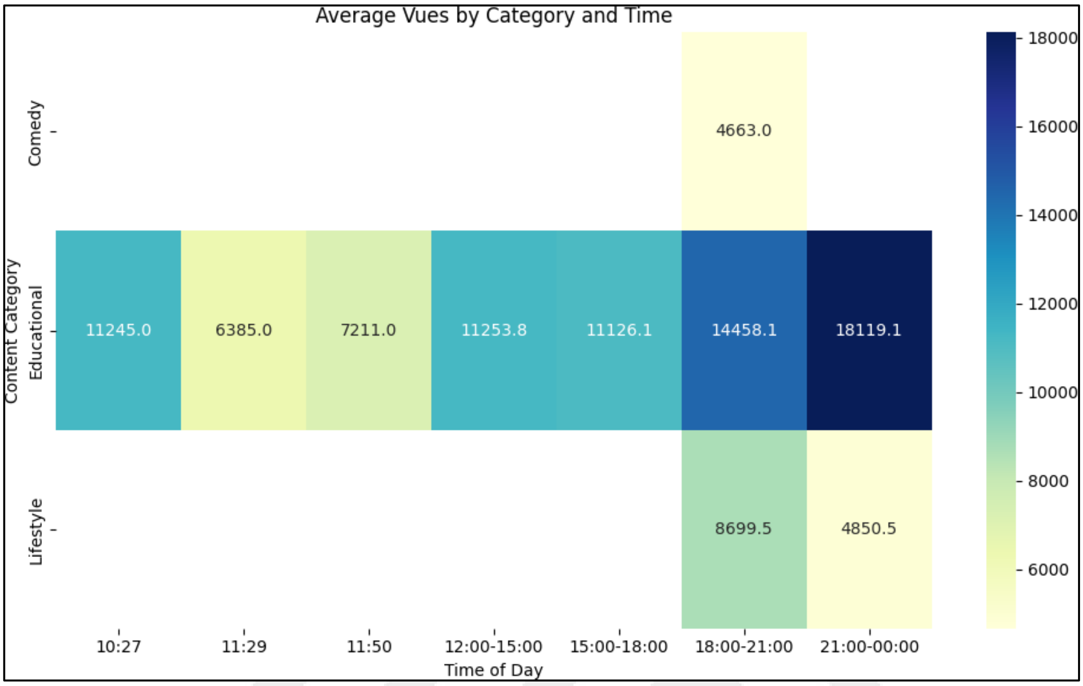

Figure 5.17: Average Views by Content Category and Time.

In Figure 5.17, we observe the distribution of average views. Educational content consistently receives high view counts across nearly all time slots, with peaks reaching over 18,000 views during the 21:00–00:00 window. Lifestyle content performs well during the 18:00–21:00 window with nearly 8,700 views, while Comedy content appears limited to a single time block with much fewer views.

This emphasizes that Educational content is not only liked and shared more, but also receives significantly more exposure, making it the most effective category overall in this dataset.

5.2.6 Comparison between Graph API Data and Survey Data

In this section, we examine both the survey data and the Instagram performance data in parallel. By combining insights from these two sources, as discussed in previous chapters, we aim to identify patterns, evaluate consistencies and discrepancies, and draw meaningful comparisons. This approach enables us to better understand the relationship between users' stated preferences and actual post performance, providing a basis for more informed future decisions.

5.2.6.1 Comparison between preferred usage time and actual engagement

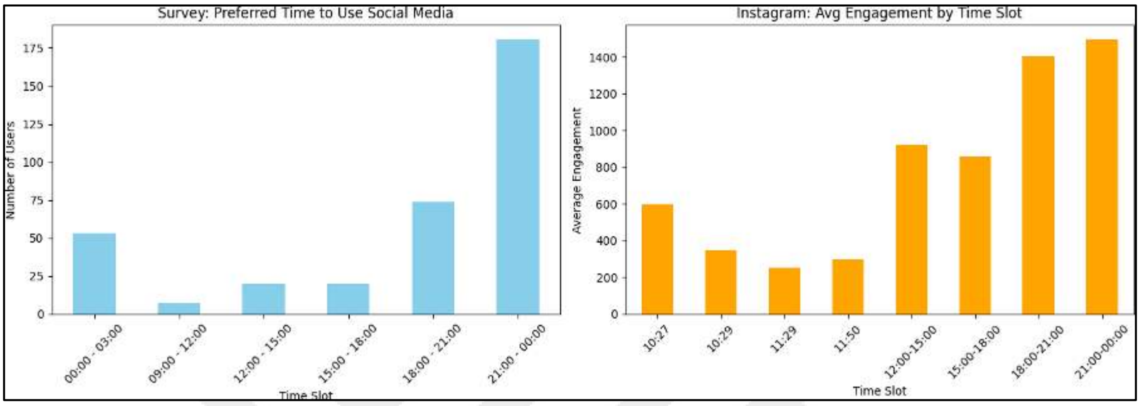

Figure 5.18: Comparison Between Survey Preferred Time and Instagram Engagement by Time Slot.

Figure 5.18 presents a side-by-side comparison between the time slots preferred by users for accessing social media (left) and the average engagement levels of Instagram posts based on those time slots (right). The analysis reveals both alignment and divergence between user preferences and actual content performance.

From the survey data, it is evident that the most preferred time slot among respondents is 21:00–00:00, followed by 18:00–21:00. These late-evening hours are when users are most likely to be active and browsing content. Interestingly, the engagement metrics on Instagram also peak during the same time periods, with 21:00–00:00 achieving the highest average engagement, followed closely by 18:00–21:00.

This overlap suggests a strong correlation between user availability and content performance, supporting the idea that posting during periods of high user activity leads to greater interaction. It is worth noting that earlier time slots, such as 09:00–12:00 and 12:00–15:00, are less preferred by users and correspond to lower average engagement rates.

The consistency between the two datasets confirms the reliability of the survey results and reinforces their potential use in content strategy. Content creators can use these insights to schedule their posts more effectively, ensuring they reach their audience during peak activity times to maximize engagement.

5.2.6.2 Comparison of preferred content type vs. performance

This section analyzes whether users' stated preferences for content types in the survey align with the types of content that actually perform best on Instagram. Figure 5.19 presents two bar charts: the left chart shows the number of users who prefer each content type based on the survey, while the right chart shows the average performance metrics (likes, shares, and views) of posts categorized by predicted content type.

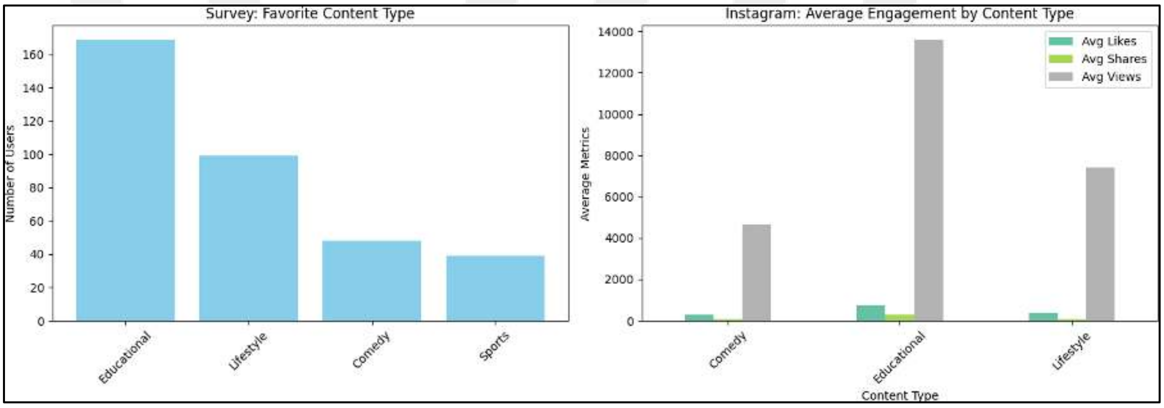

Figure 5.19: Comparison of Preferred Content Type (Survey) vs. Actual Performance (Instagram).

From the left chart, we observe that Educational content is the most preferred among users, followed by Lifestyle, Comedy, and Sports. This indicates a strong interest in informative and value-driven content.

The right chart confirms that Educational content also performs best in terms of engagement. It receives the highest number of average likes, shares, and views compared to other categories. This alignment between user preference and actual performance highlights a valuable insight for content creators: investing in high-quality educational content is likely to meet audience expectations and generate strong engagement.

On the other hand, while Lifestyle content is the second most preferred category, its engagement metrics are lower than those of Educational content, especially in terms of views. Comedy performs relatively well in terms of views but has lower likes and shares. These results demonstrate a partial correlation between survey-based preferences and content performance, suggesting that user-stated preferences can be a good indicator of what content will engage them most, though other factors like presentation, timing, and trends may also play a role.

5.2.6.3 Comparison between self-reported and actual reactions

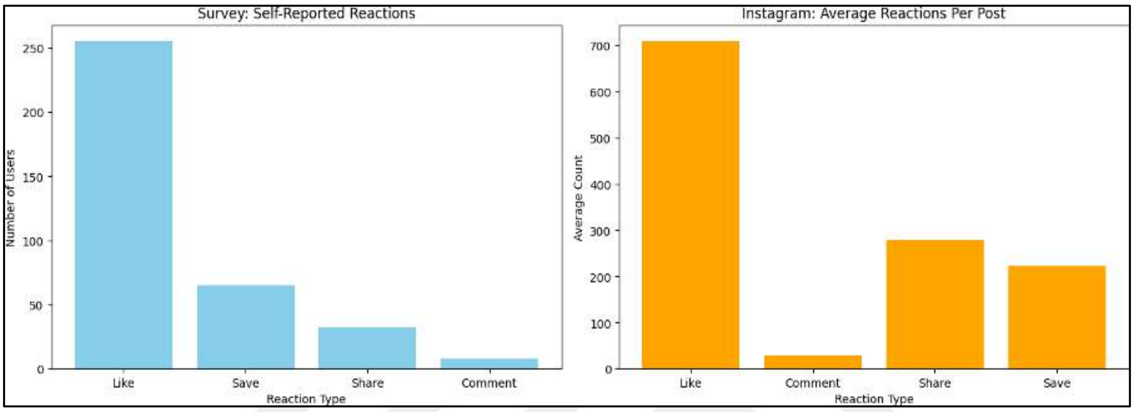

Figure 5.20: Comparison of Survey Reactions vs. Actual Instagram Reactions.

Figure 5.20 displays a comparison between the reactions that users claim they tend to perform on social media (left) and the actual average reactions received per post from Instagram data (right).

The left chart shows that the overwhelming majority of users report that they typically engage with content by liking it, followed by saving, sharing, and finally commenting, which has the lowest self-reported interaction rate. This aligns closely with the right chart, which confirms that likes are indeed the most frequent reaction type recorded per post, followed by shares and saves. Interestingly, comments are again the least frequent, consistent with user self-reporting.

This consistency between self-reported and actual behavior reinforces the reliability of the survey data in this dimension. However, the discrepancy in scale between user intention and actual metrics—for instance, the number of users who claim to save posts versus the actual average saves per post—may indicate a difference in perception versus real interaction behavior. This could also be affected by the algorithmic distribution of content and content type (e.g., educational vs. entertainment) that may trigger more passive interactions like saves and shares.

Overall, the comparison supports the idea that likes remain the dominant form of engagement, both perceived and actual, while other forms like comments and shares are less common but still significant in measuring deeper engagement.

5.3 CONCLUSION OF THE DISSCUSION

The results and discussion chapter brought together insights from both the survey data and the Instagram post analytics retrieved via the Meta Graph API. By analyzing these two datasets independently and comparatively, we were able to identify patterns, alignments, and divergences between self-reported user behavior and actual engagement metrics.

First, the multi-variable audience segmentation revealed meaningful combinations of age groups, content preferences, and interaction types, which can inform future content strategies. Educational content consistently emerged as the most favored and highest-performing type across various metrics—including likes, views, and shares—validating the preferences expressed in the survey data.

Secondly, the BERT model used to classify post content performed with high accuracy (97%), especially in identifying Educational, Comedy, and Sports categories. The classification and confusion matrix showed strong generalization, though performance was limited by the imbalance of classes, such as the underrepresentation of Lifestyle content.

The integration of both datasets allowed us to draw powerful comparisons:

i. Time-based Analysis: The time slots that users prefer to use social media (e.g., 21:00–00:00 and 18:00–21:00) were also the periods in which posts performed best in terms of engagement, suggesting a strong correlation between audience presence and content visibility.

ii. Content Type Comparison: Educational content not only dominated user preferences but also recorded the highest average performance metrics on Instagram. This confirms that aligning content creation with user preferences can enhance engagement.

iii. Reactions Analysis: There was a clear consistency between the types of reactions users claim to perform (mostly likes and saves) and the actual data from posts. Likes remained the most dominant engagement metric across both sources.

These findings confirm that audience feedback collected through surveys can be a reliable predictor of content performance, especially when supported by machine learning classification and social media analytics. The merged analysis also provides actionable insights for content creators, marketers, and researchers aiming to optimize content delivery strategies.

Future work could focus on expanding the dataset, incorporating additional platforms, and exploring real-time predictive models for content performance based on current audience sentiment and behavior.

6. CONCLUSION

6.1 SUMMARY OF FINDINGS

This study set out to explore how artificial intelligence (AI) can support content creators in making smarter decisions on social media. By analyzing engagement patterns, segmenting audiences, and predicting how well posts perform, the research brought together two powerful sources of information: what people say they prefer (through a detailed survey), and how they actually behave online (through real insights from Meta's Graph API).

The findings revealed a strong connection between users' self-reported preferences and the content that actually performs best. Educational content stood out across the board—it was the most liked by survey participants and also the most engaging on Instagram. Time of posting also mattered, with the 21:00–00:00 window not only being the most preferred by users but also delivering some of the highest engagement metrics. This highlights how critical timing is in getting content seen and interacted with.

A key part of the project was building and fine-tuning a multilingual BERT model to automatically categorize social media posts. The model achieved 97% accuracy, showing it can be trusted to classify content in real-time. Alongside this, a Random Forest model was used to analyze structured data, resulting in a complete AI pipeline for content strategy optimization.

What made this research especially valuable was the combination of survey responses and Instagram data. Together, they offered two perspectives: what users believe they want, and how they truly behave. This helped uncover not just which content performs well, but *why* providing clearer guidance for content creators to connect with their audience in more meaningful ways.

6.2 LIMITATIONS AND FUTURE WORK

While the results were promising, the study has a few limitations. The survey sample, for instance, may reflect some bias due to its limited demographic reach. On the technical side, some engagement data from the Graph API may have been influenced by platform-specific algorithms that weren't directly observable.

Future work could include more diverse data sources, like TikTok or YouTube, and add variables such as user sentiment or video watch time for richer analysis. Additionally, while the current AI models are effective, there's room for improvement. Future systems could benefit from real-time learning and better adaptation to changing trends. It would also be useful to explore visual factors—like thumbnails or facial expressions—which often influence engagement just as much as the caption or timing.

6.3 FINAL THOUGHTS

In today's digital world, where attention is everything, content creators need to understand both their audience and how platforms work behind the scenes. This thesis provides a framework that brings both of those pieces together, using AI to bridge the gap between audience insights and performance strategy. Whether you're an individual creator or a marketing team, the findings here offer a clearer path toward higher engagement, better planning, and more effective content.

REFERENCES

[1] M. Ali and L. Chen. Clustering algorithms for audience segmentation in social media advertising. *Springer Proceedings in Mathematics Statistics*, 2022.

[2] Violet Barman. Harnessing artificial intelligence for enhancing digital content production. *International Journal of Creative Research Thoughts (IJCRT)*, 11(4):147–160, 2023.

[3] R. Chen and L. Huang. Leveraging ai for social media trend prediction. *Journal of Machine Learning Applications in Social Media*, 10(4):112–128, 2022.

[4] John W. Creswell and Vicki L. Plano Clark. *Designing and Conducting Mixed Methods Research*. SAGE Publications, 3rd edition, 2017.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2019.

[6] Paul M. Di Gangi and Molly McCord. Social media engagement theory: Exploring the influence of user engagement on social media usage. *Journal of Organizational and End User Computing*, 28(1):53–73, 2016.

[7] R. Burke Johnson, Anthony J. Onwuegbuzie, and Lisa A. Turner. Toward a definition of mixed methods research. *Journal of Mixed Methods Research*, 1(2):112–133, 2007.

[8] Jackson P.-Cao L. et al. Kasneci, G. Media and artificial intelligence: Current perceptions and future outlook. *Journal of AI and Media Integration*, 2023.

[9] J. Kim. Predicting social media popularity with machine learning. *arXiv Preprint*, 2021.

[10] R. Kumar and P. Sharma. Natural language processing for audience profiling and segmentation in social media. *Journal of Big Data*, 8(1):123, 2021.

[11] Alexander Lang. Audience segmentation in social media. *Proceedings of the European Chapter of the Association for Computational Linguistics (EACL)*, 2017.

[12] Author not specified. Exploring the impact of artificial intelligence on content creation: A comprehensive study. *International Journal of Research Publications and Reviews*, 5(7):123–135, 2023.

[13] Author not specified. Forwarding in social media: Forecasting popularity of public opinion with deep learning. *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

[14] Author not specified. Impact of ai on social marketing and its usage in social media: A review. *International Journal of Social Media Studies*, 2023.

[15] Author not specified. The impact of predictive analytics and ai on digital marketing performanc. . *Journal of Digital Marketing Insights*, 2023.

[16] Author not specified. Mining social media data for sentiment analysis and trend prediction. *Journal of Big Data Research*, 2023.

[17] Author not specified. Predicting social media popularity with large language models. *IEEE Access*, 2023.

[18] Author not specified. Predicting popularity trend in social media networks with multi-layer temporal graph neural network. *Applied Intelligence*, 2024.

[19] Christian Peukert and Tim Wichmann. Generative artificial intelligence, content creation, and platforms. *Journal of Economics Management Strategy*, 2024.

[20] Abbas Tashakkori and Charles Teddlie. *SAGE Handbook of Mixed Methods in Social & Behavioral Research*. SAGE Publications, 2nd edition, 2010.

[21] H. Tran and D. Nguyen. Artificial intelligence in audience segmentation: Enhancing personalization in digital marketing. *IEEE Access*, 2021.

[22] Patti M. Valkenburg. Theoretical foundations of social media uses and effects. In *The Handbook of Media Effects*, pages 39–60. Cambridge University Press, 2021.

[23] Malte Wattenberg. Theoretical foundations of customer engagement in social media – implications for businesses. *Advances in Digital Marketing and E-Commerce*, 2024.

[24] Cheng W.-H. Zhang Y. Huang Q. Li J. Wu, B. and T. Mei. Sequential prediction of social media popularity with deep temporal context networks. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 18(7):567–589, 2022.

[25] Jiayang Wu, Wensheng Gan, Zefeng Chen, Shicheng Wan, and Hong Lin. Ai-generated content (aigc): A survey. *arXiv Preprint*, 2023.

[26] Li J. Yu, E. and C. Xu. Popalm: Popularity-aligned language models for social media trendy response prediction. *arXiv Preprint*, 2024.

[27] Y. Zhang and F. Li. Deep learning for social media audience segmentation: A case study. *Neural Computing and Applications*, 32(12):4589–4603, 2020.