

Multilingual Multimodal Content Mining for Hashtag Recommendation and Popularity Prediction in Social Networks

Ph.D. Thesis

By

Shubhi Bansal



DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY INDORE

April 2025

Multilingual Multimodal Content Mining for Hashtag Recommendation and Popularity Prediction in Social Networks

A THESIS

submitted to the

INDIAN INSTITUTE OF TECHNOLOGY INDORE

in partial fulfillment of the requirements for

the award of the degree

of

DOCTOR OF PHILOSOPHY

By

Shubhi Bansal



DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY INDORE

April 2025



INDIAN INSTITUTE OF TECHNOLOGY INDORE

CANDIDATE'S DECLARATION

I hereby certify that the work which is being presented in the thesis entitled **Multilingual Multimodal Content Mining for Hashtag Recommendation and Popularity Prediction in Social Networks** in the partial fulfillment of the requirements for the award of the degree of **Doctor of Philosophy** and submitted in the **Department of Computer Science and Engineering, Indian Institute of Technology Indore**, is an authentic record of my own work carried out during the time period from December 2020 to April 2025 under the supervision of Dr. Nagendra Kumar, Assistant Professor, Indian Institute of Technology Indore, India.

The matter presented in this thesis has not been submitted by me for the award of any other degree of this or any other institute.

07/07/2025

Signature of the Student with Date
(Shubhi Bansal)

This is to certify that the above statement made by the candidate is correct to the best of my knowledge.

07/07/2025

Signature of Thesis Supervisor with Date

(Dr. Nagendra Kumar)

Shubhi Bansal has successfully given her Ph.D. Oral Examination held on
04/07/2025

ACKNOWLEDGEMENTS

I would like to take this opportunity to express my heartfelt gratitude to a number of persons who in one or the other way contributed by making this time as learnable, enjoyable, and bearable. At first, I would like to thank my supervisor **Dr. Nagendra Kumar** who was a constant source of inspiration during my work. Without his constant guidance and research directions, this research work could not be completed. His continuous support and encouragement has motivated me to remain streamlined in my research work.

I am thankful to **Dr. Ranveer Singh** and **Dr. Gourab Sil**, my research committee members for taking out some valuable time to evaluate my progress all these years. Their good comments and suggestions helped me to improve my work at various stages. I am also grateful to **Dr. Ranveer Singh**, HOD of Computer Science and Engineering for his help and support.

My sincere acknowledgement and respect to **Prof. Suhas Joshi**, Director, Indian Institute of Technology Indore for providing me the opportunity to explore my research capabilities at Indian Institute of Technology Indore.

I extend my sincere thanks to the **Prime Minsiter Research Fellwoship, an intiatiave by the Government of India** for funding the PhD research. This work was supported by PMRF under Grant ID: 2101704.

I would like to appreciate the fine company of my dearest colleagues, labmates, friends, undergraduate students who have also supported me in my research work. I am also grateful to the institute staff for their unfailing support and assistance.

I would like to express my heartfelt respect to my parents for their love, care and support they have provided to me throughout my life.

Finally, I am thankful to all who directly or indirectly contributed, helped, and supported me.

Shubhi Bansal

To my family

Abstract

Social networking services have profoundly reshaped human interaction and information exchange transcending geographical, cultural, and temporal boundaries. These platforms empower users to both consume and produce content, expressing themselves in diverse languages and modalities, referred to as multilingualism and multimodality, respectively. This resulting surge in user-generated content, characterized by multilingualism and multimodality, has introduced a significant challenge of information overload, which hinders content discoverability and reachability. To mitigate these challenges and manage content efficiently, this thesis investigates hashtag recommendation and popularity prediction as effective solutions.

Hashtag recommendation is the process of assigning hashtags to uploaded content, thereby facilitating thematic organization of vast volumes of content. However, existing methods overlook crucial aspects such as multilingualism and multimodality. In this thesis, we propose hashtag recommendation methods tailored for various linguistic contexts and content modalities. These methods encompass monolingual content, multilingual content, multimodal content comprising texts and images, and micro-videos, acknowledging the diverse levels of user engagement they elicit.

We first address hashtag recommendation for monolingual content, leveraging language-specific features to understand text and suggest pertinent hashtags. Existing retrieval-based approaches struggle with rapid information flow in social networks, while generation-based methods lack sufficient contextual understanding. To meet this critical need for organized information in social networks, we propose a retrieval-augmented diffusion-based sequence-to-sequence framework to recommend hashtags. Furthermore, recognizing the subsequent growth of multilingual content, we then propose a framework to recommend personalized and language-specific hashtags for multilingual content. However, the development of comparable approaches is impeded for low-resource languages, due to limited data availability and significant linguistic heterogeneity. To address these limitations, we propose a novel framework to recommend hashtags for multilingual posts, explicitly considering linguistic diversity,

intra-relatedness among language groups, and users’ topical and linguistic preferences.

The rise of multimodal content on social networks, integrating visual and textual modalities, necessitates leveraging both modalities in hashtag recommendation to effectively capture user interests and their influence on content consumption. However, existing methods, relying on single modalities, fail to capture these multimodal relationships and user preferences. To address this, we propose a novel method that mines deep interactions between textual and visual modalities. We also incorporate users’ historical tagging behavior to yield personalized hashtags. Micro-videos, a dominant and highly engaging form of user-generated content, present a further challenge for hashtag recommendation due to their concise duration and high information density. Existing approaches neglect users’ modality-specific tagging preferences and the collective tagging behavior of similar users. Moreover, the cold-start user issue prevails in hashtag recommendation systems. In view of the above, we introduce a hybrid filtering method that leverages interrelationships among modalities and users to recommend hashtags for existing users. We also propose an innovative social influence and content-based solution to alleviate the cold-start user problem. Our findings demonstrate that the proposed method recommends relevant hashtags for micro-videos posted by existing and cold-start users, thereby boosting content discoverability.

The pervasive nature of social networks has empowered users to disseminate their perspectives and experiences across diverse topics through posts integrating multiple modalities, such as texts and images, leading to a significant surge in multimodal content. This growing prominence of multimodal content necessitates accurate estimation of its popularity. However, predicting the popularity of such posts remains a considerable challenge, as only a small fraction gains wider visibility, while the majority experience limited reachability. To address this critical gap, we introduce a multifaceted framework for multimodal content popularity prediction. We derive visual demographic features, sentiment from hashtags and captions, and model intricate relationships among images, texts, and hashtags to determine post popularity.

Keywords: Social Network Analysis, Multilingualism, Multimodality, Hashtag Recommendation, Popularity Prediction, Information Retrieval, Data Mining

List of Publications

A. Published

A1. In Refereed Journals

1. **S. Bansal**, K. Gowda, and N. Kumar. *A Hybrid Deep Neural Network for Multimodal Personalized Hashtag Recommendation*, IEEE Transactions on Computational Social Systems, pp. 2439-2459, 2022 (IEEE), DOI = “10.1109/TCSS.2022.3184307”. (SCIE)
2. **S. Bansal**, K. Gowda, and N. Kumar. *Multilingual Personalized Hashtag Recommendation for Low Resource Indic languages using Graph-based Deep Neural Network*, Expert Systems with Applications, vol. 236, pp. 121188, 2024 (Elsevier), DOI = “https://doi.org/10.1016/j.eswa.2023.121188”. (SCIE)
3. **S. Bansal**, K. Gowda, M.Z.U. Rehman, C.S. Raghaw, and N. Kumar. *A Hybrid Filtering for Micro-video Hashtag Recommendation using Graph-based Deep Neural Network*, Engineering Applications of Artificial Intelligence, vol. 138, pp.109417, 2024 (Elsevier), DOI = “https://doi.org/10.1016/j.engappai.2024.109417”. (SCIE)
4. **S. Bansal**, K. Gowda, C.S. Raghaw, and N. Kumar. *Sentiment and Hashtag-aware Attentive Deep Neural Network for Multimodal Post Popularity Prediction*, Neural Computing and Applications, vol. 37, pp. 2799-2824, 2025 (Springer), DOI=“https://doi.org/10.1007/s00521-024-10755-5”. (SCOPUS)

A2. In Refereed Conferences

1. **S. Bansal**, S. Parimala, and N. Kumar. *Retrieval Augmented Encoder-Decoder with Diffusion for Sequential Hashtag Recommendation in Disaster Events*, In Proceedings of the AAAI International Conference on Web and Social Media (ICWSM), March (2025) (accepted).

Contents

Abstract	i
List of Publications	iii
List of Figures	ix
List of Tables	xi
List of Abbreviations and Acronyms	xiii
1 Introduction	1
1.1 Hashtag Recommendation	3
1.1.1 Monolingual Content	5
1.1.2 Multilingual Content	6
1.1.3 Multimodal Content	7
1.1.4 Micro-videos	8
1.2 Popularity Prediction	9
2 Literature Review	13
2.1 Hashtag Recommendation	13
2.1.1 Content-based Hashtag Recommendation	13
2.1.2 Personalized Hashtag Recommendation	20
2.2 Popularity Prediction	21
3 Hashtag Recommendation for Monolingual Content	25
3.1 Introduction	25

3.2	Methodology	29
3.2.1	Diffusion	31
3.3	Experimental Evaluations	37
3.3.1	Experimental Setup	37
3.3.2	Experimental Results	44
3.4	Conclusion	51
4	Hashtag Recommendation for Multilingual Content	53
4.1	Introduction	53
4.2	Problem Definition	57
4.3	Methodology	59
4.3.1	Feature Extraction	60
4.3.2	Feature Refinement	63
4.3.3	Feature Interaction	65
4.3.4	Hashtag Recommendation	69
4.4	Experimental Evaluations	70
4.4.1	Experimental Setup	70
4.4.2	Experimental Results	76
4.5	Conclusion	82
5	Hashtag Recommendation for Multimodal Content	83
5.1	Introduction	83
5.2	Problem Definition	87
5.3	Methodology	91
5.3.1	Feature Mining	91
5.3.2	User Preference Mining	99
5.3.3	Hashtag Prediction	102
5.3.4	Candidate Hashtag Recommendation	106
5.4	Experimental Evaluations	109
5.4.1	Experimental Setup	109
5.4.2	Experimental Results	113

5.5	Conclusion	123
6	Hashtag Recommendation for Micro-videos	125
6.1	Introduction	125
6.2	Problem Definition	129
6.3	Methodology	130
6.3.1	Feature Mining	131
6.3.2	Feature Refinement	136
6.3.3	Hashtag Recommendation	144
6.4	Experimental Evaluations	145
6.4.1	Experimental Setup	145
6.4.2	Experimental Results	149
6.5	Conclusion	161
7	Popularity Prediction of Multimodal Content	163
7.1	Introduction	163
7.2	Problem Definition	167
7.3	Methodology	167
7.3.1	Feature Extraction	169
7.3.2	Feature Interaction	177
7.3.3	Feature Fusion	179
7.3.4	Popularity Prediction	181
7.4	Experimental Evaluations	183
7.4.1	Experimental Setup	183
7.4.2	Experimental Results	188
7.5	Conclusion	197
8	Conclusion and Future Work	199
8.1	Summary of the Thesis	199
8.1.1	Hashtag Recommendation for Monolingual Content	199
8.1.2	Hashtag Recommendation for Multilingual Content	200

8.1.3	Hashtag Recommendation for Multimodal Content	201
8.1.4	Hashtag Recommendation for Micro-videos	201
8.1.5	Popularity Prediction of Multimodal Content	202
8.2	Future Work	202

Bibliography	205
---------------------	------------

List of Figures

1.1	User-Generated Content	2
3.1	Effectiveness comparison curves. The proposed method significantly outperforms compared methods.	45
3.2	Example of a tweet from test dataset depicting hashtags recommended by various methods. Generated hashtags that match user-assigned hashtags are marked with green, while relevant but non-matching hashtags are marked with blue, and irrelevant predictions are marked with red.	50
4.1	Tweets of a user	55
4.2	Overall architecture of TAGALOG	59
4.3	Graph AutoEncoder	67
4.4	Example posts showing hashtags recommended by different methods . .	80
5.1	Example posts from Instagram	84
5.2	Overall architecture of DESIGN	91
5.3	Feature mining module	92
5.4	Candidate hashtag recommendation module	106
5.5	Effectiveness comparison curves on MMP-INS dataset	115
5.6	Example post depicting hashtags recommended by different methods .	120
6.1	Visual representation of problem definition	130
6.2	Overall architecture of MISHON	131
6.3	Graph construction in MISHON	137
6.4	Effectiveness comparison curves on TMALL dataset	152

6.5	Example post showing hashtags recommended by different methods . .	160
7.1	Example social media post	164
7.2	System architecture of NARRATOR	168
7.3	Posts depicting demographic features	172
7.4	Deep feedforward neural network for popularity score prediction	182
7.5	Effectiveness comparison curves on TPIC dataset	190
7.6	Performance comparison curves on SMP dataset	191
7.7	Posts depicting popularity scores predicted by different methods	194

List of Tables

3.1	Statistics of the dataset. h_t denotes number of hashtags per tweet. . . .	38
3.2	Dataset distribution by number of tweets and percentage of total tweets for each disaster type.	39
3.3	Effectiveness comparison results of ASSIGNER with existing methods for hashtag recommendation (top-2). The best result is highlighted in bold , while the second-best is <u>underlined</u>	44
3.4	Effect of individual component ablation on hashtag generation perfor- mance of ASSIGNER. The best result is highlighted in bold , while the second-best is <u>underlined</u> . Here, w/o refers to without.	46
3.5	Performance comparison of ASSIGNER with various noise scheduling algorithms for disaster-related hashtag recommendation, showing opti- mal results with the token-level adaptive sigmoid scheduler. The best result is highlighted in bold , while the second-best is <u>underlined</u>	49
4.1	IndicHash dataset statistics	72
4.2	Effectiveness comparison results with existing research works	76
4.3	Effectiveness comparison results with pre-trained models	77
4.4	Performance of TAGALOG with different attention techniques	78
4.5	Performance comparison of TAGALOG with different components	79
5.1	Effectiveness comparison results on MMP-INS dataset	114
5.2	Effectiveness comparison results on HARRISON dataset	116
5.3	Effectiveness comparison results on T-INS dataset	117
5.4	Performance of DESIGN with different modality combinations	118
5.5	Performance of DESIGN with different attention mechanisms	119

5.6	Performance Comparison of Modules	123
6.1	Statistics of different datasets	145
6.2	Effectiveness comparison results on TMALL dataset	150
6.3	Effectiveness comparison results on INSVIDEO dataset	152
6.4	Effectiveness comparison results on YFCC dataset	153
6.5	Ablation studies	154
6.6	Performance comparison on cold-start users	157
6.7	Sensitivity analysis of popular user selection ratio (α)	158
7.1	Effectiveness comparison results on different datasets	188
7.2	Feature ablation study	191
7.3	Effectiveness of attention mechanisms	193
7.4	Feature ranking and importance	196

List of Abbreviations and Acronyms

SNS Social Network Services

UGC User-Generated Content

TINS Text Dataset from Instagram

SVR Support Vector Regression

LSTM Long Short-Term Memory

SA-RO Self-Adaptive Rain Optimization

ANN Approximate Nearest Neighbor

LRL Low-resource Languages

RAG Retrieval Augmented Generation

GCN Graph Convolutional Network

RNN Recurrent Neural Network

NLP Natural Language Processing

seq2seq Sequence-to-sequence

TAM Tweet Attention Module

TAM Entity Attention Module

MTL Multi-Task Learning

ELMO Embeddings from Language Model

POS Parts of Speech

MBR Minimum Bayes Risk

BERT Bidirectional Encoder Representations from Transformers

BLEU Bilingual Evaluation Understudy

GeLU Gaussian Error Linear Units

BART Bidirectional Autoregressive Transformer

GRU Gated Recurrent Unit

Bi-LSTM Bidirectional Long Short Term Memory

SOTA State-of-the-art

RA Retrieval Augmentation

UGA User-guided Attention

LGA Language-guided Attention

mBERT Multilingual Bidirectional Encoder Representations from Transformers

MLP MultiLayer Perceptron

GraphSAGE Graph Sample and Aggregate

GAE Graph AutoEncoder

MLM Masked Language Modeling

FR Feature Refinement

FI Feature Interaction

MLC Multi-Label Classification

SG Sequence Generation

CNN Convolutional Neural Networks

VGG Visual Geometry Group

NSP Next Sentence Prediction

WA Word-level Attention

PCO Parallel Co-attention

FRM Feature Refinement Module

TFIDF Term Frequency Inverse Document Frequency

ReLU Rectified Linear Unit

PCA Principal Component Analysis

CFFN Cascade Feed-Forward Network

MSE Mean Squared Error

MAE Mean Absolute Error

DNN Deep Neural Network

SMP Social Media Prediction

Chapter 1

Introduction

Social Network Services (SNS) have fundamentally reshaped human interaction and global communication by establishing unprecedented connectivity that transcends geographical, cultural, and temporal boundaries. These platforms are dynamic, undergoing continuous evolution driven by the introduction of novel features, adaptations to evolving user behaviors, and ongoing technological advancements. The pervasive adoption of mobile devices and the increasing accessibility of Internet have fueled the exponential growth of SNS since their inception, culminating in over 5.22 billion users globally, representing 67.5% of the world's population as of October 2024¹. This dynamic environment supports a multitude of applications, from dissemination of information, formation of personal and professional networks, to the vast world of online entertainment and content sharing, which forms the engine of SNS.

The ease with which users can now both consume and generate content, facilitated by technological advancements, has empowered users to express themselves and engage with information through increasingly diverse combinations of texts, images, audios, and videos. This democratization of content creation has transformed SNS into dynamic ecosystems where users function as both producers and consumers of User-Generated Content (UGC), which constitutes the vast majority of content on these platforms and drives online discourse. Consequently, the UGC landscape exhibits a substantial prevalence of multilingual content, a direct result of SNS increasingly

¹<https://www.demandsage.com/internet-user-statistics/>

supporting vernacular languages, thereby enabling global participation in the digital sphere using native tongues. While enriching online communication, these intertwined

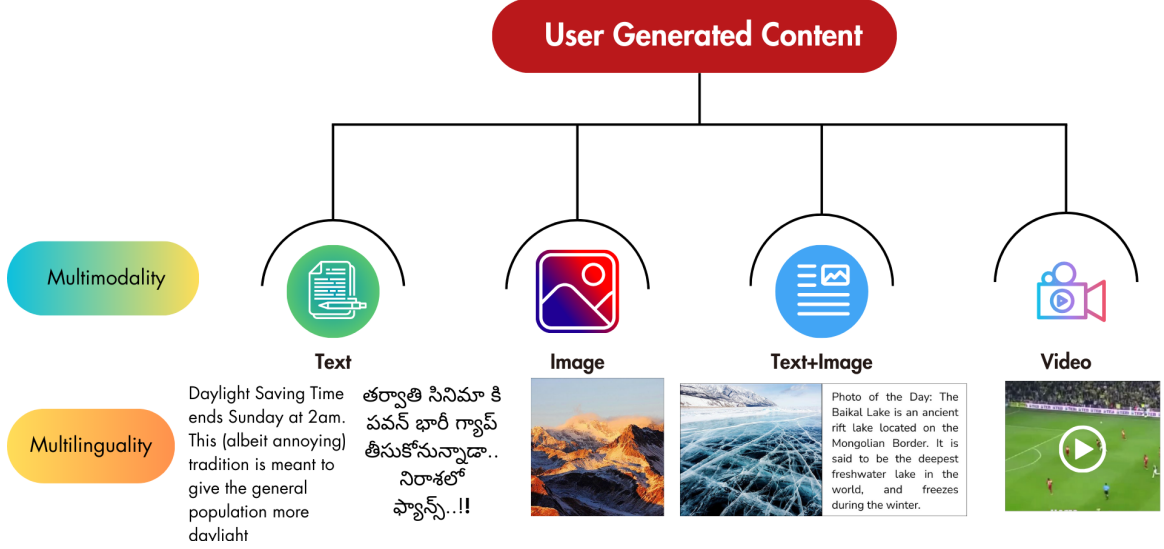


Figure 1.1: User-Generated Content

trends of multimodality and multilingualism nature of UGC, as shown in Figure 1.1 present considerable challenges for effective content management and analysis.

The sheer volume of diverse UGC, coupled with the constant stream of updates inherent in SNS, precipitates significant information overload [1]. Users are inundated with data beyond their processing capacity, resulting in cognitive overload, shortened attention spans, and difficulty discerning relevant content. This consequently diminishes user experience, reduces platform engagement, and hinders users from finding the content they seek. To effectively manage this overwhelming information landscape, it is crucial to address the underlying limitations within current SNS functionalities. The challenges hindering effective information management on SNS arise from two key limitations. First, current functionalities lack robust mechanisms for categorizing multilingual and multimodal UGC. This leads to users being overwhelmed with irrelevant content and struggling to discover pertinent communities, experiences, and information. Second, there is a lack of precise methodologies for predicting content reach across diverse modalities, consequently affecting the ability of content to reach its desired audience, leading to the disproportionate amplification of content. This dis-

torts the information ecosystem, prioritizing ephemeral content over content possessing societal significance, thereby undermining SNS as equitable spaces for knowledge exchange and meaningful interaction. To address these critical limitations, this thesis, entitled “Multilingual Multimodal Content Mining for Hashtag Recommendation and Popularity Prediction in Social Networks”, focuses on two interconnected solutions. First, it explores the development of automated hashtag recommendation systems. These systems leverage linguistic and multimodal features of UGC to enhance content categorization and searchability, thereby enabling content discovery and creator reach. Second, the thesis investigates the creation of automated popularity prediction models. These models utilize multimodal signals to forecast engagement trends, aiming to optimize content distribution and ensure equitable visibility for high-impact posts while maintaining the overall quality of user feeds.

1.1 Hashtag Recommendation

Due to the massive influx of UGC, hashtag recommendation systems have become crucial for enhancing content categorization on SNS. These systems facilitate the organization of vast volumes of UGC into manageable thematic groups through intelligent assignment of relevant hashtags. Posts annotated with hashtags then appear in dedicated feeds, enabling users to discover and engage with content aligned with their interests, even if they do not directly follow the content creator, thereby improving content discoverability and facilitating participation in topical discussions and events.

A hashtag, defined as a keyword or phrase prefixed with octothorpe symbol (#) serves as metadata and topic indicator, enabling efficient content categorization and retrieval. Since its introduction by Chris Messina in 2007 [2] for keyword-based virtual discussion groups, hashtags have become ubiquitous across major SNS platforms, underscoring their fundamental role in modern digital communication. However, the effectiveness of hashtags relies on users selecting relevant hashtags, a task complicated by informal language, idiosyncratic behavior, and platform-specific characteristics. Research indicates that a substantial proportion of UGC lacks hashtags or utilizes

suboptimal hashtags [3, 4], underscoring the need for automated hashtag recommendation systems that analyze content and user behavior to suggest relevant hashtags during content creation.

Automated hashtag recommendation systems address information overload by algorithmically suggesting relevant and impactful hashtags. By analyzing content and context, these systems can resolve incompleteness, mitigate inconsistency, correct spelling errors, map slang to standardized hashtags, and counteract lexical variability and semantic ambiguity. This automation streamlines content categorization, thereby reducing noise in feeds. Furthermore, by surfacing precise hashtags, these systems improve content discoverability and reduce cognitive load for users. Notably, research indicates that tweets with hashtags receive 12.6% more engagement than those without [5], demonstrating the impact of effective tagging. These systems also amplify high-quality content and further promote engagement.

The format of UGC on SNS plays a crucial role in shaping user interaction. Different content formats elicit varying levels of engagement within social media feeds. Notably, short-form video demonstrates the highest engagement², followed by images, indicating a user preference for visually dynamic and concise content. Conversely, unimodal textual content exhibits comparatively lower engagement, suggesting a trend towards shorter, visually-driven content consumption patterns. This can be attributed to the fact that visuals are processed significantly faster than text [6]. Therefore, considering the diverse nature of UGC, which can be unimodal such as text or images or multimodal such as combinations of text and images, or micro-videos, this thesis examines four critical aspects that significantly influence the effectiveness of hashtag recommendation systems: monolingual textual content, multilingual textual content, multimodal content comprising texts and images, and micro-videos, as shown in Figure 1.1.

²<https://sproutsocial.com/insights/types-of-social-media-content/>

1.1.1 Monolingual Content

SNS microblogging platforms, characterized by their brief and informal posts, have significantly reshaped modern communication and culture. While information on X rapidly disseminates across linguistic and geographical boundaries, English has emerged as a dominant language, constituting approximately 53%³, of its total content. This substantial prevalence underscores the importance of developing robust hashtag recommendation methods specifically tailored for English text. Hashtags play a crucial role in augmenting content discoverability, facilitating information filtering, and contributing to the systematic organization of online conversations, thereby enhancing user experience and information retrieval efficacy. The inherent lack of predefined organization or formatting in textual content, particularly the short-form text on X, renders it unstructured. Unlike structured data with clear rows and columns, textual data exists as free-flowing sequences of words, where meaning and context are embedded within the language itself. Consequently, extracting meaningful information and automatically categorizing or tagging such content necessitates sophisticated natural language processing techniques. Effective hashtag recommendation, therefore, becomes a crucial tool for imposing a degree of structure onto this otherwise unstructured data. Focusing on monolingual English text allows for the development of more precise and effective hashtag recommendation systems. By concentrating on a single language, we can leverage language-specific linguistic features, patterns of expression, and common vocabulary to better understand the nuances of the text and suggest pertinent hashtags. Recognizing the critical need for organized information to enhance accessibility, we propose a novel three-stage framework to address this challenge of intelligent hashtag recommendation. First, a retriever identifies potential candidate hashtags from a vast collection of tweets annotated with hashtags. Next, a selector refines these candidates by analyzing the input tweet, ensuring only the most relevant hashtags are retained. Finally, a diffusion-based sequence-to-sequence encoder-decoder generates informative hashtags by leveraging the refined set of candidate hashtags and the original input tweet.

³<https://semiocast.com/top-languages-on-twitter-stats/>

1.1.2 Multilingual Content

The proliferation of SNS has resulted in an exponential increase in multilingual content, particularly with the integration of support for vernacular languages, enabling users from diverse linguistic backgrounds to express themselves on trending topics. This surge is evident in nations with prevalent low-resource languages, exemplified by India, which constitutes the third largest consumer for X with an impressive daily active user count of 22.1 million⁴. Despite statistics showing that just 55% of posts are written in English, even though 99% of users have their devices set to English⁵, users tend to tweet in their regional languages when expressing their thoughts. The support for vernacular languages, allowing users to converse in Indic languages, has demonstrably transformed content dissemination and reachability. Research conducted in 2019 on X shows that 51% of Indian users tweet in English and 49% in other languages [7], with a growing trend of engaging with trending topics in native tongues. According to the census of 2001 [8], India encompasses 1,635 rationalized mother tongues, 234 identifiable mother tongues, and 22 major languages, presenting a challenge in directly matching semantically related content across languages due to script and morphological variations. While hashtags offer a potential solution for linking thematically similar multilingual posts, their infrequent use limits their efficacy. The overwhelming volume of event-related posts further complicates the retrieval of pertinent information, especially for non-English content where relevant hashtags are scarce. This poses challenges for local content creators, brands, language learners, and researchers studying multilingualism. An automated multilingual hashtag recommendation system is therefore crucial for enhancing content discoverability, facilitating connections across linguistic communities, and supporting research endeavors. The statistics on our collected dataset for posts in multiple low-resource Indic languages indicate that up to 24.16% of the 3,107,866 posts have less than two hashtags [7], underscoring the necessity and research merit of developing such a system. While extensive research exists for hashtag recommendation leveraging textual content, re-

⁴<https://backlinko.com/twitter-users#twitter-users>

⁵<https://techcrunch.com/2010/02/24/twitter-languages/>

searchers have primarily focused on high-resource languages, namely English [5, 9] and Chinese [10, 11]. However, recommending hashtags for content generated in low-resource Indic languages on SNS is mainly unexplored due to the unavailability of substantial amount of written texts, audio recordings, or other digital resources, resulting in noisy or incomplete data. Existing methods for high-resource languages cannot be directly applied due to the specialized linguistic expertise required. In light of the above, we devise an automatic hashtag recommendation system for orpheline tweets posted in low-resource Indic languages dubbed as TAGALOG. We refine tweet representations in line with user’s topical and linguistic preferences by devising novel attention mechanisms. Our graph-based neural network mines users’ historical posting behavior and language relatedness by linking tweets according to language families, namely, Indo-Aryan and Dravidian. Results from Chapter 4 shows that recommended hashtags can be used to identify the main content for specific topics regardless of the language, thereby aiding regional language users on X to effectively retrieve content and keep up to date with the latest information.

1.1.3 Multimodal Content

The continuously evolving landscape of SNS has fostered multimodal communication, wherein users integrate diverse data modes such as texts and images to exchange information, offering complementary perspectives that enrich content understanding. Multimodality is increasingly prevalent, with over a third of microblogs combining textual and visual modalities [12, 13]. These modalities provide unique insights, as text contextualizes images while images convey supplementary details. Photo-sharing platforms such as Flickr and Instagram exemplify this trend, enabling users to share photographs accompanied by textual descriptions and optional hashtags. These hashtags serve as valuable metadata facilitating tasks such as sentiment analysis [14], information retrieval [15], and topic extraction [16]. By comprehending multimodal information and user history, effective multimodal hashtag recommendation systems can enhance SNS platform quality, user engagement, and browsing experiences. While multimodal microblogs with hashtags are increasingly available on photograph sharing

services, obtaining a comprehensive latent representation from the complementary yet variably correlated visual and textual modalities necessitates effective information fusion. However, the inherent limited interaction between modalities hinders capturing their interrelations, complicating the learning of robust multimodal representations. To enhance the discoverability of prevalent multimodal content, a substantial proportion of which currently lacks effective hashtag annotations, we introduce a novel multimodal personalized hashtag recommendation method. Our proposed method captures associations between textual and visual modalities within microblogs and incorporates user tagging preferences.

1.1.4 Micro-videos

Video content elicits the highest level of user engagement, evidenced by superior information retention, as individuals recall approximately 95% of a message when viewing a video compared to just 10% when reading⁶. The rapid proliferation of micro-videos or short-form videos, a prevalent form of UGC, has been observed across video-sharing platforms such as TikTok, YouTube, and Instagram, with platform-specific naming conventions including TikTok videos, Instagram Reels, and YouTube Shorts, encapsulating substantial information within brief durations. These bite-sized clips have varying time constraints, such as six seconds on Vine, 60 seconds on YouTube, and a maximum of 90 seconds on Instagram⁷. The ease of creation and consumption of these short-form video clips aligns with the diminishing attention span of contemporary digital users. The consequent surge in micro-video data, exemplified by TikTok’s substantial user base of 689 million monthly active users and daily viewership of over a million videos [17], underscores the pressing need for effective content management and retrieval solutions. Hashtags function as crucial metadata for organizing and accessing this expanding corpus of micro-videos, facilitating efficient search and discovery for specific topics, interests, or events. Despite their importance, a significant

⁶<https://www.forbes.com/sites/yec/2017/07/13/how-to-incorporate-video-into-your-social-media-strategy/>

⁷<https://www.demandsage.com/instagram-reel-statistics/>

proportion of users neglect to incorporate hashtags into their micro-videos. Empirical evidence shows that over 33 million hashtag-devoid micro-videos posted daily on Instagram alone [18]. Given the escalating volume of micro-video data and the imperative for efficient content retrieval, we develop an automated hashtag recommendation system tailored for micro-videos that leverages content-based, collaborative filtering, and user’s historical and tagging behavior to recommend hashtags. Moreover, we devise a content-based and social influence method to recommend hashtags for micro-videos posted by cold-start user. Chapter 6 shows that our proposed social influence and content-based technique recommends both popular and content-relevant hashtags, aiding cold-start users to gain visibility on SNS.

1.2 Popularity Prediction

Information overload intensifies competition for user attention, leading to a “winner-take-all” [19] dynamic where only a small fraction of UGC captures the majority of user attention [20]. Consequently, a significant amount of UGC remains unnoticed, hindering the reach of potentially valuable UGC to their intended audiences. Popularity prediction has emerged as a crucial solution, aiming to forecast the level of public engagement UGC will receive early in its lifecycle. The popularity of UGC is operationalized through various metrics, including direct measures of user interaction such as the number of likes, shares, and comments, as well as engagement rate, audience growth rate, and view counts or watch time for video content. Forecasting UGC popularity provides insights into user interests, altering user comprehension and interaction with the digital world, and enabling content creators to optimize their output for maximum impact. By identifying UGC likely to garner substantial attention, platforms can enhance user experience by prioritizing engaging content and facilitating the discovery of relevant information. This capability also allows for efficient resource allocation in content delivery and identification of emerging trends by analyzing popular content patterns. Beyond platform optimization, accurate popularity prediction has broad applications, including enhancing recommender systems [21],

online advertising campaigns [22], sentiment analysis [23], and digital marketing [24].

Given the increasing prevalence of multimodal content, where SNS users articulate opinions and share experiences through posts comprising multiple modes of expression, understanding the impact of multimodality on popularity prediction is crucial. On X, posts containing images receive 18% more clicks, 89% more likes, and 150% more retweets compared to text-only posts⁸. Similarly, on Facebook, the inclusion of images in posts leads to a 53% increase in likes and a 104% surge in comments. Prevailing methods primarily center on the content itself, thereby overlooking information encapsulated within alternative modalities. In this thesis, we propose a novel method that leverages visual demographic features of faces in images and sentiment derived from associated hashtags [25]. Moreover, we devise a hashtag-guided attention mechanism that leverages hashtags as navigational cues to focus on the most pertinent features of textual and visual modalities.

Both hashtag recommendation and popularity prediction represent critical tools in navigating the complexities of the UGC landscape and fostering a more efficient and rewarding experience for users. The key contributions of the thesis are as follows:

- [1] We propose a retrieval augmented diffusion-based sequence-to-sequence framework to recommend hashtags for monolingual posts related to disaster events. We leverage the synergy of retrieval with the generative power of diffusion models to improve content categorization, thereby enhancing content retrieval and information dissemination on social media.
- [2] We devise a deep learning-based graph neural network to recommend hashtags for UGC in low-resource Indic languages. Our approach refines post content locally by attending to users' topical interests and language usage styles. Globally, we construct a graph to model users' long-term posting behavior and their interactions with past content. Furthermore, our framework leverages linguistic relatedness within Indo-Aryan and Dravidian language families by mining inter-language correlations. This system aims to mitigate language barriers, enhance

⁸<https://www.adweek.com/performance-marketing/twitter-images-study/>

organization and discoverability of multilingual content, and promote universal content accessibility.

- [3] We propose a hybrid deep neural network to address hashtag recommendation for multimodal microblogs by jointly formulating the task as multi-label classification and sequence generation problems. The proposed model capitalizes on the complementary strengths of both techniques. Furthermore, we leverage users’ hashtagging behavior and preferences, derived from their historical posts and associated hashtags, to recommend personalized hashtags.
- [4] We present a novel hybrid filtering approach that leverages micro-video content, users’ modality-specific tagging preferences, and community interests to facilitate context-aware, user-aware, and community-aware hashtag recommendation. To this end, we construct a heterogeneous graph capturing user-modality interactions, user-user interactions based on tagging patterns, and modality-modality interactions to capture explicit and implicit collaborative signals. To address the cold-start user problem, we propose a content-based filtering and social influence method that analyzes micro-video content and mimics tagging behavior of popular users to recommend hashtags.
- [5] We propose a deep neural network that leverages sentiment derived from hashtags, visual demographic information, and a novel hashtag-guided attention mechanism to comprehensively forecast post popularity. The hashtag-guided attention mechanism utilizes hashtags to direct the model’s focus toward content features most relevant to the intended audience and context.
- [6] This thesis also contributes two novel text-based hashtag recommendation datasets. The first, designated IndicHash, comprises posts in a diverse range of low-resource Indic languages, specifically Bangla, Marathi, Gujarati, Telugu, Tamil, Kannada, and Hindi, in addition to English. The second dataset, designated Text Dataset from Instagram (TINS), is a personalized dataset of English-language posts collected from Instagram. These curated datasets serve as valu-

able resources to facilitate further research in hashtag recommendation for Indic regional languages and English.

The organization of this thesis is as follows. Chapter 2 presents a comprehensive survey of related work. Chapter 3 details the proposed methodology for hashtag recommendation in monolingual text-based content. Chapter 4 develops a multilingual hashtag recommender to address the complexities of multilingual text. Chapter 5 investigates hashtag recommendation for multimodal content, specifically texts and images. Chapter 6 explores the unique challenges of hashtag recommendation for micro-video content. Chapter 7 presents a novel framework for predicting the popularity of multimodal content. Finally, Chapter 8 concludes the thesis and outlines directions for future research.

Chapter 2

Literature Review

In this chapter, we first present the related work on hashtag recommendation and then review literature on popularity prediction in social networks.

2.1 Hashtag Recommendation

In this section, we cover the substantial volume of research work in the domain of hashtag recommendation categorized into content-based and personalized filtering.

2.1.1 Content-based Hashtag Recommendation

UGC [26] appears in a wide spectrum of formats across social media platforms including textual data, visual media, and micro-videos. Content-based hashtag recommendation focuses on explicitly representing the core attributes of the content to suggest relevant hashtags. This approach has been rigorously investigated across diverse content formats, such as texts [7, 27, 28, 29], images [30, 31, 32], multimodal microblogs [33, 34], and micro-videos.

2.1.1.1 Text-centric Hashtag Recommendation

The textual content has dominated social media research. Hence, extensive research has been carried out to study the problem of hashtag recommendation using textual post content in monolingual and multilingual scenarios.

2.1.1.1.1 Monolingual Text-centric Hashtag Recommendation Ding et al. [35] merged topic model with translation model, positing that tweet content and associated hashtags serve as parallel descriptions of the same topic. The authors employed a topic-specific word trigger to minimize linguistic differences between tweets and hashtags. The authors first identify topics for each tweet and then generate candidate hashtags according to the learned topical translation model. Sedhai and Sun et al. [36] presented a two-phase approach to suggest hashtags for hyperlinked tweets. Their method involved gathering hashtags from similar tweets, the tweet’s domain, named entities, and hyperlinked documents, subsequently using RankSVM to rank and suggest the most relevant hashtags. Kumar et al. [28] proposed using word embeddings and external knowledge from Wikipedia and the Web to bridge the gap between tweet content and hashtags. They retrieved semantically related keywords using word2vec and integrated topical, lexical, semantic features of tweets, along with user influence, employing Learning-to-rank to aggregate different hashtag generation methods [28].

A common approach in hashtag recommendation involves directly extracting keyphrases from the source text [37, 38, 39]. Zhang et al. [40] identified hashtags as valuable keywords for extracting keyphrases from X. However, this method inherently limits hashtag generation to terms already present in the text, overlooking the creative aspect of hashtag usage. Users can create novel hashtags owing to their backgrounds, proficiency levels and linguistic styles, leading to suboptimal results.

The task of hashtag recommendation has been approached as both a classification and a generation problem. Classification-based methods typically predefine a limited set of candidate hashtags [41] and utilize a softmax layer for prediction. However, this reliance on a fixed vocabulary is restrictive, particularly in dynamic environments where new and relevant terms frequently emerge. The computational cost of continuously updating these models renders them less practical for rapidly evolving social media landscapes. In contrast, sequence generation approaches [9, 11, 42] enable the creation of more diverse and expressive hashtags. By considering the sequential nature of hashtags, these models can capture dependencies among them. While hashtags in

the output set might be correlated, they do not always follow a strict sequential order akin to words in a sentence. Although classification treats hashtags as independent categories, they can also be generated sequentially. To model sequential relationships and implicit correlations among hashtags, we interpret hashtag recommendation as a generation task to better reflect natural user behavior.

Addressing the limitations of existing monolingual hashtag recommendation methods, this thesis proposes a novel retrieval-augmented diffusion-based sequence-to-sequence framework for monolingual content (refer to Chapter 3). By leveraging the synergy of retrieval with the generative power of diffusion models, we aim to improve content retrieval and the quality of recommended hashtags. Empirical evaluations demonstrate the superior performance of our proposed method compared to state-of-the-art approaches in terms of both hashtag quality and training efficiency.

2.1.1.1.2 Multilingual Text-centric Hashtag Recommendation The task of suggesting hashtags for textual content can be posed using one of the traditional problems in Natural Language Processing (NLP), i.e., text categorization [43, 44, 45, 46]. As far as we are aware, although many works have been carried out for classifying text in low-resource Indic languages [47, 48, 49], there is only one work that predicts hashtags for multilingual content [50]. Low-resource languages (LRLs), also known as “less studied, under-resourced, low density” languages are languages with limited linguistic resources, such as textual material, language processing tools, grammar and speech databases, dictionaries, and human competence [51]. These languages are frequently spoken by small groups, lack standardized writing systems, and have a scarce digital presence. Researchers in NLP distinguish LRLs based on the availability of data and NLP tools. LRLs have a relatively small amount of data, i.e., text corpora, parallel corpora, and lack language-specific tools such as spell checkers and grammar checkers, and manually crafted linguistic resources for training NLP models. There are a number of advantages to working with low-resource languages that have the potential to impact the lives of people who speak these languages, the opportunity to develop new NLP techniques that can be applied to other languages, and the chal-

lenge of working with limited data. Efforts are being made by linguists, researchers, and organizations to document languages, construct corpora, develop technology and tools, and community-driven language revival campaigns for LRLs since LRLs offer humongous benefits some of which are enlisted below.

Due to small corpora and unseen scripts, labeled data for diverse Indic languages is sparse or nonexistent in real applications compared to high-resource languages such as English and Chinese. To get beyond corpus restrictions inherent in low-resource languages, Khemchandani et al. [52] proposed RelateLM to effectively customize language models for low-resource languages. Since numerous Indic scripts descended from Brahmi script, the authors take advantage of script relatedness through transliteration. RelateLM artificially translates relatively well-known language content into low-resource language corpora using comparable sentence structures to get around corpus limitations. Aggarwal et al. [53] performed zero-shot text classification for Indic languages by leveraging lexical similarity. To this end, the authors performed script conversion to Devanagari and divided words into sub-words to optimize the vocabulary overlap among the related Indic languages datasets. Khatri et al. [54] investigated the influence of sharing encoder-decoder parameters between related languages in Multilingual Neural Machine Translation. They developed a system trained from the languages by grouping them based on language family i.e., Indo-Aryan group to English and Dravidian group to English. Then, the authors convert the entire language data to the same script, which helps the model learn better translation by utilizing shared vocabulary. This approach obscures the underlying structural similarities between languages. Language families are typically defined based on shared ancestry and historical relationships between languages. Transliteration-based methods may not accurately capture these relationships between languages, as they focus primarily on the surface features of languages which amounts to inaccurate results for downstream tasks. Marreddy et al. [55] put forward a supervised approach to rebuild graph called as Multi-Task Text GCN. This method utilizes a Graph AutoEncoder (GAE) [56] to learn the latent word and sentence embeddings from a graph which is employed to carry out Telugu text categorization for various downstream tasks. Zhang

et al. [50] proposed a Twitter Heterogeneous Information Network (TwHIN-BERT) to anticipate hashtags for multilingual content. The authors employ Approximate Nearest Neighbor (ANN) search to identify pairs of socially appealing tweets. This method falls short of capturing the user’s language and topical choices. Furthermore, it does not take linguistic relatedness within language groups into account to address the low-resource nature of numerous languages featured in the dataset.

Therefore, a quick assessment reveals that research has primarily focused on text-only, image-only, or multimodal information posted in high-resource languages i.e., English, and Chinese. These studies do not consider recommending hashtags for content posted in low-resource languages. To tackle this issue, we propose a novel multilingual system i.e., TAGALOG, which extracts content-based, user-based, and language-based features to recommend personalized and language-specific hashtags for content created in low-resource Indic languages (refer to Chapter 4). Experimental evaluations on the curated dataset from X demonstrate that the proposed model outperformed recognized pre-trained language models and extant research, showing significant improvements in F1-score. TAGALOG recommends hashtags that align with the user’s interests and linguistic predilections, leading to a heightened level of tailored and engaging user experience. Personalized and multilingual hashtag recommendation systems for low-resource Indic languages can help to improve the discoverability and relevance of content in these languages.

2.1.1.2 Image-centric Hashtag Recommendation

Convolutional Neural Networks (CNN) have exhibited good generalization power in visual recognition applications. Sigurbjornsson et al. [57] developed a system to recommend tags to users for photos by utilizing combined knowledge of the entire Flickr community users. The authors first studied user tagging behavior, then image content, and accordingly suggested potential tags for image annotation. Gong et al. [58] leveraged the feature representations extracted from Deep Convolutional Neural Networks. The authors also made use of top-k ranking objectives to solve hashtag suggestion in terms of multi-label classification. Gong et al. [59] investigated novel

attention-based CNNs for performing hashtag recommendation to avoid hand-crafted features. The proposed CNN architecture contained two attention channels, i.e., local and global, to process the input microblog and select trigger words. This method yielded superior performance compared to those considering only global or local information. Wu et al. [60] devised a neural network that captures correlations among tags and photos. The authors applied an attention mechanism to retrieve valuable information from images and suggest hashtags accordingly. The aforementioned systems utilize information from either texts or images for recommending tags. However, text or image alone cannot capture the entire post information. Our approach aims to suggest hashtags by utilizing various modalities together to provide richer contextual information.

2.1.1.3 Multimodal Hashtag Recommendation

Social media data is primarily inclined to multimedia content. The enormous amount of available information comprises both visual and textual modalities. Zhang et al. [12] integrated visual and textual information for hashtag recommendation. The authors used an alternative co-attention network in which tweet is used to obtain visual attention. Then the obtained image representation is used to generate the textual attention and later obtain a more informative post representation. When tagging an image, a user considers both content and context of image. Rawat et al. [61] devised a Deep Neural Network that leverages content and context of an image to recommend tags to the user. The authors regarded hashtag recommendation as a multi-label classification problem and used AlexNet to retrieve visual features aggregated with ContextNet. Zhang et al. [5] used text and image to recommend hashtags for posts on photo sharing services. The authors designed a co-attention mechanism in which image and text co-guide each other, to obtain an informative post representation. Since many users prefer to create a post with texts and images, it is essential to find an effective way to incorporate both modalities. In this thesis, we propose a hybrid deep neural network to automatically recommend hashtags for multimodal unannotated social media content (refer to Chapter 5). Our proposed

method predicts suitable hashtags for posts by mining information from textual and visual modalities. We apply word-level attention on textual content to learn those words in the text that are more closely related to hashtags followed by a parallel co-attention mechanism to model deep interactions between the two modalities.

2.1.1.4 Micro video-centric Hashtag Recommendation

Unlike [62] who proposed a Guided Generative Model to generate hashtags from multimodal inputs and guided signals from a Visual Language Model-based Hashtag Retriever, we capture both individual and community interests to recommend hashtags for micro-videos. Cao et al. [4] obtained feature representations from micro-video modalities and integrated them via a multi-view representation learning framework. The regularized projections and hashtag embeddings are fed to a customized neural collaborative filtering framework to yield hashtags. Building on this, studies [63] further incorporated sentiment analysis, integrating sentiment features, content features, and semantic embeddings of hashtags using weighted concatenation. While these methods are capable of recommending sentiment hashtags and utilize all three modalities of micro-videos and employ sequential modeling, they primarily rely on concatenating modality-specific features before projection into latent space. Mehta et al. [64] built a heterogeneous graph connecting hashtags based on semantic co-occurrence, videos based on shared hashtags, with direct links between videos and assigned hashtags. They employed a Graph Convolutional Network (GCN)-based node update scheme to generate micro-video embeddings for hashtag recommendation. Chen et al. [30] created an image similarity graph to illustrate the relationship between posts, assuming visually comparable images use similar hashtags. We extend this idea to micro-videos and compute intramodality similarity, assuming that similar modalities in micro-videos share similar hashtags.

Despite their effectiveness in analyzing micro-video content, content-based methods often overlook individual tagging behavior, the tagging patterns of like-minded users, and the social context of popular and trending hashtags. This disregard limits personalization and the ability to capture the dynamic nature of hashtag usage within

social networks.

To address these limitations, this thesis presents a novel approach for micro-video hashtag recommendation (refer to Chapter 6). Our method leverages micro-video content, user’s modality-specific tagging preferences, and community interests to facilitate context-aware, user-aware, and community-aware hashtag recommendations. Specifically, our approach explicitly models user’s modality-specific interest using a GCN-based message passing strategy, thereby aligning recommendations with both individual and community preferences.

2.1.2 Personalized Hashtag Recommendation

Personalized hashtag recommendation aims to tailor suggestions to individual user preferences and behaviors, addressing the limitations of non-personalized methods [63, 65, 66, 67] that focus solely on content. Early personalized approaches leveraged user history. Van et al. [68] retrieved hashtags from the most similar past tweets based on content and user features. Durand et al. [69] employed an open vocabulary model, learning user characteristics from past images and hashtags to suggest user-aware tags. Zhang et al. [5] also considered users’ historical tagging habits alongside content-based attention mechanisms. However, relying solely on past content similarity [5] overlooks the broader social network influence. To address this, Peng et al. [70] introduced a neural memory network to model extensive user histories, incorporating a gating mechanism for unrelated hashtag usage. Jeong et al. [71] utilized demographic data alongside content, while Padungkiattawattana et al. [72] proposed PAC-MAN, integrating high-order user-hashtag relations with content-based BERT.

Other deep learning-based approaches aimed to handle the long-tail distribution of hashtags. Li et al. [73] used external knowledge and a pairwise interactive embedding network but averaged user representations across modalities, neglecting fine-grained preferences. Liu et al. [74] used metadata to guide attention, but demographic reliance can be inaccurate for idiosyncratic users. MISHON addressed this by modeling user-user interactions to incorporate community preferences. Graph Neural Networks have also been explored: Wei et al. [18] used GCNs to model user-micro-video-hashtag

relationships but overlooked user-user interactions and modality-specific preferences. Shuai et al. [75] modeled user interests at a topic level using topic and rating graphs.

In this thesis, we propose a personalized hashtag recommendation system tailored for multilingual, multimodal, and micro-video content. For multilingual content (Chapter 4), we capture users’ topical and linguistic preferences locally using attention mechanisms and model long-term behavior globally with user interaction graphs. For multimodal content (Chapter 5), our system leverages users’ historical hashtagging behavior and preferences. For micro-videos (Chapter 6), we capture modality-specific tagging preferences by linking users to the modalities of their past micro-videos. This multi-faceted approach aims to provide effective personalized hashtag recommendations by considering individual preferences and the specific characteristics of different content types.

2.2 Popularity Prediction

Popularity prediction of online content has attracted significant research interest, with scholars proposing diverse methodologies. These methodologies leverage a variety of techniques, datasets, model structures, and problem formulations i.e., either classification or regression. Kumar et al. [76] focused on predicting the popularity of news articles, particularly their ability to attract user comments, offering the opportunity for informed content modifications using various features extracted from article content. Lin et al. [77] devised a framework by stacking multiple regression models across several layers, fostering synergies among diverse models to anticipate engagement of posts on a platform similar to Flickr. Purba et al. [78] devised a novel approach for predicting Instagram post engagement rates on a global dataset. It utilizes Support Vector Regression (SVR) and incorporates features from hashtags, image analysis, user history, and manual image assessment. Cao et al. [79] proposed CoupledGNN, a novel graph neural network framework designed for predicting popularity with network awareness on social platforms. This framework leverages two interconnected GNNs to capture the propagation of influence. One GNN models the user activation status

within the network, while the other GNN models the dissemination of information itself. Mannepli et al. [80] leveraged a Long Short-Term Memory (LSTM) network to predict popularity by combining features extracted from text content, user data, time series information, and user sentiment analysis. Furthermore, Self-Adaptive Rain Optimization (SA-RO) is employed to fine-tune LSTM weights, enhancing the prediction accuracy. Tan et al. [81] leveraged transformers for time series feature extraction and CatBoost for feature selection, enabling comprehensive multi-view feature extraction and achieving superior prediction accuracy on the Social Media Prediction Dataset.

The effectiveness of these approaches depends on the selection and quality of features used to represent the content. These models may not capture several other modalities comprising the social media post, and their interrelationships might have been overlooked. Prior studies have not employed hashtags used in a post as the main feature in popularity prediction tasks. According to Zappavigna [82], hashtags have a variety of uses on social media, with subject markers serving as the most important ones. Hashtag efficiently defines the topic of the post and help the user to easily identify if the post is related to them or not. Few researchers have also explored the representation of hashtags in the form of graphs. Liu et al. [83] proposed a network framework where hashtags serve as nodes. Each node is linked to a collection of tweets, which themselves are comprised of individual words. To capture the inherent relationships within this hierarchical, heterogeneous network, they introduce the Hashtag2Vec embedding model. This model extends its embedding capability beyond hashtags to encompass short social text elements by simultaneously considering relationships between hashtags themselves (hashtag-hashtag), hashtags and tweets (hashtag-tweet), tweets and their constituent words (tweet-word), and finally, word-to-word relationships within tweets. Liao et al. [84] proposed a multimodal framework that analyzes hashtag network structure semantics, and topic modeling besides captions and images to predict popular influencer posts in Taiwan. Chakrabarti et al. [85] addressed the challenge of maximizing social media popularity by recommending context-relevant hashtags. The proposed framework leveraged post keywords, user popularity, and trending hashtags to recommend effective hashtags. However, these approaches focus

on the structural and textual aspects of hashtags ignoring the sentiment information embedded in them. Posts with positive or sentimental hashtags tend to perform better than those with neutral or negative hashtags. This indicates that sentiment analysis of hashtags can be employed to forecast which posts are more inclined to become viral.

Many researchers have stated that using the attention mechanism as a part of the popularity prediction model can significantly enhance performance. As many models use images and texts as primary features to predict the popularity score, the attention mechanism can help represent these features in a better way. Xu et al. [86] proposed a regression model to predict popularity where they used an attention layer at the top of the model and showed a significant improvement when compared to models without attention. Lin et al. [87] utilized a self-attention mechanism to fuse semantic and numeric features effectively for social media popularity prediction, outperforming other methods. Bansal et al. [88] devised a word-level parallel co-attention mechanism to derive an enriched representation of multimodal social media posts by capturing the mutual influence of text and image on each other. Wang et al. [89] presented a novel multimodal popularity prediction model grounded in hierarchical fusion, where extracted features encompass visual elements, textual content, along with attributes extracted from both modalities. The model innovatively integrates three integration stages namely, early integration, representation integration, and modality integration - culminating in a fully fused vector inputted into an XGBoost regression model for effectively estimating the virality of posts. However, extant approaches prioritize local features, neglecting a holistic understanding of the content’s multimodal nature.

To overcome this limitation, we propose a novel multimodal deep learning model with a novel hashtag-guided attention mechanism (refer to Chapter 7). This method incorporates diverse feature types, encompassing content characteristics, sentiment analysis, hashtag information, and user demographics. The proposed model uses transfer learning, deep learning, attention mechanism, and graph neural networks to learn fine-grained representations of these features and then feed these enhanced feature representations into a unified framework to estimate the post popularity.

Chapter 3

Hashtag Recommendation for Monolingual Content

3.1 Introduction

This chapter focuses on hashtag recommendation for monolingual textual content as a crucial first step within the broader scope of this thesis. Recognizing the complexities inherent in diverse languages and content formats, establishing robust techniques for a prevalent language such as English provides a necessary bedrock for subsequent research. Given that English constitutes a substantial proportion of content on platforms such as X, it represents a critical domain for developing effective hashtag recommendation systems that enhance content discoverability and thematic organization within the extensive landscape of UGC.

Hashtags serve as vital metadata for structuring short textual posts prevalent on microblogging platforms. By introducing a degree of organization, they facilitate efficient information filtering and retrieval, significantly improving user experience. However, recommending appropriate hashtags even within the confines of a single language such as English presents distinct challenges that necessitate sophisticated methodologies. Accurately capturing semantic clues embedded within brief and informal posts requires deep contextual understanding, requiring systems to discern subtle meanings, figurative language, and domain-specific terminology. Moreover, recommendation systems must effectively address the considerable linguistic variability inherent even in

monolingual communication, encompassing diverse vocabulary, syntax, slang, abbreviations, and stylistic variations. Furthermore, the sheer volume of daily posts creates information overload, underscoring the necessity of intelligent hashtag recommendation for filtering relevant content and maximizing its visibility.

Addressing these multifaceted complexities is paramount for enhancing information access and user engagement on SNS. The effective application of hashtags within monolingual text is not without these inherent challenges, necessitating the development of robust hashtag recommendation systems. Understanding these complexities is particularly critical in various application scenarios, one salient example being crisis communication, specifically concerning disaster events. During disasters, people increasingly rely on social media for updates, assistance, and vital information sharing [90, 91]. However, this valuable resource remains underutilized due to the sheer volume of information, making it difficult to identify critical updates [92]. Monolingual text-based posts, such as English tweets, play a critical role in disseminating real-time updates, coordinating relief efforts, and providing situational awareness. In this context, accurate and consistent hashtagging is paramount. It facilitates the rapid identification and dissemination of crucial information to emergency responders. Furthermore, it keeps the affected public informed about the evolving situation, evacuation routes, and available resources. Finally, it enables the aggregation and analysis of on-the-ground reports to assess damage and identify areas needing immediate assistance. Therefore, this chapter analyzes the intricacies involved in monolingual text-based hashtag recommendation and introduces a novel three-stage framework specifically designed to generate relevant and informative hashtags for English textual content, thereby laying the groundwork for the subsequent exploration of multilingual and multimodal contexts central to this thesis.

It was found that FEMA’s initial damage estimates for Hurricane Harvey overlooked nearly half of the relevant online data [93], resulting in a significant underestimation of the total cost. This instance highlights the potential consequences of overlooking online information and the need for tools to effectively filter and utilize it. Hashtags are vital for organizing and disseminating critical information on social

networks during disasters. They facilitate effective communication and coordination among emergency responders, government agencies, and the public, improving real-time situational awareness. Accurately tagged tweets identify the disaster’s nature, location, affected areas, severity, and specific needs of those on the ground. However, approximately half of disaster-related tweets lack informative hashtags [93], hindering effective response. Therefore, automated hashtag recommendation systems are essential for optimizing information accessibility, efficient filtering of critical updates, improving disaster response, efficient resource allocation, and mitigation efforts.

Recommending hashtags for disaster-related tweets presents unique challenges. The information landscape during disasters is highly dynamic and noisy, with new needs and challenges constantly emerging. Existing retrieval-based and generation-based methods struggle to keep pace with rapidly changing environment. Retrieval-based methods [4, 18, 7], relying on fixed hashtag lists, cannot keep pace with the rapidly changing information. Generation-based methods [42, 11], though better at understanding new information, may produce inaccurate hashtags without additional guidance. Therefore, disaster-related tweets necessitate a system that can accurately capture evolving needs and challenges faced by affected communities as the situation unfolds, effectively filter and process information from social media data containing informal language and misspellings, and generate hashtags that not only reflect the current situation but also anticipate future needs. Retrieval-Augmented Generation (RAG) techniques provide an effective solution by capitalizing on retrieval and generation approaches. This enables RAG models to leverage existing knowledge while adapting to new information, crucial for hashtag recommendation in disaster scenarios.

Existing hashtag generation methods, predominantly based on encoder-decoder frameworks with Recurrent Neural Network (RNN) [9, 94] or transformers [11, 95] struggle to perform effectively in disaster scenarios. Though tweets have a character limit, RNNs, while capable of capturing sequential information, struggle with long-range dependencies, hindering their ability to process the full context of a tweet. Transformers, while robust, generate generic or repetitive hashtags when faced with the noisy and informal language usage, spelling, and grammar mistakes by users com-

mon in disaster situations. Consequently, these methods fail to generate hashtags that accurately reflect the dynamic nature of needs during a disaster, hindering effective communication and response efforts. Inaccurate and irrelevant hashtags can lead to misdirection of resources and delay critical assistance. Diffusion models, which learn to reverse a progressive noising process, have successfully generated high-quality synthetic data across multiple domains [96, 97]. This success extends to various Natural Language Processing (NLP) tasks such as unconditional [98] and controlled text generation [99]. However, their application to sequence-to-sequence (seq2seq) text generation, particularly for hashtag recommendation in disaster scenarios, remains largely unexplored. Inspired by their potential, we propose the use of diffusion models to recommend hashtags for disaster-related tweets.

To address these challenges, we propose retrieval **A**ugmented encoder-decoder with diffu**S**ion for **S**equen**I**al hashta**G** recommen**D**ation in disast**ER** events (ASSIGNER). The retriever identifies candidate hashtags by searching a large tweet-hashtag corpus for similar tweets and associated hashtags. By comparing the input tweet to retrieved tweets and hashtags, the selector narrows down this collection ensuring that the generator receives the most pertinent hashtags. Our novel diffusion-based generator leverages this refined set and input tweet to generate informative hashtags. It utilizes an encoder-decoder architecture, where the continuous diffusion framework is integrated within the seq2seq generation process. Further, we incorporate self-conditioning and an adaptive non-linear noise scheduler for improved performance. Extensive evaluations on a dataset of disaster-related tweets demonstrate enhanced performance in hashtag generation, achieving superior results in both hashtag quality and training time compared to existing state-of-the-art approaches. Ablation studies confirm the benefits of self-conditioning and the adaptive non-linear noise schedule, highlighting their complementary nature in seq2seq settings.

Our key contributions can be summarized as enlisted below.

- We propose a retrieval augmented diffusion-based seq2seq framework to recommend hashtags for disaster-related tweets. We leverage the synergy of retrieval with the generative power of diffusion models to improve communication and

response effectiveness during crises.

- As far as we know, this work presents the first application of diffusion models to hashtag recommendation in disaster scenarios. We adapt the continuous text diffusion model to generate hashtags sequentially using an encoder-decoder transformer architecture.
- We leverage retrieved hashtags from similar tweets to provide contextual information and guide the generation of relevant hashtags for disaster-related tweets.
- Our newly proposed adaptive non-linear noise scheduler significantly improves the quality of generated hashtags by allowing for finer-grained control over the generation process.
- Experiments show that the proposed model performs competitively compared to existing methods in generating high-quality and informative hashtags for tweets about disaster events.

The subsequent sections of this chapter are structured as follows. Section 3.2 details the proposed novel methodology for generating hashtags for monolingual content. Following this, Section 3.3 presents the experimental setup and a comprehensive analysis of the obtained results. Finally, Section 3.4 concludes this work.

3.2 Methodology

To improve the relevance of generated hashtags, we incorporate a retrieval mechanism in our framework that leverages existing knowledge from a curated hashtag-tweet corpus. This module is composed of retriever and selector.

3.2.0.1 Retriever

The retriever module identifies relevant hashtag-tweet pairs from a knowledge database by utilizing the embedding of the input tweet. This approach, inspired by [100], leverages the observation that semantically similar tweets often share similar

hashtags, reflecting common usage patterns. We construct a knowledge base \mathcal{D} composed of tweet-hashtag pairs (d_i, H_i) where d_i represents a tweet and H_i represents its corresponding set of hashtags. For a new input tweet d_q , the retriever \mathcal{F} compares its embedding with the embedding of every other tweet in the corpus. It then retrieves the top-N most semantically similar tweet-hashtag combinations with corresponding similarity scores.

$$(d_1, H_1, s(d_q, d_1)), \dots, (d_N, H_N, s(d_q, d_N)) = \mathcal{F}(d_q | \mathcal{D}) \quad (3.1)$$

where, $s(d_q, d_i)$ denotes the similarity between the query tweet d_q and the i^{th} retrieved tweet d_i and each H_i contains a set of hashtags $\{h_{i1}, \dots, h_{i|H_i|}\}$. This retrieval process provides a candidate pool of potentially relevant hashtags based on similar tweets in the knowledge base.

3.2.0.2 Selector

The selector module refines hashtag recommendations by filtering out low-quality and less prevalent hashtags from retrieved pairs. We leverage two key indicators of hashtag prominence to refine the selection process: the semantic relatedness between the input tweet and the retrieved tweet, and the relevance of retrieved hashtags to the input tweet. This multifaceted selection process ensures that chosen hashtags are not only semantically relevant but also reflect popular and widely used terms. The selector is trained using a dataset comprising positive and hard negative samples. Each hashtag in a tweet is considered a positive sample (h^+). To construct hard negative samples (h^-), we employ a BERT-inspired perturbation strategy, where labeled hashtags are modified without altering their semantic meaning. This involves randomly selecting a word within the hashtag to replace it with a synonym, delete it, swap it with an adjacent word, or insert a synonym after it. The resulting training dataset consists of tuples (d_i, h_i^+, h_i^-) , where $i = 1, \dots, N$. The training of the selector module involves

minimizing a contrastive loss function, defined as follows:

$$\mathcal{L}_C = -\log \frac{e^{\text{sim}(E_{d_i}, E_{h_i^+})/\tau}}{\sum_{j=1}^L (e^{\text{sim}(E_{d_i}, E_{h_j^+})/\tau} + e^{\text{sim}(E_{d_i}, E_{h_j^-})/\tau})} \quad (3.2)$$

where, sim represents cosine similarity, E_d denotes the embedding of d , L is the mini-batch size, and τ is a temperature hyperparameter.

3.2.1 Diffusion

This section describes the core diffusion model employed for hashtag generation.

3.2.1.1 Input Encoding

Given an input tweet d_q and top-p hashtags $\{\tilde{h}_1, \dots, \tilde{h}_p\}$ selected by the selector module, we concatenate these hashtags to the input tweet:

$$\tilde{d}_q = \tilde{h}_1 \oplus \dots \oplus \tilde{h}_p \oplus d_q \quad (3.3)$$

where, \oplus denotes the concatenation operation. This concatenated input \tilde{d}_q is then fed into BART encoder to obtain its contextualized representation x_e .

$$x_e = \text{BART}_{\text{enc}}(\tilde{d}_q) \quad (3.4)$$

This embedding (x_e) captures information from relevant hashtags and the input tweet, providing richer context for the diffusion model.

3.2.1.2 Forward Diffusion Process

Our approach involves a forward process that gradually introduces noise into the target output sequence $y_{w_{i=1}}^l$, where l represents its maximum length. This noise diffusion process is independent of the input sequence x_e . We first represent the output sequence y_w using an embedding function $f_\phi(\cdot)$ which maps individual word tokens y_w^i to continuous embeddings $f_\phi(y_w^i) \in \mathbb{R}^m$, where m is the embedding di-

mension and ϕ represents parameters of $f(\cdot)$. The overall output sequence embedding is obtained by concatenating individual token embeddings and is denoted as $f_\phi(y_w) \in \mathbb{R}^{l \times m}$. The forward process begins by applying a Markovian transition parameterized by $q_\phi(v_0|y_w) = N(v_0; f_\phi(y_w), \gamma_0 I)$ is added. The forward process is augmented by $q_\phi(v_0|y_w)$, which incrementally introduces diffusion to the continuous features of v_0 . At each time step t , we apply the diffusion distribution $q(v_t|v_{t-1})$ to generate noisier samples. Finally, the original output sequence y_w is converted into v_T which closely resembles random noise drawn from a standard Gaussian distribution.

3.2.1.3 Reverse Process

When reversing the noise injection, diffusion models synthesize data points by progressively drawing samples from the noise-reducing distribution p_θ parameterized by θ . This process transforms noisy samples v_t into progressively clearer samples v_{t-1} . In seq2seq setting, the noise reduction distribution depends on the input representation x_e which is augmented with selected hashtag embeddings, represented as $p_\theta = p_\theta(v_{t-1}|v_t, x_e)$. When the reverse process reaches $T = 0$, the generated output \hat{v}^0 is mapped to its closest word in the embedding space. This mapping is achieved using a rounding distribution $\hat{p}_\phi(y_w|\hat{v}_0)$, ultimately producing the final sequence of hashtags.

3.2.1.3.1 Self-Conditioning Through a series of iterative refinements, the reverse process transforms a noisy representation into the final output sequence. At each iteration t , the function $v_\theta^0(v_t, x_e, t)$ takes the current noisy sample v_t and the input embedding x_e to predict a less noisy version of the output, gradually revealing the true sequence. This process inherently discards some information from the previous prediction \hat{v}_t^0 . To address this information loss, Bit-Diffusion [101] introduced a self-conditioning technique that incorporates previous sequence predictions as additional input to the denoising function, formulating it as $v_\theta^\theta(v_t, \hat{v}_t^0, x_e, t)$. Self-conditioning allows the denoising function to refine previous sequence predictions instead of entirely generating new ones. A study [102] has shown that self-conditioning enables text diffusion models to perform better. To integrate self-conditioning technique, we con-

catenate sequence features \hat{v}_0^t from previous predictions with noisier sequence features v_t , increasing the decoder input dimension to $l \times 2m$. To improve training efficiency, we adopt a strategy where, with 70% probability, $v_\theta^0(v_t, \hat{v}_0^t, x_e, t)$ is trained with the input \hat{v}_0^t set to 0. Alternatively, \hat{v}_0^t is initially approximated using $v_\theta^0(v_t, 0, x_e, t)$, and this estimate is then employed for self-conditioning during training, thus bypassing the need for backpropagation through initial forward pass.

3.2.1.3.2 Denoising with Encoder-Decoder Framework For $v_\theta^0(v_t, x_e, t)$, we use the encoder for modeling the input sequence x_e and the decoder for handling the noisy output sequence v_t , augmented with time step embeddings. This encoder-decoder structure provides computational efficiency during generation by allowing the encoder to process the input sequence x_e just a single time during the entire reverse procedure, which may require numerous iterations to produce high-quality output. Our denoising function (v_θ^0) produces samples at the sequence level throughout both training and generation phases, consistent with non-autoregressive approaches to natural language generation. The decoder utilizes an attention mechanism that can attend to all positions within the output sequence at once. This differs from causal attention, which is restricted to attending only to preceding positions. By having access to the full context of the output sequence, the decoder can generate more informed predictions.

3.2.1.4 Adaptive Sigmoid-based Non-Linear Noise Scheduler

We put forward a novel approach for adjusting noise non-linearly at the token level during training in diffusion models. This dynamic adjustment modulates the difficulty of the denoising process for the predicted output sequence, focusing on challenging tokens and thereby improving overall performance. Here, v_θ^0 represents the predicted output sequence at timestep 0 so that it increases sigmoidally with respect to timestep t . This aims to create a smooth progression of difficulty in denoising, making it easier at the start and end, facilitating initial stability and fine-grained refinement, respectively. This refined control over the noise injection process helps in generating

high-quality hashtags. Recognizing that different token positions within a sequence hold varying levels of semantic and syntactic importance, we propose individual noise schedules for each token. This is motivated by the observation that inherent properties of token embeddings vary significantly across different positions.

We estimate the complexity of denoising process by examining the training loss at each time step (t) and token position (i).

$$\mathcal{L}_t^i = \mathbb{E}_{q_\phi(x_e, y_w, v_t, v_0)} |v_0^\theta(v_t, \hat{v}_t^0, x_e, t)^i - v_0^i|^2 \quad (3.5)$$

We utilize $\beta_t^i \in [0, 1]$ to regulate the noise intensity at each step. β_t^i variable meticulously determines the noise level at each time step t for each token position i . To achieve an adaptive noise schedule for each token position i , we employ a mapping $\beta^i = \Phi_i(\mathcal{L}^i)$, which connects \mathcal{L}_t^i and β_t^i using a sigmoid function.

$$\beta_t^i = \Phi_i(\mathcal{L}_t^i) = \frac{1}{1 + \exp(-a_i(\mathcal{L}_t^i - b_i))} \quad (3.6)$$

where, $^*\beta_t^i$ is the noise level at time step t for token position i , $^*\mathcal{L}_t^i$ is the denoising loss at time step t for token position i . *a_i and b_i are learnable parameters that control the shape of the sigmoid for token position i . This function provides a smooth and flexible way to modify the noise intensity according to the observed complexity of denoising at each token position. We begin by initializing a noise schedule (using a standard cosine schedule) and tracking the loss, \mathcal{L}_t^i , at each step. This data is then used to establish the mapping function, Φ_i , which is updated periodically throughout training. In an ideal scenario, the training loss would consistently increase with each time step (t). This is because a larger t indicates a higher level of noise in the input characteristics (v_t) provided to denoising function. Nonetheless, given that total number of time steps (T) is typically very large, we end up with a highly detailed discretization of β^i . This fine granularity, combined with variations in empirical loss estimation, can lead to inconsistencies where the training loss does not strictly increase with each successive time step.

To address this and achieve a smoother mapping function (Φ_i), we employ a coarser

discretization (s) for both β^i and \mathcal{L}^i . This strategy helps to smooth out minor fluctuations in the observed loss and ensures a more stable and reliable mapping.

$$\mathcal{L}_s^i = \frac{1}{K} \sum_{t=s \times K}^{s \times (K+1)} \mathcal{L}_t^i, \beta_s^i = \frac{1}{K} \sum_{t=s \times K}^{s \times (K+1)} \beta_t^i, s = \left\lfloor \frac{t}{K} \right\rfloor \quad (3.7)$$

where K represents the stride to uniformly downsample t and $\lfloor \cdot \rfloor$ denotes the floor function. Using the learned sigmoid mapping $\beta_s^i = \Phi_i(\mathcal{L}_s^i)$, we can derive an updated

Algorithm 3.1 Adaptive Sigmoid Noise Schedule

- 1: **Input:** Losses \mathcal{L}_t^i and noise schedules β_t^i accumulated over each diffusion iteration t and sequence index i .
 - 2: **if** Step counter % Update interval == 0 **then**
 - 3: **for** each sequence index i **do**
 - 4: Fit the sigmoid function $\Phi_i(\mathcal{L}_t^i) = \frac{1}{1 + \exp(-a_i(\mathcal{L}_t^i - b_i))}$ by minimizing the error between β_t^i and $\Phi_i(\mathcal{L}_t^i)$, updating parameters a_i and b_i .
 - 5: Generate new loss values $\mathcal{L}_t^{i, \text{new}}$ sampled at uniform intervals between the minimum and maximum observed losses, $\min_t(\mathcal{L}_t^i)$ and $\max_t(\mathcal{L}_t^i)$.
 - 6: Compute the updated noise schedule $\beta_t^{i, \text{new}} = \Phi_i(\mathcal{L}_t^{i, \text{new}})$.
 - 7: **end for**
 - 8: **end if**
 - 9: **Return:** Noise schedule $\beta_t^{i, \text{new}}$ for each diffusion iteration t and sequence index i .
-

discretized noise schedule $\beta_t^{i, \text{new}}$ by $\beta_t^{i, \text{new}} = \Phi_i(\mathcal{L}_t^{i, \text{new}})$ where $\mathcal{L}_t^{i, \text{new}}$'s are evenly taken ranging from the minimum to maximum recorded values. Throughout the training process, we dynamically adjust β^i by repeating this procedure with each training update. This iterative process ensures that the noise schedule remains aligned with the evolving complexity of the denoising task. Algorithm 3.1 presents the pseudo-code for configuring adaptive sigmoid noise schedule during the training process. The learnable parameters (a_i and b_i) allow the sigmoid to adapt to different loss distributions and token positions.

3.2.1.5 Training Objective

The model parameters θ and ϕ are learned through a variational approximation for the data likelihood to reduce the difference between the learned denoising distribution

$p_\theta(v_{t-1}|v_t, x_e)$ and the true posterior distribution $q(v_{t-1}|v_t, v_0)$ from the forward process.

$$\mathcal{L}_{VB} = \mathbb{E}_q \left[\log \frac{q(v_T|v_0)p(v_T)}{p_\theta(v_0|v_1, x_e)} + \sum_{t=2}^T \log \frac{q(v_{t-1}|v_0, v_t)p_\theta(v_{t-1}|v_t, x_e)}{p_\theta(v_0|v_1, x_e)} + \log \frac{q_\phi(v_0|y_w)}{\tilde{p}_\phi(y_w|v_0)} \right] \quad (3.8)$$

Since $q(v_{t-1}|v_t, v_0)$ has a Gaussian distribution, we parameterize the denoising distribution inside the Gaussian distribution family $p_\theta(v_{t-1}|v_t, x_e) = \mathcal{N}(v_{t-1}; \tilde{\mu}_\theta(v_t, x_e, t), \tilde{\gamma}_t \mathbf{I})$ where

$$\tilde{\mu}_\theta(v_t, x_e, t) = \sqrt{\frac{\bar{\beta}_{t-1}\gamma_t}{1-\bar{\beta}_t}} v_0^\theta(v_t, x_e, t) + \sqrt{\frac{\beta_t(1-\bar{\beta}_{t-1})}{1-\bar{\beta}_t}} v_t \quad (3.9)$$

$v_0^\theta(v_t, x_e, t)$ denotes the function that predicts the output representation at each iteration of the reverse pass. Under the Gaussian noise assumption, the objective can be expressed more concisely as:

$$\begin{aligned} \mathcal{L}_{\text{simple}} = \mathbb{E}_{q_\phi(v_0, x_e, y_w)} \left[\sum_{t=2}^T \mathbb{E}_{q(v_t|v_0)} \left\| v_0^\theta(v_t, x_e, t) - v_0 \right\|^2 \right. \\ \left. + \left\| \tilde{\mu}(v_T, v_0) \right\|^2 + \left\| v_0^\theta(v_1, x_e, 1) - f_\phi(y_w) \right\|^2 - \log \tilde{p}_\phi(y_w|v_0) \right] \quad (3.10) \end{aligned}$$

where, $q(v_t|v_0) = \mathcal{N}(v_t; \sqrt{\beta_t}v_0, (1-\beta_t)\mathbf{I})$ for efficient sampling of v_t during training, and $\mu_T(v_0) = \sqrt{\beta_T}v_0$ and the denoising function $v_0^\theta(v_t, x_e, t)$, which is modeled using a transformer network with separate components for encoding the input and generating the output. During training, the distribution used for drawing samples q_ϕ includes learnable parameters from token representation model. We utilize the reparameterization trick [103] to enable backpropagation through these parameters.

3.2.1.6 Inference

Given an input tweet d_q and the output from the retriever $\{(d_1, H_1, s(d_q, d_1)), \dots, (d_N, H_N, s(d_q, d_N))\}$, the selector aggregates retrieved hashtags into a set $\{h_1, \dots, h_M\}$, where M is the number of unique hashtags. For each hashtag h_m , the selector \mathcal{C} computes relevance score between the input tweet

and each unique hashtag.

$$r(d_q, h_m) = (C)(d_q, h_m) \quad (3.11)$$

Here, $r(d_q, h_m)$ denotes relevance score between the input tweet d_q and hashtag h_m , as computed by the selector \mathcal{C} . Lastly, we calculate the average similarity between tweets for each hashtag and incorporate the semantic relatedness between the tweet and the hashtag.

$$\rho_i = \left(\frac{1}{n_i} \sum_{j=1}^{n_i} s(d_q, d_j) \right) + r(d_q, h_i) \quad (3.12)$$

Here, n_i is the frequency of hashtag h_i in retrieved tweets, $s(d_q, d_j)$ denotes the similarity score between d_q and j^{th} retrieved tweet containing h_i , and h_i is obtained from the retriever. We then rank hashtags in descending order according to final scores ρ_i and select top-p hashtags $\{\tilde{h}_1, \dots, \tilde{h}_p\}$. Since we do not have ground-truth hashtags for the test tweet, only the reverse step of the diffusion process is performed. Starting from random noise v_T , the model iteratively denoises samples to obtain v_0 . This denoised output is then passed through the rounding distribution to obtain the predicted hashtag sequence:

$$\hat{y}_w = \hat{p}_\phi(y_w | v_0) \quad (3.13)$$

Finally, the decoder of BART model generates final hashtag recommendations based on predicted sequence \hat{y}_w and the input embedding x_e .

$$\text{Hashtags} = \text{BART}_{\text{dec}}(\hat{y}_w, x_e) \quad (3.14)$$

3.3 Experimental Evaluations

This section details the experimental configurations followed by a presentation and analysis of results obtained.

3.3.1 Experimental Setup

Table 3.1: Statistics of the dataset. h_t denotes number of hashtags per tweet.

#Tweets	# Hashtags	#Avg. h_t	#Max. h_t	#Min. h_t
26,665	12,230	2	23	1

3.3.1.1 Dataset

This study uses a disaster-related tweet dataset, originally presented by [93], to investigate hashtag recommendation. The dataset contains tweets during Harvey, Irma, Maria, Mexico earthquake, Chiapas earthquake, and California wildfire crawled using Twitter streaming API and tweets sourced from publicly available datasets [104, 105, 106]. To ensure data quality, [93] implemented a rigorous filtering process removing uninformative, non-English, and duplicate tweets that contained a total of 67,288 tweets spanning a total of 37 types of disasters. We further removed tweets with invalid links and taken down from X. The final dataset used in our study contains 26,665 tweets, 12,230 unique hashtags with an average of 2 hashtags per tweet as depicted in Table 3.1. The dataset and code for ASSIGNER has been made publicly available¹.

3.3.1.1.1 Disaster Type Distribution To evaluate potential biases in our dataset, we provide a detailed breakdown of the distribution of tweets across different disaster categories. Table 3.2 presents the number and percentage of tweets associated with each disaster type. As shown in Table 3.2, the dataset exhibits a diverse representation of disaster types, with floods being the most prevalent (29.64%), followed by hurricanes (16.44%) and earthquakes (10.80%). Conversely, disasters such as tornadoes (2.35%), typhoons (2.42%), and viruses (1.25%) are less represented. This imbalance reflects the frequency and visibility of disasters in social media discourse, where high-impact events such as floods and hurricanes naturally generate more engagement. While this skew mirrors real-world attention patterns, it may limit the model’s generalizability to less frequent disasters. Further, we discuss implications of this imbalance in the limitations section.

¹<https://github.com/abcd3007/ASSIGNER>

Table 3.2: Dataset distribution by number of tweets and percentage of total tweets for each disaster type.

Disaster	Number of Tweets	(in %)
Tornado	627	2.35
Hurricane	4386	16.44
Fire	1567	5.87
Earthquake	2882	10.8
Flood	7904	29.64
Haze	1830	6.86
Typhoon	645	2.42
Virus	332	1.25
MERS	735	2.76
Cyclone	2486	9.32
Tsunami	1861	6.98
Explosion	1410	5.56

3.3.1.2 Compared Methods

The performance of our proposed model is evaluated against extant hashtag recommendation methods. These include sequence generation models such as AMNN [94], SEGTRM [11], and HashTation [95]; keyphrase extraction methods such as LSTM-MTL [93]; retrieval-augmented generation methods such as RIGHT [107]; and diffusion models including Diffuseq [108] and SeqDiffuSeq [109].

- [1] AMNN [94] employed a seq2seq encoder-decoder architecture with CNN for visual and Bi-LSTM for textual feature extraction from multimodal microblogs. An attention mechanism identifies salient information, and a GRU-based decoder generates the hashtag sequence.
- [2] SEGTRM [11] proposed a model for microblog hashtag generation that operates in three phases. The encoder processes the input text using segment tokens and various attention mechanisms. A segment-selector block identifies important segments based on semantic similarity, while the decoder generates hashtags sequentially using selected segmental representations. The model efficiently determines the number of hashtags required and learns to generate hashtags based on post content.

- [3] HashTation [95] The authors propose a multi-component framework for hashtag recommendation and tweet classification, with four main modules namely, Hashtag Generator, Tweet Attention Module (TAM), Entity Attention Module (EAM), and Tweet Classifier. It begins with Hashtag Generator using self-attention mechanism to create hashtags of an input tweet. TAM is combined with a cross-document attention network to capture latent topics in relevant tweets within a collection and thus improve hashtag generation. EAM employs a graph attention network to extract and utilize semantic information at the entity level, thereby constructing an entity graph from named entities present in tweets. The Tweet Classifier then utilizes a transformer-based encoder equipped with a classification head to classify tweets.
- [4] LSTM-MTL [93] developed a joint-layer LSTM trained using Multi-Task Learning (LSTM-MTL) to recommend hashtags for disaster-related tweets. The authors incorporate features capturing informal writing and identify relevant hashtags based on disaster name, location, and situational awareness. The model’s variant, utilizing ELMo embeddings, Parts of Speech (POS) tags, and CNN-encoded phonetic features, achieves the best overall performance.
- [5] RIGHT [107] proposed a mainstream hashtag recommendation framework comprising a retriever, selector, and generator. The retriever employs sparse (BM25) and dense (SimCSE) retrieval techniques to identify relevant tweet-hashtag pairs. The selector utilizes a contrastive learning approach with three features (hashtag similarity, frequency, and positive/negative samples) to filter non-mainstream hashtags. The generator combines the input tweet with selected hashtags and employs a generative model fine-tuned with cross-entropy loss to recommend final set of hashtags, ranked by similarity and frequency.
- [6] Diffuseq [108] a diffusion-based model for conditional text generation adapted for seq2seq tasks. The model employs a partial noising strategy, injecting noise only into the target sequence during the forward process. In the reverse process, a transformer architecture predicts the mean and standard deviation of the

distribution at each step to denoise the target sequence. The training objective is designed as a variational lower bound with regularization terms. The model also utilizes importance sampling to address training inefficiencies. During inference, sequences are generated by sampling from a learned diffusion process and employing Minimum Bayes Risk (MBR) decoding for quality enhancement.

- [7] SeqDiffuseq [109] proposed a diffusion-based model for seq2seq language generation. In the forward process, the target output sequence is transformed into random noise. The reverse process utilizes a denoising function, conditioned on the input sequence, to iteratively reconstruct the sequence. The model employs an encoder-decoder transformer architecture and incorporates self-conditioning to improve text quality. It also features an adaptive noise schedule that adjusts the denoising difficulty at each token position and time step, enhancing generation performance.

3.3.1.3 Evaluation Metrics

To evaluate the quality and diversity of generated hashtags, we employed four metrics namely, BERTScore, dist.1, ROUGE-1, and BLEU. BERTScore [110] assesses the semantic similarity with ground-truth hashtags, ROUGE-1 [111] quantifies unigram overlap, BLEU [112] determines the precision of generated hashtag sequences, and distinct uni-gram (dist. 1) measures the diversity of words within generated sequences. Higher scores of dist. 1 indicate less repetition. We utilize the official ROUGE script² (version 0.3.1) for calculating ROUGE scores.

- **BERTScore:** BERTScore leverages pre-trained contextual embeddings from Bidirectional Encoder Representations from Transformers (BERT) [113] to assess the semantic similarity between generated and reference hashtag sequences. It computes a similarity score by comparing contextualized representations of

²<https://pypi.org/project/pyrouge/>

corresponding tokens in both sequences.

$$BERTScore(G, R) = \frac{1}{|G|} \sum_{g \in G} \max_{r \in R} \cos(c, r) \quad (3.15)$$

Here, G represents the generated hashtag sequence, R denotes the reference hashtag sequence, g and r are contextualized embeddings of each hashtag in G and R , respectively, and $\cos(g, r)$ denotes the cosine similarity between embeddings g and r .

- Distance-1 (dist. 1): This metric assesses the similarity between two sequences based on the minimum number of edits required to make them identical.

$$Distance - 1(G, R) = \text{minimum number of edits}(G \rightarrow R) \quad (3.16)$$

High values of Distance-1 implies high diversity.

- BLEU: Bilingual Evaluation Understudy (BLEU) measures the precision of n-gram matches between generated and reference hashtag sequences. It calculates the overlap of n-grams of varying lengths (typically 1 to 4) and combines them with a brevity penalty to discourage overly short generations.

$$BLEU(G, R) = BP \cdot \exp \left(\sum_{n=1}^N w_n \log p_n \right) \quad (3.17)$$

Here, BP represents the brevity penalty that penalizes generated sequences that are shorter than the reference sequence. The variable N denotes the maximum n-gram order considered in the calculation, typically set to 4. The n-gram precision accounting for length differences is represented by p_n . BP is calculated as follows:

$$BP = \begin{cases} 1, & \text{if } c > r \\ e^{1-(r/c)}, & \text{otherwise} \end{cases} \quad (3.18)$$

where, c and r represent the number of tokens in the generated and ground-truth

sequences, respectively.

- ROUGE-L (Longest Common Subsequence) measures the longest common subsequence of words between the generated and reference hashtag sequences. It focuses on recall, assessing the extent to which the generated sequence covers the reference sequence.

$$ROUGE - L(G, R) = \frac{LCS(G, R)}{|R|} \quad (3.19)$$

Here, $LCS(G, R)$ is the length of the longest common subsequence between sequences G and R . $|R|$ is the length of actual sequence i.e., ground-truth hashtags (R). We utilize ROUGE-1 to assess the similarity between the generated hashtag sequence and ground-truth sequence by measuring the overlap of unigrams. This metric is widely used for sequence generation tasks and can identify relevant hashtags even if they are not identical to the ground-truth, which is crucial in hashtag recommendation where multiple hashtags can contribute to conveying the overall topic. Additionally, we examine n-gram overlaps between generated and ground-truth hashtags to evaluate the model’s ability to identify and utilize salient information from text for hashtag generation.

3.3.1.4 Implementation Details

The proposed model, ASSIGNER was implemented using the PyTorch framework. It employs a 6-layered encoder-decoder transformer with Gaussian Error Linear Units (GeLU) activation functions. The model utilizes a diffusion-based approach with 200 iterative steps, where noise is incrementally added and then removed. We employed AdamW optimizer with a learning rate of 10e-4, incorporating a warm-up period of 500 steps followed by a linear decay. The model was trained for 15 epochs with a batch size of 64. The dataset was split into 75% for training, 15% for validation, and 10% for testing, with tweets truncated to a maximum length of 128 tokens. A threshold of 0.7 was used in the selector module. All hyperparameters were optimized based on the validation data. To ensure reproducibility, we set a random seed of 101.

Table 3.3: Effectiveness comparison results of ASSIGNER with existing methods for hashtag recommendation (top-2). The best result is highlighted in **bold**, while the second-best is underlined.

Methods	BERTScore	dist. 1	ROUGE-1	BLEU
<i>Sequence Generation</i>				
AMNN	0.359	<u>0.892</u>	0.001	0.002
SEGTRM	0.255	<u>0.777</u>	0.001	0.002
HashTation	0.246	0.514	0.001	0.001
<i>Keyphrase Extraction</i>				
LSTM-MTL	0.355	0.700	0.001	0.001
<i>Retrieval-Augmented Generation</i>				
RIGHT	<u>0.389</u>	0.877	0.003	0.001
<i>Diffusion</i>				
Diffuseq	0.344	0.799	0.014	0.012
SeqDiffuseq	0.235	0.522	0.003	0.001
ASSIGNER	0.458	0.987	<u>0.008</u>	<u>0.011</u>

3.3.2 Experimental Results

To analyze our proposed model, we conducted quantitative analysis, ablation studies, and qualitative case studies detailed below.

3.3.2.0.1 Quantitative Analysis Table 3.3 demonstrates that ASSIGNER significantly outperforms established methods across all assessment criteria. ASSIGNER outperforms AMNN by incorporating a retrieval mechanism to focus on relevant candidate hashtags and diffusion-based encoder-decoder architecture (BART) to capture linguistic characteristics of disaster-related tweets. Unlike AMNN, which relies on RNN-based encoder-decoder (BiLSTM-GRU) with softmax layer and produces generic hashtags, ASSIGNER leverages the strength of transformer and diffusion model in capturing complex data distributions, allowing it to grasp the dynamic nature of disaster-related language, leading to accurate hashtag recommendation. While SEGTRM uses segment selection to identify important parts of text, ASSIGNER’s retrieval and selector components provide a more focused set of candidate hashtags. Additionally, the diffusion-based generator in ASSIGNER generates diverse and creative hashtags compared to SEGTRM’s transformer decoder. ASSIGNER surpasses LSTM-MTL by moving beyond keyphrase extraction and utilizing a generation-based approach. This

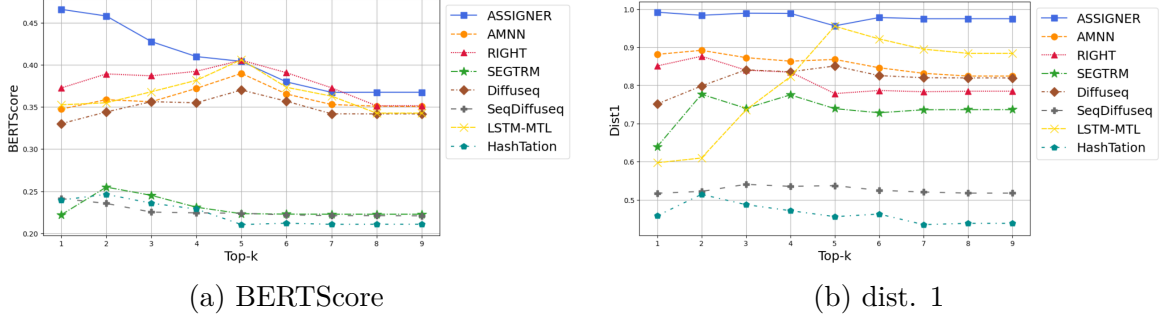


Figure 3.1: Effectiveness comparison curves. The proposed method significantly outperforms compared methods.

allows ASSIGNER to generate novel hashtags not limited to those present in text, unlike LSTM-MTL, which extracts keyphrases directly. While RIGHT incorporates a retrieval mechanism, ASSIGNER further enhances this with a diffusion-based generator and adaptive sigmoid noise scheduling. This allows ASSIGNER to generate diverse and relevant hashtags compared to RIGHT, which relies on a standard generative model. ASSIGNER builds upon DiffuSeq and SeqDiffuSeq to improve hashtag generation. It incorporates an encoder-decoder architecture (BART), self-conditioning, and adaptive sigmoid noise scheduling for enhanced efficiency and performance compared to DiffuSeq’s encoder-only framework. Additionally, ASSIGNER extends SeqDiffuSeq by adding a retrieval and selection mechanism to focus on relevant candidate hashtags besides adaptive sigmoid noise scheduling algorithm. This combined approach leads to more accurate and diverse hashtag generation.

3.3.2.1 Visualisation of Quantitative Results

To enhance the readability of our experimental results, Figure 3.1 presents a comparative analysis of hashtag recommendation models’ performance across varying recommendation set sizes ($top - h$), ranging from 1 to 9. As depicted in Figure 3.1(a), ASSIGNER consistently achieves the highest BERTScore values across all $top - h$ settings, indicating superior semantic similarity between its generated and ground-truth hashtags. The observed stability of ASSIGNER’s BERTScore, even with increasing recommendation set sizes, demonstrates its robustness. In contrast, AMNN, RIGHT, SeqDiffuseq, LSTM-MTL, HashTation, and SEGTRM, exhibit lower BERTScore val-

Table 3.4: Effect of individual component ablation on hashtag generation performance of ASSIGNER. The best result is highlighted in **bold**, while the second-best is underlined. Here, w/o refers to without.

Methods	BERTScore	dist. 1	ROUGE-1	BLEU
w/o Self-conditioning	0.145	0.310	0.003	0.003
w/o RAG	0.421	0.987	0.003	0.003
w/o Noise Scheduling	0.412	0.783	0.020	0.030
w/o Diffusion	0.409	0.844	0.002	0.002
w/o Selector	0.416	0.974	0.003	0.002
ASSIGNER	0.458	0.987	0.008	0.011

ues, highlighting their limitations in capturing semantic relevance. Figure 3.1(b) further illustrates ASSIGNER’s superior performance in terms of hashtag diversity, achieving dist. 1 scores close to 1.0 across all $top-h$ values. While AMNN and RIGHT show relatively better diversity compared to other state-of-the-art (SOTA) methods, their scores remain significantly lower than that of ASSIGNER. SeqDiffuseq, LSTM-MTL, HashTation, and SEGTRM exhibit substantially lower dist. 1 scores, indicating limited diversity. Thus, the graphical representation in Figure 3.1 visually confirms ASSIGNER’s competitive advantage in both semantic similarity and hashtag diversity. The consistent and significant gaps in both BERTScore and dist. 1 between ASSIGNER and existing methods underscore the efficacy of our proposed approach.

3.3.2.2 Ablation Studies

- **w/o Self-conditioning:** Our ablation study highlights the crucial role of self-conditioning in ASSIGNER. Removing it drastically reduces performance across all metrics (BERTScore: 0.4581 to 0.1446, dist.1: 0.9872 to 0.3096, ROUGE-1: 0.0078 to 0.0028, BLEU: 0.0113 to 0.0029) as evident from Table 3.4. This decline is attributed to information loss inherent in standard diffusion process where each denoising step relies solely on the current noisy input, neglecting refined information from previous predictions. Self-conditioning mitigates this by incorporating the previous prediction into the denoising function, allowing the model to refine its estimations and generate contextually relevant hashtags that capture the evolving event-specific language prevalent in disaster-related

tweets.

- **w/o RA:** To assess the impact of Retrieval Augmentation (RA) mechanism in ASSIGNER, we conducted an ablation study where RA was removed. We replaced the retriever (which selects relevant candidate hashtags), with a random selection of top-k hashtags from the training dataset. This modification resulted in a noticeable performance drop across all metrics (BERTScore: 0.458 to 0.421, ROUGE-1: 0.008 to 0.003, BLEU: 0.011 to 0.003) as evident in 3.4. This result highlights the crucial role of RA in providing the diffusion-based encoder-decoder framework with a focused set of relevant candidate hashtags. By leveraging information from similar tweets and their associated hashtags, RA effectively guides the generator towards more informative and accurate hashtag recommendations. These retrieved hashtags serve as guiding signals and a starting point for generating final hashtags, ultimately enhancing their quality and effectiveness.
- **w/o Noise Scheduling:** As shown in Table 3.4, removing the adaptive sigmoid noise scheduler from diffusion pipeline significantly hinders performance (BERTScore: drops from 0.458 to 0.412, dist. 1 from 0.9872 to 0.783). This underscores the importance of a well-designed noise scheduling for guiding the diffusion process towards meaningful outputs. By applying this scheduler at the token level, ASSIGNER achieves two key advantages namely, contextual adaptation, enabling the model to adjust noise based on each token’s specific context, crucial in dynamic disaster situations, and enhanced learning that captures complex inter-token dependencies to recommend pertinent hashtags. This precise control over noise introduction and reduction empowers effective learning and generation of contextually relevant hashtags, supporting information dissemination during critical events.
- **w/o Diffusion** These ablation results underscore the significant contribution of the diffusion-based encoder-decoder framework to ASSIGNER’s strong performance. Removing this component and replacing it with a standard encoder-

decoder framework leads to a substantial drop in performance across all metrics (BERTScore: 0.458 to 0.409, dist.1: 0.987 to 0.844, ROUGE-1: 0.008 to 0.002, BLEU: 0.011 to 0.002) as can be seen in Table 3.4. This decline is attributed to limitations of standard encoder-decoder models in capturing the complex and dynamic language characteristic of disaster-related tweets. These models tend to produce generic hashtag recommendations due to their reliance on maximizing training data likelihood. In contrast, the diffusion-based generator in ASSIGNER leverages a gradual noising process, enabling it to explore a wider range of possibilities and generate diverse and informative hashtags. This approach is well-suited for capturing the evolving and informal language used in disaster situations, where new terms and expressions may emerge rapidly, leading to improved performance.

- **w/o Selector** In this ablated variant, the selector is omitted and we directly choose top-p retrieved hashtags. Removing the selector leads to a significant decrease in performance across all metrics. BERTScore drops from 0.458 to 0.416, dist.1 from 0.987 to 0.974, ROUGE-1 from 0.008 to 0.003, and BLEU from 0.011 to 0.002, as can be seen in 3.4. This decline highlights that the selector plays a vital role in refining candidate hashtags identified by the retriever. By analyzing how closely the input tweet matches the retrieved tweet and hashtags in terms of meaning, the selector ensures that only the most relevant hashtags are passed to the diffusion-based generator. This filtering step is essential, as simply relying on top-p hashtags from the retriever, based on similarity to the input tweet, proves insufficient for generating accurate and informative hashtag recommendations. The selector thus acts as a quality control mechanism, guiding the generator towards optimal hashtag selection and improving overall performance.

3.3.2.3 Performance Comparison with Noise Scheduling Algorithms

To investigate the impact of different noise scheduling algorithms on performance of ASSIGNER in recommending hashtags for disaster-related tweets, we conducted

Table 3.5: Performance comparison of ASSIGNER with various noise scheduling algorithms for disaster-related hashtag recommendation, showing optimal results with the token-level adaptive sigmoid scheduler. The best result is highlighted in **bold**, while the second-best is underlined.

Noise Scheduler	BERTScore	dist. 1	ROUGE-1	BLEU
Gaussian	0.167	0.302	0.001	0.001
Adaptive Linear	0.414	0.846	0.007	0.010
Adaptive Quadratic	0.202	0.565	0.002	0.002
Adaptive Cubic	0.171	0.427	0.001	0.0004
Adaptive Fibonacci	0.383	0.733	0.004	0.002
Adaptive Cosine	0.210	0.405	0.005	0.006
Adaptive Exponential	0.310	0.602	0.012	0.014
Adaptive Sigmoid	0.458	0.987	0.008	0.011

experiments with various schedulers. As demonstrated in Table 3.5, the token-level adaptive sigmoid scheduler achieves the highest BERTScore (0.4581), dist. 1 (0.9872), ROUGE-1, and BLEU scores. This scheduler outperforms others, including token-level adaptive (Fibonacci, exponential) and non-adaptive Gaussian scheduler. The success of the token-level adaptive sigmoid scheduler is attributed to its precise and dynamic noise control, which is crucial for disaster-related hashtags where keyword relevance can fluctuate rapidly. The sigmoid curve introduces noise gradually to each token, promoting exploration of hashtag spaces, and then reduces it sharply for fine-grained refinement. This dynamic approach enhances contextual sensitivity by adjusting noise based on specific context of each token within the input tweet and generated sequence. Furthermore, the sigmoid curve, combined with token-level adaptation, allows the model to effectively learn intricate inter-token relationships, crucial for generating pertinent hashtags.

3.3.2.4 Qualitative Analysis

This section presents a qualitative evaluation of the performance of our proposed method i.e., ASSIGNER and contrasts it with well-established extant methods. Figure 3.2 presents a qualitative comparison of generated hashtags for an example tweet from the test dataset, alongside ground-truth hashtags and hashtags generated by various methods. As illustrated in Figure 3.2, ASSIGNER demonstrates a superior ability to suggest relevant hashtags for the given tweet. Notably, ASSIGNER is the

Input Tweet

Please help all NGOs and all the people who are going to Kerala by either giving food or water or supplies or money for the people in Kerala.

Ground-truth Hashtags: #KeralaFloodRelief #KeralaRains #KeralaReliefFund #KeralaFoodRescue

AMNN: #flood #medical #relief #NGO #Kerala #support #cyclone

LSTM-MTL: #KeralaFlood #NGO #help #aid #quake

RIGHT: #KeralaFlood #food #relief #help #aid #storm

SEGTRM: #flood #reach #help #KeralaSupport #aid #rescue #hurricane

DiffuSeq: #KeralaFloodRelief #rescue #supply #NGO #Kerala #protest

HashTation: #flood #food #aid #rescue #fire

SeqDiffuSeq: #KeralaFlood #cyclone #rescue #wildfire

ASSIGNER: #KeralaFloodRelief #help #KeralaRains #support

Figure 3.2: Example of a tweet from test dataset depicting hashtags recommended by various methods. Generated hashtags that match user-assigned hashtags are marked with green, while relevant but non-matching hashtags are marked with blue, and irrelevant predictions are marked with red.

only model that correctly identifies two ground-truth hashtags: #KeralaFloodRelief and #KeralaRains. While DiffuSeq also generates #KeralaFloodRelief, ASSIGNER is unique in its ability to simultaneously identify both of these crucial hashtags. Moreover, ASSIGNER effectively captures general terms such as #flood, #medical, and #relief and specific terms such as #Kerala and #NGO, which are relevant to the given tweet. In contrast, other methods exhibit varying degrees of success, but none achieve the same level of accuracy as ASSIGNER. HashTation primarily focuses on generic terms such as #flood, #food, and #aid, lacking the specificity of ASSIGNER. SeqDiffuSeq incorrectly predicts hashtags such as #cyclone and #wildfire, which are not relevant to Kerala floods. AMNN and LSTM-MTL incorrectly predict hashtags #cyclone #quake, respectively. These errors highlight the difficulty other methods have in capturing the specific context of the input tweet. The enhanced performance of ASSIGNER can be attributed to the effective integration of retrieval augmentation with a diffusion-based encoder-decoder framework. The retriever module pro-

vides valuable context to the diffusion-based generator by identifying similar tweets and their corresponding hashtags. The selector module further refines these retrieved hashtags, ensuring that only the most relevant candidates are passed to the generator. Furthermore, ASSIGNER’s self-conditioning mechanism and adaptive sigmoid noise scheduler contribute to generating high-quality hashtag sequences by exploring a broader spectrum of possibilities. Overall, the qualitative analysis demonstrates ASSIGNER’s ability in leveraging existing knowledge, capturing contextual information, and generating diverse hashtag sequences, significantly outperforming other methods.

3.4 Conclusion

This chapter introduces ASSIGNER, a novel retrieval-augmented encoder-decoder with diffusion for sequential hashtag recommendation in disaster events. ASSIGNER extends continuous text diffusion model to generate hashtags sequentially and a retrieval mechanism that leverages existing knowledge from semantically similar tweets and hashtags. This approach addresses the limitations of existing methods by capturing both semantic relationships among hashtags and contextual information embedded in tweets. Furthermore, a novel adaptive sigmoid noise scheduler is proposed to improve the quality of generated hashtags. Experimental results validate the capability of ASSIGNER in generating relevant and informative hashtags for disaster-related tweets, with the potential to improve information dissemination and response efforts during crises.

Chapter 4

Hashtag Recommendation for Multilingual Content

4.1 Introduction

The pervasive influence of platforms such as X on contemporary discourse is evident in the rapid cross-lingual and cross-geographical dissemination of information. The increasing engagement with regionally specific content, particularly in nations with substantial low-resource language user bases such as India, underscores this trend. The platform’s support for vernacular languages has empowered users to express themselves in their native tongues, leading to a transformation in content dissemination and reach. While the platform’s support for vernacular languages facilitates broader expression, connecting semantically related multilingual content remains challenging due to linguistic variations.

Hashtags offer a potential solution to bridge these linguistic divides by serving as common semantic anchors. However, the infrequent or suboptimal use of hashtags, particularly in low-resource language content, limits their efficacy in cross-lingual content discovery. Furthermore, the sheer volume of event-related posts on SNS platforms can overwhelm users seeking relevant information. This challenge is amplified for non-English content, where the scarcity of relevant hashtags makes it difficult to filter and retrieve pertinent discussions. This information gap impacts various stakeholders, including local content creators seeking broader reach, brands aiming to engage with

regional audiences, language learners looking for authentic content, and researchers studying multilingualism and language contact. An effective automated multilingual hashtag recommendation system is therefore crucial for enhancing content discoverability and fostering connections across linguistic communities. An analysis of our curated dataset of low-resource Indic language posts revealed that a significant proportion of posts upto 24.16% contain fewer than two hashtags, underscoring the pressing need for developing such a system.

Prior research has explored hashtag recommendation for various content modalities, including textual [11, 28, 29], visual [114, 115, 116], and multimodal content [117, 118, 119, 120]. Some studies have focused on personalized hashtag suggestions by incorporating content, user characteristics, and metadata [18, 72]. Despite considerable attention to text-based hashtag recommendation, the primary focus has been on high-resource languages such as English [5, 9] and Chinese [10, 11, 121]. Recommending hashtags for content in low-resource Indic languages on SNS remains largely underexplored. Indic languages are categorized as low-resource due to the limited availability of written texts, audio recordings, and other digital resources, resulting in noisy or incomplete datasets. The direct application of existing high-resource language hashtag recommendation methods to low-resource scenarios is challenging due to the specialized linguistic knowledge or native speaker expertise required for these languages.

For instance, Zhang et al. [5] utilized a parallel co-attention mechanism to model interplay between visual and textual modalities of a post. The authors also considered the similarity between the current post and a user’s past posts to infer their tagging behavior and suggest relevant hashtags. However, relying on historical post similarity might overlook evolving user interests or changes in posting habits, potentially leading to suboptimal recommendations. Jeong et al. [71] proposed a hashtag recommendation approach based on post content and user demographics, computing the similarity between these feature sets. A potential limitation of solely relying on demographic data is the failure to capture individual user preferences that may deviate from broader demographic trends. These unique, individual patterns in user behavior can be cru-

Tweet	Hashtags
'आप सभी को फूल देइ पर्व की हार्दिक शुभकामनाएं। 🌻🌻🌻🌻 सुख और समृद्धि से आपका जीवन संपन्न रहे। 🙏😊	#फूलदेई, #phooldei, #uttarakhand, #nature, #flowers
'आने वाला पल, जाने वाला है। हो सके तो इस में ज़िंदगी बिता दो, पल जो ये जाने वाला है। 🌸❤️	#present, #moment, #nature, #gratitude, #flowers, #bees, #life, #songs, #lovelive

Figure 4.1: Tweets of a user

cial for accurate recommendation and can reveal insights beyond what demographic or profile data alone can provide. Zhang et al. [50] constructed a bipartite graph of tweets and users to identify socially similar tweets for multilingual hashtag prediction. Nevertheless, TwHIN-BERT does not adequately account for users' specific interests and language usage patterns. The content creator's context, including their interests, preferences, expertise, language choice, and communication style, offers valuable information about the post. An illustrative example from X is seen in Fig. 4.1 where a user employs similar hashtags across thematically distinct tweets, revealing his underlying interests. In the first tweet, the user wishes Happy Flowers Day and annotates it with #phooldei. Phooldei is a festival of flowers and springtime celebrated in Uttarakhand. According to tweet content, he assigned #flowers, #Uttarakhand and #nature. In the second tweet, he emphasizes living in the present through lines of a Hindi Bollywood song. According to the tweet content, he annotates #present, #moment, and #songs to his tweet. The tweet has no relation with flowers, yet he assigns #flowers and #nature to the second tweet, reflecting his interest in topics, i.e., nature and flowers. Mining information from a user's posting history can therefore provide insights into their personal preferences and posting patterns, leading to a richer understanding of their content engagement. Furthermore, users often develop idiosyncratic language patterns on X, characterized by their unique vocabulary, punctuation, and emoji usage, influenced by their personality and communication style. Capturing these highly individual linguistic traits is essential for a comprehensive understanding of language

use variations among users.

Additionally, the user from Fig. 4.1 demonstrates a tendency to recommend hashtags in the same language as the post (Hindi). He transliterates `#फूलदेई`, to its English equivalent `#phooldei`. This emphasizes that user tends to take language into consideration when posting tweets and annotating hashtags. In contrast, TwHIN-BERT does not explicitly model users’ linguistic preferences or the relationships between languages. Language relatedness, referring to the similarities between languages in terms of grammar and vocabulary, can be particularly useful in low-resource settings by leveraging shared linguistic features. Modeling these relationships within language families can help overcome data scarcity by leveraging shared knowledge and resources. This approach is particularly valuable in multilingual settings, where users speak multiple languages within the same language family.

In this chapter, we devise an automatic **hashtag** recommendation system for orphan tweets in **low-resource** Indic languages dubbed as TAGALOG. It leverages tweet content, language relatedness, and user preferences to suggest topic-relevant, personalized and language-aware hashtags. We refine tweet representations using language-guided and user-guided attention mechanisms to capture language usage style and user interests. We employ a graph neural network to model the relatedness between languages from different families namely, Indo-Aryan and Dravidian and user posting behavior. The recommended hashtags effectively identify the core content across languages, aiding regional language users in retrieving relevant information and staying informed.

Below are the key highlights of our contributions.

- [1] We devise a deep learning-based graph neural network to suggest semantically related, personalized, and language-specific hashtags for tweets posted in low-resource Indic languages.
- [2] We not only capture the distinct topical and linguistic inclinations of individual users on a local scale but also their long-term behavior and global interests.
- [3] On a local scale, we refine the content of tweets by devising a novel way of

attending to users' topical interests and language usage style.

- [4] Globally, we construct a graph to model users' interactions with tweets by considering their historical tweets and capturing the long-term posting behavior.
- [5] We also leverage relatedness among languages belonging to the same language family. The framework can mine correlation among languages of the same family group, i.e., Indo-Aryan and Dravidian.
- [6] We have constructed a new text-based hashtag recommendation dataset containing tweets in Indic languages called Indic Hash. The collected tweet samples span various low-resource languages: Bangla, Marathi, Gujarati, Telugu, Tamil, Kannada, and Hindi besides English. Our curated dataset can be a primary resource to recommend hashtags for tweets posted in Indic regional languages.
- [7] Our experimental findings show that the proposed hashtag recommendation model performs well in a low-resource environment with a minimal amount of labeled data.

The remainder of this chapter is organized as follows. We define the problem in Section 4.2. Section 4.3 elucidates the novel approach to recommend hashtags for multilingual content. Subsequently, Section 4.4 describes the experimental framework and offers a thorough analysis of the findings. Lastly, Section 4.5 summarizes this work and presents concluding thoughts.

4.2 Problem Definition

Let us consider a dataset with a tweet set $T = \{t_i\}_{i=1}^{|T|}$, a set of users $U = \{u_j\}_{j=1}^{|U|}$, a set of hashtags $H = \{h_k\}_{k=1}^{|H|}$ and a set of languages $L = \{IA(Hindi, Gujarati, Marathi, Bangla), D(Kannada, Tamil, Telugu), English\}$. Here, $|T|, |U|, |H|$ denotes the cardinality of the tweet set, user set, and hashtag set. *IA* and *D* refer to Indo-Aryan and Dravidian family groups.

Given a user $u \in U$ who uploads a tweet t written in language $l \in L$, we aim to recommend a personalized and language-specific set of hashtags $RH \subset H$ that are relevant to users' posting and language usage behavior.

Our objective is to develop a customized hashtag recommendation model for tweets in low-resource Indic languages that can automatically recommend hashtags from H to a new tweet t uploaded by a user u .

Given a tweet written in l by a user u , we intend to learn a function $f(.)$ that can capture his topical and linguistic preferences.

$$t_u, t_l = f(UGA(t, u), LGA(t, l)) \quad (4.1)$$

Here, UGA refers to the user-guided attention and LGA refers to the language-guided attention mechanisms that yield latent user and language representations denoted by t_u and t_l . Hashtags are a potent tool for self-expression because they allow users to succinctly and rapidly communicate their interests, thoughts, feelings, and views on a certain topic. To address the variances in hashtag labels that result from how individuals express themselves and their unique language usage style, we devise two attention mechanisms to fine-tune user and language representations. To further enhance tweet representation, we aim to learn a function $g(.)$ to model various types of interactions.

$$t'_u, t' = g(t_u, t) \quad (4.2)$$

Here, t'_u, t' denote the enhanced user and tweet representation derived from the graph, and $g(.)$ resembles a graph neural network. We employ a graph neural network to model tweet-tweet interactions based on language relatedness and user-tweet interactions. We construct a heterogeneous graph $G = (V, E)$ such that $V = (U, T)$ where V is the set of nodes comprising users and tweets, and E is the set of edges. Each edge $e \in E$ is based on either the relatedness of the language in which the tweet is written with tweets published in other languages within the same language group or whether the user created that tweet in the past. Hashtag recommendations can then

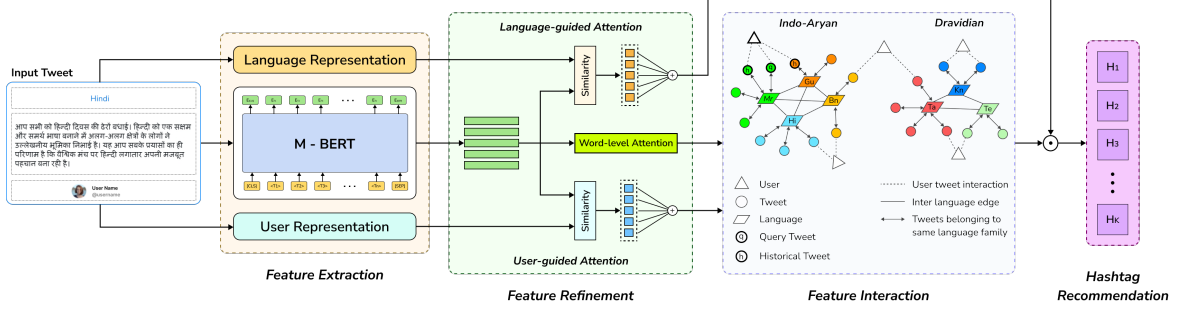


Figure 4.2: Overall architecture of TAGALOG

be formulated as given in Equation 4.3.

$$RH = HASH - REC(t'_u, t_l) \quad (4.3)$$

Here, $HASH - REC$ refers to the hashtag recommender that resembles a deep neural network. It takes enhanced tweet representation derived from the graph denoted by t'_u and language-guided tweet representation i.e., t_l to recommend a reasonable collection of hashtags denoted by RH . We posit that TAGALOG encodes not only the user’s topical and linguistic preferences but also relatedness among languages of a family group pertaining to the language in which a tweet is written. The following sections provide more information on the UGA , LGA , $f(\cdot)$, $g(\cdot)$, and $HASH - REC$.

4.3 Methodology

In this section, we present a detailed overview of our proposed approach. Fig. 4.2 showcases the overview of our innovative polyglot hashtag recommender. We propose a deep neural network based on graphs to recommend hashtags for tweets posted in multiple Indic languages. Our system receives a tweet as input, together with information on the language used in the tweet and the user who posted it. The proposed system first retrieves features from a tweet’s textual modality to obtain its low-dimensional feature vector representation. Then we use attention techniques to mimic how language and user affect the representation of a tweet. We create a graph to capture the correlation between tweets and the interaction between tweets and users.

The node embeddings which are modified in response to information dissemination and neighborhood aggregation are fed into the hashtag recommendation module. After assessing the plausibility of each hashtag, this module yields a sorted list of hashtags for polyglot tweets. As demonstrated in Fig. 4.2, our proposed framework comprises four components: (a) feature extraction; (b) feature refinement; (c) feature interaction, and (d) hashtag recommendation. Each component is discussed in profundity below.

4.3.1 Feature Extraction

In this section, we elucidate the textual, linguistic, and user feature retrieval from tweets.

4.3.1.0.1 Textual Feature Retrieval We encode tweets written in various resource-scarce Indic languages using Multilingual Bidirectional Encoder Representations from Transformers [122], abbreviated as the mBERT model. Wikipedia articles written in 104 different languages serve as the training data for the multilingual variant of BERT. Since mBERT shares a common input space at the sub-word level, this pre-trained neural language model is utilized to generate context-aware embeddings of tweets posted in different languages. The input tweet t is enclosed within two special tokens, class (CLS) and separator (SEP) to signal its start and endpoints. We pass the raw tweet through mBERT’s tokenizer to produce the corresponding set of tokens as shown in Equation 4.4.

$$M = mBERT_Tokenizer([CLS] + t + [SEP]) \quad (4.4)$$

Here, M represents the created collection of tokens. The number of tokens in the sequence denoted by S is capped at 50. We shorten or lengthen the token sequence derived from the tweet to S if it is greater or lesser than S to construct a uniform-sized token sequence for all tweets. Then, we encode tokens using an mBERT encoder to

generate token representations according to Equation 4.5.

$$T_f = mBERT(M) \quad (4.5)$$

The derived textual feature matrix is denoted by $T_f \in \mathbb{R}^{S \times D}$, where $S = 50$ denotes the number of tokens derived from the tweet, and $D = 768$ denotes the embedding size for every token. The textual feature matrix of the encoded tweet is passed to the feature refinement module.

4.3.1.0.2 Language Feature Retrieval Social media language is often informal, abbreviated, and contains hashtags, emojis, and other elements that are specific to these platforms. By learning language embeddings from a large corpus of social media data, we can better capture these unique linguistic characteristics and represent them in a way that captures their meaning. Language embeddings are vector representations of words or phrases that are learned through training on large amounts of text data. It consists of two steps namely language identification, and language embedding generation.

4.3.1.0.2.1 Language Identification We used the `langdetect`¹ library to identify the language in which tweet t is published. About 50 languages can be recognized by this package, which is a direct transfer of Google’s language-detection library from Java to Python. Nakatani Shuyo created the software at Cybozu Laboratories, Inc. We determine the language used to write the tweet t as depicted in Equation 4.6.

$$l = \text{langdetect}(t) \quad (4.6)$$

Here, l is the language identified for tweet t .

4.3.1.0.2.2 Language Embedding Generation Language embeddings are used for tweet representation because they enable us to capture the meaning and

¹<https://pypi.org/project/langdetect/>

context of words used in tweets. They capture the semantic and syntactic relationships between words, which allows us to understand the meaning of individual words and the overall context. Using language embeddings to represent tweets allows us to capture the nuances of language used on social media platforms. After identifying the language in which the tweet was written, we generate the feature vector for language using the Keras embedding layer² as discussed in Equation below.

$$l_f = \text{Embedding}(l) \quad (4.7)$$

Here, $l_f \in \mathbb{R}^D$ refers to a feature vector to represent language, with a dimensionality (D) of 768.

4.3.1.0.3 User Feature Retrieval User embeddings can be useful in deriving post features because they capture information about the users who created the posts. In many cases, the user who creates a post can provide important contextual information about the post, such as the user’s interests, preferences, or expertise. By incorporating this information into post features, models can improve their ability to understand and analyze posts. This can help the model make personalized recommendations that are more relevant to the user’s interests. The publisher of the tweet t is expressed as u . We encode u into a low-dimensional embedding vector (u_f) by employing the Keras embedding layer as demonstrated by the following Equation.

$$u_f = \text{Embedding}(u) \quad (4.8)$$

Here, $u_f \in \mathbb{R}^D$ refers to a feature vector to represent the user, with a dimensionality of 768. Users’ hidden features, such as preferences, may theoretically be captured by user embeddings and used to direct how the tweet representation is learned.

²https://keras.io/api/layers/core_layers/embedding/

4.3.2 Feature Refinement

The cornerstones of the feature refinement module comprising our proposed model are language-guided and user-guided attention mechanisms that successfully capture the topical and linguistic inclinations of individual users at a local level to enrich the tweet representation. We discuss these two mechanisms below.

4.3.2.1 Language-guided Attention Mechanism

We devise a novel language-specific attention block that selectively attends to language-oriented information in the tweet and filters out unnecessary information thus, enriching its representation. For the tweet embedding obtained using the mBERT encoder, we denote it as $T_f = \{e^s\}_{s=1}^S$. We use an attention technique to identify key terms, then aggregate the acquired word representations to create a comprehensive representation of the tweet’s textual content with respect to the linguistic preferences of the user. To this end, we feed the token-based embedding matrix T_f through a dense layer to create its hidden representation, as illustrated in the equation below.

$$h^l = \tanh(T_f W_l + b_l) \quad (4.9)$$

Here, $h^l = \{h_s^l\}_{s=1}^S$, where h_s^l is the hidden representation of e^s . We then determine how closely the token’s latent representation (h_s^l) resembles the language embedding vector (l_f) and run the outcome through a softmax algorithm to generate attention scores (α^s) using the formula presented in Equation 4.10.

$$\alpha = \text{softmax}(h^l l_f) \quad (4.10)$$

Here, $\alpha = \{\alpha^s\}_{s=1}^S$, where α^s designates a word’s significance with respect to language. The language-guided tweet representation is then derived by computing the weighted sum of token embeddings with attention scores α^s serving as weights as presented below.

$$t_l = \sum_{s=1}^S \alpha^s h_s^l \quad (4.11)$$

Here, t_l represents the language-guided tweet representation.

4.3.2.2 User-guided Attention Mechanism

Users tend to express their interest in the semantic attributes of a tweet's text. Thus, exploring users' attention to words appearing in tweets towards recommending hashtags is crucial. By using user-guided attention, the model can capture the user's unique perspectives, which can provide additional context and improve the accuracy of post features. We utilize a user-guided attention mechanism for identifying salient words and combining their corresponding representations to obtain a comprehensive representation of the tweet's textual content with respect to the user. To achieve this, we first process the mBERT-based token embedding matrix (T_f) using MultiLayer Perceptron (MLP) to derive h^u as illustrated in the subsequent equation.

$$h^u = \tanh(T_f W_u + b_u) \quad (4.12)$$

Here, $h^u = \{h_s^u\}_{s=1}^S$, where h_s^u serves as the covert way of representing e^s . We first calculate how comparable h_s^u and u_f are, then run the resulting through a softmax function to produce normalized weight β^s as demonstrated below.

$$\beta = \text{softmax}(h^u u_f) \quad (4.13)$$

Here, $\beta = \{\beta^s\}_{s=1}^S$, where β^s signifies the relevance of a term with respect to a user. The user-guided tweet representation is determined by summing the weighted word annotations i.e., β^s . as shown.

$$t_u = \sum_{s=1}^S \beta^s h_s^l \quad (4.14)$$

Here, t_u denotes the user-guided tweet representation. The obtained representations are forwarded to the feature interaction component.

4.3.3 Feature Interaction

The feature interaction module employs a graph neural network to capture global interests by analyzing long-term user behavior and preferences, in addition to tweet correlation. It comprises two major stages namely, graph construction and feature encoding. We discuss these two stages in detail below.

4.3.3.1 Graph Construction

Algorithm 4.1 Graph Construction

```
Input:       $T$ : Tweets  
              $U$ : Users  
Output:     $G(V, E)$ : User Tweet Graph  
function get_graph( $T, U$ )  
1:  $V = T \cup U$   
2:  $E = []$   
3: for  $(t1, t2) \in T \times T$  do  
4:    $sim\_score = cos\_sim(t1, t2)$   
5:   if  $langdetect(t1)$  and  $(langdetect(t2) = 'bn'$  or  $langdetect(t2) = 'hi'$  or  
      $langdetect(t2) = 'mr'$  or  $langdetect(t2) = 'gu')$  then  
6:      $E = E \cup (t1, t2, sim\_score)$   
7:   else if  $langdetect(t1)$  and  $(langdetect(t2) = 'kn'$  or  $langdetect(t2) = 'te'$  or  
      $langdetect(t2) = 'ta')$  then  
8:      $E = E \cup (t1, t2, sim\_score)$   
9:   else if  $langdetect(t1) = 'en'$  then  
10:     $E = E \cup (t1, t2, sim\_score)$   
11:   end if  
12: end for  
13: for  $t \in T$  do  
14:    $u = get\_user(t)$   
15:    $E = E \cup (t, u, 1)$   
16: end for  
17:  $G = (V, E)$   
18: return  $G$ 
```

To mine the correlation between tweets and the interaction between tweets and users, we create an undirected heterogeneous graph as illustrated in Algorithm 4.1. Here, $G = (V, E)$ is the resultant user-tweet graph, and V and E denote the collection of vertices and edges between them, respectively. We construct a graph with two different kinds of nodes, as shown in Line 1 of Algorithm 4.1. The total number of nodes

in the graph is I where $I = |T| + |U|$ and $E \subset V \times V$ is a set of relationships among nodes to model tweet-tweet correlations and user-tweet interactions. The edges constructed based on tweet-tweet correlations are weighted, whereas those corresponding to user-tweet interactions are unweighted. First, we compute the pairwise similarity between tweets appearing in the tweet set T , as depicted in Line 4. We then assign an edge between tweets of related language families corresponding to the language in which the tweet under consideration is written, as shown in Lines 5-8, corresponding to the Indo-Aryan and Dravidian family groups. The tweets not falling under these two groups imply they are written in English, as shown in Lines 9-10. The edge weight is the similarity score between mBERT-based embeddings of a tweet with tweets written in related languages comprising the language group. Grouping posts concerning their language family, like Indo-Aryan and Dravidian, can help in recommendations by personalizing content and recommendations based on the user’s linguistic and cultural background. Language families are a collection of languages that share the same ancestor. Languages in the same family often share similar grammatical structures, vocabulary, and cultural contexts. By grouping posts based on a language family, we identify posts that are likely to be relevant and exciting to users with a particular linguistic background. For example, suppose a user writes tweets in a language from the Indo-Aryan family. In that case, we can group posts that are written in languages from this family, such as Bangla (Bn), Hindi (Hi), Marathi (Mr), and Gujarati (Gu), and recommend hashtags to the user. Similarly, suppose a user uses a language from the Dravidian family. In that case, we can group posts that are written in languages from this family, such as Kannada (Kn), Telugu (Te), and Tamil (Ta), and recommend them to the user. By personalizing recommendations in this way, we can increase the relevance and engagement of content for users. Furthermore, as depicted in Lines 13-16, for every tweet, we retrieve its corresponding user. We then create an edge to connect the user to his uploaded tweets. By capturing the user-tweet relationship through edge creation, tweet representations can be enriched with the contextual information of the associated user, such as the user’s topical interests and historical posting patterns. Incorporating the user context allows for more contextualized and person-

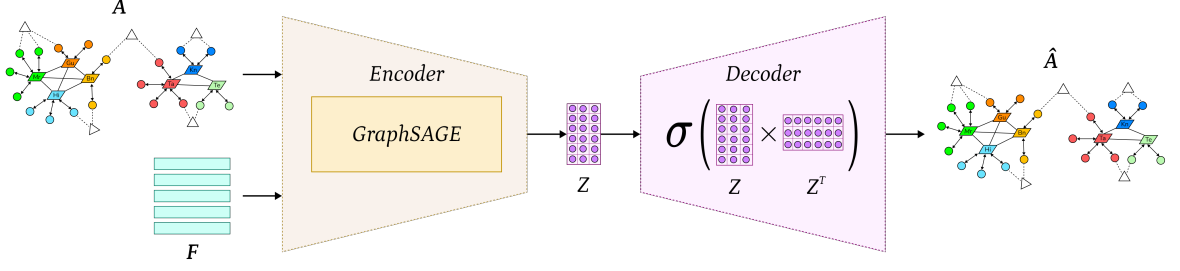


Figure 4.3: Graph AutoEncoder

alized tweet representations. It considers the relationship between the user and his tweets, allowing for a more nuanced understanding of their behavior and motivations. Unlike similarity-based analysis [5] that overlook the unique context and significance of individual posts, treating them as isolated entities, the edge-based approach explicitly models the relationship between a user and his tweets within the graph structure, thus enabling a comprehensive analysis of interdependencies and interactions between users and their tweeted content. The edge connecting a user to their tweets indicates the range and diversity of their topical interests. We utilize this edge information to identify patterns and recommend accurate hashtags.

4.3.3.2 Graph Feature Encoding

Our primary goal is to create and train a model to learn tweet and user embeddings given an input graph G in order to perform hashtag recommendations. GAE is a type of unsupervised learning model used for graph representation learning. GAE can capture complex, non-linear relationships between nodes in a graph, which cannot be easily captured by traditional graph embedding techniques such as DeepWalk [123] or node2vec [124]. GAE preserves the structural properties of nodes even when the data is noisy. GAE can be used for hashtag recommendation, where the input data consists of both user-tweet interaction data and tweet features represented as a graph. This allows for a more comprehensive recommendation system that takes into account both user behavior and tweet attributes. The proposed GAE pipeline is shown in Fig. 4.3. Let $G = (V, E)$ represent a graph with N nodes and A be its adjacency matrix. Let F be the feature matrix with N rows, where each row represents the feature vector of a

vertex. The goal of GAE is to acquire a reduced-dimensional latent representation Z that encompasses the structural and semantic information of the graph. The adjacency and feature matrices, when combined (AF), are the encoder's input. Graph Sample and Aggregate (GraphSAGE) [125] can be used as the encoder in the GAE by adapting it to aggregate information from the entire graph. GraphSAGE is a neural network that is designed to learn node embeddings by compiling information from its immediate surroundings. The input for the GraphSAGE encoder is F_v which is the feature vector that node v is initialized with, and $N(v)$ is the set of neighboring nodes of node v in the graph. The tweet node is initialized by employing word level attention [126] over the textual feature matrix of tweet t since tweets contain noisy user-generated text. User nodes are initialized with a feature vector obtained as depicted in Section 4.3.2.2. Generally, h_v^k is the embedding vector of node v at the k^{th} layer of the GraphSAGE encoder and N_L is the number of layers in the encoder. We adopt the mean aggregator in GraphSAGE as evident in Equation 4.15.

$$h_v^k = GraphSAGE_{mean}(h_v^{k-1}, A) \quad \forall k \in [1, N_L] \quad (4.15)$$

The updated feature matrix Z is obtained from the last layer as shown in Equation 4.16.

$$Z = h_v^{N_L} \quad (4.16)$$

Here, Z consists of the updated user representation (t'_u) and text feature (t'). The decoder maps this latent representation back to the original graph structure. It consists of a sigmoid activation function as shown in Equation 4.17.

$$\hat{A} = sigmoid(Z.Z^T) \quad (4.17)$$

Here, \hat{A} is the reconstructed adjacency matrix.

4.3.4 Hashtag Recommendation

By considering both the user and the language used in a tweet, we can better capture the user’s intent, perspective, language usage style, and the meaning of the words they use. To this end, we derive the overall tweet representation by concatenating the updated tweet embedding obtained from GAE and language-guided tweet representation as shown below.

$$t_f = \text{concat}(t'_u, t_l) \quad (4.18)$$

Here, t_f is the overall tweet representation. The hashtag recommendation module receives t_f as input and outputs a reasonable set of hashtags Rh as given in Equation 4.19.

$$Rh = \text{HASH} - \text{REC}(t_f) \quad (4.19)$$

The hashtag suggestion task is structured as a multilabel classification problem. Given that a tweet can belong to numerous classes simultaneously, this formulation procedure can assist in forecasting labels for non-exclusive classes. A pool of preconfigured hashtags H is employed to assign suitable hashtags to the multilingual tweet as exhibited in Equation 4.20.

$$y_{pred} = \text{softmax}(\text{Dense}(\text{units} = |H|)(t_f)) \quad (4.20)$$

Here, the symbol $y_{pred} \in \mathbb{R}^{|H|}$ refers to the softmax probabilities of the supplied hashtags, $|H|$ is the cardinality of the set of hashtags. These probabilities are used to rank hashtags and generate the final set of predicted hashtags (RH).

$$RH = \text{argsort}(y_{pred}) \quad (4.21)$$

The objective loss function for training TAGALOG can be seen in Equation 4.22.

$$L = L_{GAE} + L_{HR} \quad (4.22)$$

Here, L is the overall loss function, L_{GAE} is the reconstruction loss of GAE, and L_{HR} is the loss function for the hashtag recommendation module. The loss function (L_{GAE}) is described in Equation 4.23.

$$L_{GAE} = ||A - \hat{A}||^2 \quad (4.23)$$

Here, A and \hat{A} represent the actual and reconstructed adjacency matrices, and $||\cdot||$ denotes the squared norm. The objective of L_{GAE} is to reduce the difference between the predicted and actual adjacency matrices across the entire training dataset, with the purpose of achieving better reconstruction accuracy. The optimization problem is solved by minimizing L_{GAE} with respect to the parameters of the encoder and decoder (θ_e and θ_d) using a gradient-based optimization algorithm. Through this process, GAE learns a compressed representation of the input graph. The training loss function for the hashtag recommendation module is described in Equation 4.24.

$$L_{HR} = \frac{1}{|M|} \sum_{(t,G) \in M} \sum_{g \in G} -\log(P(g|t)) \quad (4.24)$$

Here, the current tweet is represented by t , the related ground-truth hashtag set is indicated by G , and the softmax probability that the ground-truth hashtag g will be used for the tweet t is given by $P(g|t)$, and variable M represents the training set of multilingual tweets.

4.4 Experimental Evaluations

In the ensuing subsections, we go over the experimental settings followed by experimental findings to validate the viability of our proposed framework.

4.4.1 Experimental Setup

Here, we present our curated dataset on which experiments were performed. Next, we go into state-of-the-art approaches and existing models for comparison, followed

by the criteria employed for evaluation.

4.4.1.1 Dataset

In our opinion, we have curated the first large-scale multilingual low-resource Indic tweets dataset dubbed as IndicHash. This dataset is designed for the task of recommending hashtags for tweets posted in multiple low-resource Indic languages. We create an exhaustive dataset from tweets published by Indian users covering seven low-resource languages besides English. Regional language tweets have increased significantly on Twitter. This served as our inspiration to broaden the endeavor to Indic languages. We chose a total of seven different Indic languages namely Bangla, Hindi, Kannada, Gujarati, Tamil, Telugu, and Kannada. This decision was primarily motivated by the widespread usage of these Indic languages across various regions of India. We now elucidate the techniques used to gather and process the independent tweets followed by a description of the dataset’s specifications.

4.4.1.1.1 Data Collection We gathered nearly equal numbers of posts for each keyword and a similar amount of keywords for each category. We first curated a generic list of categories namely technology, business, education, environment, gadgets, sports, festivals, people’s movement, politics, cricket, entertainment, movies, music, news, culture, food, military, career, fashion, fitness, gaming, nature, weather, emotions, pets, hobbies, astrology, and crisis. The total number of keywords considered for data collection is 213. For example, keywords under the education category: *education*, *ed-tech*, *ParikshaPeCharcha*, *teacher*, *learning*, *school*, *university*, *neweducationpoilcy*, *students*, and *exams*. Likewise, under the category of people’s movements which is a hot topic on Twitter, we included keywords such as *StudentLivesMatter*, *ShaheenBagh*, *FarmersProtest*, *KisaanAndolan*, *metoo*, *BlackLivesMatter*, *pride*, *feminism*, *NeverAgain*, and *EnoughIsEnough*. We used Scraper for SNS abbreviated as *snsrape*³ to download tweets. We scraped attributes like user IDs, and hashtags, and retrieve the relevant tweets using keywords as a search query. We gathered user tweet

³<https://github.com/JustAnotherArchivist/snsrape>

Characteristic	Original	Pre-processed	Final
No.of tweets	31,07,866	10,65,848	81,944
No. of users	4,78,120	1,36,348	17,660
No. of keywords	213	213	205
No. of hashtags	9,17,833	45,535	37,151
No. of tweets/keyword	14,591	5,004	400
Average no. of hashtags/tweet	5	8	8
Average no. of tweets/user	7	8	5

Table 4.1: IndicHash dataset statistics

data in a variety of languages since people use hashtags regardless of their language of origin. The dataset collection comprises a total of 31,07,866 tweets, and 9,17,833 hashtags posted by 4,78,120 users for a total of 8 languages. The average number of tweets per keyword and tweets per user in the collected dataset amounts to 14,591 and 7 whereas the average number of hashtags per tweet is 5.

4.4.1.1.2 Data Pre-processing The subsequent measures were adopted to ensure a high-quality input for our model. We removed tweets that contain less than three words. The acquired data was noisy due to Twitter’s quick and erratic nature. The data was sanitized by deleting duplicate posts with null values. The pre-processed data underwent several modifications, including the removal of links, conversion of text to lowercase, and exclusion of all non-alphanumeric characters except space and full stop. Hashtags were also collected from these pre-processed posts. Post information such as the content of the original post, hashtags used, and the user id of the user who created that tweet was extracted. To balance the dataset, we randomly sampled an equal number of tweets from each language. The final dataset collection comprises a total of 81,944 tweets, 17,660 users, and 37,151 hashtags. Table 4.1 provides a summary of the dataset’s statistics.

4.4.1.2 Compared Methods

In order to assess the efficacy of the suggested model, we conducted a comparative analysis against prior research endeavors in the domain of hashtag recommendation

as well as established language models based on transformer architecture.

4.4.1.2.1 Existing Research Works To evaluate the efficiency of the proposed model, we contrast our approach with the recent research works on hashtag recommendation.

- [1] AMNN [94] generated hashtags by developing a sequence-to-sequence encoder–decoder framework. The encoder retrieves visual and textual embeddings individually which are then subjected to an attention technique. The attended visual and textual features upon concatenation are fed into GRU, which generates hashtags sequentially according to softmax probabilities.
- [2] TwHIN-BERT [50] developed the Twitter Heterogeneous Information Network which is a polyglot language model that frames the objective of predicting hashtags as a problem of multi-class classification. It is trained with a vast volume of tweets and rich social interactions in order to emulate the brief and noisy nature of user-generated content.
- [3] SEGTRM [11] introduced a transformer-based model which produces hashtags in a sequential manner. SEGTRM consists of three steps: a hashtag generator, a segments-selector, and an encoder. The encoder removes extraneous data at various granularities within text, segments, and tokens in order to derive global textual representations. The segments-selector selects many segments and reorganizes them into a novel sequence to serve as an input to the decoder, enabling end-to-end hashtag construction. To predict hashtags in terms of both quality and quantity concurrently, the authors employ a sequential decoding algorithm.

4.4.1.2.2 Existing Models We discuss various transformer-based models against which we compare the performance of our devised framework. To derive features of tweets in our dataset, we investigated different transformer-based models. These models can be perfectly tailored for classification tasks after being trained on general

tasks. [113] introduced BERT, a transformer-based approach for pre-training NLP models and learn contextual representations during pre-training. It is a deep bidirectional and flexible model that can be fine-tuned by appending a few output layers. Consequently, BERT serves as the underlying architecture for all fundamental models.

- [1] mBERT: Devlin et al. [122] devised mBERT, which stands for multilingual BERT. It is a transformer-based model trained on and usable with 104 languages with Wikipedia (2.5B words) with 110 thousand shared word-piece vocabulary using a masked language modeling (MLM) objective. The input is transformed into vectors with BERT’s capability of bidirectionally training the language model which captures a deeper context and flow of the language.
- [2] mBERT with Transliteration: We used IndicTrans⁴ package released by AI4Bharat to transliterate the text of tweets. We employ transliteration (script conversion) for Indic languages since it helps in reducing the lexical gap among different Indic languages. After transliteration, we obtain embeddings for transliterated tweets using mBERT which in turn are employed to recommend suitable hashtags.
- [3] IndicBERT: Kakwani et al. [127] introduced an ALBERT-based multilingual model featured in AI4Bharat’s IndicNLP Suite. This model was trained on a massive corpus containing over 9 billion tokens in 12 major Indian languages. IndicBERT is capable of extracting sentence and word embeddings.
- [4] XLMR: Conneau et al. [128] proposed the multilingual RoBERTa variant called XLM-RoBERTa which is used to carry out various NLP tasks. It has been pre-trained on an enormous amount of multilingual data with 100 languages using MLM objective. More intriguingly, cross-lingual instruction on a big scale has a major positive impact on languages with few resources. Sentencepiece tokenization is used by XLM-RoBERTa on raw text without any performance loss. Since it uses the same training program as the RoBERTa model, the moniker “Roberta” was incorporated.

⁴<https://ai4bharat.org/indic-trans>

[5] DistilBERT: Sanh et al. [129] developed a condensed adaptation of mBERT with the objective of reducing its size, cost, processing time, and computational load. It contains a reduced number of parameters, up to 40% less than Bert-base-uncased, and it guarantees a faster runtime of 60% while maintaining 97% of the original performance. Furthermore, it is trained on Wikipedia texts in 102 distinct languages. There are 134M parameters in all. DistilBERT is typically twice as quick as mBERTbase.

4.4.1.3 Evaluation Metrics

To evaluate the performance of our suggested hashtag recommendation system, we use assessment criteria from the literature on multi-label classification. The standard evaluation metrics for analyzing the performance of hashtag recommendation methods are Hit rate, Precision, Recall, and F1-score. These metrics are computed by comparing predicted hashtags and ground-truth hashtags for each tweet. We describe each evaluation metric below. The occurrence of at least one common hashtag ($GH \cap RH$) between the set of recommended hashtags (RH) and ground-truth hashtags (GH) accounts for the hit-rate metric when dealing with hashtag recommendation systems. Hit rate is described in the following equation.

$$Hitrate(HR) = \min(|GH \cap RH|, 1) \quad (4.25)$$

The division of the number of hashtags that are present in the set of both ground-truth and recommended hashtags by the cardinality of the set of recommended hashtags yields precision. The following is the formula for precision.

$$Precision(P) = |GH \cap RH| / |RH| \quad (4.26)$$

Recall is the ratio between the number of hashtags shared between ground-truth and recommended hashtags set with the quantity of ground-truth hashtags. The recall is

Technique	Hit rate	Precision	Recall	F1-score
AMNN [94]	0.489	0.195	0.210	0.202
SEGTRM [11]	0.520	0.211	0.228	0.219
TwHIN-BERT [50]	0.600	0.179	0.194	0.187
TAGALOG	0.824	0.334	0.366	0.349

Table 4.2: Effectiveness comparison results with existing research works

computed as given in Equation 4.27.

$$Recall(R) = |GH \cap RH|/|GH| \quad (4.27)$$

To compute F1-score, we derive the harmonic average of precision and recall measures as shown in Equation 4.28.

$$F1 - score(F1) = 2 * P * R / (P + R) \quad (4.28)$$

The outcome of each evaluation metric is denoted as HR@K, P@K, R@K, and F1@K, where K denotes the number of recommended hashtags. Note that larger values imply better performance.

4.4.2 Experimental Results

In this segment, we present an exposition of the empirical findings resulting from the comparison of the proposed framework to state-of-the-art approaches and extant models, analyzing performance enhancement, and examination of visual representations of recommendations.

4.4.2.1 Effectiveness Comparisons

We begin by outlining TAGALOG’s overall benefits, particularly its superiority in outperforming the previous research works and various transformer-based models. We regard the top- K hashtags as the recommended ones, with K being 8, since the mean number of hashtags per tweet is 8. As can be seen in Table 4.2, the performance

Technique	Hit rate	Precision	Recall	F1-score
mBERT [122]	0.757	0.261	0.286	0.273
mBERT with transliteration	0.715	0.240	0.263	0.251
IndicBERT [127]	0.637	0.213	0.229	0.221
XLMR [128]	0.655	0.200	0.221	0.210
DistilmBERT [129]	0.549	0.147	0.159	0.153
TAGALOG	0.824	0.334	0.366	0.349

Table 4.3: Effectiveness comparison results with pre-trained models

gain achieved by TAGALOG is 33.5%, 13.9%, 15.6%, and 14.7% over AMNN, 30.4%, 12.3%, 13.8%, and 13.0% over SEGTRM, 22.4%, 15.5%, and 17.2%, and 16.2% over TwHIN-BERT in terms of hit-rate, precision, recall, and F1-score respectively. The improvement in performance achieved by TAGALOG over AMNN is due to the superiority of mBERT over LSTM [130]. The bidirectional and multilingual nature of the BERT-based feature extractor helps to capture the multilingual context in a better way. Further, TAGALOG considers language and user characteristics when creating the tweet representation to recommend high-quality hashtags in contrast to content-based information used by AMNN. The reason behind performance enhancement over SEGTRM is that SEGTRM filters text at different granularities, whereas TAGALOG adopts language-guided and user-guided attention mechanisms to filter content with respect to the user’s topical and linguistic interests. The remarkable improvement of TAGALOG over TwHIN-BERT is due to modeling user preferences besides user interaction with tweets and language relatedness through graph construction.

Table 4.3 shows the performance comparison of TAGALOG with extant transformer-based models. The performance gain achieved by TAGALOG is 6.7%, 7.3%, 8.0%, and 7.6% over mBERT without transliteration, 10.9%, 9.4%, 10.3%, and 9.8% over mBERT with transliteration, 18.7%, 12.1%, 13.7%, and 12.8% over IndicBERT, 16.9%, 13.4%, 14.5%, and 13.9% over XLMR, 27.5%, 18.7%, 20.7%, and 19.6% over DistilmBERT in terms of four performance measures. The reasons behind this gap are the incorporation of a novel language-guided attention mechanism in addition to user-guided attention, the construction of a user-tweet graph to capture interactions among tweets belonging to languages of the same family, and user-

Mechanism	Hit rate	Precision	Recall	F1-score
$TAGALOG_{NA}$	0.784	0.285	0.313	0.299
$TAGALOG_{LGA}$	0.783	0.292	0.321	0.306
$TAGALOG_{UGA}$	0.824	0.330	0.361	0.345
$TAGALOG_{UGA+LGA}$	0.824	0.334	0.366	0.349

Table 4.4: Performance of TAGALOG with different attention techniques

tweet interaction to enrich user and tweet embeddings. These procedures help in constructing an effective tweet representation which in turn recommends high-quality and relevant hashtags for tweets posted in low-resource Indic languages.

4.4.2.2 Performance Gain Analysis

We analyze the performance pickup of the suggested approach in this section. Following a performance comparison with various model components, we examine how TAGALOG performs using various attention techniques.

4.4.2.2.1 Attention Techniques We discuss how TAGALOG performs with diverse attention strategies in this part. The variants of TAGALOG that use no attention, language-guided attention, user-guided attention, and user-guided along with language-guided attention are $TAGALOG_{NA}$, $TAGALOG_{LGA}$, $TAGALOG_{UGA}$, and $TAGALOG_{UGA+LGA}$ respectively. Here, $TAGALOG_{UGA+LGA}$ refers to our devised system. Table 4.4 illustrates the performance obtained on eliminating attention mechanisms that comprise the feature refinement module. Here, UGA and LGA refer to user-guided attention and language-guided attention mechanisms. The performance difference when TAGALOG is implemented without any attention mechanism is 5.0% in terms of the F1-score. To derive the overall tweet representation in the case of the no-attention model, we compute the average of mBERT-based token embeddings. The performance of TAGALOG is the lowest in the absence of any attention mechanism. The drop in the F1-score on eliminating UGA from TAGALOG, termed as $TAGALOG_{LGA}$, is 4.3%, while the difference in excluding LGA from TAGALOG, abbreviated as $TAGALOG_{UGA}$, is 0.4%. UGA helps to learn the context in which a user

Technique	Hit rate	Precision	Recall	F1-score
$TAGALOG_{FI}$	0.784	0.285	0.313	0.299
$TAGALOG_{FR}$	0.806	0.314	0.342	0.328
$TAGALOG_{FR+FI}$	0.824	0.334	0.366	0.349

Table 4.5: Performance comparison of TAGALOG with different components

created a post and LGA assists in learning the user’s language choice and usage style. UGA is typically used to improve the relevance and usefulness of tweets for individual users and to enhance the overall user experience, while LGA focuses on modeling idiosyncratic language behavior. The above-mentioned performance gap demonstrates the significance of language-guided and user-guided attention techniques.

4.4.2.2.2 Model Component Analysis We conduct model component analysis to emphasize the significance of various components constituting the proposed model. Below, we put forth the performance of Feature Refinement (FR) and Feature Interaction (FI) components comprising TAGALOG. We eliminate the feature refinement component to stress its pertinence. The resultant model is referred to as $TAGALOG_{FI}$. Similarly, the model obtained on the exclusion of feature interaction from TAGALOG is referred to as $TAGALOG_{FR}$. We use acronyms $TAGALOG_{FR+FI}$ and $TAGALOG$ in tandem since $TAGALOG_{FR+FI}$ is the model we have developed. Table 4.5 shows the performance of TAGALOG on eliminating its different components. The performance gap in terms of evaluation metrics on the exclusion of FR is 4.0%, 4.9%, 5.3%, and 5.0% respectively, while that on the exclusion of FI is 1.8%, 2.0%, 2.4%, and 2.1%, which demonstrates the significance of these components. Additionally, the performance of the proposed model which includes both FR and FI beats the performance of individual components. This implies these components complement each other when recommending hashtags. FR captures local topical and linguistic interests of individual users through UGA and LGA, while FI captures global interests by analyzing the long-term behavior and preferences of the user besides tweet correlation based on language relatedness. Overall, the experimental results show that each component contributes positively to TAGALOG’s performance.

4.4.2.3 Qualitative Analysis

We conduct qualitative investigations to demonstrate how effective our framework is. We show user-created tweets together with hashtags proposed by different mod-

Actual Tweet: Punjab election 2022: আজ পাঞ্জাবে প্রচারে মোদী, নিরাপত্তায় ত্রুটি কাণ্ডের পর প্রথম সফর প্রধানমন্ত্রীর	Language: Bangla	Actual Tweet: રશિયા-યુક્રેન યુદ્ધ બાબેલું વિનાશની સુનામી: અબજો ડોલર ખર્ચાયા, સંકડો મૃત્યુ	Language: Gujarati
Translated Tweet: Punjab election 2022: Modi to campaign in Punjab today, Prime Minister's first visit after security lapses		Translated Tweet: The Russia-Ukraine war brought a tsunami of destruction: billions of dollars spent, hundreds of deaths.	
#punjabelection2022 #narendramodi #congress #bjp #politics #কংগ্রেস (Congress) #বিজেপি (BJP) #पांजावविधानसभा (PunjabVidhanSabha) #नरेन्द्रमोदी (NarendraModi)		#internationalnews #middaynews #ukraine #middaygujarati #russia #war #russiaukraine #russianarmy	
<div>Predictions</div> <div> Accurate Pertinent Erroneous </div>			
TAGALOG: #punjabelection2022 #punjab #politics #bjp #narendramodi #congress #pmmodi #কংগ্রেস (Congress) #বিজেপি (BJP) #রাহুলগান্ধী (RahulGandhi)		TAGALOG: #ukraine #middaygujarati #russia #middaynews #war #internationalnews #russiaukrainewar #russiaukraine #vladimirputin #gujaratinews	
SEGTRM: #punjab #punjabelection2022 #congress #navjotsinghsidhu #amritsar #candidate #পঞ্জাব (Punjab) #প্রার্থী (Candidates) #politics		SEGTRM: #ukrainerrussiawar #russia #ukraine #war #pmmodi	
AMNN: #puberkalom #uttarpradesh #farmerprotest #india #lockdown3 #ভারত (India) #নরেন্দ্রমোদী (NarendraModi)		AMNN: #russia #ukraine #nato #pmnarendramodi #advice	
TwHIN-BERT: #india #socialmedia #news #actor #covid19 #politics #cricket #bollywood #entertainment #twitter		TwHIN-BERT: #middaynews #middaygujarati #bollywoodgossip #gujarat #socialmedia #india #bollywood #actor #mumbainews #covid19	
(a) Post 1		(b) Post 2	

Figure 4.4: Example posts showing hashtags recommended by different methods

els. For sample tweets chosen from test data, accurate hashtags are shown in green, pertinent in blue, and erroneous in red. The hashtags that models recommend and

are consistent with hashtags that reflect the actual situation are considered accurate. On the other hand, pertinent hashtags do not belong to the category of ground-truth hashtags but are compatible with the tweet’s content.

The tweet given in Fig. 4.4(a) is in context with the Punjab elections held in 2022, written in Bangla. As can be seen, the user assigns a few hashtags to the tweet in his native language. It indicates that these hashtags used are wildly trending about Punjab elections among Bangla Twitter users. The user assigns #congress and #bjp not only in English but also in Bangla. Besides assigning hashtags in English, users tend to assign topics of their interests with hashtags in their native language. Users are more inclined to adopt hashtags in their native language to connect with others who share their cultural background or interests. Hashtags in different languages can also promote diversity and inclusivity on social media platforms, allowing users to find content and connect with others from a broader range of backgrounds and perspectives. Hashtags recommended in Bangla indicate the ability of our model in recommending language-specific topical hashtags. This implies our model recommends multilingual hashtags and learns the user’s language usage style by adopting his linguistic behavior. The hashtag #punjab is directly related to the event of the Punjab Elections; #pmmodi and #rahulgandhi are prominent political figures and therefore deemed pertinent. TAGALOG recommends seven accurate and three pertinent hashtags. DESIGN recommends four accurate, five pertinent, and one erroneous hashtag. SEGTRM recommends three accurate and six pertinent hashtags. AMNN recommends one accurate, five pertinent, and one erroneous hashtag. TwHIN-BERT recommends one accurate, four pertinent, and five erroneous hashtags. Our model recommends the highest number of accurate hashtags indicating that mining users’ posting and linguistic behavior help suggest plausible hashtags.

The tweet in Fig. 4.4(b) is written in Gujarati in the context of the global event, the Russia-Ukraine war. TAGALOG recommends seven accurate and three pertinent hashtags; DESIGN recommends five accurate, four pertinent, and one erroneous hashtag; SEGTRM recommends three accurate, one pertinent, and one erroneous hashtag; AMNN recommends two accurate, one pertinent, and two erroneous hashtags, TwHIN-

BERT recommends two accurate, one pertinent, and seven erroneous hashtags. The example posts demonstrate how, by suggesting customized hashtags based on users’ thematic and linguistic preferences, TAGALOG surpasses earlier research methods.

4.5 Conclusion

In this chapter, we have tackled hashtag recommendations to facilitate multilingual content retrieval and break through language barriers inherent in social media platforms. The proposed polyglot model, TAGALOG, can recommend personalized and language-specific hashtags for online content generated in various low-resource Indic languages. The system proposed in this study comprises feature extraction, refinement, and interaction modules. We first extract content-based, linguistic, and user-based features using a transformer and deep learning-based models. We then employ language-guided and user-guided attention mechanisms to fine-tune tweet representation in line with users’ linguistic and topical preferences. In the feature interaction module, we connect the historical tweets of a particular user to mine his posting behavior. Furthermore, we group tweets written in various languages concerning their families, i.e., Indo-Aryan and Dravidian, to capture their interrelatedness. Extensive experiments conducted on the curated Twitter dataset reveal that our proposed model is superior in performance to language models that have been trained and state-of-the-art methods.

Chapter 5

Hashtag Recommendation for Multimodal Content

5.1 Introduction

The dynamic nature of SNS fosters diverse modes of communication and information sharing. As a result, SNS users share microblogs that consist of texts and images, occasionally elucidated with hashtags. Platforms such as Instagram, boasting a substantial user base of one billion [131] and a daily upload of approximately 95 million photos, exemplify the widespread creation of such multimodal UGC. While the presence of even a single hashtag has been shown to boost user participation by 12.6% [132], a considerable volume of this content remains unannotated due to user reluctance. This under-tagging underscores the imperative need for automated hashtag recommendation processes to suggest relevant hashtags for social media posts. Given the prevalence of multimodal content, leveraging the complementary information from both textual and visual sources is crucial for enhancing hashtag recommendation accuracy, a capability lacking in existing unimodal approaches focused solely on text [28, 133] or image [58, 59].

Consider examples in Figure 5.1 to illustrate the importance of multimodal information. In the first post (Figure 5.1(a)), the user’s philosophical reflection on a sunset is evident through Richie Norton’s quote, leading to the hashtag `#positive-quotes`, which is derivable from the text. Conversely, hashtags such as `#sunset` and

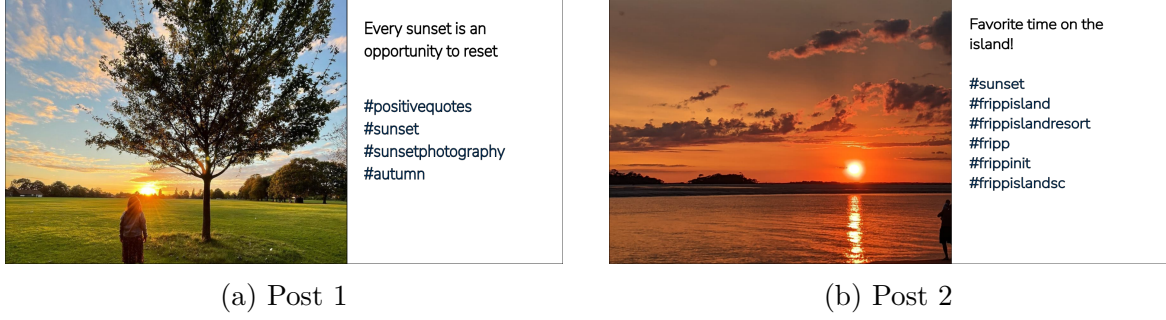


Figure 5.1: Example posts from Instagram

`#autumn` are directly related to the visual content and cannot be solely inferred from the text. This demonstrates that text and images convey distinct yet complementary information about a post, highlighting the necessity of learning features that capture information across multiple modalities. Furthermore, hashtags offer valuable insights into users' interests, and individual tagging habits significantly influence content consumption and discovery. As seen in Figure 5.1, different users posting distinct content might still employ the same hashtags (`#sunset`), indicating shared interests. Conversely, users with different interests might tag similar content with different hashtags (`#positivequotes` in Figure 5.1(a) and hashtags `#frippisland`, `#frippislandresort` in Figure 5.1(b)). This underscores the limitations of purely content-based hashtag recommendation methods, which often fail to capture these crucial user-specific interests. Consequently, these methods may not be directly suitable for personalized hashtag recommendation. Therefore, to address these limitations and contribute to the overarching goal of this thesis, this chapter proposes a novel multimodal personalized hashtag recommendation system.

Numerous approaches have been put forward to formulate hashtag recommendation task. Most of the prior works employing Deep Learning techniques have modeled hashtag recommendation as a multiclass classification problem [5] while others have considered it as a sequence generation problem [94, 65]. Classification-based approaches for hashtag recommendation suggest hashtags from a pre-defined list with a limited categories of hashtags. These approaches do not consider the interrelationship among hashtags. Word by word sequence generation considers the sequential nature

of hashtags, thus capturing the dependencies among generated hashtags. Sequence generation approaches perform better on sparse and infrequent hashtags. In sequence generation[9, 94], the following hashtag to be generated highly depends on the preceding hashtag as hashtags are considered as an ordered sequence. Sequence generation can model the interdependencies among words present in a sentence. The output of hashtag recommendation is a set of hashtags that may be correlated but may not follow a strict order as exhibited by words present in a sentence. Classification-based approaches treat hashtags to be recommended as independent categories. While hashtags can be viewed as predetermined categories, they can also be generated sequentially, similar to how words in natural language sentences are generated. Intending to model the sequential relationship among hashtags and investigate the implicit correlation among them, we interpret hashtag recommendation as a sequence generation problem. To this end, we employ a novel personalized sequence generation framework based on an encoder-decoder architecture that can generate correlated and personalized hashtags for social media posts. Recommended hashtags are generated in the form of a sequence where previously generated hashtags are trusted to be relevant and used for generating the following most relevant hashtag. The sequence generation-based framework fully exploits the multimodal information of microblog posts and models correlations between hashtags, multimodal post content, and users’ historical posts.

Traditional methods for hashtag recommendation majorly formulate it as a classification problem which neglects correlations among hashtags. Few works model hashtag recommendation as Sequence Generation. In Sequence Generation, recommended hashtags are supposed to be ordered as words in the sentence. However, the output of hashtag recommendation is a set of hashtags that may be correlated but may not follow a strict order as exhibited by words present in a sentence. The recommended candidate hashtags may not exhibit any ordered sequence yet match ground truth hashtags. The existing works formulate hashtag recommendation either in terms of Multi-Label Classification (MLC) or Sequence Generation (SG). None of the existing works formulate this task from both perspectives. To the best of our knowledge, we are the first to propose a unified hybrid model that casts the task of

hashtag recommendation to both MLC and SG. We propose an integral model that encodes different aspects of hashtag recommendation in a coherent encoder-decoder framework. The two strategies play a mutually complementary role in recommending personalized hashtags to multimodal content. The proposed model capitalizes on benefits of both approaches to suggest better and relevant hashtags to users in contrast to hashtags predicted by individual approaches of MLC and SG.

In this chapter, we propose a hybrid deep neural network for multimodal personalized hashtag recommendation system that can automatically recommend hashtags to unannotated social media content. We consider textual and visual modalities available in social media posts to improve hashtag recommendation. The proposed hybrid Deep Neural Network uses two different formulation procedures i.e., classification and generation. Since both approaches captures different key aspects of social media posts, we capitalize hashtags predicted from both approaches to recommend more relevant hashtags. Our method recommends suitable and correlated hashtags for text-only, image-only, and multimodal social media posts. Our proposed approach takes users' preferences and tagging behavior into account to recommend personalized hashtags. Further, we employ word-level and parallel co-attention mechanisms. The word-level attention captures the importance of different words in the text, and parallel co-attention learns the mutual influence of one modality on the other. Parallel co-attention jointly models the text-image inter-relationship to enrich the contextual information.

Our contributions are summarized as follows:

- We propose a hybrid Deep Neural Network to address the problem of hashtag recommendation by jointly formulating it as MLC and SG problems. To the best of our knowledge, we are the first to propose a hybrid model that capitalizes on benefits of both MLC and SG techniques, boosting the performance of hashtag recommendation. These two strategies play a mutually complementary role in recommending personalized hashtags to multimodal social media posts.
- We devise a novel personalized generative framework to suggest personalized and correlated hashtags for multimodal social media posts. Our system recommends

hashtags based on the user’s hashtagging behavior and preferences derived from his historical posts and associated hashtags.

- Our proposed method predicts suitable hashtags for posts by mining information from textual and visual modalities. We apply word-level attention on textual content to learn those words in the text that are more closely related to hashtags followed by a parallel co-attention mechanism to model deep interactions between the two modalities.
- We have constructed a new dataset named TINS (Text dataset from INSta-gram), consisting of 23,868 posts associated with at least one hashtag crawled from Instagram. This dataset can be used to carry out text-based hashtag recommendation research. Extensive experimental results on three datasets show that our proposed method surpasses current state-of-the-art methods by incorporating information from textual and visual modalities and the user’s hashtagging behavior. Furthermore, our model beats other baselines in image-based and text-based hashtag recommendations when only image or text information is provided.

The structure of the remainder of this chapter is as follows. Section 5.2 formally defines the problem under investigation. The proposed methodology is then detailed in Section 5.3. Subsequently, Section 5.4 presents and discusses the experimental evaluations conducted. Finally, Section 5.5 offers concluding remarks and a summary of this work.

5.2 Problem Definition

In this section we present the problem definition and formulation.

Problem 1 (*Hashtag Recommendation*) Suppose there is a social media dataset with a post set $P = \{p_i\}_{i=1}^N$, a hashtag set $H = \{h_j^g\}_{j=1}^J$, and a user set $U = \{u_k\}_{k=1}^K$, in which the post $p_i \in P$ composed of the image (I) and the text (T), is created by a user $u_k \in U$ with some hashtags $h_j^g \in H$.

Given a test post (p_i) created by a user (u_k) , we aim to automatically recommend a set of hashtags $Rh = \{rh_r\}_{r=1}^R$, such that set of recommended hashtags Rh for the given post p_i match to ground truth hashtag set $Gh = \{gh_g\}_{g=1}^G$.

Here, N denotes the number of posts in the dataset, J is the number of unique hashtags, K denotes the number of users, i, j, k are used to index the post, hashtag, and user, respectively. For a post p_i , R denotes the number of recommended hashtags, G denotes the number of ground-truth hashtags, and r, g are used to index the recommended hashtag, and ground-truth hashtag respectively.

Problem 1 is the hashtag recommendation for a social media post. In this problem, we automatically recommend a good quality of hashtags that can be used by social media users to annotate their content. Good quality of hashtags increase audience engagement [134] and help to search, and categorize social media posts. We can model hashtag recommendation using multi-label classification or sequence generation. We discuss these two problems below.

Problem 1.1 (*Hashtag Recommendation using Multi-Label Classification*) To tackle the task of hashtag recommendation, we formulate it as a multi-label classification problem. Assume that there is a post set $P = \{p_i\}_{i=1}^N$, user set $U = \{u_k\}_{k=1}^K$ and a predefined set of candidate hashtags $H = \{h_j^g\}_{j=1}^J$, where N denotes the cardinality of set P , K denotes the cardinality of set U and J denotes the cardinality of set H .

Given a test post p_i such that $p_i \in P$, the goal of multi-label classification is to recommend a plausible set of hashtags $Rh = \{rh_r\}_{r=1}^R$ that the creator of the test post p_i is likely to assign to (p_i) from the pool of predefined hashtags (H).

Social media posts generally contain multiple hashtags pointing towards specific portions of associated texts and images. The hashtags are the various class labels to which a post belongs. The membership of a post in these classes is not mutually exclusive as the post usually belongs to several classes simultaneously. We can formulate hashtag recommendation as a multi-label classification problem and assign multiple hashtags to a post. To this end, we solve hashtag recommendation in terms of multi-label classification to recommend a plausible set of hashtags for a given post (p_i) .

Problem 1.2 (*Hashtag Recommendation using Sequence Generation*) Though we have formulated our problem as multi-label classification, another efficient way to address this problem is to model it as a sequence generation problem. Suppose there is a social media dataset as discussed in Problem 1.1.

Given a test post (p_i) , our goal is to output hashtags $Rh = \{rh_r\}_{r=1}^R$ represented by a sequence of words, which a user (u_k) can annotate to post (p_i) .

Here, Rh denotes the set of hashtags recommended for a post p_i . Hashtags for social media posts are often strongly correlated with each other. It is essential to capture the underlying structure among hashtags based on their semantics. Due to the limited data to learn from, classification-based approaches suffers from data sparsity problem. The generation of hashtags in a word-by-word manner enables the internal structure of hashtags to be exploited, thus capturing the semantic dependencies among them. We solve hashtag recommendation as a sequence generation problem to recommend semantically related hashtags for a post (p_i) . We propose a hybrid model that integrates the multi-label classification-based and sequence generation-based approaches to take their combined advantages into account.

Problem 2 (*Feature Learning from Posts*) Social media users create multimodal posts containing different facets such as texts and images. Each facet is highly informative in providing hashtags to the user-created post. A large number of posts containing such facets are generated that cannot be directly utilized by hashtag recommendation methods. We are required to derive a good set of features that can serve as input to hashtag recommendation methods. We examine different approaches to model the intrinsic multi-modality of social media posts and extract the fundamental features that can be leveraged to propose a good quality of hashtags. We define the feature extraction problem below.

Given a post set P and a post $p_i \in P$, we aim to extract meaningful numerical representations from the post’s textual modality denoted as p_i^t , and post’s visual modality denoted as p_i^v .

Firstly, for the visual modality of i^{th} post p_i represented as p_i^v , our goal is to obtain a visual feature matrix $V = \{v_y\}_{y=1}^Y$, where, v_y corresponds to the visual feature vector

of the y^{th} region of the image, and $v_y \in \mathbb{R}^D$, where D is the embedding size. Secondly, for the textual modality p_i^t appearing as a sequence of words $W_i = \{w_i^x\}_{x=1}^W$, our goal is to embed the words present in p_i^t into low dimensional real-valued vectors and obtain a text feature matrix $T = \{e^x\}_{x=1}^X$. Here, W represents the number of words appearing in the textual modality p_i^t , e_x corresponds to the text feature vector of the word w_i^x , and $e_x \in \mathbb{R}^D$, where D is the embedding size and X represents the length of token sequence obtained after passing W_i to the text encoder.

Since raw texts and images cannot be used directly, they need to be converted into appropriate embeddings. To obtain text and image distributed representations, we segment the multimodal social media posts into their constituent modalities. We employ attention mechanisms to distill the significant parts of textual and visual information present in the post and jointly model the retrieved information conducive to hashtag recommendation.

Problem 3 (*Personalised Hashtag Recommendation*) Suppose there is a social media dataset with a post set $P = \{p_i\}_{i=1}^N$, a hashtag set $H = \{h_j^g\}_{j=1}^J$, a user set $U = \{u_k\}_{k=1}^K$ and for a post p_i created by user (u_k) , the historical post sequence of user (u_k) is denoted as $HP = \{hp_l\}_{l=1}^L$.

Given the historical posts of a user (u_k) denoted as $HP = \{hp_l\}_{l=1}^L$, our goal is to recommend hashtags for the new, i.e., i^{th} post created by user (u_k) based on the user's preferences.

Social media users have a unique pattern of assigning hashtags to their created posts. For making personalized recommendation, it is useful if the suggested hashtags reflect the users' preferences. Analysis of users' preferences, which helps obtain a deeper insight into their interests, requires the implicit modeling of the user's tagging behavior. To achieve this goal, we attempt to recommend relevant hashtags for a new post (p_{l+1}) posted by a user (u_k) based on the user's tagging patterns and vocabulary choices, which are extracted from the user's history. We recursively access l historical posts $HP = \{hp_l\}_{l=1}^L$ of a user with the current post content (p_i) to make personalized hashtag recommendation.

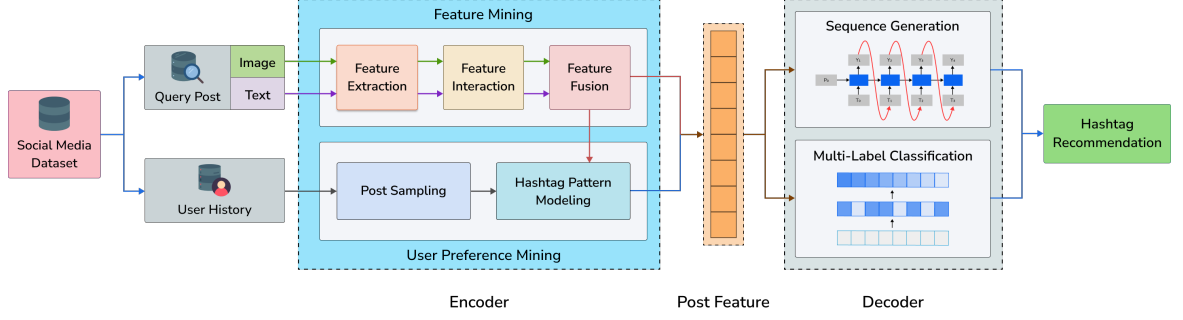


Figure 5.2: Overall architecture of DESIGN

5.3 Methodology

In this section, we present our proposed methodology. Figure 5.2 depicts the overall architecture of our proposed hashtag recommendation system. The proposed system first retrieves the coherent features from visual and textual modalities to obtain a joint feature vector representation of a social media post. User habits are learned and integrated with joint features to get a post feature vector influenced by user’s tagging behavior. Hashtags are predicted from post features using different hashtag prediction methods namely, MLC and SG. The predicted hashtags are then effectively sampled and ranked to recommend a relevant set of hashtags for social media posts. We perform the following four steps to automatically recommend a set of hashtags for the social media posts: (a) feature mining; (b) user preference mining; (c) hashtag prediction; and (d) candidate hashtag recommendation.

5.3.1 Feature Mining

The feature mining module is shown in Figure 5.3. It comprises three submodules: (a) feature extraction, (b) feature interaction, and (c) feature fusion. We first extract visual and textual features from social media posts. The distributed representations of the constituent texts and images are co-attended to model their interaction to obtain a richer contextual representation of the post. We describe the details of each submodule in the following sections.

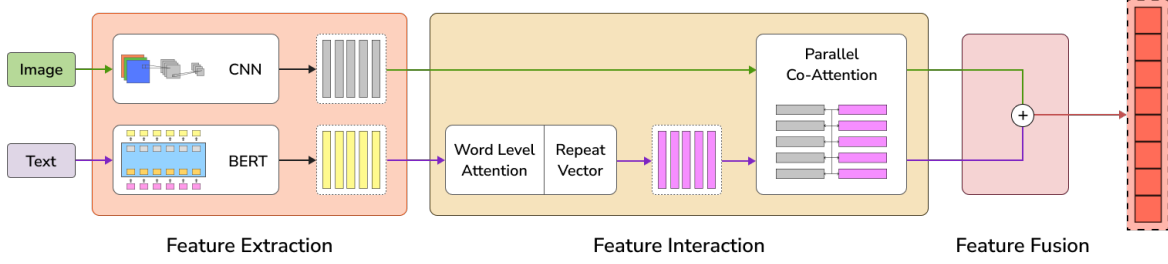


Figure 5.3: Feature mining module

5.3.1.1 Feature Extraction

As a post usually contains textual and visual content, it is essential to extract the features from these contents. In the following sections, we will discuss visual and textual feature extraction.

5.3.1.1.1 Visual Feature Extraction Image content is one of the vital modalities prevalent in social media posts. It is therefore essential to retrieve the features that capture the essence of the visual content. We use different transfer learning models that have exhibited exceptional performance in visual recognition tasks to derive useful information from posts' visual content. We experimented with two different Convolutional Neural Networks (CNN) to extract visual features, namely VGG-16 [135] and ResNet-50 [136]. Image content is one of the vital modalities prevalent in social media posts. In order to retrieve the features that capture the essence of the visual content, we use different transfer learning models that have exhibited exceptional performance in visual recognition tasks. We experimented with two different CNN to extract visual features, namely VGG-16 and ResNet-50. First, we rescale the input image to dimensions 224×224 . An image can be represented as a 3D matrix consisting of 3 primary color channels, i.e., red, green, and blue. An image of height h and width w is denoted as I where $I \in \mathbb{R}^{h \times w \times 3}$. Here, \mathbb{R} represents the set of real numbers. We adopt widely used Convolutional Neural Network-based models (i.e., VGG-16 and Resnet-50). We first pass the input image I to CNN to derive the visual features of an image as given in Equation 5.1.

$$I_f = CNN(I) \quad (5.1)$$

The image feature vectors thus obtained are stacked horizontally to form the visual feature matrix V , given by Equation 5.2.

$$V = Reshape(I_f) \quad (5.2)$$

Here, $V \in \mathbb{R}^{Y \times 512}$, $Y = 7 \times 7 = 49$ in case of VGG-16, $V \in \mathbb{R}^{Y \times 2048}$ in case of ResNet-50 and $Y = 7 \times 7 = 49$ represents the number of regions in the image. Since the spatial features describe the image precisely, we divide an image into 7×7 regions to construct a 512-dimensional feature vector for each region in case of VGG-16 in contrast to the 2048 dimensional regional feature vector in case of ResNet-50.

Finally, we use a dense layer to transform the visual feature matrix V that we have obtained from one of the above-mentioned CNN to a matrix having the same embedding size as the text feature matrix T which is shown in Equation 5.3.

$$V = Dense(units = D)(V) \quad (5.3)$$

Here, $V \in \mathbb{R}^{Y \times D}$, $Y = 49$ and $D = 768$. We explain the textual feature extraction in the next section.

5.3.1.1.2 Textual feature extraction Text is an integral part of social media posts. To extract the features from textual content embedded in a social media post, we employ a transformer-based Deep Learning model, BERT. Bidirectional Encoder Representations from Transformers or BERT in short, captures the bidirectional context of the input. This nature of BERT is attributed to the fact that it reads all the input words simultaneously. BERT is a context-aware model which focuses on the surrounding words before generating the embeddings. Unlike word2vec, which is context-independent and does not consider homonyms, BERT is context-dependent and takes homonyms into account. For example, “fair” could refer to some event related to entertainment, or “fair” could refer to impartial behavior. Based on their usage, the words are represented by different vectors.

BERT is pre-trained on extensive unlabeled data from Wikipedia and Book Corpus

using two unsupervised methods i.e., Masked Language Modelling (MLM) and Next Sentence Prediction (NSP) [113]. In MLM, 15% of the words in each word sequence are substituted with a [MASK] token before being fed into BERT. The model tries to generate a prediction for the masked word by comprehending the context of surrounding words. For NSP, BERT is fed with pairs of sentences and learns to predict whether the second sentence in the pair follows the first sentence.

The transformer is made up of attention-based components: encoder and decoder. The encoder reads the input sentence to generate an abstract continuous vector representation, and the decoder uses the generated representation to predict the output. BERT is built upon the transformer architecture. Since BERT is a language representation model for generating word embeddings, it employs stacked layers of transformer encoder to represent each input token. Given a sequence of W words $W_i = \{w_i^x\}_{x=1}^W$ representing the textual modality (p_i^t) of the post (p_i), our model begins with inserting two unique tokens in each post’s textual content, a class [CLS] token at the beginning and a separator [SEP] token at the end. These tokens are used to mark the beginning and end of the sentence, respectively. Next, we use BERT’s tokenizer to generate a set of integer-based tokens B as shown in Equation 5.4.

$$B = BERT_Tokenizer(W) \quad (5.4)$$

In our case, we restrict the maximum length of the token sequence to X . For sequences with a length greater than X , we perform truncation; otherwise, we insert empty tokens to perform padding. Finally, we employ BERT to construct the embedding vector for each token, which captures its syntax and semantics. BERT receives a sequence of tokens that moves up the stack. Each layer employs self-attention and routes its output through a feed-forward network before passing it to the next encoder. We feed B into the BERT model, as given in Equation 5.5, which in turn generates a 768-dimensional vector for each token.

$$T = BERT(B) \quad (5.5)$$

Here, $T = \{e^x\}_{x=1}^X$, $T \in \mathbb{R}^{X \times D}$ is the textual feature matrix and e^x is the BERT embedding for a token. mBERT has been pretrained on 104 different languages with MLM objective. Since we have social media posts written in diverse languages, we use the Multilingual BERT model to generate embeddings for the tokens obtained from text appearing in the posts. For our task, we use the base version of Multilingual BERT having 12 encoder layers, with each layer having 12 self-attention heads.

5.3.1.2 Feature Interaction

We apply different feature interaction mechanisms on features extracted from visual and textual modalities. To perform interaction, we use attention techniques. John Robert Anderson defined attention as a process that allows humans to concentrate more on a certain piece of information to derive conclusions [137]. Similarly, neural networks perform better by focusing more on relevant parts of the input. Attention refers to concentrating specifically on certain vital parts of data and generating feature vectors based on these specific parts. In the following sections, we will present how we first apply word-level attention to enrich the textual representation of the post. We also employ parallel co-attention to model the interaction between visual and textual features. At last, we integrate these two attention mechanisms to perform word-level and parallel co-attention.

5.3.1.2.1 Word-level Attention A social media post contains multiple words. Various words appearing in the post’s textual content vary in importance when providing information about the post. The basic intuition behind the word-level attention mechanism is that words constituting the post contribute differentially to its semantics. The importance of words depends heavily on context, i.e., the same word may be differentially important in some other context. Attention offers insight into which words deliver essential information for generating relevant hashtags, resulting in improved performance. The word-level attention allows the model to pay varying degrees of attention to individual words by assigning them different weights. We employ a word-level attention mechanism to model the post’s textual content [138, 139]. The words

appearing in the post are first encoded into low-dimensional vectors using the encoder. Then a word-level attention mechanism is used to retrieve the words that contribute significantly to the meaning of the post. Given the sequence of different words denoted as $W_i = \{w_i^x\}_{x=1}^W$, representing the textual content of the post p_i^t denoted as (p_i^t) , we use BERT, a pre-trained transformer-based model to derive word embeddings by merging input from both left and right sides for each word besides including the contextual information. We employ an attention mechanism to extract important words and aggregate the obtained word representations to derive an overall representation of textual content of the post. Specifically, we first feed the token annotation e^x through MLP to get h^x as a hidden representation of e^x as shown in Equation 5.6.

$$h^x = \tanh(We^x + b_w) \quad (5.6)$$

Here, h^x as a hidden representation of e^x . We compute the importance of word α^x as shown in Equation 5.7.

$$\alpha^x = \text{softmax}((h^x)^T u_w) \quad (5.7)$$

Here, α^x indicates the importance of a word. First, we compute the similarity of h_x with u_w and pass the product through a softmax function to obtain normalized weight α^x . After that, we compute the textual post vector as follows:

$$t = \sum_{x=1}^X \alpha^x e^x \quad (5.8)$$

Here, t denotes the textual post vector which is computed as a weighted sum of the word annotations based on the weights α^x .

5.3.1.2.2 Parallel Co-Attention Description of a single social media post exists in multiple modalities, e.g., texts contain natural language words, images have visual signals, and objects of different attributes such as size, color, and position. Since these modalities describe the same content from different perspectives, they exhibit varying degrees of correlation at specific levels. We need to focus on multimodal information

fusion when obtaining a latent representation of the post (p_i). Learning multimodal representations involves integrating information from multiple data sources comprising the post.

The various modalities constituting the post barely interact with one another. As a result, interrelation among different modalities cannot be addressed. To comprehend how modalities interact, the two must be combined so that the resultant vector can convey joint reasoning across the visual and textual modalities. Algorithm 5.1 shows

Algorithm 5.1 Parallel Co-Attention

Input: T : Text feature matrix
 V : Image Feature Matrix
Output: \tilde{t} : Text feature vector
 \tilde{v} : Image feature vector
function Para_Co-Attention(T, V)

- 1: $C \leftarrow \tanh(TW_bV^T)$
- 2: $F^t \leftarrow \tanh(W_tT^T + (W_vV^T)C^T)$
- 3: $a^t \leftarrow \text{softmax}(W_{ht}^T F^t + b_{ht})$
- 4: $\tilde{t} \leftarrow \sum_{i=1}^X a_i^t t_i$
- 5: $F^v \leftarrow \tanh(W_vV^T + (W_tT^T)C)$
- 6: $a^v \leftarrow \text{softmax}(W_{hv}^T F^v + b_{hv})$
- 7: $\tilde{v} \leftarrow \sum_{i=1}^Y a_i^v v_i$
- 8: **return** \tilde{t}, \tilde{v}

the parallel co-attention mechanism that attends to image and text simultaneously. We model their association based on similarity between visual and textual features computed for all combinations of image and text locations. Line 1 shows how to compute the affinity matrix. Given an image feature matrix $V \in \mathbb{R}^{Y \times D}$, and the text feature matrix $T \in \mathbb{R}^{X \times D}$, the affinity matrix $C \in \mathbb{R}^{X \times Y}$ is calculated as given in Line 1, where $W_b \in \mathbb{R}^{D \times D}$ denotes the correlation matrix to be learned. To capture the correlations between text and image features, we transfer the image and text feature space into each other. The affinity matrix C maps text-based attention to image-based attention (vice versa for C^T). We can define the new text feature matrix ($F^t \in \mathbb{R}^{D \times X}$) as given in Line 2. Here, the visual feature matrix V is multiplied by C^T and then integrated into the textual features, and $W_t, W_v \in \mathbb{R}^{D \times D}$ are the parameters. The image features guide the attention learning of text. Similarly, we compute the

new visual feature matrix ($F^v \in \mathbb{R}^{D \times Y}$) as given in Line 5. Here, $W_t, W_v \in \mathbb{R}^{D \times D}$, $T \in \mathbb{R}^{X \times D}$, $V \in \mathbb{R}^{Y \times D}$, $C \in \mathbb{R}^{X \times Y}$.

Next, we use the new feature matrices to compute the attention weights as shown in Lines 3 and 6 where, $W_{ht}, W_{hv} \in \mathbb{R}^D$ and $b_{ht}, b_{hv} \in \mathbb{R}$ are the parameters. The dimensions of the resultant attention weights are given as $a^t \in \mathbb{R}^{1 \times X}$ and $a^v \in \mathbb{R}^{1 \times Y}$. The global text and image feature vectors are calculated as the weighted sum of the textual and visual feature vectors respectively with the above attention weights as shown in Lines 4 and 7. Here, $\tilde{t} \in \mathbb{R}^D, \tilde{v} \in \mathbb{R}^D, a_i^t$ and a_i^v represent the attention weights corresponding to a certain word and an image region, respectively. The correlation between the two can help filter out noisy data and provide richer semantic representation as it focuses only on relevant multimodal features.

5.3.1.2.3 Word-level and Parallel Co-Attention Different modalities depict the intrinsic content of the social media post from different angles. In order to learn the importance of different words representing the textual content of a post (p_i), we apply word-level attention to associated text i.e., p_i^t which is shown in Equation 5.9.

$$t = \text{Word-level}(p_i^t) \quad (5.9)$$

Next, we use the `Repeat_Vector` function to transform the text feature vector t into a matrix T' as follows:

$$T' = \text{Repeat_Vector}(t) \quad (5.10)$$

Finally, we employ the parallel co-attention mechanism as given in Equation 5.11 to model the interrelationship between textual and visual modalities.

$$\tilde{t}, \tilde{v} = \text{Para_Co-Attention}(T', V) \quad (5.11)$$

The two feature matrices T' and V obtained from textual and visual modalities are co-attended together to obtain the global feature representation of texts and images comprising the social media post.

5.3.1.3 Feature Fusion

The adopted fusion strategy considers the interaction among different modalities. The content-based post feature vector representation (\tilde{p}) of the social media post is obtained as shown in Equation 5.12.

$$\tilde{p} = \tilde{v} + \tilde{t} \quad (5.12)$$

Here, \tilde{p} denotes the content-based post feature vector representation. The global image feature vector \tilde{v} and the global text feature vector \tilde{t} are summed together to obtain \tilde{p} . This representation is then passed to the user preference mining module to generate plausible hashtags.

5.3.2 User Preference Mining

Social media users engage in diverse tagging practices. Users tend to create posts comprising texts, images and sometimes assign hashtags to their posts. Distinct users interpret the same hashtag in different ways. For hashtag recommender systems to suggest user-aware hashtags, it is critical for these systems to understand the user behavior and interaction with hashtags. These interactions act as clues for learning users' tagging preferences and provide crucial information for tailoring personalized hashtag suggestions. However, modeling user preferences in hashtag recommender systems has received minimal attention. We aim to identify the users' tagging pattern on their created posts as described in the following sections. We first randomly choose some posts from the current user's posting history. Then, we use these historical posts to learn and relate the tagging habits with the current post to be tagged.

5.3.2.1 Post Sampling

To improve user experience in hashtag recommendation systems, we attempt to model users' preferences from their historical posts and associated hashtags. The main idea is to learn about users' interests and model their tagging behavior by mining

information from previous posts. Post sampling is employed to select prior posts of users to learn their tagging behavior and hashtags usage style. We randomly sample L historical posts for each user to understand their tagging patterns. Since users may have created a limited number of posts, we limit L to a reasonably small value.

5.3.2.2 Hashtag Pattern Modeling

Social media users may spontaneously assign hashtags to their created posts; therefore, attaching hashtags to the user-generated content is a social behavior. It is challenging to automatically generate hashtags for social media content because the associated hashtags are related to user preference besides exhibiting relation to the content of the social media post. Motivated by the intuition that user tagging behavior should impact the recommendations, in this work, we attempt to model users' interests by incorporating information from their historical posts.

Our proposed method suggests hashtags based on the current post's content and users' tagging behavior. It considers the historical posts of the given user to learn user preferences and accordingly assign hashtags to his newly created posts. For modeling user behavior, we apply the techniques mentioned above to extract features from the given user's historical posts and then compare the current post p_i to his historical posts hp_l . Finally, the influence vector \tilde{u} is estimated by taking the weighted sum of the hashtags in the database, where the similarities between the posts determine the weights. Algorithm 5.2 shows the procedure for modeling user tagging behavior. Line 1 presents the equation to extract the post feature vector \tilde{p} for i^{th} post p_i . Lines 4-15 reveal the equations for obtaining hashtag attention and post similarity matrix for L historical posts. Line 5 presents the equation to extract features for the l^{th} historical post hp_l . Line 8 shows the equation to obtain hashtag embedding g_e of the particular hashtag h^g contained in the hashtag set of the l^{th} historical post i.e., hp_l^{hg} . BERT is used to embed the hashtags into low dimensional real-valued vectors. The hashtag embeddings g_e are stacked together to obtain a matrix of hashtag embeddings denoted by G^l for all hashtags appearing in the l^{th} historical post hp_l as depicted in Line 9. Line 11 indicates the hashtag attention mechanism adopted to generate a hashtag

Algorithm 5.2 Modeling User Tagging Behavior

Input: HP : Set of historical posts
 p_i : Test post of user $u_k \in U$
Output: \tilde{u} : Influence vector of user $u_k \in U$
function $\text{getInfluenceVector}(p_i, HP)$
 $\tilde{p} \leftarrow \text{FeatureExtraction}(p_i)$
 $S \leftarrow []$
 $\tilde{G} \leftarrow []$
 for $hp_l \in HP$ **do**
 $\tilde{hp}_l \leftarrow \text{FeatureExtraction}(hp_l)$
 $G^l \leftarrow []$
 for $h^g \in \tilde{hp}_l^{hg}$ **do**
 $g_e \leftarrow \text{Embedding}(h^g)$
 $G^l.append(g_e)$
 end for
 $\tilde{g}^l \leftarrow \text{Attention}(G^l)$
 $\tilde{G}.append(\tilde{g}^l)$
 $s^l \leftarrow \tanh(p \odot \tilde{hp}_l^f)$
 $S.append(s_l)$
 end for
 $a^s \leftarrow \text{softmax}(W_s^T \mathbf{S}^T + b_s)$
 $\tilde{u} \leftarrow \sum_{l=1}^L a_l^s \tilde{g}^l$
 return \tilde{u}

attention vector \tilde{g}^l which converts each hashtag set G^l into a single hashtag influence vector \tilde{g}^l . The attention mechanism mentioned in Line 11 can be summarised as given in Equations 5.13-5.15.

$$H^g = \tanh(W_g G^l) \quad (5.13)$$

$$a^g = \text{softmax}(W_{hg}^T H^g + b_{hg}) \quad (5.14)$$

$$\tilde{g}^l = \sum_{k=1}^{N_g} (a_k^g g_k^l); k = 1, 2, \dots, N_g \quad (5.15)$$

Here, N_g is the fixed length of hashtag sequence, $D = 768$, $W_g \in \mathbb{R}^{D \times D}$, $H^g \in \mathbb{R}^{D \times N_g}$, $W_{hg} \in \mathbb{R}^D$ and $\tilde{g}^l \in \mathbb{R}^D$; $l \in [1, 2, \dots, L]$. The hashtag attention vectors \tilde{g}^l are stacked together to obtain the hashtag attention matrix \tilde{G} , as depicted in Line 12. Here, $\tilde{G} \in \mathbb{R}^{L \times D}$ represents the hashtag attention matrix for L historical posts, $\tilde{g}^l \in \mathbb{R}^D$ indicates the hashtag attention vector corresponding to l^{th} historical post

hp_l . The equation in Line 13 is used to compute the similarity vector s_l by measuring the similarity of l^{th} historical post feature vector $\tilde{h}p_l$ with the current post feature vector \tilde{p} . Historical post features are neither pre-trained nor independently trained. Instead, they are trained with the current post features. The similarity vectors s_l are stacked together to obtain the similarity matrix S as shown in Line 14. We use the similarity matrix S to compute the attention weights denoted by a^s for each historical post hp_l as shown in Line 16. Here, $W^s \in \mathbb{R}^D$ and $b_s \in \mathbb{R}$ are parameters, $a^s \in \mathbb{R}^L$. Line 17 shows the computation of the influence vector \tilde{u} . Here, \tilde{u} is calculated by taking the weighted sum of attention weights of historical post denoted by a_l^s with the hashtag embedding vector \tilde{g}^l corresponding to l^{th} historical post hp_l . The hashtags are assigned weights on the basis of similarities between historical posts and the current post. The final feature vector q is obtained by concatenating these two feature vectors as follows.

$$q = \tilde{p} \oplus \tilde{u} \quad (5.16)$$

where, \oplus represents the concatenation operator, $\tilde{p} \in \mathbb{R}^D$ and $\tilde{u} \in \mathbb{R}^D$. The final feature vector q is then fed in the hashtag prediction module to recommend quality hashtags by considering the post's content and the user's tagging behavior. In the following section, we will go through the hashtag prediction module in detail.

5.3.3 Hashtag Prediction

This section discusses MLC and SG techniques to predict hashtags for the post p_i .

5.3.3.1 Multi-label Classification

We generate hashtags for the post p_i by formulating the hashtag recommendation task as a MLC problem. As a social media post can belong to several classes simultaneously, this technique helps to predict the mutually non-exclusive class labels. It assigns hashtags to the post from a pool of predefined hashtags $H = \{h_j^g\}_{j=1}^J$ where J is the cardinality of set H . Given the final feature vector q , we first use dense layer of size J and then a softmax activation function to obtain softmax scores of hashtags as

shown in Equation 5.17.

$$y_{pred} = (\text{softmax}(\text{Dense}(\text{units} = J))(q)) \quad (5.17)$$

Here, $y_{pred} \in \mathbb{R}^J$ represents the softmax scores of the predefined hashtags, J is the total number of hashtags in our dataset. We then sort the hashtags based on these scores to get the final set of predicted hashtags as given in Equation 5.18.

$$Rh_{MLC} = \text{argsort}(y_{pred}) \quad (5.18)$$

Here, argsort is used to get the corresponding indices of softmax scores sorted in descending order, and Rh_{MLC} denotes the hashtags predicted when hashtag recommendation is modeled as a Multi-Label Classification problem. The training objective loss function is given in Equation 5.19.

$$J = \frac{1}{|Z|} \sum_{(p_i, Gh_i) \in Z} \sum_{gh \in Gh_i} -\log(\text{Prob}(gh|p_i)) \quad (5.19)$$

Here, Z ($Z \subset P$) denotes the training post set, p_i and Gh_i represent the current post and corresponding hashtag set, and $\text{Prob}(gh|p_i)$ is the probability of choosing hashtag gh for the post p_i .

5.3.3.2 Sequence Generation

We adopt an encoder decoder-based model to formulate hashtag recommendation tasks in terms of sequence generation. The encoder extracts the visual, textual features and user features from a social media post. We then obtain its hybrid vector representation as mentioned in Equation 5.16. In this section, we first explain the decoder, then we highlight the procedure for hashtag generation and training of the SG framework. GRU is employed to model interrelationships between hashtags and multimodal data. The update gate is determines how much of previous state information to retain in the current state. In contrast, the reset gate limits the extent to

which previous hidden state information can be neglected. GRU makes use of these two gates to generate the next hidden state h_t conditioned on the previous hidden state h_{t-1} . We now discuss the hashtag generation procedure. Given a social media post p_i comprising image and text, the hybrid encoder in our proposed model first retrieves textual and visual feature vectors separately. The distributed feature vector representations are combined with a word-level and parallel co-attention mechanism. This representation is then combined with information mined from user's historical posts to obtain the overall post feature vector representation.

$$x_t = \text{Embedding}(hg_t^{inp}) \quad (5.20)$$

Here, x_t is the embedding of hashtag obtained at time step t . This embedding along with the post feature vector is fed into the GRU network to generate a hidden state vector h_t as follows:

$$h_t = \text{GRU}(x_t) \quad (5.21)$$

Subsequentl, we calculate probabilities of each hashtag at time step t by employing a dense layer and a softmax function as shown in Equation 5.22.

$$y_t = \text{softmax}(W_h^T h_t + b_h) \quad (5.22)$$

Finally, we employ greedy search to get the predicted hashtag hg_t^{pred} as follows:

$$hg_t^{pred} = \text{argmax}(y_t) \quad (5.23)$$

Since there is a possibility that the output hashtags may repeat, we filter out the redundant hashtags at each step. $Rh_{SG} = \{hg_i^{pred}\}_{i=1}^{N_g}$ is the final set of hashtags generated by the sequence generation technique.

In general, SG models utilize the previous time step's output hg_{t-1}^{pred} as the model's input at the current time step t . This is a typical approach in language models that output a single word at a particular point of time, where the current word is

dictated by the ones preceding it. During the early phases of training, the model's predictions are extremely poor. A series of incorrect predictions updates the model's hidden states, and the model finds it challenging to learn from this. This technique may result in slower convergence and model instability. Hence, we employ teacher forcing, an approach used to boost the model's learning capabilities. Teacher forcing is a strategy used for training recurrent neural networks in a fast and effective way by feeding the actual output at the current time step gh_t^i as input to the next time step hg_{t+1}^{inp} as shown in Equation 5.24, in place of the current output hg_t^{pred} produced by the network.

$$hg_{t+1}^{inp} = gh_t^i \quad (5.24)$$

When we employ teacher forcing, the model learns the statistical features quickly and predicts the correct sequence. Unlike the training procedure, we cannot access the ground truth hashtags during testing. Hence, the hashtag predicted at timestep t is fed as the input to the GRU unit in the next timestep, as shown in Equation 5.25.

$$hg_{t+1}^{inp} = hg_t^{pred} \quad (5.25)$$

Note that we use two special tokens [START] and [END]. These tokens signal the beginning and end of the hashtag sequence respectively. Our framework employs greedy search to recommend the sequence of relevant hashtags. The greedy search decoder generates the hashtag sequence by selectively choosing the most probable hashtags. Hashtags are ranked based on decreasing order of probabilities. At time step t , the greedy search algorithm selects the most probable hashtag. It is desirable to output the most probable hashtag at every step when adopting the generation framework for the recommendation task. This decoding technique suggests plausible hashtags for a given social media post. The training objective loss function is shown in Equation 5.26.

$$J = - \sum_{t=1}^c \log(Prob(gh_t|q, gh_1, \dots, gh_{t-1}; \theta)) \quad (5.26)$$

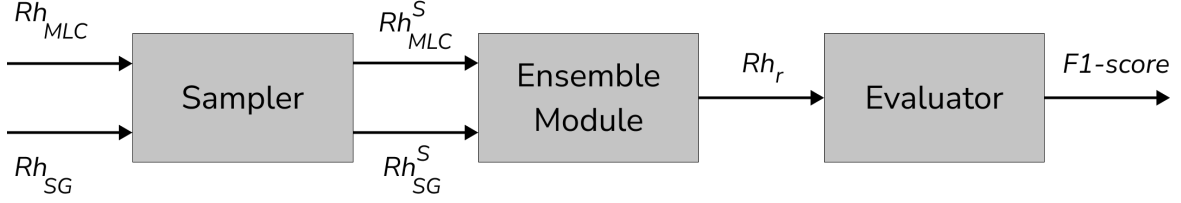


Figure 5.4: Candidate hashtag recommendation module

Here, $gh_{\{1,\dots,t\}} \in Gh_i$, where Gh_i corresponds to the ground-truth hashtag set for the current post p_i , gh_t denotes the ground-truth hashtag at time step t , q is the vector representation of the current post p_i , and θ denotes all the parameters of the SG component. The hashtags predicted by MLC and SG models are aggregated to recommend high-quality hashtags.

5.3.4 Candidate Hashtag Recommendation

To capitalize on the hashtag prediction techniques mentioned above, we devise a candidate hashtag recommendation module that will suggest relevant hashtags for the given social media post p_i . The candidate hashtag recommendation module is shown in Figure 5.4. It consists of 3 components: sampler, ensemble module, and evaluator.

5.3.4.1 Sampler

This component effectively samples the hashtags from MLC and SG models' predictions. We leverage the validation set for this procedure. Given the inputs from the validation set, the hashtags predicted by both MLC(Rh_{MLC}) and SG(Rh_{SG}) are fed into the sampler. It samples $s1$ hashtags from MLC and $s2$ hashtags from SG as follows:

$$Rh_{MLC}^S = \{Rh_{MLC}^s\}_{s=1}^{s1} \quad (5.27)$$

$$Rh_{SG}^S = \{Rh_{SG}^s\}_{s=1}^{s2} \quad (5.28)$$

where, $s1, s2 \in \{1, 2, \dots, 20\}$. The sampled hashtags are then fed into the ensemble module.

5.3.4.2 Ensemble Module

In the ensemble module, we aggregate the sampled hashtags which are provided by the sampler. The aggregation procedure is as follows: Firstly, we compute the intersection between hashtags sampled from both the procedures to retrieve common hashtags as given in Equation 5.29.

$$Rh_i = Rh_{MLC}^S \cap Rh_{SG}^S \quad (5.29)$$

Here, Rh_i represents the obtained set. We then obtain the unique hashtags from MLC set as shown in Equation 5.30.

$$Rh_{MLC}^U = Rh_{MLC}^S - Rh_{SG}^S \quad (5.30)$$

Here, Rh_{MLC}^U refers to those hashtags that are exclusively present in MLC but not in SG. Similarly, we obtain the unique hashtags from SG as shown in Equation 5.31.

$$Rh_{SG}^U = Rh_{SG}^S - Rh_{MLC}^S \quad (5.31)$$

Here, Rh_{SG}^U refers to those hashtags that are exclusively present in SG but not in MLC. Lastly, we compute the union of three sets as follows:

$$Rh_u = Rh_i \cup Rh_{MLC}^U \cup Rh_{SG}^U \quad (5.32)$$

Here, Rh_u denotes the combined set. From the combined set (Rh_u) we recommend top- K hashtags as given in Equation 5.33.

$$Rh_r = Rh_u[1, 2, \dots, K] \quad (5.33)$$

These hashtags (Rh_r) are then fed into the evaluator.

5.3.4.3 Evaluator

Given the ground truth hashtags(Gh) from the validation set and the hashtags generated by the ensemble module (Rh_r), we compute the F1-score as shown in Equation 5.34.

$$score = F1\text{-score}(Gh, Rh_r) \quad (5.34)$$

We keep track of the maximum F1-score generated for the validation set and the corresponding values for $s1$ and $s2$. This is explained in the following algorithm. Algorithm 5.3 shows the procedure to sample the number of hashtags from MLC and

Algorithm 5.3 Hashtag Sampling Algorithm

Input: *score*: F1-score generated by the evaluator
 s1, s2: Number of hashtags sampled by the sampler
 maxF1: maximum F1-score recorded, which is initialised to 0

Output: *maxF1, bestS1, bestS2*

function HashtagSampling(*score, s1, s2, maxF1*)

1: **if** *score* > *maxF1* **then**

2: *maxF1* \leftarrow *score*

3: *bestS1* \leftarrow *s1*

4: *bestS2* \leftarrow *s2*

5: **end if**

6: **return** *maxF1, bestS1, bestS2*

SG strategies i.e., $s1$ and $s2$ that yield the highest F1-score. Here, $bestS1$, $bestS2$ represent the values of $s1$ and $s2$ corresponding to the highest F1-score recorded. This procedure continues until all possible combinations of $s1$ and $s2$ are exhausted.

The task of obtaining values for $bestS1$ and $bestS2$ in order to maximise the F1-score can be thought of as a state space search problem, where the set of states is given by $\{(s1, s2) : 1 \leq s1, s2 \leq 20\}$. One simple technique would be to choose a value at random, however this method is inefficient. We computed our algorithm's efficacy, which can be defined as the ratio of the number of states with an F1 score smaller than $bestF1$ to the total number of possible states. We experimented with the above mentioned datasets and the algorithm achieved an efficiency of 98.25%, 91.3125% and

99.75% on MMP-INS, T-INS and HARRISON respectively. This justifies the relevance and effectiveness of the hashtag sampling algorithm. Finally, to recommend hashtags for the query post p_i , we sample $bestS1$ and $bestS2$ hashtags from the predictions of MLC and SG models, respectively as shown in Equations 5.35 and 5.36.

$$Rh_{MLC}^{Final} = \{Rh_{MLC}^s\}_{s=1}^{bestS1} \quad (5.35)$$

$$Rh_{SG}^{Final} = \{Rh_{SG}^s\}_{s=1}^{bestS2} \quad (5.36)$$

These hashtags are fed into the ensemble module to generate candidate hashtag recommendation as follows:

$$Rh_{pred} = Ensemble_Module(Rh_{MLC}^{Final}, Rh_{SG}^{Final}) \quad (5.37)$$

where, Rh_{pred} represents the recommended hashtags. The recommended hashtags capture not only the multimodal aspects of the current post but also the preferences of the user who created that post.

5.4 Experimental Evaluations

In this section, we first describe the experimental settings and then present the experimental results to show the effectiveness of our method.

5.4.1 Experimental Setup

In this section, we present the different datasets on which experiments have been carried out. Subsequently, we discuss the baseline methods for comparison followed by evaluation metrics.

5.4.1.1 Datasets

We perform hashtag recommendation on three different datasets, namely MMP-INS, HARRISON, and T-INS. MMP-INS is a multimodal personalized dataset from

Instagram. HARRISON is a publicly available benchmark dataset for image-based hashtag recommendation. T-INS is a text-based dataset that we have crawled from Instagram. We discuss these datasets in detail below.

5.4.1.1.1 MMP-INS In this chapter, we use Multi-Modal Personalised INSTagram dataset abbreviated as MMP-INS. This dataset was originally presented by Zhang et al. [5]. We have used a subset of the dataset after applying different pre-processing techniques. To pre-process the original dataset, we performed lemmatization on the hashtags appearing in the crawled posts. Next, we remove the low-frequency hashtags. The final dataset for the usage of our experiments contains 20,790 posts. There are a total of 3,636 unique hashtags, with an average of 6.95 hashtags per post. The minimum number of hashtags associated with a post is one whereas the maximum number of hashtags related to any post is 30. The dataset contains 3,153 unique users where a user has an average of 6.59 posts.

5.4.1.1.2 HARRISON HARRISON is a popular benchmark dataset for image-based hashtag recommendation. This dataset was created by Park et al. [115] in the year 2016 to recommend hashtags for Instagram photos. The raw dataset comprises 57,383 images and a mean of 4.5 hashtags per image. To pre-process the raw dataset, we first drop the low-frequency hashtags followed by removal of posts without hashtags. The final dataset for the usage in our experiments contains 36,428 images and an average number of 4.64 hashtags per image. The minimum number of hashtags associated with an image in HARRISON dataset is one, whereas the maximum number of hashtags for an image is 10.

5.4.1.1.3 TINS TINS is a novel dataset that we have created by crawling public posts from Instagram. We randomly selected 1,649 users and crawled an average of 15 posts per user. The collected dataset is cleaned to carry out the experiments. First, we lemmatize hashtags appearing in the crawled posts. Then, we remove the posts that do not contain any hashtags and retain the text-only portion of the collected posts. The resultant dataset contains 23,868 posts with only text and at least one

hashtag. It has 1,597 users, with an average of 14.94 posts per user and 9,780 unique hashtags with an average number of 12.13 hashtags per post. The minimum number of hashtags associated with a post in the resulting dataset is one whereas the maximum number of hashtags for a post is 149.

5.4.1.2 Compared Methods

To evaluate the effectiveness of the proposed model, we compare our method with the following methods for hashtag recommendation.

- Attention based Multimodal Neural Network(AMNN) [94]: The authors convert the hashtag recommendation task to a sequence generation problem. They adopt a sequence to sequence architecture with a softmax mechanism. The hybrid encoder decouples the feature extraction process of multimodal microblogs by separately retrieving the visual and textual features using CNN and BiLSTM. The attention mechanism is applied independently to visual and textual features to learn the most important parts of texts and images. These features are concatenated to obtain the overall post representation. GRU, which functions as a decoder, receives the combined representation of the post and generates the hashtag sequence based on the probability scores of hashtags.
- Image Attention (ImgAtt) [140]: ImgAtt was initially formulated for visual question answering. It uses a Stacked Attention Network (SAN) which comprises two attention layers that generate the visual attention distribution to pinpoint the most indicative regions to infer the answer. The first attention layer focuses on the portion of the image most relevant to the question. The second attention layer uses the fine-grained query vector representation obtained from the first attention layer to attend to the most relevant portions of image that correspond to the answer. Since it comprises both textual and visual modalities, this model can be easily adapted to recommend hashtags for multimodal social media posts.
- Co-Attention (CoA) [12]: Co-Attention is one of hashtag recommendation methods for multimodal posts. It converts the hashtag suggestion task into a multi-

label classification problem. The co-attention network generates text attention and image attention sequentially. Since it lays more emphasis on the textual information contained in the post, this method first computes text-guided visual attention. It uses the obtained representation to generate image-guided textual attention. This feature representation is passed into a single-layer softmax classifier to predict the hashtags.

- Memory Augmented Co-attention Model (MACoN) [5]: MACoN is a recent multimodal hashtag recommendation method. It adopts a parallel co-attention mechanism to extract textual and visual features from multimodal posts simultaneously. Since this method considers both image and text as equally important for tagging in the social media platforms, it generates the textual and visual attention co-guided by each other. It also learns the user’s tagging habits to make personalized recommendation.
- Triplet-Attention Graph Networks for Hashtag Recommendation (TAGNet) [30]: The authors construct a visual similarity graph considering that images that are similar are annotated with similar hashtags. The node features are computed using textual and user features to enhance the performance of hashtag recommendation. Triplet attention module is employed to incorporate the mutual influence of textual, visual and user features on each other. Aggregated graph convolution rule is used to disseminate information over the graph for predicting hashtags.

5.4.1.3 Evaluation Metrics

Hashtag recommendation methods have been designed to suggest a good quality of hashtags for a user’s post. To measure the effectiveness of these methods, we need to evaluate their performance. The parameters widely employed for assessing the performance of hashtag recommendation systems are hit rate, precision, recall, and F1-score. Hence, we have used these four evaluation metrics in this chapter. Let Rh denote the set of recommended hashtags, Gh represent the set of ground-

truth hashtags, and Ch represents the set of common hashtags between the top- K recommended hashtags(Rh) and ground-truth hashtags (Gh), i.e., $Ch = Rh \cap Gh$.

5.4.2 Experimental Results

We evaluate the proposed method by comparing its performance to the existing methods on different datasets, analyzing performance gain, visualizing the recommendation, analyzing the computation time of different models and identifying the sensitivity of various parameters.

5.4.2.1 Effectiveness Comparisons

We compare the performance of the proposed method with the existing methods on different datasets.

5.4.2.1.1 Performance on MMP-INS dataset To validate the effectiveness of our proposed model in hashtag recommendation task, we carry out its comparison with existing methods. Following the prior research in this line, we analyze the performance of various methods in terms of hit rate, precision, recall, and F1-score. The comparison results with existing methods on the MMP-INS dataset are presented in Table 5.1. Both CoA and ImgAtt consider the textual and visual modalities. Yet CoA yields superior performance as compared to ImgAtt. The poor performance of ImgAtt is ascribed to the fact that it was specifically designed to infer answers for text-based queries about an image instead of hashtag recommendation and that it does not implement a co-attention mechanism. Table 5.1 shows that our model outperforms the previous state-of-the-art approaches on the mentioned dataset significantly. As can be seen from Table 5.1, DESIGN achieves an absolute improvement of 42.5%, 20.4%, 25.8%, and 22.8% in terms of accuracy, precision, recall, and F1-score, respectively, over AMNN. The performance improvement is because our model considers the user’s historical posts, hashtagging history, and multimodal information. In contrast, AMNN considers only the multimodal information of the microblog. DESIGN comprises a novel personalized generative framework to generate the hash-

Technique	Hit rate	Precision	Recall	F1-score
AMNN	0.226	0.063	0.062	0.062
ImgAtt	0.286	0.074	0.074	0.074
CoA	0.411	0.122	0.125	0.124
MACoN	0.541	0.185	0.206	0.195
TAGNet	0.575	0.190	0.224	0.205
DESIGN	0.651	0.266	0.320	0.291

Table 5.1: Effectiveness comparison results on MMP-INS dataset

tag sequence, significantly boosting hashtag recommendation performance. DESIGN also employs classification based formulation of hashtag recommendation. In addition to that, DESIGN employs a word level attention on textual modality followed by a parallel co-attention on visual and textual modalities as opposed to self attention mechanisms employed by AMNN on textual and visual features. The improvement of DESIGN is 36.5%, 19.2%, 24.6%, 21.6% over ImgAtt, 23.9% 14.4%, 19.5% and 16.7% over CoA in terms of hit rate, precision, recall and F1-score respectively. Further, our model achieves an improvement of 11.0%, 8.1%, 11.4% ,9.6% in terms of hit rate, precision, recall and F1-score respectively over the MACoN model. Both DESIGN and MACoN have been designed to maximize the ability to correctly match the hashtags that users may assign to posts based on their tagging behavior. DESIGN achieves an improvement of 7.6%, 7.6%, 9.7% and 8.5% over TAGNet in terms of hit rate, precision, recall and F1-score respectively over the MACoN model. TAGNet considers the userid only as the user feature whereas DESIGN mines user tagging behavior from a user’s historical posts to recommend personalised hashtags for a new post created by the user. From Table 5.1, we can see that our proposed model, i.e., DESIGN performs substantially better than MACoN and TAGNet. The reasons for the improvement in performance are robust heterogeneous features, various attention mechanisms, and different approaches of MLC and SG. We determine the robust features using state-of-the-art textual and visual feature extractors. We employ two attention mechanisms. Firstly, we apply the word-level attention on text to effectively extract the important features from textual modality followed by a parallel co-attention mechanism to learn the joint feature representations of the visual and textual modalities in which the two

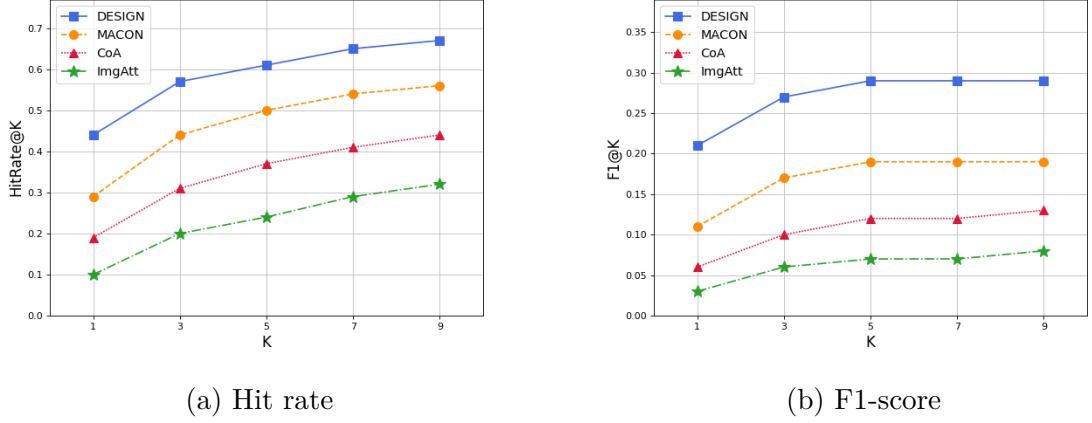


Figure 5.5: Effectiveness comparison curves on MMP-INS dataset

modalities co-guide each other. Besides that, we also model the user’s tagging behavior to learn his preferences. Our model incorporates both MLC and SG techniques that complement each other and predict good quality of hashtags.

Figure 5.5 shows the performance comparison of hashtag recommendation models in terms of hit rate and F1-score on MMP-INS. The x-axis indicates the number of hashtags recommended by different methods, and the y-axis represents the hit rate and F1-score, respectively. The number of recommended hashtags lie in the range of 1 to 9. As the number of recommended hashtags increases, hit rate increases. DESIGN model’s curves are always the highest in all metrics compared to the existing models, suggesting that our proposed model outperforms other models even when the number of recommended hashtags vary. Furthermore, the gaps in hit rate and F1-score curves are all widening. The significant improvements in all four metrics over the existing methods demonstrate our proposed model’s competitive advantage and efficacy.

5.4.2.1.2 Performance on HARRISON dataset In this section, we show the performance of different models on the publicly available HARRISON dataset. As HARRISON is an image-only dataset, when performing the experiments of MACoN, CoA, and ImgAtt on this dataset, we pad a special token to account for the missing text. We consider top- K hashtags to be recommended, where $K = 5$, as the average number of hashtags per image is 4.64. The comparison results of different methods on

Technique	Hit rate	Precision	Recall	F1-score
AMNN	0.125	0.027	0.035	0.030
ImgAtt	0.517	0.135	0.186	0.157
CoA	0.570	0.146	0.205	0.171
MACoN	0.605	0.160	0.223	0.186
TAGNet	0.612	0.161	0.225	0.187
DESIGN	0.634	0.179	0.247	0.208

Table 5.2: Effectiveness comparison results on HARRISON dataset

HARRISON dataset are shown in Table 5.2. As can be seen in Table 5.2, our proposed model achieves better performance than the existing models on the image-only dataset. DESIGN shows an absolute improvement of 50.9%, 15.2%, 21.2%, 17.7% over AMNN, 11.7%, 4.4%, 6.1% and 5.1% over ImgAtt, 6.5%, 3.3%, 4.2% and 3.7% over CoA, 3.0%, 1.9%, 2.4%, 2.2% over MACoN and 3.6%, 1.7%, 1.4%, 1.6% over TAGNet in terms of hit rate, precision, recall and F1-score respectively. The results show that our model outperforms existing methods on the dataset containing only images. One of the reasons for the improvement of DESIGN is that it formulates hashtag recommendation in terms of MLC and SG. DESIGN makes use of an effective visual feature extractor to capture information embedded in images, classification-based and generation-based approaches that assist in recommending relevant and correlated hashtags.

5.4.2.1.3 Performance on T-INS Dataset To validate the effectiveness of our proposed model on text-only hashtag recommendation, we compare it with different hashtag recommendation methods on the Text dataset from INStagram termed as T-INS. It can be observed from the given Table 5.3 that DESIGN outperforms the existing methods while recommending hashtags solely based on the text. All the results shown in Table 5.3 are at top- K . Here, the value of K is 12 as the average number of hashtags per image is 12.13. The proposed model shows a relative improvement of 50.3%, 27.9%, 29.0%, 28.6% over AMNN, 8.9%, 5.5%, 8.1%, 6.7% over ImgAtt, 5.3%, 2.4%, 3.0%, 2.7% over CoA, 1.0%, 1.5%, 2.3%, 1.9% over MACoN in terms of hit rate, precision, recall, and F1-score respectively. The results indicate the competitive advantage of our proposed model in text-based hashtag recommendation. The factors

Methods	Hit rate	Precision	Recall	F1-score
AMNN	0.172	0.056	0.078	0.065
ImgAtt	0.586	0.280	0.288	0.284
CoA	0.622	0.311	0.338	0.324
MACoN	0.664	0.320	0.346	0.332
DESIGN	0.675	0.335	0.368	0.351

Table 5.3: Effectiveness comparison results on T-INS dataset

responsible for DESIGN’s high performance in making text-only hashtag recommendation are the word-level attention and sequence generation technique. When applied to the textual modality, word-level attention retrieves the significant words appearing in the post’s textual content. The semantically relevant hashtags yielded by SG are further complemented by hashtags predicted from MLC, resulting in DESIGN predicting better hashtags.

5.4.2.2 Performance Gain Analysis

In this section, we analyze the performance gain of the proposed method. We first examine the performance of DESIGN with different modality combinations followed by different attention mechanisms.

5.4.2.2.1 Modality Combinations In this section, we analyze the performance of DESIGN with different modality combinations. We aim to recommend hashtags for social media posts by incorporating multiple modalities. The importance of each modality varies a lot. To distinguish the effect of different modalities in our suggested multi-modal approach, we conducted a micro-level study by taking different modality combinations. The inputs to our presented framework are threefold, i.e., images (i), texts (t), and user features (u). The performance comparison of different modality combinations is shown in Table 5.4.

In Table 5.4, DESIGN (t+i+u) represents our proposed DESIGN model. To demonstrate the usability of text input, we remove it while retaining image and user features. The resulting model is denoted as DESIGN (i+u). Similarly, DESIGN (t+u)

Methods	Hit rate	Precision	Recall	F1-score
DESIGN (t+i)	0.459	0.138	0.155	0.146
DESIGN (t+u)	0.518	0.206	0.220	0.213
DESIGN (i+u)	0.540	0.187	0.205	0.196
DESIGN (t+i+u)	0.651	0.266	0.320	0.291

Table 5.4: Performance of DESIGN with different modality combinations

takes text and user features as input. DESIGN (t+u) beats DESIGN (i+u) in terms of precision, recall and F1-score by 1.9%, 1.5% and 1.7% respectively. This can be attributed to the fact that we have adopted word-level attention on the textual modality, which effectively encodes the important information from the textual content of the post. Our proposed DESIGN framework comprising textual, visual modalities, and user preferences outperforms all the variants in terms of all performance metrics. DESIGN shows an improvement of 13.3%, 6.0%, 10.0%, 7.8% over DESIGN (t+u) and 11.1%, 7.9%, 11.5% and 9.5% over DESIGN (i+u) in terms of hit rate, precision, recall and F1-score respectively. Substantial improvement of DESIGN over its variants lies in utilizing the multimodal information contained in the post and incorporating user’s tagging behavior. The performance of the overall system improves by incorporating diverse modalities. The obtained results demonstrate the efficacy of our proposed multimodal approach, particularly in utilizing disparate modalities.

5.4.2.2.2 Attention Mechanisms In this section, we present the experimental analysis of the proposed model with different attention mechanisms. DESIGN-WAPCO, DESIGN-PCO, and DESIGN-WA are the variants of our proposed model by employing word-level and parallel co-attention, parallel co-attention, and word-level attention respectively. As DESIGN-WAPCO is our proposed model, we use terms DESIGN-WAPCO and DESIGN interchangeably.

- **DESIGN-WA:** DESIGN-WA refers to the implementation of DESIGN with word-level attention applied to the textual content of social media posts. This variant applies attention to the word embeddings obtained from the transformer-based BERT model to identify the essential words in the posts’ textual modal-

Methods	Hit rate	Precision	Recall	F1-score
DESIGN-WA	0.541	0.205	0.225	0.215
DESIGN-PCO	0.628	0.238	0.281	0.258
DESIGN-WAPCO	0.651	0.266	0.320	0.291

Table 5.5: Performance of DESIGN with different attention mechanisms

ity. Since, DESIGN-WA only models the interaction among words in the textual modality, it yields a hit rate of 54.14%, 20.54% precision, 22.50% recall, 21.48% F1-score, which are significantly lower than DESIGN-PCO and DESIGN-WAPCO.

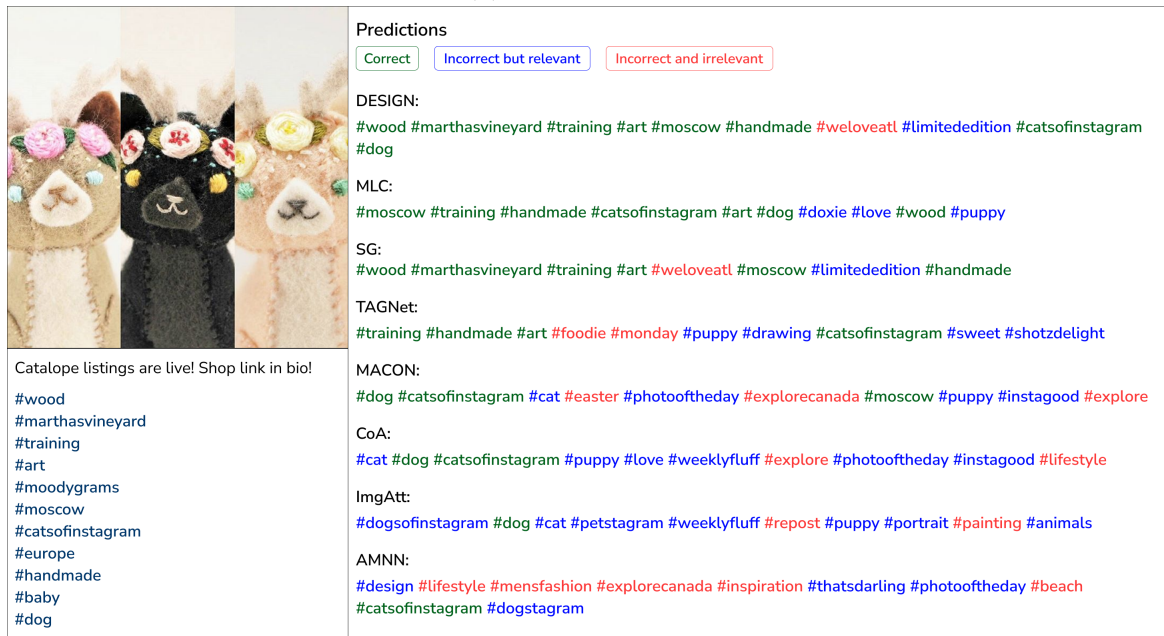
- **DESIGN-PCO:** DESIGN-PCO refers to implementing DESIGN with a parallel co-attention (PCO) mechanism. This mechanism learns not only the importance of different feature vectors in both modalities but also the influence of one modality on the other. It computes the similarity between visual and textual features for all combinations of image regions and text portions in order to attend to the image and text simultaneously. DESIGN model based on this attention mechanism achieves a hit rate of 62.80%, 23.75% precision, 28.11% recall, and F1-score of 25.75%.
- **DESIGN-WAPCO:** DESIGN-WAPCO refers to the implementation of DESIGN with a Word-level and Parallel Co-Attention mechanism. First, we apply word-level attention to the textual content of the posts. Then, we employ parallel co-attention on visual and textual modalities to model the interaction between the two. WAPCO captures the interaction of different words in the textual modality. Along with that, it models the interaction between textual and visual modalities. Our model utilizes the co-attended feature representation of social media posts to perform hashtag recommendation. Table 5.5 presents the performance comparison of DESIGN (or DESIGN-WAPCO) with different attention mechanisms.

DESIGN-WAPCO achieves an improvement of 10.9%, 6.1%, 9.5%, 7.6% over DESIGN-WA, 2.3%, 2.9%, 3.9%, and 3.3% over DESIGN-PCO in terms of hit rate,

precision, recall, and F1-score respectively. This improvement is due to the effectiveness of WAPCO mechanism. It not only captures important information within the textual modality but also models the interaction among the textual and visual modalities to learn the joint representation of social media posts. This representation is used for recommending relevant hashtags for social media posts.



(a) Historical posts



(b) Current post

Figure 5.6: Example post depicting hashtags recommended by different methods

5.4.2.3 Qualitative Analysis

The common evaluation protocol in hashtag recommendation methods is to assess effectiveness by accurately predicting hashtags. In this section, we present a qualitative analysis to assess how different methods recommend hashtags. To this end, we show a social media post as an example to investigate the hashtags recommended by different

models. In Figure 5.6, we first show the user’s historical posts together with text and hashtags given by the user for each historical post. The current post of that user along with hashtags recommended by different methods is shown below the user’s historical posts. The example post has been selected from the test data, with green indicating the correct hashtags, blue signifying the relevant hashtags, and red indicating the hashtags that are neither relevant nor correct. Here, correct refers to the recommended hashtags that match the ground-truth hashtags, and relevant refers to those that are coherent with the post’s content but not listed in ground-truth hashtags. The blue-colored hashtags demonstrate that DESIGN can identify some good hashtags, although these hashtags are not listed as ground-truth hashtags for that post.

As can be seen in Figure 5.6, DESIGN recommends content-based hashtags i.e., #catsofinstagram, #art, and #handmade. These hashtags are related to the current post’s content. We can see hashtags #marthasvineyard, #handmade, and #moscow appear in the predictions made by DESIGN and ground-truth hashtags. These hashtags help to visualize the capability of DESIGN in making personalized recommendations. The hashtags #marthasvineyard and #moscow are not related to the content of the current post. We can observe that these hashtags appear in both the historical posts of that user. These hashtags have been assigned to the current post, signaling that our model can recommend hashtags according to the tagging behavior and vocabulary choices of the user who has created the post. Similarly, we can see that hashtags #wood, #training, and #art have been assigned to the current post. These hashtags also appear in the second historical post. Hashtag #limitededition is recommended by SG component of DESIGN. It might be interpreted as relevant since the post shows some handmade articles which might be a limited creation. Hashtag #puppy can be considered relevant because the current post shows some handmade cats and #puppy could refer to the related pet category i.e., puppy. Similarly, hashtag #doxie which is predicted by MLC component of DESIGN, might be related to a dog. The number of green hashtags in DESIGN is higher than its components i.e., MLC and SG. MLC shows seven green hashtags and SG shows six green hashtags. Moreover, #weloveatl recommended by SG component is incorrect since it does not exhibit

relevance with the content or the user’s tagging behavior. MACoN recommends three green-colored hashtags (hashtags which appear in ground-truth hashtags) i.e., #catsofinstagram, #dog, #moscow, CoA recommends two correct hashtags #dog, and #catsofinstagram and ImgAtt recommends only one correct hashtag #dog. We can see that DESIGN recommends a higher number of correct hashtags compared to the existing methods.

5.4.2.4 Model Component Analysis

In this section, we discuss the different components of our proposed model. We first show two key components of DESIGN namely, MLC and SG. The experiments in this section have been conducted on MMP-INS dataset.

5.4.2.4.1 Multi-Label Classification and Sequence Generation DESIGN has two significant components namely, MLC and SG. Table 5.6 shows the performance of different components (modules) MLC, SG with DESIGN.

Multi-Label Classification (MLC): Multi-Label Classification is an extension of multiclass classification. In multiclass classification, the data sample can exclusively belong to one class. In contrast, there is no restriction on how many classes the data sample can be assigned to in multi-label classification. We formulate hashtag recommendation as MLC. The encoder comprises feature mining and user preference mining submodules. The decoder is a fully connected layer followed by softmax activation.

Sequence Generation (SG): To model interdependencies among hashtags, we attempt to solve hashtag recommendation in terms of sequence generation. We deploy a neural network based on encoder-decoder architecture. The encoder comprises feature mining and user preference mining submodules. The decoder consists of GRU units that generate hashtags for the social media posts in a sequence. We can observe from Table 5.6 that SG shows an improvement of 1.9%, 3.9%, and 2.8% over MLC in precision, recall, and F1-score, respectively. MLC treats hashtags as independent categories. It neglects the semantic relationship between hashtags. Hashtags for a particular piece of content are usually strongly correlated with each other. GRU units

Module	Hit rate	Precision	Recall	F1-score
MLC	0.600	0.219	0.244	0.231
SG	0.594	0.238	0.283	0.258
DESIGN	0.651	0.266	0.320	0.291

Table 5.6: Performance Comparison of Modules

in the decoder section of SG predict hashtags in the form of a sequence. It can be seen from Table 5.6 that DESIGN achieves better performance than MLC and SG. DESIGN shows an improvement of 4.8%, 4.7%, 7.6%, 6.0% over MLC and 5.7%, 2.8%, 3.7% and 3.2% over SG in terms of hit rate, precision, recall and F1-score respectively. This improvement is due to the integration of different recommended hashtags that complement each other. MLC recommends hashtags from a predefined set of classes, whereas SG recommends a semantically correlated sequence of hashtags.

5.5 Conclusion

In this chapter, we present a hybrid deep neural network for multimodal personalized hashtag recommendation. Our method is built upon encoder-decoder architecture. The encoder’s feature mining module extracts features from visual and textual modalities. We employ a word-level and parallel co-attention mechanism to coherently learn textual and visual features in order to obtain a richer post representation. User preference mining module utilizes users’ historical posts to learn their tagging behavior and models its influence in assigning hashtags to a newly created post. The hybrid decoder consists of two neural networks that simulate the task as MLC and Sequence SG paradigms. We design a hybrid module to capitalize on the hashtags predicted by MLC and SG in order to recommend a final plausible set of hashtags. The set of recommended hashtags not only captures the contextual information from the post’s content but also follows the user’s tagging behaviour. Experiments are conducted on multiple social media datasets containing visual, textual, and user information. Experimental results show that the proposed method achieves superior performance to existing methods.

Chapter 6

Hashtag Recommendation for Micro-videos

6.1 Introduction

This chapter addresses the critical research problem of automated hashtag recommendation, specifically tailored for micro-video content. In micro-videos, which are characterized by their short duration and ease of consumption, the challenge of effective content management and retrieval is particularly pronounced due to the sheer volume of daily uploads. For instance, Instagram alone sees approximately 95 million photos and videos posted daily¹. Hashtags [141] serve as vital metadata for categorizing and accessing these micro-videos, facilitating efficient search and user discovery. However, a substantial portion of micro-videos, estimated at 65% [4], remains untagged, significantly impeding information access. This research aims to contribute to the field by investigating novel approaches for automated hashtag recommendation that account for the specific characteristics of micro-video content and user interaction patterns. Addressing this gap has significant implications for enhancing content discoverability, improving user experience, and optimizing platform efficiency within the dynamic landscape of short-form video.

Numerous methods have been proposed for micro-video hashtag recommendation that leverage features such as content [4, 62, 117, 118], sentiment [63], user metadata

¹<https://www.marketingscoop.com/blog/how-many-posts-are-posted-on-instagram-per-day/>

[74] and user history [18, 73]. Li et al. [73] derived features from a user’s historical micro-videos and hashtags and combined them using average pooling to generate a user representation. However, this approach treats all micro-videos and hashtags equally, obscuring users’ tagging preferences for specific modalities. We et al. [18] constructed a heterogenous graph comprising users, hashtags, and micro-videos. Each user is connected to hashtags they used and micro-videos they uploaded, creating a rich network of interactions. Although user-to-micro-video and user-to-hashtag edges capture the overall theme of a micro-video and provide information about user’s general interests, they lack the granularity and contextual understanding necessary to capture a user’s fine-grained preferences. These edges inform us about the user’s interests but not necessarily how they prefer to express themselves and engage with those interests. These approaches fail to account for the interplay between user preferences and the specific modalities within each micro-video. Users upload micro-videos that reflect their interests, and their assigned hashtags provide valuable clues about their specific focus within those micro-videos. A user may be interested in the visual appeal of one micro-video, while the same user may be drawn to soundtrack or narrative elements in another micro-video. Furthermore, hashtag usage patterns can vary. When tagging, some users consistently emphasize visual elements, while others might switch between acoustic and textual elements depending on what appeals them in the micro-video. **Challenge 1:** How to capture user’s modality-specific tagging preferences on micro-videos?

Existing collaborative filtering approaches [142, 143, 144] for hashtag recommendation rely on features such as shared topics [145], hashtag usage [143], time [146], and social network information such as followers [147] and mentions [145] to find like-minded users for a given user and provide recommendations. While effective for established users, these methods struggle with cold-start users who lack historical data and social connections. Content-based methods offer a partial solution by analysing the content of the micro-video. While valuable for understanding the topical relevance of hashtags, content-based methods struggle to capture the dynamic social trends and community preferences crucial for effective hashtag recommendation. This is espe-

cially true on fast-paced platforms where popular hashtags and user interests evolve rapidly, potentially leading to recommendations that lack social resonance and engagement. Hashtags, however are not mere content descriptors, they also serve as tools for boosting social engagement and content visibility [148]. Studies have shown that tweets with hashtags receive twice the level of engagement than those without [5]. Moreover, hashtags used by popular users tend to be popular and impactful [28]. As such, cold-start users, seeking recognition and community integration, are naturally inclined to adopt hashtags used by popular and influential users. **Challenge 2:** How to recommend context-aware and popular hashtags for micro-videos posted by cold-start users, thus increasing their visibility and reach?

Prior research studies have leveraged user metadata [74] and historical information [18, 73] to recommend personalized hashtags for micro-videos. Although Wei et al. [18] emphasized personal preferences of a user by modeling user to hashtag and user to micro-video interaction, it neglects the user-to-user interactions within the broader social network. These interactions reveal community trends, hashtag usage dynamics within specific communities and aid in identifying like-minded users. Incorporating the behavior of similar users enhances personalization beyond past behavior, enabling the discovery of relevant hashtags that users might not have explicitly engaged with but are likely to find relevant. Prior research has demonstrated that visually similar images often share common hashtags [30]. Extending this principle to micro-videos, we recognize that micro-videos with similar visual characteristics, music genres, or textual content are likely to appeal to the same audience and thus can be assigned similar hashtags, even if the overall themes differ. Capturing similarities within each modality (visual, acoustic, and textual) can uncover valuable connections between micro-videos, leading to refined hashtag recommendations. User-user interactions capture explicit collaborative signals, such as shared interests and preferences based on past hashtag usage. Meanwhile, modality-modality interactions can reveal implicit collaborative signals, suggesting underlying connections and themes across different content types. Existing hashtag recommenders for micro-videos based on content and personalisation, despite their effectiveness, underutilize user-to-user interactions and modality to

modality, limiting their ability to capture the rich social dynamics and content similarity that shape hashtag usage and influence recommendations. **Challenge 3:** How to incorporate user-user and modality-modality interactions to further enhance hashtag recommendations for micro-videos?

In response to the above mentioned challenges, we propose a hybrid filtering graph-based deep neural network for **MI**cro-video **haSH**tag recommendati**ON**, i.e., **MIS-HON**. To tackle **Challenge 1**, we construct a heterogeneous graph comprising micro-video modalities and users as nodes. We connect the user to the constituent modalities of his historical micro-videos which enables to capture fine-grained modality-specific preferences that align with each user’s unique taste and creative expression. The node representations are refined by leveraging the message passing strategy. The enriched micro-video representation thus derived is utilized to recommend pertinent hashtags for micro-videos. To tackle **Challenge 2**, we introduce a hybrid system that integrates content analysis with social influence. By emulating tagging patterns of popular users while maintaining relevance to the user’s content, our approach empowers recommends popular hashtags to cold-start users aiding them to expand their network and engagement within the community. To tackle **Challenge 3**, we capture user-user interactions based on shared hashtags, and modality-modality interactions based on modality similarity. Furthermore, experiments conducted on three real-world datasets show the encouraging performance of our proposed framework. Our proposed model can recommend relevant hashtags and has a significant gain in performance.

We present the notable contributions of our work below.

- We present a novel hybrid filtering approach that leverages micro-video content, user’s modality-specific tagging preferences and community interests to facilitate context-aware, user-aware as well as community-aware hashtag recommendations.
- We model users’ modality-specific tagging preferences by linking them to the constituent modalities of their previously tagged micro-videos, enabling more personalized hashtag recommendations.

- We tackle the cold-start user problem with a hybrid approach, combining the strengths of content-based filtering and social influence. This strategy analyzes micro-video content and observes popular user hashtag trends, generating initial recommendations that exhibit topical relevance with community engagement potential.
- We construct a heterogeneous graph encompassing user-to-user and modality-to-modality interactions to capture explicit and implicit collaborative signals.
- Extensive experiments performed on three real-world datasets demonstrate the competitive advantage of the proposed framework against the state-of-the-art methods, as demonstrated through quantitative metrics and qualitative analysis.

The remainder of the chapter is organized as follows. Section 6.2 defines our problem setting and formulation. In Section 6.3, we present our technique. Section 6.4 then shows the experimental evaluations. Finally, in Section 6.5, we conclude our work.

6.2 Problem Definition

Consider a social media dataset D with the following attributes: a micro-video set $M = \{mv_i\}_{i=1}^{|M|}$, a hashtag set $H = \{hg_j\}_{j=1}^{|H|}$, and a user set $U = \{u_k\}_{k=1}^{|U|}$. Given a micro-video post (mv_i) uploaded by a user (u_k), we seek to automatically recommend a collection of hashtags $R = \{rh_r\}_{r=1}^{|R|}$ that is credible and corresponds to the set of ground-truth hashtags $G = \{gh_g\}_{g=1}^{|G|}$.

Here, $|\cdot|$ stands for the cardinality of a set. The variables i , j , and k serve as indices for the micro-video post, hashtag, and user correspondingly, while r and g function as indices for the recommended and ground-truth hashtags. Hashtag recommendation for an existing user and the cold-start user is defined as follows.

Problem 1 (*Hashtag Recommendation for an Existing User*) Given a new micro-video post (mv_{L+1}), where $mv_{L+1} \in M$ created by a user (u_k) who posted L micro-videos in the past, we intend to suggest appropriate hashtags for the new micro-video post (mv_{L+1}) automatically by using collaborative signals and user’s modality-specific

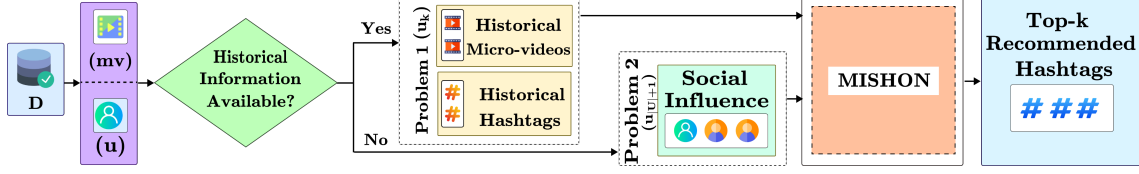


Figure 6.1: Visual representation of problem definition

preferences. Hashtags function as abstract labels that represent the information contained in each modality. Given an existing user, we aim to recommend a plausible set of hashtags for a new micro-video posted by that user. To this end, we model the user’s modality-specific tagging preferences (Challenge 1) along with user to user and modality to modality interactions thereby, capturing explicit and implicit collaborative signals (Challenge 3).

Problem 2 (*Hashtag Recommendation for a Cold-start User*) *Given a target micro-video post (mv_i) created by a cold-start user $(u_{|U|+1})$, our aim is to automatically recommend a relevant set of hashtags for the micro-video posted by that user.*

Here, we solve the cold-start problem inherent in hashtag recommendation systems by recommending appropriate hashtags for a new micro-video (mv_i) created by a cold-start user $(u_{|U|+1})$. To this end, we devise a social influence and content-based technique to recommend contextually relevant as well as popular hashtags, thereby empowering them to broaden their network and content visibility (Challenge 2).

The above-mentioned problems have been pictorially depicted in Figure 6.1.

6.3 Methodology

In this section, we elucidate our proposed approach. We propose an integrated model that incorporates micro-videos and users to tackle the task of micro-video hashtag recommendation. We begin with a high-level overview of our system as shown in Figure 6.2 before delving into its components. The input to our system is a micro-video post and the corresponding user who created that micro-video post. The micro-video post is divided into its constituent modalities. The modality-specific features of the micro-video are extracted through respective feature extractors. We then employ an

attention mechanism on the modality-specific features to find information that is most apt to recommend hashtags. We enhance micro-video representation by constructing an interaction graph that comprises four types of nodes and seven types of edges to capture modality-to-modality, user-to-user, and user-to-modality interactions. The

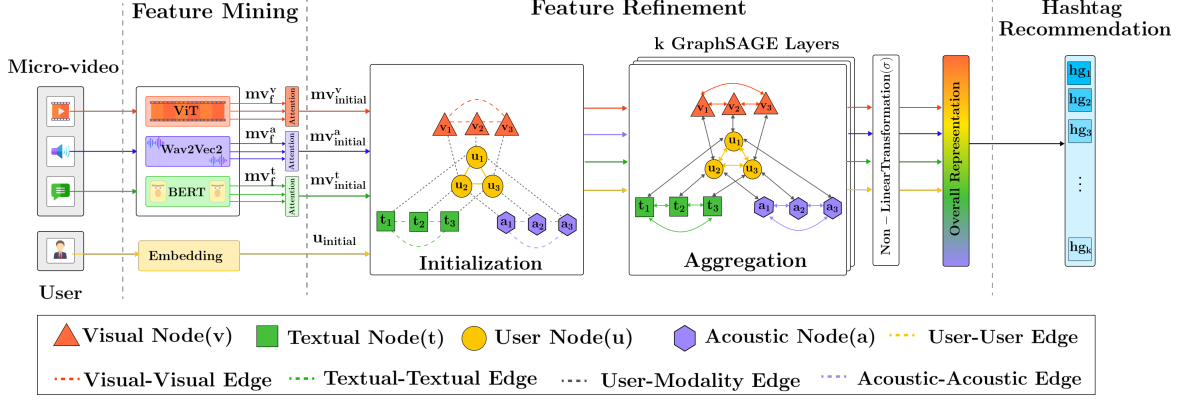


Figure 6.2: Overall architecture of MISHON

initial node embeddings are updated based on information propagation and neighborhood aggregation. The overall representation thus obtained, is fed into the hashtag recommendation module as input. After taking into account the likelihood of each hashtag, the hashtag suggestion module produces a sorted list of “top-K” hashtags. As demonstrated in Figure 6.2, our proposed framework comprises three components: (a) feature mining (b) feature refinement, and (c) hashtag recommendation. Below, we go through each component in further detail.

6.3.1 Feature Mining

This section describes the feature mining module that is made up of two submodules: (a) feature extraction and (b) attention modeling. We first extract features of modalities constituting the micro-video. We endeavor to enrich modality-specific representations followed by an attention mechanism to filter out noisy information. We describe the details of each submodule below.

6.3.1.1 Feature Extraction

In this, we elaborate on details of feature extraction from modalities constituting the micro-video. We first segment every micro-video into visual, acoustic, and textual modalities denoted by mv_i^v , mv_i^a , and mv_i^t respectively. We then retrieve features corresponding to each modality.

Visual Feature Extraction: Micro-videos are only a few seconds long. Due to the concise nature of micro-videos, a limited number of keyframes can effectively encapsulate the entirety of the visual content. To extract the micro-video frames, we use FFmpeg² and extract 12 frames for each micro-video at equally spaced intervals. We employ Vision Transformer [149] to derive visual attributes for every frame of the micro-video. Vision Transformer, abbreviated as ViT, is a variant of the language-based transformer model that takes an image as an input, uses the image structure to learn meaningful embeddings, and performs image classification. We employ the basic Vision Transformer architecture with a 16×16 input patch size for the frame feature extraction of every micro-video. We rescale every frame to meet the model’s requirements for input size. ViT creates a grid of square patches to split the frame. The channels of all pixels in a patch are concatenated, and the patch is then linearly projected to the chosen input dimension, flattening each patch into a single vector. Since transformers do not take into account the input element’s structure, therefore we give each patch learnable position embeddings so the model can pick up on the image’s structure. There are 12 attention modules in total. It is important to note that ViT does not contain any convolutional layers.

Given a sequence of frames representing the visual modality mv_i^v of the micro-video mv_i , we employ the pretrained ViT to extract frame features. We regard the penultimate layer of ViT to obtain visual features.

$$mv_f^v = ViT(mv_i^v) \quad (6.1)$$

Here, $mv_f^v \in \mathbb{R}^{N_f \times D}$ represents the resultant visual feature matrix, where $N_f = 12$

²<https://www.ffmpeg.org/>

stands for the number of frames and $D = 768$ for the concealed size of each frame.

Acoustic Feature Extraction: The acoustic modality, as a crucial supplement to the visual modality, is especially effective when the visual content is too diversified or conveys inadequate information. To capture the acoustic characteristics, we separate the audio channel from the video and subsequently divide it into equidistant segments of uniform duration using FFmpeg. Following that, we employ wav2vec2.0 [150] to extract features for each audio clip. Wav2vec2.0, a self-supervised speech representation model, aims to capture essential characteristics of unprocessed audio files by harnessing the strength of transformers and contrastive learning. This method is comparable to the masked language modeling used in Bidirectional Encoder Representation from Transformer, abbreviated as BERT [113]. Wav2vec2.0 can obtain high-level contextual representation and learn basic units for less labeled data. There are two stages to wav2vec2.0 training procedure: in the first stage, the model is trained on hundreds of unlabelled data, and in the second stage, it is fine-tuned on a small dataset for certain tasks. The wav2vec2.0 model has the following components: convolutional layers that turn the raw waveform input into latent representation (Z); transformer layers that produce contextualized representation (C) and linear projection produces the output (Y). Wav2vec2.0 uses a multilayer CNN to extract latent audio representations of 25 ms each from raw audio data. For feature extraction and selection, the representations are contained in a quantizer and a transformer. Gumbel and K-means are used to quantify data. Every 20 ms, the wav2vec2.0 toolkit extracts a 768-dimensional feature vector from the voice stream for a certain encoder layer. Each layer produces a new representation, which can vary in suitability for a job than a preceding or succeeding layer. We convert the raw audio from .mp3 to .wav format to satisfy the model’s input requirements, and the sampling rate is preserved at 16,000 Hz. We employ the base version of the wav2vec2.0 model that is pre-trained on 960 hours of unlabelled speech from the LibriSpeech [151] corpus. Given the acoustic modality of the micro-video denoted by mv_i^a , we apply wav2vec2.0 to extract acoustic features.

$$mv_f^a = Wav2vec2.0(mv_i^a) \quad (6.2)$$

Here, mv_f^a represents the resultant acoustic feature matrix where $mv_f^a \in \mathbb{R}^{299 \times D}$ and $D = 768$ represents the embedding size. We chose to extract acoustic features from the penultimate layer. We obtain a 768-dimensional feature vector for the entire audio segment of a given micro-video.

Textual Feature Extraction: Textual descriptions play a pivotal role in providing information about the micro-video post from a different perspective. Textual modality has established its importance for hashtag recommendation as demonstrated by previous works [28, 152, 153, 154]. The text encoder generates final text representations from the natural language sentence i.e., a textual description of the micro-video post. For the text encoder, we employ a Transformer-based model i.e., BERT. For the textual modality of micro-video denoted by mv_i^t which comprises a word sequence, we add two tokens: class (CLS) and separator (SEP) that mark the start and end of the input text, respectively. We generate the corresponding set of tokens T using BERT tokenizer.

$$T = BERT_Tokenizer(mv_i^t) \quad (6.3)$$

We set a 30-token cap on the length of the token sequence S . For textual descriptions less than S , we apply padding, otherwise, we perform truncation to make all textual descriptions of uniform size. Finally, we create token-based embeddings using BERT, as depicted in Equation 6.4.

$$mv_f^t = BERT(T) \quad (6.4)$$

The final textual feature matrix $mv_f^t \in \mathbb{R}^{S \times D}$, where $S = 30$ denotes the maximum length of the associated text for the micro-video post, and $D = 768$ denotes the embedding size.

6.3.1.2 Attention Modeling

Hashtags are typically used to emphasize significant information in micro-videos. As a result, eliminating noisy information and determining the importance of each unit constituting the modality-specific representation is critical in the hashtag suggestion task. Due to the effectiveness of attention mechanism [155], we apply it individually

on every modality as given in Equation 6.5.

$$mv_{initial}^m = Attention(mv_f^m) \quad (6.5)$$

The enriched modality-specific feature vector in this instance, $mv_{initial}^m$, was acquired via an attention method. The modality-specific embedding can be thought of as a sequence of feature vectors as shown in Equation 6.6.

$$mv_f^m = \{mv_x^m\}_{x=1}^X \quad (6.6)$$

where X denotes the number of units in every modality. We feed each unit mv_x^m to MLP to get h_x^m as a hidden representation of mv_x^m , as shown in Equation 6.7. By assigning an attention weight to every unit in each modality, we explicitly represent its varied relevance. To create an enriched representation of the constituent modality, we extract key units in each medium and combine the resultant unit representations.

$$h_x^m = tanh(Wmv_x^m + b_w) \quad (6.7)$$

Here, h_x^m is the concealed representation of mv_x^m . We compute each unit's relevance (α^x) as shown in Equation 6.8.

$$\alpha^x = softmax\left((h_x^m)^T u_w\right) \quad (6.8)$$

In this case, α^x symbolizes the importance of a unit while u_w denotes the context vector. To obtain the standardized coefficient (α^x), we initially calculate the resemblance between h_x^m and the contextual vector (u_w). We subsequently subject the result to a softmax function for normalization. The enriched modality-specific feature vector is then calculated, as shown in Equation 6.9.

$$mv_{initial}^m = \sum_{x=1}^X \alpha^x h_x^m \quad (6.9)$$

Here, $mv_{initial}^m$ denotes the enriched modality-specific feature vector which is obtained by aggregating annotations using the coefficients α^x . Further, the user-to-modality and user-to-user interactions can help obtain a better micro-video representation. To facilitate the learning of these interactions, we employ a graph neural network to refine modality-specific and user representations.

6.3.2 Feature Refinement

In this section, we elaborate on the feature refinement module. It consists of three steps namely: (1) graph construction; (2) information propagation and neighborhood aggregation; and (3) micro-video representation. We discuss these steps below.

6.3.2.1 Graph Construction

We construct an undirected graph $G = (N, E)$ as illustrated in Figure 6.3, where N and E denote the collection of vertices and edges, respectively. The total number of nodes in the graph is I where $I = 3M + U$ and $E \subset I \times I$ is a set of relationships containing their interdependencies. We further elaborate on graph construction in the following sections. *Node Settings:* We construct a graph with four different kinds of nodes as shown below.

$$N = V \cup A \cup T \cup U \quad (6.10)$$

Specifically, N comprises four different types of entities: V , A , T , and U , where V , A , and T represent the set of visual, acoustic, and textual modalities constituting the micro-videos contained in the micro-video set (M), and U represents the set of corresponding users. The micro-video (mv) is represented by three nodes v, a, t corresponding to three modalities, initialized with $mv_{initial}^v, mv_{initial}^a, mv_{initial}^t$ respectively. The user who created micro-video is considered the fourth type of node in the graph. The user id is transformed into a fixed-size vector representation $u_{initial} \in \mathbb{R}^D$, which is randomly initialized and refined throughout the training.

Edge Generation: When two nodes n_i and n_j interact, an edge $e_{ij} = (n_i, n_j) \in E$ is formed to link two nodes in the network. To exploit dependency amongst different

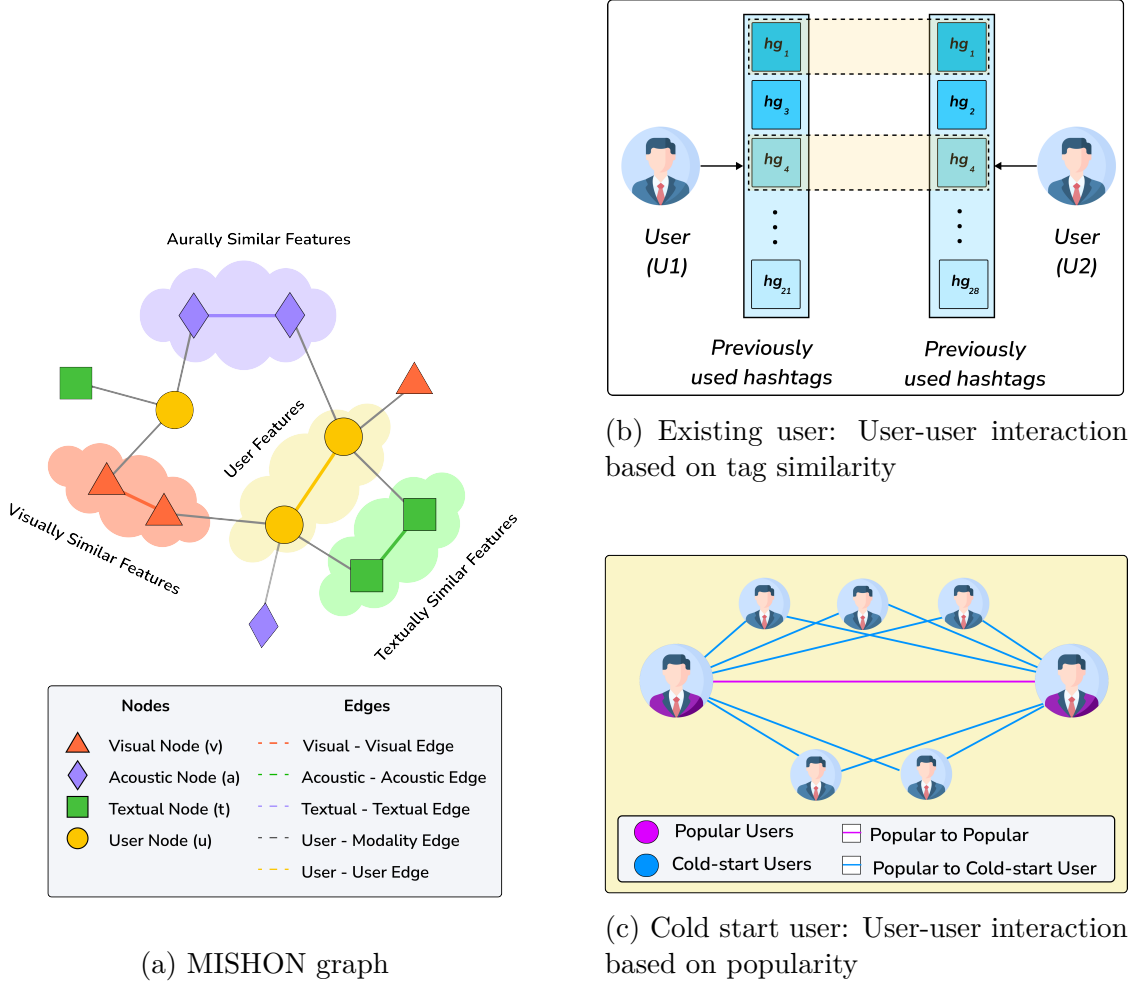


Figure 6.3: Graph construction in MISHON

kinds of nodes, we consider homogeneous and heterogenous edges. To differentiate them, we use distinct edge weighting strategies.

(i) **Homogeneous Edges:** These edges connect the same type of nodes. There are four types of interactions: video-to-video, audio-to-audio, text-to-text referred to as modality-modality edges, and user-user edges. We create weighted edges to link two nodes in the graph and set a threshold to filter out low-weighted edges. We discuss edge construction criteria for modeling modality-to-modality and user-to-user interaction below.

- **Modality-Modality Edges:** We create edges to connect nodes from the same modality of different micro-videos as shown in Figure 6.3. Given the modality-specific feature representations of two micro-videos denoted by $(mv_f^m)_i$ and

$(mv_f^m)_j$, where $m = \{v, a, t\}$ is the modality indicator for micro-video (mv) , the edge weight $e(m_i, m_j)$ is assigned as shown in Figure 6.11.

$$sim(m_i, m_j) = cs((mv_f^m)_i, (mv_f^m)_j) \quad (6.11)$$

Here, $sim(m_i, m_j)$ refers to the similarity score between the same modality of different micro-videos, $cs((mv_f^m)_i, (mv_f^m)_j)$ is the cosine similarity value between m^{th} modality-specific feature representations of i^{th} and j^{th} micro-videos. We create intramodality edges as exhibited in Figure 6.12.

$$e(m_i, m_j) = \begin{cases} sim(m_i, m_j), & \text{if } sim(m_i, m_j) \geq \theta \\ 0, & \text{if } sim(m_i, m_j) < \theta \end{cases} \quad (6.12)$$

Here, $e(m_i, m_j)$ refers to the weight assigned to edges connecting the same modality of different micro-videos denoted by m_i and m_j and θ refers to the threshold. We use cosine similarity value to assign weight to the edge between two nodes of the same modality. We assume that if two nodes have higher similarity, they contain rich information. To this end, we set threshold θ to 0.5 to filter out edges with low semantic similarity.

- **User-User Edges:** Under user-to-user interaction, we first discuss the correlation of an existing user followed by a cold-start user with other users on that micro-video sharing platform.

Existing User: The user who has already posted micro-videos on the video-sharing platform is considered an existing user. We model co-occurrence relationships among existing users based on common historical hashtags as depicted in Figure 6.3. The historical hashtag set of an existing user comprises all hashtags used by him/her in previously posted micro-videos. We take collaborative filtering into account which assumes that users who have had similar interests in the past will have similar interests in the future. Users with similar interests tend to assign similar hashtags to their micro-videos since hashtags reflect user

preferences from different granularities. The tagging behavior of each user is hidden in user co-occurrence relationships. We compute the similarity among users as illustrated below.

$$sim(u_i, u_j) = |H_i \cap H_j| / |H_i \cup H_j| \quad (6.13)$$

Here, $sim(u_i, u_j)$ is the similarity score of two users, \cap , \cup , and $|\cdot|$ denotes intersection, union operators, and set cardinality. H_i and H_j where $H_i \subset H$ and $H_j \subset H$ is the set of historical hashtags of user u_i and user u_j respectively. The numerator in Equation 6.13 denotes the number of common hashtags of two different users for their uploaded micro-videos and the denominator denotes the total number of hashtags contained in the set of their historical hashtags. The interaction modeling between two users is carried out as depicted in Equation 6.14.

$$e(u_i, u_j) = \begin{cases} sim(u_i, u_j), & \text{if } sim(u_i, u_j) \geq \gamma \\ 0, & \text{if } sim(u_i, u_j) < \gamma \end{cases} \quad (6.14)$$

Here, $e(u_i, u_j)$ denotes the weight assigned to the edge connecting two different users i.e., u_i and u_j . We assign an edge weight to model the degree of relatedness of users. We also assign a threshold γ to filter out edges with a low weight. Here, the threshold γ is set to 0.5.

Cold-start User: Users who are new to the system and lack any historical and social network information are called cold-start users. The user's historical interactions contain his interests based on which recommendations can be made. However, such interactions are often sparse, leading to cold-start user problems where a user has no historical posts and hashtags. Owing to the unavailability of user history on posted micro-videos and hashtags used, we employ a social influence technique to model the interaction of cold-start users with other users on that platform. People having high popularity are conceived as influential and more credible. The hashtagging patterns of influential users are mimicked by other users to garner social attention. We devise Algorithm 6.1 to construct

user-user edges and associate cold-start users with popular users as shown in Figure 6.3, assuming that cold-start users tend to utilize hashtags as used by the most popular users to increase their content’s visibility and garner attention from other users on that platform. To determine a user’s popularity, we compute the engagement rate (Line 2), which is the ratio of the total number of likes on the user’s profile to his total number of followers. For a cold-start user, we set this value to 0. We then sort these users based on their engagement rate (Line 4), and only the top 10% (Line 5, 6) are considered popular. Here, *argsort()* returns the corresponding users with engagement rates sorted in descending order. Edges are constructed between popular users and cold-start user (Lines 8-13). The final set of obtained edges is denoted by (E_{user}) as shown in Line 14.

Algorithm 6.1 Addressing cold-start user problem

Input: U : List of users
 α : Popular user selection ratio
 Metadata of user $u_i \in U$:
 Number of likes (l_i)
 Number of followers (f_i)

Output: User-user edges (E_{user})

```

1: for  $i = 1$  to  $|U|$  do
2:    $Er_i = l_i / f_i$ 
3: end for
4:  $users\_sorted = argsort(Er)$ 
5:  $top\_p = \alpha * |U|$ 
6:  $popular\_users = users\_sorted[1 \dots top\_p]$ 
7:  $E_{user} = []$ 
8: for  $pu_i \in popular\_users$  do
9:   for  $u_i \in U$  do
10:     $E_{user}.append(pu_i, u_i)$ 
11:     $E_{user}.append(u_i, pu_i)$ 
12:   end for
13: end for
14: return  $E_{user}$ 

```

(ii) Heterogeneous Edges: These edges connect different types of nodes. To interchange high-level semantic information among users and micro-video modalities, we model interactions among them. User-to-modality edges are drawn between users and

constituent modalities of their uploaded micro-videos. An edge exists between user node u_i and modality node m_i , where $m = \{v, a, t\}$ if the modality m_i constituting the micro-video mv_i was posted by user u_i . All edges are undirected with weights assigned to one for convenience. To explicitly model the user’s preference in modalities of his created micro-video, we construct the following edges: u-v, u-t, and u-a. We represent the interaction between the user and constituent modalities of his uploaded micro-video as $e(u_i, m_i)$ if $m_i \in mv_i$ and mv_i is a micro-video created by user u_i . Here, m is the modality indicator of micro-video mv_i .

6.3.2.2 Information Propagation and Neighborhood Aggregation

We leverage GraphSAGE [125], a powerful graph neural network technique, to refine the representations of micro-video modalities and users. GraphSAGE operates on the principle that nodes in the same neighborhood should have similar embeddings. It achieves this by iteratively aggregating and transforming feature information from neighboring nodes. In our model, we initialize modality-specific nodes with their respective feature representations and user nodes randomly. By applying GraphSAGE, we can capture contextual information and semantic relationships between micro-videos based on the similarity of their modalities, leading to more informative and contextually aware embeddings. This approach enhances the performance of hashtag recommendation by enabling the model to capture cross-modality relationships and nuanced user preferences. The method for refining node embeddings is described in Algorithm 6.2. The input consists of the whole graph, $G = (N, E)$, where N represents the set of all nodes and E represents the set of edges connecting these nodes.

- [1] Initialization (Line 1): The initial node representations h_n^0 are set to their corresponding input feature vectors denoted by $x_n, \forall n \in N$. Here x_n consists of visual features $mv_{initial}^v$, acoustic features $mv_{initial}^a$, textual features $mv_{initial}^t$, and user features $u_{initial}$.
- [2] Iteration over j (Line 2): The algorithm performs J iterations of message passing and node update. Here j denotes the current step and h^j denotes a node’s

Algorithm 6.2 Feature refinement

Input: $G(N, E)$: Graph
 $x_n, \forall n \in N$: Input features
 K : Depth
 $W^j, \forall j \in \{1, \dots, J\}$: Weight matrices
 σ : Non-linearity
 $MEAN_j, \forall j \in \{1, \dots, J\}$: Aggregator function
 $F : n \rightarrow 2^N$: Neighborhood function

Output: $z_n, \forall n \in N$: Vector representations

```
1:  $h_n^0 \leftarrow x_n, \forall n \in N$ 
2: for  $j = 1$  to  $J$  do
3:   for  $n \in N$  do
4:      $h_n^j \leftarrow \sigma(W^j \cdot (\{h_{n'}^{j-1}\} \cup MEAN_j(\{h_{n'}^{j-1}, \forall n' \in F(n)\})))$ 
5:   end for
6:    $h_n^j \leftarrow h_n^j / \|h_n^j\|_2, \forall n \in N$ 
7: end for
8:  $z_n \leftarrow h_n^J, \forall n \in N$ 
9: return  $z_n$ 
```

representation at j^{th} step. The parameter J regulates the method's neighborhood depth considered during the refinement process. A higher J allows the algorithm to incorporate information from more distant nodes in the graph.

- [3] Message Passing and Aggregation (Line 4): Each node n aggregates information from its neighbors n' defined by the neighborhood function $F(n)$ using an aggregation function (MEAN in our case). This can be mathematically represented as:

$$\mathbf{m}_n^{j-1} = \text{AGGREGATE}(\{\mathbf{h}_{n'}^{j-1}, \forall n' \in F(n)\}) = \frac{1}{|F(n)|} \sum_{n' \in F(n)} \mathbf{h}_{n'}^{j-1} \quad (6.15)$$

where m_n^{j-1} is the aggregated message from the neighborhood of node n at layer $j - 1$, *AGGREGATE* is the aggregation function (MEAN), which computes the average of the representations of all neighbors of node n' at the previous iteration.

- [4] Node Update (Line 4): The node's representation is then updated by combining its previous representation h_n^{j-1} with the aggregated message m_n^{j-1} using a

learnable transformation. This is mathematically expressed as:

$$\mathbf{h}_n^j \leftarrow \sigma(W^j \cdot [\mathbf{h}_n^{j-1} || \mathbf{m}_n^{j-1}]) \quad (6.16)$$

Here, h_n^j is the updated representation of node n at layer j , $||$ denotes concatenation, W^j is the weight matrix at layer j , σ is the non-linear activation function.

- [5] Final Representations (Line 8): The final node representations z_n are obtained from the last iteration J .

These refined node representations are then utilized for subsequent hashtag recommendation tasks.

$$z_n = \{mv_{final}^v, mv_{final}^a, mv_{final}^t, u_{final}\} \quad (6.17)$$

Here, $mv_{final}^v, mv_{final}^a, mv_{final}^t, u_{final}$ denotes the refined visual modality, acoustic modality, textual modality, and user feature vectors respectively.

6.3.2.3 Micro-video Representation

In the proposed framework, we jointly consider the modality-specific and user representations to investigate the impact that different modalities and users have on the overall micro-video representation. The content-based micro-video representation (mv_{final}) is obtained by concatenating modality-specific representations.

$$mv_{final} = concat(mv_{final}^v, mv_{final}^a, mv_{final}^t) \quad (6.18)$$

We employed the concatenation operator since it helps to preserve the features in every modality. Subsequently, we concatenate the derived content-based micro-video embedding with user embedding to obtain the overall enriched micro-video representation as shown in Equation 6.19.

$$mv_{overall} = concat(mv_{final}, u_{final}) \quad (6.19)$$

Here, $mv_{overall}$ is the derived micro-video representation. The hashtag recommendation module then uses this representation to anticipate hashtags for the given micro-video.

6.3.3 Hashtag Recommendation

The hashtag recommendation module takes the features extracted from the feature refinement module as input and yields a reasonable set of hashtag recommendations for a micro-video. Using the comprehensive feature vector ($mv_{overall}$) as input, we employ a dense layer of size $|H|$ followed by a softmax activation function to derive softmax scores for hashtags, as depicted in Equation 6.20.

$$y_{pred} = softmax\left(Dense(units = |H|)(mv_{overall})\right) \quad (6.20)$$

Here, softmax probabilities of specified hashtags are represented by $y_{pred} \in R^{|H|}$. The final collection of anticipated hashtags is then obtained by using $argsort()$ that sorts hashtags according to softmax scores in descending order, as given in Equation 6.21.

$$R = argsort(y_{pred}) \quad (6.21)$$

Here, R denotes the recommended hashtags. The training objective loss function is given in Equation 6.22.

$$J = \frac{1}{|Z|} \sum_{(mv_i, G_i) \in Z} \sum_{g \in G_i} -\log\left(P(g|mv_i)\right) \quad (6.22)$$

Here, J is the loss function, $Z(Z \subset M)$ denotes the training set of micro-videos, mv_i represents the current micro-video, G_i denotes the corresponding ground-truth hashtag set, and $P(g|mv_i)$ is the likelihood of selecting ground-truth hashtag (g) for the micro-video (mv_i).

Table 6.1: Statistics of different datasets

Datasets	Micro-videos	Hashtags	Users	A_h	A_{mv}
TMALL [156]	13140	3354	839	44.24	15.66
INSVIDEO [73]	30083	19930	2847	195.69	10.56
YFCC [157]	16611	16354	1455	138.80	11.41

6.4 Experimental Evaluations

To demonstrate the efficiency of our methodology, we first provide a description of the experimental conditions in this section, followed by the experimental findings.

6.4.1 Experimental Setup

In this section, we showcase various datasets utilized for conducting experiments. Afterward, we delve into distinct approaches employed for comparison, followed by evaluation metrics.

6.4.1.1 Datasets

We assess our devised framework on three real-world micro-video datasets namely, TMALL [156], INSVIDEO [73], and YFCC [157]. We customized each dataset to match our needs for the task of hashtag recommendation for micro-videos. First, we conducted lemmatization on hashtags and later removed the low-frequency hashtags, i.e., hashtags appearing less than 50 times. Next, we removed those micro-videos that lacked any modality or hashtags. Further, we retain users who have posted at least four micro-video postings. Table 6.1 contains statistical information for all datasets after pre-processing. In Table 6.1, A_{mv} denotes the average number of micro-videos per user, and A_h denotes the average number of hashtags per micro-video. TMALL, INSVIDEO, and YFCC datasets were collected from Vine³, Instagram, and Flickr⁴ platforms, respectively. Below, we go over these datasets in further detail.

³<https://vine.co/>

⁴<https://www.flickr.com/>

- TMALL: Chen et al. [156] created this dataset for micro-video popularity prediction. Initially, there were 1.6 million video postings in the crawled dataset, including 3,03,242 distinct micro-videos with a combined runtime of 499.8 hours. After carrying out pre-processing steps, the dataset used in our research contained 13,140 micro-video posts and 3,354 distinct hashtags. The minimum, average, and maximum hashtag count per post is 4, 44.64, and 1,424, respectively. The dataset includes 839 unique individuals, each posting an average of 15.66 micro-videos. For every micro-video, the complete user profile and associated metadata are also available.
- INSVIDEO: Li et al. [73] created INSVIDEO dataset to advocate hashtags for micro-videos. The authors crawled micro-videos from Instagram with associated descriptions and hashtags. The crawled dataset contained 3,34,826 micro-videos and 9,170 users. The dataset used by [73] contains 2,13,847 micro-videos, 15,751 hashtags, and 6,786 users. Following pre-processing, the dataset contains 30,083 micro-video postings from 2,847 users, with a mean of 10.56 posts per user. The dataset contains micro-video posts with a range of hashtag counts, including a minimum of 4, an average of 13.4, and a maximum of 1,494.
- YFCC: The Yahoo Flickr Creative Commons 100M, dubbed as YFCC100M [157] dataset is a comprehensive publicly accessible multimodal dataset tht contains nearly 99.2 million photos and 0.8 million micro-videos from Flickr. To perform the task of micro-video hashtag recommendation, we crawled micro-videos, user profiles, and annotated hashtags. Finally, the collected dataset contained 1,34,992 micro-videos, 8,126 users, and 23,054 hashtags. Following data cleaning methods, the dataset used in our studies included 16,611 micro-videos and 16,354 unique hashtags, 1,455 unique users, and an average of 11.41 micro-videos per user. The micro-videos in the resulting dataset has a minimum of 4 and an average of 138.8 linked hashtags.

6.4.1.2 Compared Methods

In this section, we outline the prevailing models that recommend hashtags for micro-videos.

- Memory Augmented Co-attention (MACON) [5]: MACON employs a mutually co-directed attention mechanism that learns from both text and images to improve hashtag suggestions for multimodal microblogs. Additionally, it tailors hashtag recommendations to individual users by analyzing their past posting behavior. We used the implementation provided by the authors.
- User-Video Co-Attention Network (UVCAN) [158]: UVCAN was originally developed for personalized micro-video recommendation. UVCAN lays more emphasis on the user’s hidden preference to obtain micro-video and user representation. UVCAN uses stacked attention techniques to learn multimodal information from both the user and micro-video. We have adapted for micro-video hashtag recommendation.
- Attention-based Multimodal Neural Network (AMNN) [94]: AMNN utilizes an encoder-decoder architecture with softmax for hashtag generation. The encoder uses CNN and Bi-LSTM to extract features from texts and images constituting multimodal microblogs followed by an attention mechanism on the constituent modalities. The attended visual and textual features upon concatenation are fed into GRU to generate hashtags sequentially based on probability scores.
- Dual Graph Neural Network (DualGNN) [159]: The two main modules that constitute DualGNN are single-modal and multimodal representation learning. The single-modal representation learning module uses the user-micro-video graph in each modality to identify unimodal user proclivities. In contrast, the multimodal representation learning module shows how the user weighs various modalities and infers the multimodal user preference. The ranking of the pertinent micro-videos for users is then done using a prediction mechanism. Originally designed

for micro-video recommendation, we adapted this system to recommend hashtags for micro-videos.

- Learning the User’s Deeper Preferences (LUDP) [160]: A user-item interaction graph, an item-item modal similarity network, and a user preference graph for each modality are the three components that make up LUDP. Through the user-item interactions matrix, the authors construct a bipartite graph of users and items. The authors leverage modal information to propagate and aggregate item ID embeddings on the similarity network in order to generate modal similarity graphs and collect structural information about items. The multi-modal attributes are combined to represent the user’s choice for the modal in the user preference graph, which is built based on the user’s prior engagement with the item. These newly discovered user and item representations are combined with representations found through collaborative signals on the bipartite network to provide multimodal recommendations. We adapt LUDP to carry out hashtag recommendations for micro-videos.
- Hashtag-guided Tweet Classification (HashTation) [95]: It is a two-stage framework for low-resource tweet classification using hashtag guidance. It features a transformer-based hashtag generator with two attention modules: one captures topical context from tweets, the other extracts entity insights from a co-occurrence graph. This enables encoding of both post-level and entity-level information, generating meaningful hashtags via latent topic embeddings and graph entity encoding. Beam search is used for sequential hashtag generation. We focus solely on the hashtag generator for comparison.
- Segments Selective Transformer (SEGTRM) [11]: SEGTRM is a transformer-based model that generates hashtags sequentially. It uses an encoder to remove extraneous data at text, segments, and token levels and a segments selector to reorganize segments. It employs a sequential decoding algorithm for hashtag prediction.

6.4.1.3 Evaluation Metrics

To gauge the capability of our devised hashtag recommendation system, we use assessment criteria from the literature on multi-label classification. The standard evaluation metrics for analyzing how well hashtag recommendation systems perform are hit rate, precision, recall, and F1-score. These metrics are computed by comparing predicted hashtags and ground-truth hashtags for each micro-video post. Note that larger values indicate better performance.

6.4.1.4 Implementation Details

For all the datasets, we partitioned them into a 70:10:20 ratio for training, validation, and testing, respectively. The model was trained for 20 epochs, and the evaluation was conducted using the top 5 recommendations. The Adam optimizer was employed for parameter updates, and the batch size was set to 32. A dropout rate of 0.5 was incorporated to mitigate overfitting. Regarding parameters in MIS-HON, we set the popular user selection ratio (α) to 0.1, thresholds for homogeneous edge filtration γ and θ to 0.5 each, and GraphSAGE aggregator function to mean. Tags occurring less than 50 times were removed from the dataset during preprocessing. The experiments were conducted on a Linux Server equipped with an Intel(R) Xeon(R) Silver 4215R CPU @ 3.20 GHz, 256 GB RAM, and a 16-GB NVIDIA Tesla T4 GPU. Additionally, the Conda environment management system was used for code execution.

6.4.2 Experimental Results

In this section, we outline the performance of our devised framework. We initially evaluate the performance of our proposed model against state-of-the-art methods on multiple datasets to determine its efficacy. Next, we analyze the performance gain, and performance of cold-start users, determine the sensitivity of various parameters, visualize the recommendations, and analyze the computational time. Note that K denotes the number of suggested hashtags and that the findings in this section are

Table 6.2: Effectiveness comparison results on TMALL dataset

Methods	Hit rate	Precision	Recall	F1-score
MACON [5]	0.458	0.156	0.291	0.202
UVCAN [158]	0.586	0.165	0.355	0.225
AMNN [94]	0.374	0.127	0.268	0.172
DUALGNN [159]	0.707	0.257	0.505	0.340
SEGTRM [11]	0.303	0.097	0.221	0.135
LUDP [160]	0.644	0.212	0.432	0.284
HashTation [95]	0.214	0.055	0.110	0.071
MISHON	0.753	0.283	0.563	0.376

expressed at $K = 5$.

6.4.2.1 Quantitative Analysis

We undergo rigorous experiments on several datasets to highlight that our suggested model is superior to state-of-the-art methods.

- Performance on TMALL Dataset: We assess the effectiveness of the proposed approach MISHON in comparison to its existing competitors. Table 6.2 highlights the experimental findings of our suggested technique against baselines on the TMALL dataset. We can observe from Table 6.2 that MISHON outperforms the compared methods on the TMALL dataset. The relative improvement of our model in terms of hit rate, precision, recall, and F1-score is 37.9%, 15.6%, 29.5%, 20.4% over AMNN, and 29.5%, 12.7%, 27.2%, 17.4% over MACON. The performance improvement over AMNN is due to taking user correlations and users’ interactions with constituent modalities of posted micro-videos whereas AMNN solely considers the content information embedded in the post’s multiple modalities. The reason behind the performance gain over MACON is that MISHON employs GraphSAGE to learn enriched embeddings of micro-video modalities and users whereas MACON relies on encoder-decoder architecture coupled with a parallel co-attention mechanism. The relative improvement of MISHON is 16.7%, 11.8%, 20.8%, 15.1% over UVCAN, 10.9%, 7.1%, 13.1%, and 9.2% over LUDP, and 4.6%, 2.6%, 5.8%, 3.6% over DualGNN in terms of hit rate, precision, recall, and F1-score respectively. UVCAN does not take the acoustic

modality into consideration. MISHON performs better than UVCAN due to the incorporation of three modalities constituting the micro-video. Although we mine collaborative information of the user and modality-specific embeddings similar to DualGNN, MISHON achieves better performance. This is due to the inclusion of modality feature similarity, user correlations, and user interactions with constituent modalities of posted micro-videos. Unlike LUDP which creates three separate subgraphs, we create one graph containing four types of nodes and seven types of edges to enrich modality-specific representations based on semantic similarity, the user representations based on similar tagging behavior, user-to-modality interactions, and derive the embedding of the micro-video. The relative improvement of MISHON is 54.5%, 22.9%, 45.4%, and 31.0% over HashTation, and 45.0%, 18.6%, 34.2%, and 24.1% over SEGTRM in accuracy, precision, recall, and F1-score, respectively. This superior performance is attributed to MISHON’s multifaceted approach, which adopts a hybrid filtering and GNNs to capture intricate user-user interactions and user to modality preferences, along with explicit handling of cold-start scenarios. Both HashTation and SEGTRM, while effective in leveraging contextual information through transformer-based generators with attention mechanisms, might not be as adept at modeling user preferences for different micro-video modalities and the complex web of user-user interactions that can significantly influence hashtag recommendations for micro-videos. These limitations highlight the strengths of MISHON’s design, allowing it to generate more accurate and personalized hashtag recommendations, particularly in scenarios involving new users or complex interaction patterns.

The performance comparison of hashtag recommendation models in terms of hit rate and F1-score for a variable number of hashtags on the TMALL dataset is shown in Figure 6.4. Hit rate and F1-score are plotted on the y-axis against the number of hashtags recommended on the x-axis. The count of suggested hashtags is between 1 and 9. Our proposed model beats state-of-the-art models despite having a variable amount of suggested hashtags since its curves are consistently the highest across all performance criteria. The improvements in each

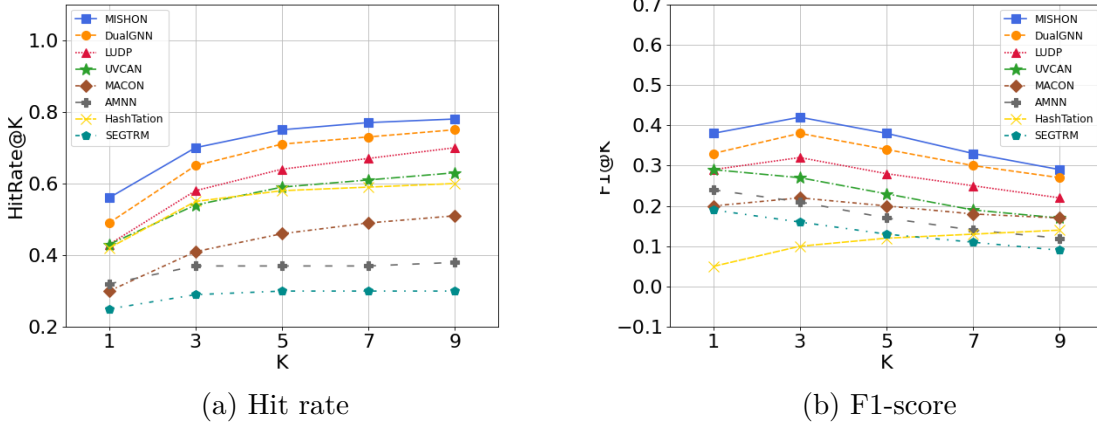


Figure 6.4: Effectiveness comparison curves on TMALL dataset

Table 6.3: Effectiveness comparison results on INSVIDEO dataset

Methods	Hit rate	Precision	Recall	F1-score
MACON [5]	0.569	0.348	0.099	0.154
UVCAN [158]	0.747	0.407	0.132	0.200
AMNN [94]	0.536	0.420	0.143	0.213
DUALGNN [159]	0.920	0.726	0.260	0.382
SEGTRM [11]	0.511	0.400	0.133	0.200
LUDP [160]	0.925	0.710	0.253	0.373
HashTation [95]	0.669	0.423	0.110	0.173
MISHON	0.941	0.764	0.280	0.410

of the four evaluation measures over extant methods demonstrate the capability and competitive advantage of our suggested model.

- Performance on INSVIDEO Dataset: To investigate the generalizability of MISHON in recommending hashtags for micro-videos on different platforms, we experimented using the INSVIDEO dataset. Table 6.3 shows the performance of our model against several other hashtag recommendation systems on INSVIDEO. Our model consistently outperforms AMNN by 40.5%, 34.4%, 13.7%, and 19.7%; MACON by 37.2%, 41.6%, 18.1%, and 25.6%; UVCAN by 19.4%, 35.7%, 14.8%, and 21.0%; LUDP by 1.6%, 5.4%, 2.7%, 3.7%; DualGNN by 2.1%, 3.8%, 2.0%, and 2.8%; HashTation by 27.2%, 34.1%, 17.0%, and 23.7%; and SEGTRM by 43.0%, 36.4%, 14.7%, and 21.0% in terms of hit rate, precision, recall and F1-score respectively. We tend to see the same performance regime

Table 6.4: Effectiveness comparison results on YFCC dataset

Methods	Hit rate	Precision	Recall	F1-score
MACON [5]	0.465	0.218	0.164	0.187
UVCAN [158]	0.543	0.188	0.176	0.182
AMNN [94]	0.527	0.330	0.284	0.305
DUALGNN [159]	0.745	0.401	0.354	0.376
SEGTRM [11]	0.450	0.282	0.249	0.264
LUDP [160]	0.464	0.213	0.171	0.190
HashTation [95]	0.240	0.125	0.117	0.119
MISHON	0.801	0.471	0.413	0.441

on the INSVIDEO dataset as observed in the case of the TMALL dataset.

- **Performance on YFCC Dataset:** We compare MISHON with other methods on YFCC dataset to illustrate its efficacy in micro-video hashtag recommendation. The performance of the suggested model is superior to that of state-of-the-art methods, as shown in Table 6.4. In terms of hit rate, precision, recall, and F1-score, MISHON exhibits relative improvements of 27.4%, 14.1%, 12.9%, and 13.6% over AMNN; 33.6%, 25.3%, 24.9%, and 25.4% over MACON; 25.8%, 28.3%, 23.7%, and 25.9% over UVCAN; 33.7%, 25.8%, 24.2%, and 25.1% over LUDP; and 5.6%, 7.0%, 5.9%, and 6.5% over DualGNN; 56.1%, 34.6%, 29.6%, and 32.2% over HashTation; and 35.1%, 18.9%, 16.4%, and 17.7% over SEGTRM. Our model, MISHON generally maintains relative performance improvements across three datasets when compared to other approaches. The results demonstrate the superiority and effectiveness of our proposed method for recommending high-quality hashtags regardless of the platform taken into consideration.

6.4.2.2 Ablation Studies

In this section, we conduct ablation experiments to assess the effectiveness of the feature refinement module, user, and attention mechanism. All experiments conducted in this section have been executed utilizing the TMALL dataset procured from Vine platform. We compare our model with five variations:

- **w/o FRM:** This variant represents the MISHON model without the Feature

Table 6.5: Ablation studies

Methods	Hit rate	Precision	Recall	F1-score
MISHON w/o FRM	0.464	0.156	0.303	0.207
MISHON+FRM (w/o Homo. Edges)	0.714	0.272	0.513	0.353
MISHON+FRM (w/o Hetero. Edges)	0.657	0.247	0.457	0.320
MISHON w/o User	0.467	0.158	0.310	0.210
MISHON w/o Attention	0.721	0.270	0.527	0.357
MISHON (Ours)	0.753	0.283	0.563	0.376

Refinement Module (FRM). In the absence of FRM, we directly utilize the modality-specific embeddings of the micro-video and its corresponding user, concatenating them to generate hashtag recommendations.

The FRM itself encompasses graph construction, incorporating both homogeneous and heterogeneous edges, followed by information propagation to refine the feature representations.

- **w/o Homo. Edges:** This ablation within the FRM removes homogeneous edges, which connect similar modalities across different micro-videos and establish connections between users.
- **w/o Hetero. Edges:** This ablation within the FRM removes heterogeneous edges, which link users to the modalities of their uploaded micro-videos.
- **w/o User:** The MISHON model without User
- **w/o Attention:** In this variant, we remove attention mechanism from MISHON. Here, we compute the average of the extracted modality-specific features and assign them as initial node embeddings.

6.4.2.2.1 Ablation on Feature Refinement: Table 6.5 demonstrates the substantial performance degradation incurred when FRM is ablated from MISHON. We observe a significant drop of 28.9%, 12.7%, 26.0%, and 16.9% in hit rate, precision, recall, and F1-score, respectively. This ablation study underscores the critical role of FRM in enhancing recommendation performance. The FRM first constructs a hetero-

geneous graph where nodes represent users, micro-videos, and their constituent modalities. The graph incorporates four nodes, homogeneous and heterogeneous edges. Subsequently, GraphSAGE propagates information across this graph structure. Through multiple layers of graph convolution, node embeddings are iteratively refined by aggregating information from neighboring nodes, capturing both local structural patterns and feature distributions. This process yields enriched modality-specific and user representations, which are then integrated to form a more comprehensive micro-video representation. The superior performance of the MISHON over this variant is a testament to the efficacy of these refined node embeddings derived from FRM.

6.4.2.2.2 Ablation on Homogeneous Edges Our empirical analysis underscores the critical role of homogeneous edges in the MISHON model, thus illustrating their ability in tackling **Challenge 3**. Excluding these edges resulted in a substantial drop in performance: 9.6% in accuracy, 3.6% in precision, 10.6% in recall, and 5.6% in F1-score. Intramodality edges enable the model to discern relationships between content sharing similar formats (visual, audio, or textual) across different micro-videos. Removing these edges deprives the model of valuable insights into how micro-videos interrelate based on their inherent content modalities. User-User edges encapsulate collaborative filtering principles, implying that users with overlapping past tag usage likely share future interests. Eliminating these edges hampers the model’s capacity to personalize recommendations by leveraging user preferences and community behavior. This performance degradation highlights that shared modalities play a pivotal role in capturing the core ideas and topics of micro-videos, aligning with the essence of hashtag descriptions. Our findings empirically support the hypothesis that users sharing common historical hashtags exhibit similarity, and micro-videos with similar latent modality representations are more likely to be associated with comparable hashtags.

6.4.2.2.3 Ablation on Heterogeneous Edges We observed a significant performance degradation when heterogeneous edges were removed from the MISHON model, thus illustrating its ability in tackling **Challenge 1**. Specifically, we saw a

decrease of 3.9% in hit rate, 1.1% in precision, 5.0% in recall, and 2.3% in F1-score. This underscores the critical role these edges play in capturing the nuanced relationship between users and the content they produce. Heterogeneous edges, which connect users to the specific modalities (video, audio, text) of their previously uploaded micro-videos, serve several key functions. First, they encode each user’s unique content creation style, providing insights into individual preferences and patterns in modality utilization. This personalized understanding allows the model to tailor hashtag recommendations that resonate with each user’s creative tendencies. Second, these edges establish strong, direct connections between users and the content they’ve created, facilitating the model’s ability to trace a user’s history and preferences when making recommendations. Furthermore, they enable indirect associations, allowing the model to infer potential interests even for content the user hasn’t directly created. Removing these edges disrupts the flow of information within the graph structure, hindering the model’s capacity to learn complex user-item relationships. Additionally, it weakens the model’s ability to leverage collaborative filtering, not only based on tag similarities but also on content creation choices. Consequently, the model struggles to retrieve relevant items and make personalized recommendations, leading to a drop in hit rate, recall, and precision.

6.4.2.2.4 Ablation on User Modeling Removing user significantly impacts the performance showing a drop of 28.6%, 12.5%, 25.3%, and 16.6% in accuracy, precision, recall, and F1-score as shown in Table 6.5, emphasizing the value of modeling user preferences and interactions. This is because it not only removes a node type but also disrupts crucial edge types, fundamentally altering the graph structure and hindering the model’s ability to capture user preferences, collaborative signals, and content-user relationships. The loss of user-user edges limits the model’s ability to leverage collaborative filtering, while the removal of user-modality edges disrupts the flow of information between users and content. Additionally, the absence of user nodes removes a critical contextual layer for understanding hashtag usage and making personalized recommendations.

Table 6.6: Performance comparison on cold-start users

Technique	Hit rate	Precision	Recall	F1-score
MISHON (C)	0.460	0.161	0.315	0.213
MISHON (SC)	0.741	0.280	0.550	0.371

6.4.2.2.5 Ablation on Attention Mechanism As can be seen from Table 6.5, excluding attention mechanism from MISHON leads to a decrease of 3.2%, 1.3%, 3.6%, and 1.9% in hit rate, precision, recall, and F1-score respectively. The attention mechanism adaptively weights information from different modalities, focusing on key details relevant to hashtag recommendation. Hashtags highlight important aspects of micro-videos, and the attention mechanism helps identify critical units in each modality for accurate hashtag suggestions. Since varied modalities have different representations, it is crucial to assign differential weights to the information contained in the constituent modalities. This underscores the significance of using attention mechanisms in learning important information from each modality to obtain the overall micro-video representation.

6.4.2.3 Performance Analysis on Cold-Start Users

In this section, we discuss how the performance of variants of MISHON differ in recommending hashtags for micro-videos posted by cold-start users, impact of number of historical posts for cold-start users, and popular user selection ratio.

6.4.2.3.1 Filtering Approaches for Cold-start Users Table 6.6 illustrates the capability of MISHON in tackling **Challenge 2**, i.e., recommending hashtags for micro-videos posted on the Vine platform by cold-start users i.e., the TMALL dataset. Here, MISHON (C) utilizes only the micro-video content and MISHON (SC) employs the social influence technique besides content features to recommend hashtags for micro-videos posted by cold-start users. The performance gain of MISHON (SC) over MISHON (C) is 28.1%, 11.9%, 23.5%, and 15.8% in hit rate, precision, recall, and F1-score respectively. We speculate the performance improvement is due to modeling

Table 6.7: Sensitivity analysis of popular user selection ratio (α)

α	Hit rate	Precision	Recall	F1-score
0.1	0.741	0.280	0.548	0.371
0.3	0.736	0.278	0.532	0.365
0.5	0.734	0.276	0.530	0.363
0.7	0.732	0.275	0.527	0.361

the influence of popular users on cold-start users. Users tend to follow hashtags used by the most popular users to gain social attention. We simulate the impact of social influence by applying GraphSAGE. MISHON (SC) employs the engagement rate of users for user-user edge construction. After information propagation and neighborhood aggregation, we can infer embeddings for cold-start users.

6.4.2.3.2 Popular User Selection Ratio We run experiments to select the optimal ratio of popular users (α). To get their content discovered on the platform and expand their audience, we assume that users tend to follow more well-known users. To this end, we first find the most popular users. Then we try to determine hashtags used by the most popular users that can be recommended to cold-start users. Popular users can be considered highly influential people and their hashtags are also adopted by other users to expand their social network, gain attention, and content visibility. Since hashtags are abstract labels to indicate topics, using popular hashtags related to that topic helps the micro-videos created by cold-start users to be included under those categories. This usually results in gaining new followers and better reachability. To determine the association between users with varied engagement rates and cold-start users, this supposition is taken into account. In accordance with this supposition, we run experiments to find the optimal popular-user selection ratio ranging from 0.1 to 0.7. As can be seen from Table 6.7, the variations in the performance metrics (hit rate, precision, recall, and F1-score) are minimal and could be within the margin of error. This consistency suggests that the choice of α does not significantly impact the performance of the hashtag recommendation system for cold-start users. However, the best F1-score was obtained when α was set to 0.1. As we increase α , the represen-

tations of users tend to be general rather than specific and inclined toward popular users. A selection ratio of 0.1 for popular users might be more realistic for many online platforms. In practice, only a small percentage of users tend to be extremely popular. This aligns with real-world usage patterns on social media platforms. We aim to build a recommendation system that performs well across different platforms or domains; a conservative selection of popular users (0.1) may provide a better generalization to various contexts than higher α values. This aligns with realistic usage patterns and provides stability and consistency in results.

6.4.2.4 Qualitative Analysis

To visually depict the quality of hashtags suggested by several methods, we present a micro-video post sourced from Vine platform in Figure 6.5. This example post has been chosen from test data, with correct, relevant, and incorrect hashtags shown in green, blue, and red respectively. The recommended hashtags matching ground-truth hashtags are called correct, relevant hashtags are consistent with the micro-video content but not specified in the set of ground-truth hashtags, and incorrect hashtags are neither correct nor relevant. As can be seen in Figure 6.5, our model recommends the highest number of correct hashtags as opposed to four, four, three, one, and one hashtags recommended by DualGNN, LUDP, UVCAN, MACON, and AMNN respectively. MISHON recommends `#vine` and `#footballvines` which are logical recommendations since this post is related to football and is uploaded on Vine. Furthermore, our model recommends `#comedy`, which is deemed relevant because it has been derived from the acoustic modality of the micro-video shown in the example post. This justifies the importance of mining information from constituent modalities of micro-videos. Further, we observed that user has previously used hashtags such as `#vineturkiye`, `#run`, `#omg`, `#6secondcover`, `#revined`, `#edits`, `#nba`, `#basketball`. We can see that MISHON also recommends `#omg` and related hashtags such as `#Vine`, `#footballvines` which are very similar to user’s historical hashtag: `#vineturkiye` and `#revined`. This further emphasizes that MISHON takes user’s historical tagging pattern into account to recommend personalised hashtags. Further, we observed that most similar users to

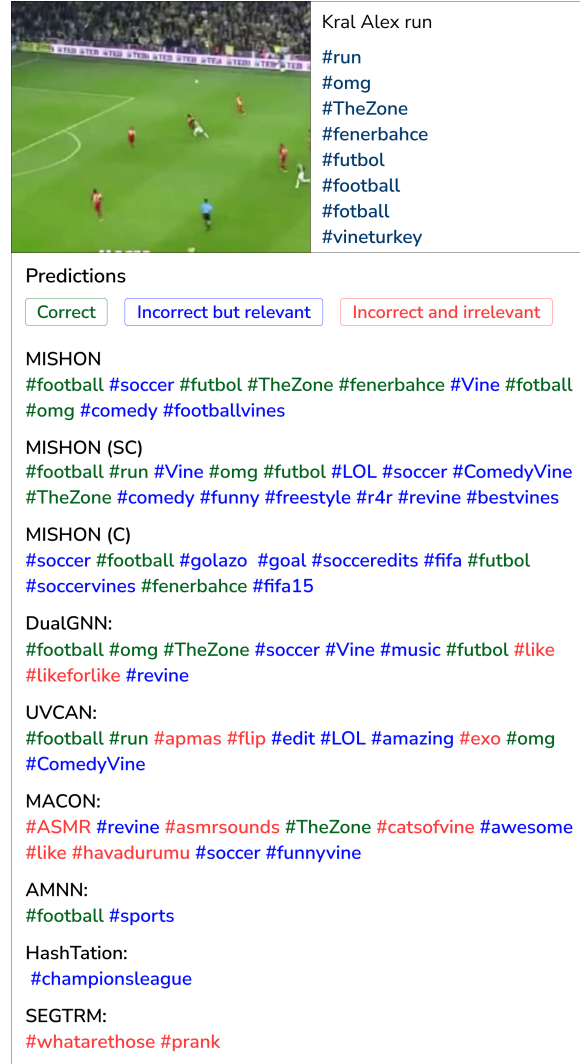


Figure 6.5: Example post showing hashtags recommended by different methods

this user employed hashtags such as #football, #TheZone, #Vine, #omg , #comedy, and MISHON is also recommending these hashtags for the micro-video uploaded by the user. This highlights that MISHON also considers community preferences by capturing user-user interactions. The higher the quality of hashtags recommended, the more likely users are to assign hashtags to micro-videos, thus enriching the user experience. As a result of better hashtag recommendations by our proposed model, more people will enter hashtag channels they actually enjoy and spend more time viewing hashtag-specific micro-videos.

Further, if the same micro-video was posted by a cold-start user lacking any histor-

ical and social network information, we have two variants of MISHON. Here, MISHON (SC) employs social influence and content features to recommend hashtags for micro-videos posted by cold-start users and MISHON (C) employs only content features. MISHON (SC) recommends more correct and relevant hashtags than MISHON (C), such as #funny, #freestyle, #r4r, #revine, and #bestvines. This demonstrates that MISHON’s hybrid approach of using both content and social influence is effective in recommending relevant hashtags for micro-videos posted by cold-start users. MISHON (SC) analyzes the micro-video’s content and leverages the tagging patterns of influential users on the platform. This helps cold-start users tap into trending hashtags and gain exposure to a wider audience.

6.5 Conclusion

This study aims to recommend pertinent hashtags for micro-videos while also alleviating the cold-start user problem. We propose effective hybrid filtering for micro-video hashtag recommendation based on Deep Learning and GNN. The proposed framework comprises three components: feature mining; feature refinement; and hashtag recommendation. The feature mining module attentively derives features from modalities constituting micro-videos. In the feature refinement module, we construct a graph using the constituent modalities of micro-videos and corresponding users as nodes. The edges are built to simulate modality-to-modality, user-to-user, and user-to-modality interactions. The user representation is derived by inductively modeling the hashtag preferences of like-minded users. The constructed graph enables learning of high-quality node embeddings based on information propagation and neighborhood aggregation. We run comprehensive experiments on three real-world datasets comprising users’ posted micro-videos with accompanying hashtags. We also alleviate the cold-start user problem by proposing a social influence and content-based technique to yield hashtags for micro-videos posted by them. Our proposed approach demonstrates superior performance compared to the existing methods both empirically and qualitatively.

Chapter 7

Popularity Prediction of Multimodal Content

7.1 Introduction

This chapter focuses on the critical task of predicting the popularity of content integrating multiple modalities. Given the increasing prevalence and demonstrated enhanced engagement rates of multimodal posts, understanding and accurately forecasting their popularity is paramount. This capability allows platforms to optimize the dissemination of diverse content formats [161] and enables the refinement of targeted advertising strategies that leverage the distinct features of various modalities [22, 162]. By analyzing patterns of public attention across combinations of texts and images, we aim to yield valuable insights into user behavior and preferences within these rich media environments, ultimately advancing areas such as recommender systems for multimodal content [21, 163] and digital marketing strategies tailored to multimodal trends [24].

Unveiling Hashtag-guided Attention Mechanism: Existing approaches to multimodal popularity prediction employ attention mechanisms [86], including self-attention [87, 164], hierarchical attention [165], and cross-modal attention [166, 167, 168, 169]), to dynamically assess the significance of features across textual and visual modalities. However, these methodologies frequently neglect the crucial contextual information embedded within hashtags [141]. These hashtag annotations,



Figure 7.1: Example social media post

prefixed with “#”, function as salient semantic indicators, revealing the content creator’s intended meaning and desired audience interpretation. Hashtags are instrumental in the online dissemination of events and topics, condensing complex ideas into concise labels and interlinking content at a granular level. The omission of this valuable contextual layer can impede the performance of popularity prediction models, given the demonstrable influence of hashtags on post engagement. Research indicates that posts including hashtags achieve double the engagement of those without [5]. Consider the example post in Figure 7.1. Current attention mechanisms might analyze visual elements of a selfie such as facial expressions, background details, and colors. However, these mechanisms overlook the contextual cues provided by hashtags such as #chasefield, #summer, #azdbacks, and #arizona, which specify location, season, and team affiliation, respectively. These hashtags guide the interpretation of visual features, highlighting the team jersey as a key element signifying user affiliation. Consequently, the integration of hashtag context is vital for developing more effective popularity prediction models within the dynamic social media landscape.

Leveraging Visual Demographics: Prior research on social media popularity prediction has incorporated demographic information through metadata [170] and user profiles. However, these methods are constrained by the incomplete or inaccurate metadata and outdated or intentionally misleading information in user profiles. Furthermore, the exclusive reliance on explicit user data can raise significant privacy

concerns. In contrast, the direct extraction of demographic attributes from visual cues, particularly facial features, presents an underexplored avenue. The findings of Bakshi *et al.* [171], which demonstrated a 38% higher likelihood of likes and a 32% higher likelihood of comments for Instagram photos containing faces among a dataset of 1.1 million images, underscore the power of facial cues in capturing attention and conveying emotions that directly impact a post’s popularity. Therefore, we aim to investigate the untapped potential of visual demographic analysis to enhance popularity prediction models for multimodal social media posts.

Harnessing Sentiment from Hashtags: Existing methodologies [172, 173, 174] predominantly focus on extracting sentiment from the textual content of social media posts, thereby overlooking the valuable sentiment information encoded within hashtags. Beyond conveying topical information, hashtags also reveal user sentiment [63] and audience perception, both of which can significantly influence a post’s popularity. While previous studies have utilized structural [84, 175] and topical information [78, 84] from hashtags for multimodal popularity prediction, the impact of hashtag sentiment remains largely unexamined. As exemplified in Figure 7.1, alongside content-related hashtags such as #baseball, #diamondbacks, and #arizona, users employ hashtags such as #smile, #funtimes, and #goodtimes to express their sentiment. The sentiment conveyed through hashtags reflect the audience’s collective emotional response to the content, providing valuable insights into prevailing trends and ongoing conversations. Given that emotional intensity within topics tends to drive greater engagement, and hashtags encapsulate sentiments potentially absent from captions, the underutilization of hashtag sentiment analysis represents a significant research gap. Addressing this gap offers an opportunity to develop more comprehensive and accurate popularity prediction methods by integrating the sentimental insights embedded within hashtags.

To bridge these identified research gaps, we propose NARRATOR, a Sentiment and hAshtag-aware deep neuRal netwoRk for multimodal posT pOpularity pRediction. NARRATOR presents a novel hashtag-guided attention mechanism that enables the model to dynamically weight the significance of different features in images and texts, informed by contextual cues provided by hashtags. This facilitates a more

holistic understanding of the interplay between content and its surrounding context. Furthermore, NARRATOR leverages visual cues within images to gain demographic insights, discerning fine-grained details such as age, gender, race, and emotions directly from faces. Moreover, NARRATOR explicitly incorporates sentiments extracted from hashtags, capturing the subtle emotional undertones that resonate with audiences and further refining our ability to predict post popularity. By combining these innovations—hashtag-guided attention, leveraging visual demographics, and analysing sentiment of hashtags, NARRATOR provides a deeper understanding of user engagement and emotional response, improving the performance of popularity forecasts.

Our major contributions are enlisted below.

- We propose a deep neural network that leverages sentiment from hashtags, visual demographic information, and employs a hashtag-guided attention mechanism to forecast post popularity comprehensively besides content-based features and sentiment from text.
- We devise a novel hashtag-guided attention mechanism that uses hashtags to guide the model’s focus on content features most relevant to the intended audience and context.
- Our work pioneers the use of visual demographic information for popularity prediction. We leverage visual demographics to identify engagement trends within specific audience contexts.
- We derive sentiment information embedded in hashtags to decipher the emotional appeal of a post and understand how it amplifies user engagement.
- Extensive experiments conducted on two real-world datasets demonstrate the superior performance of our proposed method over existing state-of-the-art methods both empirically and qualitatively.

The rest of the chapter is structured as follows. Section 7.2 formally defines the problem under investigation. We elaborate on our methodology in Section 7.3. The

evaluations of experiments are then covered in Section 7.4. Section 7.5 presents the concluding remarks of our research.

7.2 Problem Definition

Suppose there are P multimodal social media posts. Let p_i denote the i^{th} multimodal post such that $p_i = \{p_i^t, p_i^v, p_i^d, p_i^h, p_i^m\}$. Here, $p_i^t = \{w_i^x\}_{x=1}^X$ denotes the textual modality of the post, X denotes the length of the post caption and w^x denotes x^{th} word appearing in p_i . Furthermore, p_i^v and p_i^h denote the visual and hashtag modality of the post such that $p_i^h = \{h_i^j\}_{j=1}^H$. Here, h_i is the set of hashtags associated with post p_i , j is used to index a hashtag in set h_i , H is the cardinality of hashtag set (h_i) assigned to (p_i). The symbol p_i^d represents the demographic information such as age, gender, race, and emotion on the faces of people appearing in images and p_i^m denotes metadata of p_i and the user (u_i) who created it which contains followers count, following count, post count, hashtag count, and caption length. We specify our problem using the notations discussed above.

Given a multimodal social media post (p_i), our aim is to train a function $f(.)$ that allows us to forecast its popularity score.

$$\hat{y}^i = f(p_i) \quad (7.1)$$

Here, \hat{y}^i represents the predicted popularity score for the post p_i . We frame the popularity prediction task as a regression problem. Our objective is to learn the enriched feature representation of (p_i) and predict its popularity score.

7.3 Methodology

We introduce our novel methodological approach within this section. Figure 7.2 illustrates the architecture of our proposed framework for the popularity prediction of multimodal posts. We analyze varied features that significantly influence the popularity prediction of social media content. First, we investigate the textual features of

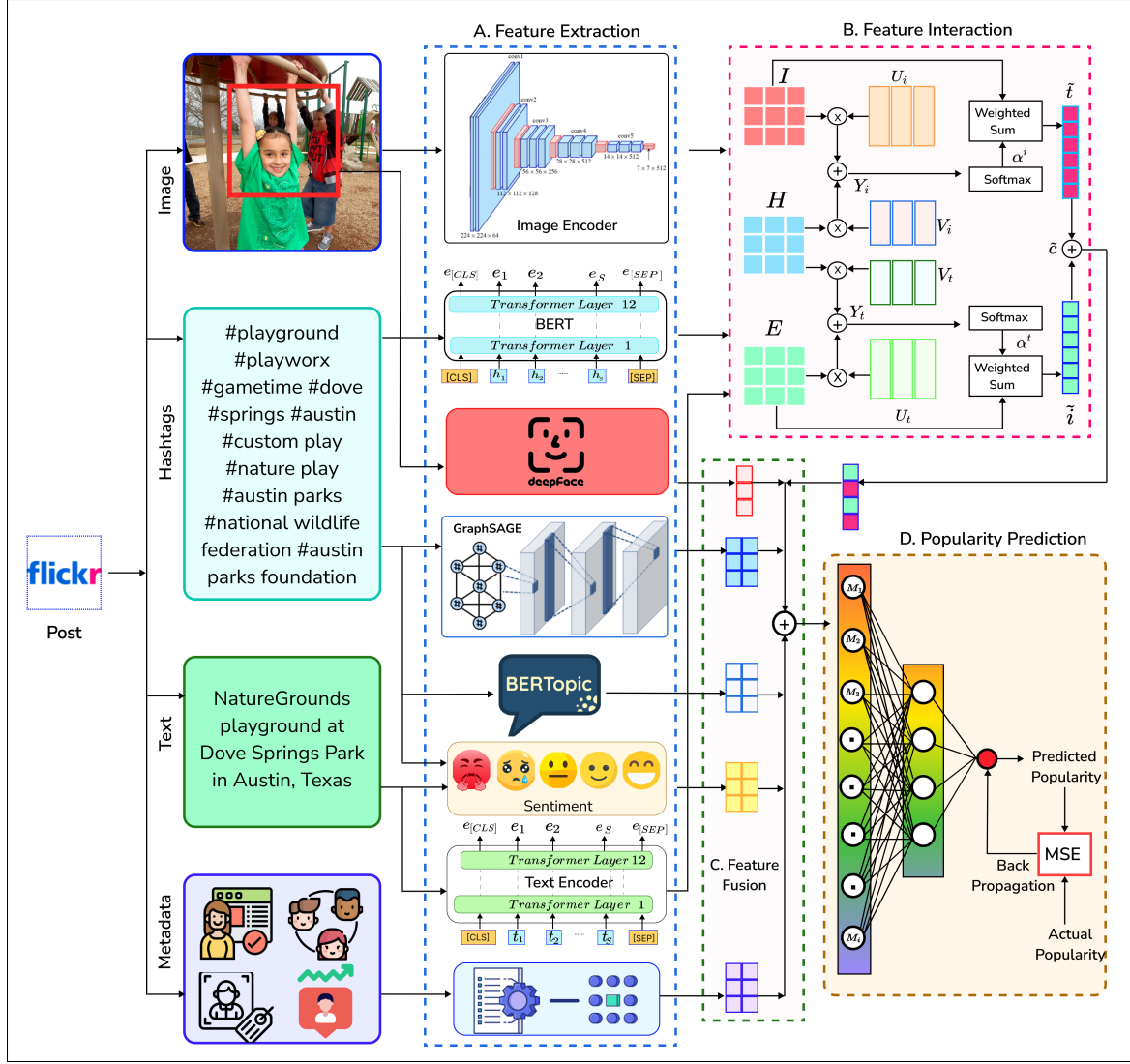


Figure 7.2: System architecture of NARRATOR

captions followed by visual features derived from images of social media posts. Additionally, we examine demographic information from the faces of people appearing in images accompanying social media posts. Then, several social features based on user and post metadata are explored. We leverage content-based and sentiment-based information from hashtags and post captions to effectively capture the rich information embedded in these posts. We derive topical and structural information from hashtags annotated to these posts. We also learn the mutual influence of hashtags on visual and textual modalities by devising a novel hashtag-guided attention mechanism. These features are then passed to several dense layers to predict the popularity score. Our

proposed framework entails a three-step process for accurately predicting the popularity score of social media posts: (1) feature extraction, (2) feature interaction, and (3) feature fusion for popularity prediction. These steps are discussed in detail below.

7.3.1 Feature Extraction

Here, we discuss the feature retrieval procedure for multimodal posts.

7.3.1.1 Textual Feature Extraction

Social media posts inherently rely on user-provided captions for context. To extract a textual feature representation from these captions, we leverage a transformer-based model, BERT [113]. Limited by its context-agnostic approach, word2vec [176] cannot effectively handle homonyms. BERT, on the other hand, prioritizes the words surrounding a target word during the embedding creation process. This enables BERT to capture the nuances of language and generate more semantically rich representations.

For the textual modality of social media post (p_i^t) which comprises a word sequence denoted by $p_i^t = \{w_i^x\}_{x=1}^W$, we add two tokens: class (CLS) and separator (SEP) to indicate the beginning and end of the input text respectively. Here, W is the number of words appearing in the post caption. We generate the corresponding set of tokens T using the BERT tokenizer as given in Equation 7.2.

$$T = \text{BERT_Tokenizer}(p_i^t) \quad (7.2)$$

We process the text sequence, denoted by T , through BERT as defined in Equation 7.3. This process yields a 768-dimensional vector representation for each token within the sequence.

$$B = \text{BERT}(T) \quad (7.3)$$

Here, $B = \{e^x\}_{x=1}^M$ is a matrix that encodes the textual features extracted from the post caption using BERT. This matrix comprises M rows, where M represents the

fixed length of the token sequence. Each row e^x contains the 768-dimensional BERT embedding for a corresponding token within the sequence (where x denotes the token index ranging from 1 to M). For textual descriptions less than M , we apply padding, otherwise, we perform truncation to make all textual descriptions of uniform size. Further, we use LSTM to model the sequential relationship among words. The LSTM unit outputs a hidden state t_i^x for the current word w_i^x by taking the embedding of the current word derived from BERT i.e., e_i^x and the hidden state of the preceding time step t_i^{x-1} as inputs as shown in Equation 7.4.

$$t_i^x = LSTM(e_i^x, t_i^{x-1}) \quad (7.4)$$

Here $t_i^x \in \mathbb{R}^D$, $x = 1, 2, \dots, M$, and $D=768$. For the sake of conciseness, we skip the specific LSTM formulae. We stack hidden state feature vectors for each word derived from LSTM to generate textual feature matrix E as given in Equation 7.5.

$$E = \{t_i^x\}_{x=1}^M \quad (7.5)$$

Here, $E \in \mathbb{R}^{M \times D}$ is the textual feature matrix, and $M = 15$ denotes the maximum length of the associated text for the post. The dimension of each t_i^x is \mathbb{R}^D where $D = 768$ denotes the embedding dimension.

7.3.1.2 Visual Feature Extraction

The image of the social media post plays a pivotal role in predicting the post's popularity. Deep learning approaches for extracting visual information have progressed remarkably in recent years. To extract visual features of the post, we employ VGG19 [135] model. VGG19 classifies 1.2 million images from the ImageNet [177] database into 1000 categories during its training process. We extracted visual features using the output of the final pooling layer of VGG19. We create several feature vectors for a picture by retaining the regional feature vectors. The feature matrix for

an image (V) can be expressed as exhibited in Equation 7.6.

$$V = \{v_i^k\}_{k=1}^K \quad (7.6)$$

Here $v_i^k \in \mathbb{R}^N$, $k = 1, 2, \dots, K$ with $N = 512$ which denotes the size of regional feature vector. We retain $K = 7 \times 7 = 49$ regional feature vectors for each image since the final pooling layer of VGG-19 is a $7 \times 7 \times 512$ tensor for 7×7 regions, each of which is represented by a 512-dimensional vector. Following the feature extraction stage, a fully connected (FC) layer is employed to project each regional feature vector into a new vector space. This transformation ensures that the dimensionality of the resulting image feature vectors aligns with the dimensionality of the text feature vectors, facilitating their subsequent concatenation and joint processing within the model architecture. The mathematical formulation for this transformation is presented in Equation 7.7.

$$I = \{v_i^k\}_{k=1}^K \quad (7.7)$$

Here, I is the visual feature matrix where $I \in \mathbb{R}^{K \times D}$ and $v_i^k \in \mathbb{R}^D$ where $D = 768$ is the embedding dimension for each regional feature vector.

7.3.1.3 Demographic Feature Extraction

We used DeepFace [178], a compact framework for identifying faces and analyzing characteristics namely age, gender, race, and emotion from faces of people appearing in images associated with the uploaded post. The VGG-Face model was used to investigate DeepFace. In DeepFace, 101 nodes are present for predicting the age between 0 to 100 years of the person present in the image. The race model predicts six different races namely Black, White, Asian, Middle Eastern, Indian, and Latino. The emotion on the users' faces is computed as one of the seven categories i.e., fear, sadness, happiness, anger, disgust, surprise, and neutral. The gender of the user is defined as male or female. Race, emotion, and gender were present as integral values. At last, we derive the demographic feature vector by concatenating gender,



Figure 7.3: Posts depicting demographic features

age, emotion, and race as given in Equation 7.8.

$$f_i^d = \{g, a, e, r\} \quad (7.8)$$

Here, f_i^d denotes the derived demographic feature vector that has a dimension of 116 and g, a, e, r represents gender, age, emotion, and race. Figure 7.3 shows four Flickr posts where we used a DeepFace model to analyze facial features and infer emotions and demographics. The first Flickr post shows an Asian woman with long hair who appears to be in her early thirties, smiling and happy. The second post shows an early middle-aged white man playing guitar and is sad. The third post shows a man with eyes frowning and trying to silence; hence, the inferred emotion from facial expression is fear. The fourth post is a man who is in his late thirties, inferred ethnicity is black, and emotion is sad because of his furrowed brows, downturned mouth, and dimly lit surroundings.

7.3.1.4 Hashtag Feature Extraction

In social media, hashtags serve as useful subject labels and search tools. Rather than attractive titles or pictures, trendy hashtags are the reason behind some social media posts getting a lot of attention. Hashtags encapsulate important information

that should be incorporated in social media post representation. However, conveying textual or topical information is just as crucial as providing the hashtag network’s structural components. Hence, we define the hashtag feature as a combination of topic embeddings and node embeddings that represent the content of the hashtag and the graphical structure between hashtags respectively. For the extraction of topic embeddings (X_h), we used the BERTopic [179]. BERTopic is a topic modeling technique that builds topic representations using the transformers framework and c-TFIDF. BERTopic first uses sentence transformers to produce a number of document embeddings. Next, it uses HDBSCAN [180] for document clustering and UMAP [181] for embedding dimension reduction. In order to determine the relevance of each word inside each subject, we compute the class-based Term Frequency Inverse Document Frequency (TF-IDF) for every cluster (topic). The average of all document embeddings inside a given subject is used to determine the topic embedding for a particular topic.

For the structure embedding, we constructed a graph-like network where we used hashtags as nodes of the network, and edges between two hashtags are created based on co-occurrence. We assign weights to the edge as the number of times the two hashtags appear in a single post. Further, we calculated structure embeddings \bar{V}_h for each node using GraphSAGE [125]. GraphSAGE is used for inductive representation learning on huge graphs. When creating low-dimensional vector representations of nodes, GraphSAGE is particularly helpful for graphs that include a wealth of node attribute data. The structural embedding of a post denoted by \bar{V}_h is determined only if there are at least two hashtags present in the post and taking the average of node embeddings of hashtags appearing in the post. On the other hand, a zero-vector is allocated as the structural embedding if a post is devoid of any hashtags. We then concatenate the topic and structure embedding to derive the overall hashtag representation for the post as shown in Equation 7.9.

$$f_i^h = \{X_h \oplus \bar{V}_h\} \quad (7.9)$$

Here, f_i^h is the resultant hashtag feature vector post p_i , X_h is the topic embedding and \bar{V}_h is the average hashtag node embedding.

7.3.1.5 Sentiment Feature Extraction

We embed each post caption into a 5-dimensional vector using Stanford's CoreNLP Sentiment Analysis tool¹. The scale for sentiment values ranges from zero to four which represents the likelihood that the sentence is extremely negative, negative, neutral, positive, or very positive. This tool was created by the Stanford NLP group as a module of the Stanford CoreNLP toolset [182]. Java is used to power Stanford's CoreNLP. Unlike Vader [183] and TextBlob which look at the sentiment of individual words, Stanford CoreNLP output the sentiment values based on the entire sentence structure resulting in improved performance. We derive the sentiment features from the post caption as shown in Equation 7.10.

$$s^t = \text{Stanford CoreNLP}(p_i^t) \quad (7.10)$$

Here, s^t denotes the sentiment feature vector derived from the textual modality of the post (p_i^t). We additionally treat hashtags as sentences and construct a 5-dimensional vector from the hashtag modality (p_i^h) of the multimodal social media post using Stanford CoreNLP.

$$s^h = \text{Stanford CoreNLP}(p_i^h) \quad (7.11)$$

Here, s^h denotes the hashtag-based sentiment feature vector having a dimension of 5. The overall sentiment feature vector for the post (p_i) is derived by concatenating the text-based and hashtag-based sentiment feature vectors as illustrated below.

$$f_i^{st} = \text{concat}(s^t, s^h) \quad (7.12)$$

Here, f_i^{st} is the resultant sentiment feature vector for the post (p_i) having a dimension of 10, s^t and s^h denote the text-based and hashtag-based sentiment feature vectors,

¹<https://stanfordnlp.github.io/CoreNLP/>

respectively.

7.3.1.6 Social Feature Extraction

The multimodal post's popularity is influenced by both its content and the user who posted it in terms of social media presence [184]. We have categorized social features into two categories i.e., user metadata and post metadata. We discuss these two below.

[1] User Metadata: The number of prior posts a user has made and their activity on the platform are both strongly connected with the popularity of their most recent post. Therefore, we have taken some user-centric features which are as follows:

- (a) User Id: A unique integer defining the user on the platform uniquely.
- (b) Average Views: It is obtained by computing the sum of all views over all the posts posted by the user in the past divided by the number of his previously uploaded posts.
- (c) Group Count: Total number of groups the user has joined on that platform.
- (d) Average Member Count: Average number of members in the group that the user joined.

[2] Post Metadata: The textual information associated with a post significantly influences the post's popularity. A post with a large title may not get huge popularity or a post with a large number of hashtags will appear more frequently so that it may gain more popularity. The post metadata consists of:

- (a) Tag Count: The number of hashtags a person used in their post.
- (b) Title Length: Word count of the caption of the post.
- (c) Description Length: Length of the description of the post
- (d) Tagged People: It is defined by a binary number 0 if people are not tagged in the post else 1.

- (e) Comment Count: The number of comments received by a post from other users.

[3] Time: Beyond user and content characteristics, predicting post popularity necessitates incorporating temporal features. Research suggests a diurnal cycle in social media activity, with weekends experiencing increased user engagement. Consequently, posts uploaded during these high-activity periods are tend to garner more views and interactions. To account for this temporal influence, we leverage the following time-based features:

- (a) Post Day: This categorical feature denotes the day of the week on which a post is uploaded. We employ one-hot encoding to represent the post-day as a 7-dimensional vector.
- (b) Post Month: This categorical feature indicates the month in which a post is uploaded. Similar to post-day, one-hot encoding is used to represent the post-month as a 12-dimensional vector.
- (c) Post Time: This categorical feature captures the time of day during which a post is uploaded. We divide the day into four distinct time segments (morning, afternoon, evening, and night), each encompassing six hours. One-hot encoding is then applied to represent the post time as a 4-dimensional vector.
- (d) Post Duration: This numerical feature represents the number of days an image remains posted on Flickr.

By incorporating these temporal features, our model can learn how time-related trends impact post popularity, potentially leading to more accurate predictions. The social feature vector (f_i^s) is obtained using the user ID, average views, group count, average member count, tag count, title length, description length, tagged people, comment count, and temporal data.

7.3.2 Feature Interaction

The feature interaction module sheds light on how hashtags interact with textual and visual modalities by devising a novel hashtag-guided attention mechanism. At its core, this mechanism utilizes hashtags as guiding signals to focus the attention of the predictive model on relevant features within the content. By considering hashtags associated with the content, the model can better understand the context in which the content is shared, leading to more accurate predictions. It allows the model to adapt its focus dynamically, making it suitable for a wide range of content types and social media platforms. Algorithm 7.1 shows how our devised hashtag-guided attention mechanism leverages hashtags to guide attention toward relevant text and image features. The use of hashtag embeddings and attention weights provides insights into the factors influencing content popularity, making the model more interpretable for users and content creators. By doing so, it aims to improve the prediction of social media post popularity. Lines 1-6 show how to compute the intermediate representation of text and image based on hashtags. We apply transformations on text and image feature matrices using learnable parameters to capture their interactions with hashtags. Here, $Y_t \in \mathbb{R}^{D \times A}$ and $Y_i \in \mathbb{R}^{D \times A}$ are the intermediate representation of the text and image feature matrix based on hashtags, respectively, A is the number of attention units set to 768, $U_t \in \mathbb{R}^{M \times A}$, $V_t \in \mathbb{R}^{L \times A}$, $U_i \in \mathbb{R}^{K \times A}$, $V_i \in \mathbb{R}^{L \times A}$ are learnable parameters. The hashtags associated with the post are embedded into a continuous vector space representation by using BERT.

$$H = BERT(p_i^h) \quad (7.13)$$

Here, $H \in \mathbb{R}^{L \times D}$ is the resultant hashtag feature matrix, $L=60$ based on the maximum number of hashtags associated with a post. To have a feature matrix of uniform dimensions across different posts in the data, we padded zeros for posts having a tag count of less than 60. We employ BERT because one hashtag can have different meanings in different posts. For example, #rock can be used to refer to stones and in other posts, the same hashtag can refer to music rock band. Therefore, it is important to capture

Algorithm 7.1 Hashtag-guided attention

Input: E : Text feature matrix
 V : Image Feature Matrix
 H : Hashtag Feature Matrix
Output: \tilde{c} : Updated Content Feature Vector
function Hashtag-guided Attention(T, V, H)
1: **for** $t = 1$ **to** M **do**
2: $Y_t[t] = \tanh(E[t] \times U_t + H \times V[t])$
3: **end for**
4: **for** $i = 1$ **to** K **do**
5: $Y_i[i] = \tanh(V[i] \times U_i + H \times V[i])$
6: **end for**
7: **for** $t = 1$ **to** M **do**
8: $\alpha^t[t] = \text{softmax}(Y_t[t] \times W_t)$
9: **end for**
10: **for** $i = 1$ **to** K **do**
11: $\alpha^i[i] = \text{softmax}(Y_i[i] \times W_i)$
12: **end for**
13: **for** $j = 1$ **to** D **do**
14: $\tilde{t}[j] = \sum_t (E[t][j] \times \alpha^t[t])$
15: **end for**
16: **for** $j = 1$ **to** D **do**
17: $\tilde{i}[j] = \sum_i (V[i][j] \times \alpha^i[i])$
18: **end for**
19: $\tilde{c} = \tilde{i} + \tilde{t}$ **return** \tilde{c}

the context in which a particular hashtag is being used. BERT-based embeddings of hashtags capture their semantic relationships, allowing the model to understand their contextual meanings. By introducing learnable parameters associated with the transformation of text/image features, the model can adaptively learn how to combine these features with hashtag features to derive meaningful representations. This allows the model to adapt to the specific characteristics of the content and the nuances of hashtag usage patterns. The resulting intermediate representations encapsulate not only the inherent characteristics of the text/image features but also their contextual relevance concerning the associated hashtags. This semantic enrichment enhances the model's ability to understand the underlying themes, topics, and sentiments expressed in the content, thereby improving the quality of feature representations. These intermediate representations Y_t and Y_i signify how text and image features are influenced

by associated hashtags. We apply the hyperbolic tangent (\tanh) activation function to a combination of text feature matrix (E) and hashtag feature matrix (H) represented by Y_t . Similarly, for image features, we apply \tanh to a combination of image feature matrix (I) and hashtag feature matrix (H) which is denoted by Y_i . The intuition here is to capture the interaction between text and hashtags (Y_t), and image and hashtags (Y_i). Lines 7-12 compute the attention weights for text and image features. Here, α^t and α^i denote attention weights for text modality and image modality, respectively. Here, $W_t \in \mathbb{R}^{A \times D}$, $W_i \in \mathbb{R}^{A \times D}$ are learnable parameters. These weights represent the importance of different features based on the associated hashtags. The idea is that certain hashtags may be more relevant to either text or image content, affecting their contribution to the overall content vector. Then, we compute the attended modality representations. Here, $\tilde{t} \in \mathbb{R}^D$ and $\tilde{i} \in \mathbb{R}^D$ denotes the attended text feature vector and attended image feature vector, respectively. In Lines 13-15, we take the weighted sum of text feature matrix (E) with the attention weights α^t to get an attended text representation \tilde{t} . This step emphasizes the text features that align with relevant hashtags. Similarly, in Lines 16-18, we multiply the image feature matrix I with the attention weights α^i to get an attended image feature vector representation (\tilde{i}). Here, the focus is on image features associated with specific hashtags. Line 20 computes the hashtag-guided content feature vector by taking the sum of the attended text feature vector \tilde{t} and the attended image feature vector \tilde{i} . Here, $\tilde{c} \in \mathbb{R}^D$ represents the comprehensive feature vector derived from hashtag guidance. This updated feature vector represents a comprehensive view of the content, incorporating information from both text and image modalities, with attention focused on the relevant features guided by hashtags. By leveraging hashtags as guiding signals, the algorithm enhances the model's ability to capture contextually relevant features, ultimately improving the performance of the model in predicting popularity.

7.3.3 Feature Fusion

In this section, we delve into the details of feature fusion, a critical step in constructing a unified representation of multimodal social media posts and ultimately

predicting their popularity within the proposed framework followed by the theoretical background for feature fusion. The feature fusion component is structured in a grid-like manner, with layers arranged horizontally. Each layer corresponds to a stage in the data processing conducted by the CNN. During the feature extraction stage, we extract demographic, sentiment, hashtag, and social context features from each post denoted by $f_i^d, f_i^{st}, f_i^h, f_i^s$, respectively. Conv1D refers to a one-dimensional convolutional layer, a common building block in CNNs for processing sequential data. Each layer has a box with parameters such as the filter size, number of filters, and activation function (ReLU). These parameters define how the layer performs its operations on the data. Dropout layers randomly set a fraction of activations to zero during training. This helps prevent the network from overfitting to the training data. After the convolutional layers, there are “Flatten” layers. These layers flatten the data from a multi-dimensional representation into a one-dimensional vector suitable for feeding into a fully connected layer.

Following the extraction of social features, we concatenate them into a single feature vector with a dimensionality of 85. To address potential issues of high dimensionality and redundancy within this feature space, we employ Principal Component Analysis (PCA) [185]. PCA is a widely recognized method for reducing dimensionality, wherein the data is projected into a lower-dimensional space to maximize the retained variance. All feature vectors thus obtained are passed to an individual network consisting of three CNN layers. The output of all CNNs along with the hashtag-guided content feature vector is fed into a fusion network. The fusion network is composed of a merged layer and a series of several Fully Connected (FC) layers termed Cascade Feed-Forward Network (CFFN). The output of all CNN networks along with the hashtag-guided content feature vector are concatenated in the merged layer. The merged layer’s concatenated output serves as the input for the CFFN. Ultimately, the output of CFFN is summed together at the final node, which gives the popularity score for the specific social media post. Mean Squared Error (MSE) is then calculated using both the ground truth and forecasted popularity scores. The computed error is backpropagated and weights are updated accordingly. For each iteration, output

vectors for the social feature, demographic feature, hashtag, and sentiment feature are calculated as illustrated in Equations 7.14, 7.15, 7.16, 7.17.

$$S_i = \text{Conv1D}_s3(\text{Conv1D}_s2(\text{Conv1D}_s1(f_i^s))) \quad (7.14)$$

$$D_i = \text{Conv1D}_f3(\text{Conv1D}_f2(\text{Conv1D}_f1(f_i^d))) \quad (7.15)$$

$$H_i = \text{Conv1D}_h3(\text{Conv1D}_h2(\text{Conv1D}_h1(f_i^h))) \quad (7.16)$$

$$St_i = \text{Conv1D}_{st}3(\text{Conv1D}_{st}2(\text{Conv1D}_{st}1(f_i^{st}))) \quad (7.17)$$

where $i = 1, 2, \dots, N$. Here, S_i, D_i, H_i , and St_i denote the social, demographic, hashtag, and sentiment feature vectors obtained after passing through CNN layers. Further, we have flattened S_i, D_i, H_i , and St_i and concatenated these feature vectors along with the hashtag-guided content feature vector (\tilde{c}) and denote the final merged vector as M_i where

$$M_i = [S_i, F_i, H_i, St_i, C_i] \quad (7.18)$$

Here, M_i denotes the concatenated feature vector which has a size of 27104.

7.3.4 Popularity Prediction

We employ a Deep Feedforward Neural Network as illustrated in Figure 7.4, consisting of 12 fully connected layers (denoted by N) with decreasing sizes (13552, 6776, 3388, 1694, 847, 424, 212, 106, 53, 27, 13, 1) to forecast the popularity score. Each hidden layer is followed by a ReLU activation function and a dropout [186] layer with a rate of 0.2. The output layer utilizes a linear activation function to directly predict the continuous star count.

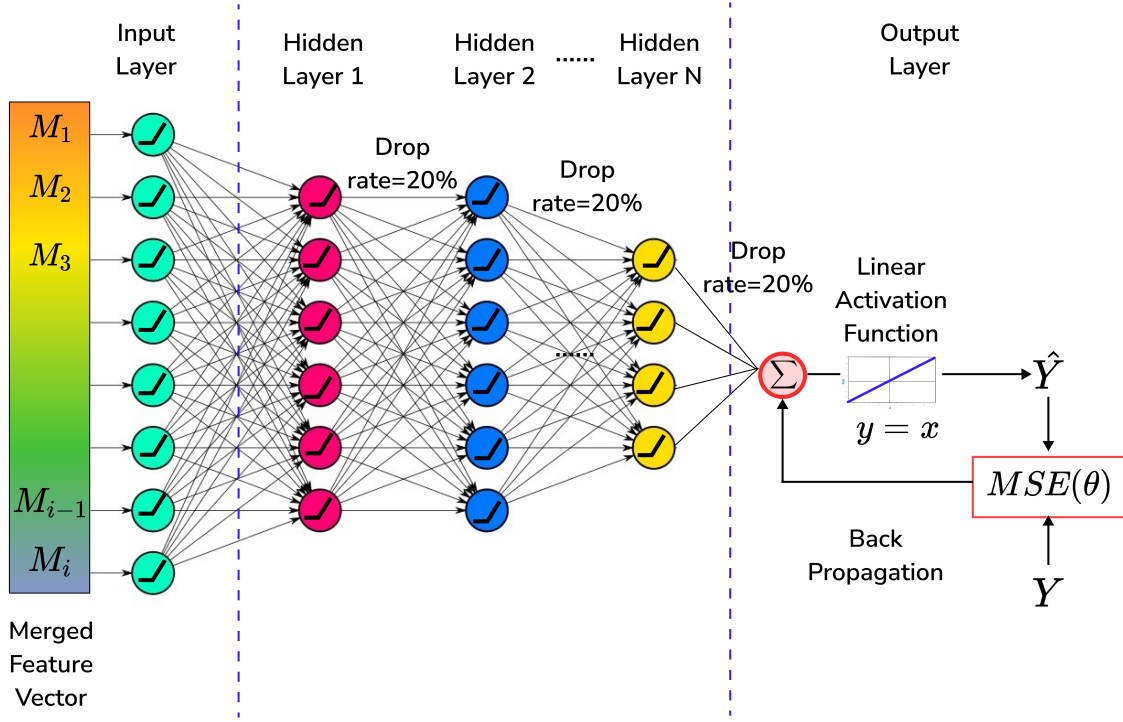


Figure 7.4: Deep feedforward neural network for popularity score prediction

The final popularity score can be calculated as given in Equation 7.19.

$$\hat{Y}_i = DNN(M_i) \quad (7.19)$$

Here, \hat{Y}_i and Y_i are predicted and ground-truth popularity score of post p_i , and, DNN is the Deep Feedforward Neural Network. The training objective is defined in Equation 7.20.

$$MSE(\theta) = \min_{\theta} \left(\frac{1}{2|P|} \sum_{i=1}^{|P|} P|(\hat{Y}_i - Y_i)^2 \right) \quad (7.20)$$

Here, $|P|$ is the number of posts in the training data and θ represent the NARRATOR's parameters. These are trained via back-propagation by maximizing MSE cost function after being set with random values between -1 and 1.

7.4 Experimental Evaluations

Following a detailed account of the experimental setup, this section presents a comprehensive analysis of the data obtained from the experiments. This analysis will provide valuable insights into the effectiveness of our proposed approach.

7.4.1 Experimental Setup

This section details the datasets used in the study and the adopted preprocessing procedures. We next go through several comparison techniques, which are then followed by evaluation metrics.

7.4.1.1 Datasets

In this section, we cover various datasets on which experiments were conducted, followed by strategies for dataset preprocessing.

- SMP: The Social Media Prediction (SMP) [187] is a real-world dataset provided by ACM Multimedia Grand Challenge in 2019. Initially, the raw data collection consisted of about 432K posts that were gathered from 135 distinct users' personal Flickr albums. In the dataset, each post has a unique picture ID that identifies the image and a user ID that identifies the person who uploaded it, the date the post was created, how many comments it received, how many hashtags were used, whether any users were tagged in the image, number of words in the title and the image caption. User-centric information such as the average view count, average member count, and group count were also collected as part of the data. Each image has a label that reflects its popularity based on the number of log-normalized views.
- TPIC: TPIC2017 [188] is a social media dataset for temporal popularity prediction that contains 680K photos and accompanying photo-sharing records on Flickr over a three-year period. The TPIC2017 dataset is diverse, containing photos, user data, and time information.

7.4.1.1.1 Dataset Preprocessing We preprocess the input data to convert it into an appropriate format to extract coherent features and accurately predict the popularity of multimodal posts. We apply several adaptations and normalization techniques to these datasets. To this end, we have taken all posts that contain hashtags, titles, and faces in the image. This left us with a total of 21,000 samples in the SMP dataset and 11,000 samples in the TPIC dataset. For our experiments, 80% of the posts were utilized for training, 10% for validation, and 10% for testing.

7.4.1.2 Compared Methods

In this part, we discuss various state-of-the-art methods for assessing the effectiveness of the suggested framework.

- [1] Leveraging Hashtag Networks for Multimodal Popularity Prediction of Instagram Posts (HashPop): Liao *et al.* [84] has predominantly used hashtags as a separate modality to gauge the popularity of a specific post. The authors have constructed a graph-like hashtag network where hashtags used in the post serve as nodes of the graph. Edges are created between hashtags if many hashtags appear in a single post. This combined vector is then represented as a hashtag feature. In addition, authors used InceptionV3 to obtain image embedding, and sentence transformer to obtain caption embedding for the post along with additional metadata. The overall post-representation is obtained by concatenating these obtained features. The output from the last layer is utilized to calculate the popularity score in a dense layer once this composite representation has been passed through.
- [2] Multimodal Deep Learning Framework for Image Popularity Prediction on Social Media (VSCNN): Abousaleh et al. [170] chose social metadata, post metadata, and time metadata as primary components for predicting post popularity. Low-level, high-level, and Deep Learning-based visual attributes are extracted from the image that constitutes a given post. Furthermore, after feature extraction, the authors employed CNNs on visual and social metadata features indepen-

dently to learn their high-level representations. At the output layer, shared multimodal properties are learned and the popularity score is determined. To this end, the output of two separate CNN networks is then integrated into a shared network which is made up of one merge layer and two dense layers to predict the popularity of a post.

- [3] Multimodal Deep Learning for Social Media Popularity Prediction With Attention Mechanism (MMAAtt): Xu et al. [86] provided an attention mechanism for forecasting post popularity. The authors retrieved four distinct features namely categorical, numerical, visual, and textual. The categorical feature consists of category, subcategory, and path alias of post embeddings which were obtained and passed to individual dense layers, the output of which was further passed to the common dense layer. The numerical information contains the number of followers, the date the post was created, and the location. The authors applied ResNet50 to extract visual features which were later sent to a dense layer. The textual features contain the title and hashtags of the post, and embeddings of both were obtained using word2vec and were passed to two-layer LSTM and individual dense layers. The output of all dense layers was concatenated and transferred to a single dense layer, after which an attention layer was utilized to compute a popularity score.

- [4] Social Media Popularity Prediction: A Multiple Feature Fusion Approach with Deep Neural Networks (FuseDNN): Ding et al. [189] retrieved visual, textual, user, temporal, and geographical location features. In visual features, the authors retrieved ResNet-101 characteristics, an intrinsic popularity score, and an aesthetics score from the post's provided image. Tag count, title length, and title embedding using BERT are all included in text features. While the time component contains the post duration and the location information provides geographic coordinates, user data includes the count of followers, followings, and posts. The remaining entities were concatenated and passed to a different dense layer. The ResNet-101 and BERT embeddings were passed to separate dense

layers. To forecast the popularity score, the output from all three dense layers was combined and transferred to three dense layers to yield the predicted popularity score.

- [5] Predicting Tweet Engagement with Graph Neural Networks (TweetGage): Arazzi et al. [175] developed a novel Graph Neural Network framework for predicting user engagement on social media. TweetGage leverages a graph-based model based on hashtag relationships, capturing semantic connections beyond individual post features.
- [6] Gradient Boost Tree Network based on Extensive Feature Analysis for Popularity Prediction of Social Posts (MFTM): Hsu et al. [190] presented a multi-modality feature mining framework for social media popularity prediction. It leverages identity-related user features alongside traditional modalities (text and image) to achieve significant performance improvements, suggesting a stronger influence of identity compared to other user metadata. LightGBM and TabNet are employed to capture complex relationships within the enriched feature set.
- [7] Enhanced CatBoost with Stacking Features for Social Media Prediction (ECSF): Mao et al. [191] proposed a novel social post popularity prediction approach utilizing enriched post and user features. ECSF employs innovative stacking features to capture higher-order interactions between text and image features, potentially leading to a more comprehensive understanding of the underlying factors that influence social media post popularity. ECSF's effectiveness is substantiated by its state-of-the-art performance on the SMP challenge dataset, surpassing prior methods that primarily relied on extracting lower-order features.

7.4.1.3 Evaluation Metrics

In this study, we adopted Mean Squared Error (MSE) and Mean Absolute Error (MAE) as primary metrics to quantify the prediction accuracy of our model.

- **MSE:** To determine the mean of the squared sum of prediction errors, MSE is frequently used. Each prediction error represents the discrepancy between a data point's actual value and the value estimated by a regression model. It is simpler to determine the gradient of MSE since it has straightforward mathematical features. Due to its computational simplicity, smooth differentiability, and greater optimization amenability, MSE is typically supplied as the default measure for the majority of predictive models. A serious flaw with MSE is that it squares large prediction errors, which strongly penalizes them. Due to the quadratic accumulation of each MSE error, the overall error is significantly influenced by outliers in data. This demonstrates that MSE undervalues the model's performance because of its high sensitivity to outliers and the disproportionate weight assigned to their effects, MSE undervalues the model's performance in this case. When outliers are present in the data, only then the disadvantage of MSE becomes obvious, making MAE an adequate replacement. MSE is defined as given in Equation 7.21.

$$MSE = \frac{1}{|N|} \sum_{i=1}^N (y_i - \hat{y}_i)^2 \quad (7.21)$$

Here, N denotes the total number of posts, \hat{y}_i denotes the predicted popularity score, and y_i denotes the actual popularity score of the i^{th} multimodal post (p_i).

- **MAE:** MAE is a straightforward metric typically used to assess the precision of a regression model. It calculates the mean absolute value of each prediction mistake made by the model overall test set samples. Unlike metrics sensitive to outliers, MAE assigns equal weight to all prediction errors. This characteristic ensures that larger deviations from the actual values contribute linearly to the overall error score. However, MAE focuses solely on the absolute magnitude of the error, without considering the direction of the discrepancy (overprediction or underprediction). Consequently, MAE provides an objective measure of the model's overall performance in terms of absolute prediction error, but it doesn't necessarily indicate whether the model consistently overestimates or underestimates the target variable. By employing the absolute value of the prediction

Table 7.1: Effectiveness comparison results on different datasets

Methods	TPIC		SMP	
	MSE	MAE	MSE	MAE
FuseDNN, Ding <i>et al.</i> '19	2.716	1.318	4.831	1.707
MMAAtt, Xu <i>et al.</i> '20	2.367	1.170	4.447	1.617
VSCNN, Abousaleh <i>et al.</i> '20	1.711	1.015	5.023	1.732
HashPop, Liao <i>et al.</i> '22	3.200	1.435	6.262	1.946
TweetGage, Arazziet <i>et al.</i> '23	2.132	1.145	4.211	1.566
MFTM, Hsu <i>et al.</i> '23	5.264	1.946	6.815	2.097
ECSF, Mao <i>et al.</i> '23	5.115	1.911	6.783	2.072
NARRATOR	1.196	0.854	2.022	0.972

MSE: Mean Squared Error, MAE: Mean Average Error

error rather than its squared value, MAE becomes more resilient to outliers than MSE because it does not penalize significant errors as heavily as MSE. Therefore, MAE has both benefits and drawbacks. While it helps in managing outliers, it does not penalize significant forecasting errors. MAE is described in Equation 7.22.

$$MAE = \frac{1}{|N|} \sum_{i=1}^N |y_i - \hat{y}_i| \quad (7.22)$$

Here, N denotes the total number of posts, \hat{y}_i denotes the predicted popularity score, and y_i denotes the actual popularity score of the i^{th} multimodal post (p_i).

7.4.2 Experimental Results

The experimental results for the proposed and current state-of-the-art approaches on various datasets, ablation investigations, visualization of predictions from both existing and proposed methods, statistical analysis, ranking of important features, implementation details, and limitations are all covered in this section.

7.4.2.1 Effectiveness Comparison

To assess the performance of our suggested technique on various datasets, we compare it to the existing methods on TPIC and SMP datasets. The performance comparison using the aforementioned datasets is given in Table 7.1. It is evident from

Table 7.1 that NARRATOR performs noticeably better than state-of-the-art methods. NARRATOR beats HashPop with a 62.625% improvement in MSE and 40.487% in MAE on the TPIC dataset and 67.709% improvement in MSE and 50.051% in MAE on the SMP dataset. The authors in HashPop have focused more on using the hashtag network besides content-based and metadata information. Our proposed model which consists of demographic and sentiment information embedded in captions and hashtags along with hashtag-guided attention content features beats HashPop. NARRATOR beats VSCNN with a relative improvement of 30.099% in terms of MSE and 15.862% in terms of MAE on the TPIC dataset and 59.745% in MSE and 78.189% in MAE on the SMP Dataset. VSCNN predicted popularity based on social and visual features whereas in our model we have incorporated four additional features namely demographic information, hashtag, textual, sentiment of caption, and hashtags besides a hashtag-guided attention mechanism employed on content-based features. NARRATOR beats MMAtt with a relative improvement of 49.471% in MSE and 27.008% in MAE on the TPIC dataset and 54.531% in MSE and 39.888% in MAE on the SMP dataset. While authors incorporated an attention mechanism within their model, NARRATOR achieves demonstrably stronger performance. This improvement can be attributed to our implementation of a hashtag-guided attention mechanism that models the influence of hashtags on linguistic and visual features. Our attention mechanism produces superior results because the hashtags are very closely related to the title and the image associated with the post, which in turn boosts the model’s performance significantly. The proposed model beats FuseDNN exhibiting a relative improvement of 55.964% and 35.204% in MSE and MAE on the TPIC dataset and 58.145% improvement in MSE and 43.057% in MAE on the SMP dataset. NARRATOR surpasses TweetGage exhibiting a relative improvement of 43.092% and 25.415% in terms of MSE and MAE on the TPIC dataset and 51.982% improvement in terms of MSE and 37.931% in terms of MAE on the SMP dataset. Unlike TweetGage which captures relationships among posts based on common hashtags, we employ the topical, structural, semantic, and sentiment information from hashtags besides image, caption, and demographics. The proposed model beats MFTM exhibiting a relative improve-

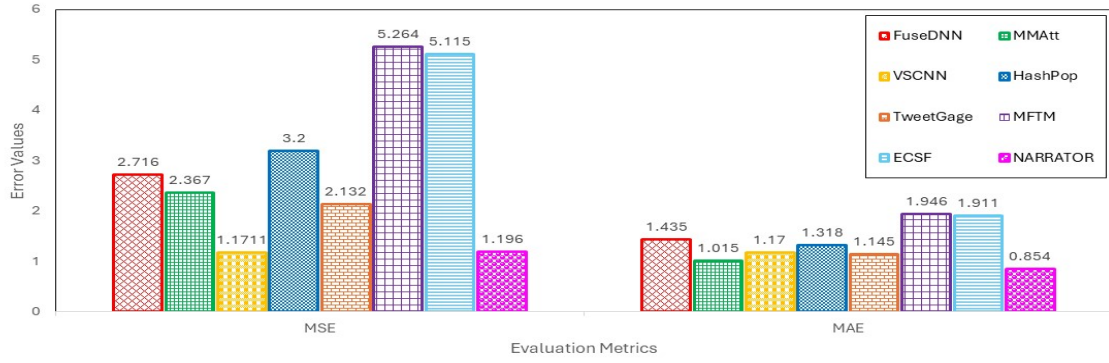


Figure 7.5: Effectiveness comparison curves on TPIC dataset

ment of 77.279% and 56.115% in terms of MSE and MAE on the TPIC dataset and 70.330% improvement in terms of MSE and 53.648% in terms of MAE on the SMP dataset. After feature extraction, MFTM employs an ensemble of TabNet and LightGBM which are Machine Learning models to predict post popularity whereas NARRATOR employs a hashtag-guided attention mechanism on content features and a deep neural network to forecast the post popularity. NARRATOR surpasses ECSF by a relative improvement of 76.617% in MSE and 55.311% in MAE on the TPIC dataset and 70.190% improvement in MSE and 53.088% in MAE on the SMP dataset. ECSF relies on feature stacking and a CatBoost model for prediction, which limits its ability to capture complex relationships between features. NARRATOR's deep neural network with a hashtag-guided attention mechanism overcomes this limitation by learning more intricate feature interactions. As illustrated in Figure 7.5 and Figure 7.6, NARRATOR demonstrates superior performance compared to all four baseline models on both datasets. This is evident in the consistently lower MSE and MAE values achieved by NARRATOR. Notably, on the TPIC dataset, NARRATOR achieves the lowest MSE of 1.196 and the lowest MAE of 0.854. Similarly, on the SMP dataset, NARRATOR outperforms other models with an MSE of 2.022 and an MAE of 0.972. These results suggest that NARRATOR effectively captures the underlying factors influencing social media post popularity across different datasets. This implies that our derived features, such as sentiments, demographics, and hashtags, are essential for determining how popular social media posts are. The fact that hashtags influence

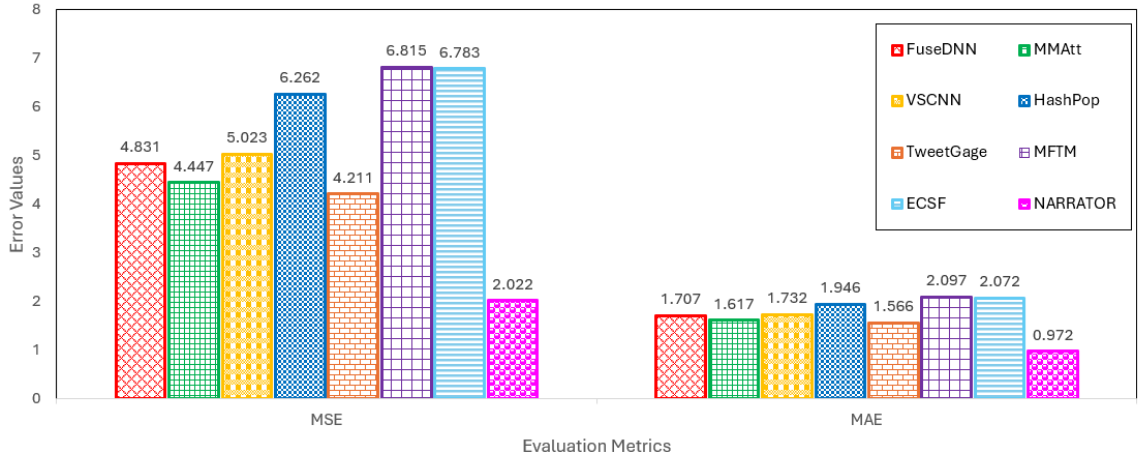


Figure 7.6: Performance comparison curves on SMP dataset

Table 7.2: Feature ablation study

Variant	TPIC		SMP	
	MSE	MAE	MSE	MAE
NARRATOR w/o (Sentiment from Hashtags+ Demographics)	1.387	0.912	2.367	1.084
NARRATOR w/o Sentiment from Hashtags	1.372	0.872	2.346	1.092
NARRATOR w/o Demographics	1.238	0.897	2.465	1.131
NARRATOR	1.196	0.854	2.022	0.972

MSE: Mean Squared Error, MAE: Mean Average Error

both textual and visual content features and enrich the overall representation of the posts in predicting post popularity on social network platforms is another important result.

7.4.2.2 Ablation Studies

In this section, we discuss the model’s performance by analyzing feature combinations, and the effectiveness of different attention mechanisms on the performance of the proposed model.

7.4.2.2.1 Feature Ablation We investigate the role of our two novel features- visual demographics and sentiment extracted from hashtags in enhancing popularity prediction. Table 7.2 demonstrates that the complete NARRATOR model, incorporating both novel features- visual demographics and sentiment extracted from hashtags, achieves superior performance compared to ablated versions. The complete model

achieves the lowest MSE and MAE values across both TPIC and SMP datasets, highlighting the significant contribution of these features to accurate popularity prediction. This degradation is particularly pronounced when both features are removed simultaneously, resulting in an absolute increase of 13.77% and 6.36% in MSE and MAE on TPIC, and 14.58% and 10.33% on SMP, respectively. Removing demographics and sentiment features from hashtags affects the model’s ability to understand the target audience and the emotional tone of the post, both of which could be relevant for popularity. The exclusion of visual demographics leads to an absolute increase of 3.39%, 4.79% in MSE and MAE on TPIC and 17.97% and 14.06% in MSE and MAE on SMP dataset, respectively. This highlights the importance of understanding the target audience. Visual demographics provide insights into the age group, gender, and other visual cues that may resonate with specific user segments, enabling the model to better predict post appeal. The exclusion of sentiments from hashtags leads to an absolute increase of 12.83%, 2.06% in MSE and MAE on TPIC and 13.81% and 10.99% in MSE and MAE on SMP dataset, respectively. This emphasizes the role of emotional tone in driving post popularity. Hashtags often encapsulate the sentiment or theme associated with a post. By incorporating hashtag sentiment, the model can gauge the emotional appeal of the content, which is a key factor influencing user engagement and sharing behavior. In essence, visual demographics help the model understand who the content might appeal to, while hashtag sentiment helps understand how the content might make the audience feel. By integrating both, NARRATOR gains a more comprehensive understanding of the factors driving popularity, enabling it to make more accurate predictions. This underscores the synergistic effect of combining visual demographics and hashtag sentiment for understanding and predicting multimodal post popularity.







7.4.2.2.2 Attention Mechanisms To demonstrate the significance of the novel hashtag-guided attention mechanism, we compare its performance with various attention mechanisms discussed below. Table 7.3 showcases the performance of NARRATOR with different attention mechanisms. Here, the variants that use no at-

Table 7.3: Effectiveness of attention mechanisms

Attention Technique	TPIC		SMP	
	MSE	MAE	MSE	MAE
$NARRATOR_{NA}$	1.510	0.917	3.715	1.417
$NARRATOR_{SA}$	1.567	0.944	3.474	1.445
$NARRATOR_{CA}$	1.486	0.916	2.904	1.218
$NARRATOR_{PA}$	1.389	0.905	3.367	1.359
$NARRATOR_{HGA}$	1.196	0.854	2.022	0.972

MSE: Mean Squared Error, MAE: Mean Average Error

tention, self-attention, cross-attention, parallel co-attention, and hashtag-guided attention are $NARRATOR_{NA}$, $NARRATOR_{SA}$, $NARRATOR_{CA}$, $NARRATOR_{PA}$, $NARRATOR_{HGA}$ respectively. The performance difference when NARRATOR is implemented without any attention mechanism is 23.68% and 9.53% in MSE and MAE on the TPIC dataset and 41.80% and 32.73% on the SMP dataset compared to hashtag-guided attention. The performance of NARRATOR is the lowest in the absence of any attention mechanism. Compared to self-attention, hashtag-guided attention shows and absolute improvement of 23.68%, 9.53% in MSE and MAE on TPIC dataset and 41.80% and 32.73% on SMP dataset. The superior performance of hashtag-guided attention over self-attention indicates that solely modeling intra-modal relationships (within text or image) is less effective than incorporating the semantic context provided by hashtags. Compared to cross-attention, hashtag-guided attention shows and absolute improvement of 19.52%, 6.77% in MSE and MAE on TPIC dataset and 30.37% and 20.20% on SMP dataset. While cross-attention improves performance by modeling inter-modal interactions (between text and image), hashtag-guided attention further refines this by leveraging the contextual cues embedded in hashtags, leading to a better understanding of content. Our hashtag-guided attention mechanism outperforms parallel co-attention, demonstrating an absolute improvement of 13.89% and 56.35% on TPIC and 39.95% and 39.81% on SMP in terms of MSE and MAE metrics, respectively. Unlike the parallel co-attention mechanism that solely focuses on the relationship between text and image features, the proposed hashtag-guided attention mechanism introduces a crucial element: the influence of hashtags on both modalities. The superior performance of hashtag-guided attention compared

 <div>  Holi Festival in Madrid </div> <div>  #holi #madrid #españa #spain #india #bollywood #bolymadrid </div>		 <div>  IMG_2130 </div> <div>  #prop8 #sanfrancisco #rally #demonstration #protest #california </div>	
Method	Predicted Popularity	Method	Predicted Popularity
NARRATOR	4.195	NARRATOR	1.580
ECSF	4.723	ECSF	4.716
MFTM	4.971	MFTM	5.048
TweetGage	6.754	TweetGage	2.457
HashPop	6.554	HashPop	4.483
VSCNN	5.325	VSCNN	2.907
MMAtt	2.975	MMAtt	6.192
FuseDNN	6.055	FuseDNN	3.733
Ground truth	4.200	Ground truth	1.600

(a) Post 1
(b) Post 2

Figure 7.7: Posts depicting popularity scores predicted by different methods

to parallel co-attention emphasizes the value of explicitly incorporating hashtag semantics into the attention mechanism. Hashtags provide a bridge between textual and visual content, allowing the model to focus on the most relevant aspects of both modalities for popularity prediction.

7.4.2.3 Qualitative Analysis

Predicting the popularity score of the given post is a common evaluation protocol in popularity prediction tasks. This section presents a qualitative evaluation aimed at understanding how accurately our proposed model predicts the popularity of social media posts. Example posts with accompanying captions, images, and associated hashtags are illustrated. We also display the ground-truth popularity scores and popularity scores predicted by state-of-the-art methods. These example posts have been chosen randomly from test data. As can be seen in Figure 7.7(a), our proposed model predicts a popularity score that is very close to the ground-truth popularity score. The caption of the first example post i.e., “Holi festival in Madrid”, bears a striking resemblance to the accompanying image, which depicts a girl celebrating the festival.

Hashtags such as #holi and #spain indicate a cultural celebration happening outside its traditional location (India), #india and #bollywood suggest the post might target the Indian diaspora or Bollywood fans in Spain. The model focuses on visual features in the image related to the Holi celebration (colors, people celebrating). Textual features in the caption (“Holi festival”) are analyzed alongside hashtags such as #india to understand the cultural significance. Hashtags such as #spain and #bollymadrid help identify a potential audience interested in Indian culture or Bollywood within Spain. By considering these contextual cues from hashtags, the model refines its understanding of the post’s content and target audience. This validates our hypothesis that modeling the interaction between the title and image under the influence of hashtags via a hashtag-guided attention mechanism effectively captured the relevant aspects for predicting engagement and considerably enhanced the model’s performance. Furthermore, our experimental results show that sentiments are related to post popularity. The hashtags #holi and #india are directly related to the post caption as Holi is the festival of colors that celebrates spring and the triumph of good over evil in India and is associated with joy, love, and new beginnings. Hashtag “#bollymadrid” combines Bollywood with Madrid, suggesting a fusion of Indian and Spanish culture, which can be seen as positive. Overall, the post promotes a positive and inclusive sentiment about cultural exchange. The positive mood communicated by captions and hashtags aids the model in anticipating better outcomes. Further, as illustrated in Figure 7.7(b), the ground-truth popularity score for the post is 1.6. Our model’s prediction closely approximates this value, achieving a score of 1.580. Overall, this research helps to visualize our model’s ability to estimate post popularity accurately by leveraging innovative features.

7.4.2.4 Feature Ranking and Importance

To assess the importance of derived features, we independently ranked the performance of each feature in terms of MSE and MAE, with rank 1 indicating optimal performance. Following that, by aggregating the preliminary ranks acquired for each parameter, the resulting average rank was used as a comprehensive assessment of over-

Table 7.4: Feature ranking and importance

Feature	TPIC		SMP	
	MSE/MAE	Rank	MSE/ MAE	Rank
w/o Sentiment (Text)	1.362 ₍₅₎ /0.867 ₍₆₎	5.5	2.367 ₍₅₎ / 1.084 ₍₆₎	5.5
w/o Sentiment (Hashtags)	4.51.372 ₍₄₎ /0.872 ₍₅₎	4.5	2.346 ₍₆₎ / 1.092 ₍₅₎	5.5
w/o Hashtags	2.553 ₍₁₎ / 1.251 ₍₁₎	1	2.458 ₍₄₎ / 1.125 ₍₄₎	4
w/o Demographics	1.238 ₍₆₎ / 0.897 ₍₄₎	5	2.465 ₍₃₎ /1.131 ₍₃₎	3
w/o Social	2.166 ₍₂₎ / 1.093 ₍₂₎	2	2.897 ₍₂₎ /1.267 ₍₁₎	1.5
w/o Content	1.435 ₍₃₎ / 0.908 ₍₃₎	3	3.021 ₍₁₎ /1.228 ₍₂₎	1.5

MSE: Mean Squared Error, MAE: Mean Average Error

all performance. Table 7.4 highlights the significance of both established and newly introduced features in predicting post popularity. While certain features such as social information and hashtags consistently rank high across both datasets, showcasing their fundamental role in capturing engagement and contextual information, respectively, the impact of other features such as sentiment and visual demographics appears to be more nuanced and context-dependent. The relatively lower individual ranks of derived features such as sentiments from hashtags and visual demographics might not fully reflect their true value. Their strength lies in their synergistic contribution to the overall model, capturing subtle nuances and previously overlooked aspects of popularity dynamics. In particular, they address specific gaps in prior research, offering a more comprehensive understanding of the factors influencing post popularity. Furthermore, the importance of features can vary depending on the specific platform and its user base. While some features might be universally influential, others might play a more significant role in specific contexts or for particular types of posts. Our analysis underscores the complex and multifaceted nature of popularity prediction, highlighting the need for a holistic approach that considers a diverse range of features and their interactions.

7.4.2.5 Implementation Details

The experiments leveraged a high-performance computing environment featuring a Linux server architecture. The server’s processing power included an Intel(R) Xeon(R) Silver 4215R CPU operating at 3.20 GHz, complemented by 256 GB of RAM and a

dedicated NVIDIA Tesla T4 GPU with 16 GB of memory. This configuration facilitated efficient model training and experimentation. To achieve optimal model performance, a rigorous hyperparameter tuning process was conducted. The learning rate was meticulously set to 0.0001, ensuring convergence without excessive learning speed. A batch size of 20 was chosen to strike a balance between computational efficiency and gradient estimation accuracy. The Adam optimizer, renowned for its adaptive learning rate capabilities, was employed to facilitate efficient optimization. Training proceeded for a maximum of 30 epochs, incorporating an early stopping mechanism with a patience of 5 epochs to prevent overfitting. Additionally, a dropout rate of 0.2 was strategically applied after each layer within the model architecture to further mitigate overfitting tendencies. To ensure a level playing field for performance evaluation, the embedding dimension (D) was consistently set to 768 for all comparative methods employed in the study. Before feeding image data into the VGG-19 network, all image samples underwent a standardized pre-processing step. This step involved rescaling each image to a uniform size of 224 x 224 pixels. This normalization ensured consistent image representation and facilitated network training. In Deep Feed Forward Network, we employ a series of 12 fully connected layers of size 13552, 6776, and so on till size 1.

7.5 Conclusion

In this chapter, we propose a novel paradigm for forecasting the popularity of social media posts by leveraging multimodal characteristics. Our approach leverages a multifaceted feature extraction process, capturing content-based information from both text and visuals, sentiment-oriented information from hashtags and text, user demographics, and social network data, along with topical and structural characteristics derived from hashtags. We propose a novel hashtag-guided attention mechanism that captures the influence of hashtags on both visual and textual content. This mechanism facilitates the model in learning the relative importance of different image regions and text segments based on their association with hashtags, leading to a more nuanced

understanding of how hashtags shape user engagement. To demonstrate the efficacy of our proposed method, we undertake quantitative and qualitative comparisons in addition to ablation and statistical investigations. Our method achieves significant performance improvements compared to existing state-of-the-art approaches, as evaluated on two real-world datasets. This finding suggests the potential effectiveness of our proposed method for predicting the popularity of multimodal posts.

Chapter 8

Conclusion and Future Work

In this chapter, we first present key findings of thesis, followed by a discussion of prospective directions for future research.

8.1 Summary of the Thesis

In this thesis, we addressed problems of content discoverability and reachability by devising automated methods for hashtag recommendation and popularity prediction in social networks. We analyzed prominent modalities of UGC, namely monolingual content, multilingual content, multimodal content comprising textual and visual modalities, and micro-videos. A brief summary of all the proposed solutions is presented in the following subsections.

8.1.1 Hashtag Recommendation for Monolingual Content

We introduce a novel retrieval-augmented diffusion-based sequence-to-sequence framework for monolingual text-based hashtag recommendation. This work pioneers the application of diffusion models to this task, strategically integrating the contextual awareness inherent in information retrieval with the generative capabilities of diffusion models. Employing an encoder-decoder transformer architecture, our framework leverages retrieved hashtags from semantically similar posts to contextually guide the sequential generation of relevant hashtags. The newly proposed adaptive non-linear

noise scheduler significantly enhances the quality of generated hashtags by providing fine-grained control over token-level diffusion process, effectively capturing dynamic nature of language. The diffusion-based generative approach overcomes the limitations of traditional encoder-decoder models that relies on maximum likelihood estimation to produce generic hashtags. By reversing a gradual noising process, our method explores a broader spectrum of hashtag possibilities, yielding more diverse and contextually appropriate recommendations. Furthermore, the integration of self-conditioning within the generator optimizes the utilization of previously predicted sequence information, leading to improved coherence. Empirical evaluations demonstrate the superior performance of our proposed framework against state-of-the-art methods in both hashtag quality and training efficiency.

8.1.2 Hashtag Recommendation for Multilingual Content

In Chapter 4, we develop a novel hashtag recommendation method to enhance content discoverability and bridge language barriers for content in low-resource languages on social networks. Our method enriches tweet representations by leveraging the user’s topical and linguistic preferences, historical posting behavior, and language relatedness, yielding pertinent hashtag suggestions. Experimental results, derived from a curated dataset from X, demonstrate the efficacy of TAGALOG. It significantly outperforms recognized pre-trained language models and existing research, with average F1-score improvements of 12.3% and 12.8%, respectively. These substantial improvements underscore its ability to recommend hashtags that resonate with individual user interests and linguistic inclinations, leading to a more tailored and engaging user experience. These findings effectively validate the potential of personalized and multilingual hashtag recommendation systems in facilitating multilingual content retrieval and improving the discoverability and relevance of content within low-resource language communities.

8.1.3 Hashtag Recommendation for Multimodal Content

In Chapter 5, we propose a hybrid deep neural network to recommend personalized and relevant hashtags for multimodal microblogs devoid of hashtags. We formulate hashtag recommendation through both classification and generation paradigms, effectively leveraging image, text, and user hashtagging behavior as crucial modalities. By learning intricate connections between information embedded within these modalities and analyzing user interests, DESIGN recommends pertinent hashtags. The architecture incorporates a word-level attention mechanism to identify significant textual segments and a parallel co-attention mechanism to bridge the semantic gap between visual and textual data through mutual representation learning. Notably, our proposed method demonstrates superior performance over existing approaches both quantitatively and qualitatively. Extensive evaluations on HARRISON, T-INS, and MMP-INS datasets underscores the efficacy of DESIGN for image-based, text-based, and multimodal hashtag recommendation.

8.1.4 Hashtag Recommendation for Micro-videos

In Chapter 6, we devise an automated system to recommend hashtags for micro-videos, a prevalent form of user-generated content requiring efficient organization. We develop a hybrid filtering approach that considers micro-video content, individual preferences, and shared interests of like-minded users. To this end, we construct a heterogeneous graph that models users' modality-specific tagging behavior by linking them to constituent modalities of their past micro-videos. This graph also incorporates user-to-user and modality-to-modality interactions to leverage explicit and implicit collaborative filtering signals. Extensive experiments across three real-world datasets demonstrate MISHON's comparative F1-score enhancement of 3.6%, 2.8%, and 6.5%, respectively, highlighting its robustness. Moreover, to address the cold-start user problem, we introduce a social influence and content-based solution. This technique recommends meaningful hashtags for cold-start users by modeling their interaction with influential users and popular tagging trends, thus overcoming initial data scarcity.

This solution exhibits a substantial relative F1-score improvement of 15.8% over a content-only approach, underscoring its effectiveness in expanding new users’ network and content visibility.

8.1.5 Popularity Prediction of Multimodal Content

We develop a sentiment and hashtag-aware attentive deep neural network for multimodal post popularity prediction, dubbed NARRATOR. It captures content information from textual and visual modalities, sentiment information from hashtags and captions, visual demographic features from faces, and social network data. Additionally, we incorporate topical and structural characteristics derived from hashtags, enabling a comprehensive analysis of factors influencing post popularity. A key contribution is a novel hashtag-guided attention mechanism. This mechanism models hashtag influence on visual and textual modalities, learning the importance of image regions and text segments related to hashtags. Experimental results on two real-world datasets demonstrate NARRATOR’s significant quantitative and qualitative outperformance of existing methods.

8.2 Future Work

Building upon the contributions of this thesis, we now present a few promising directions for subsequent research.

- [1] To recommend hashtags for monolingual content, we devised a retrieval augmented framework with a diffusion-based generator, comprising a retriever for candidate identification, a selector for relevance filtering, and a generator for informative hashtag creation. To further enhance the framework’s capabilities, future work will focus on refining the selector module’s ability to capture linguistic landscape of social media with greater accuracy, particularly in the face of vocabulary, syntactic, and semantic variations. We will also focus on optimizing the retriever’s efficiency to ensure scalability.

- [2] Our proposed method focuses on Indo-Aryan and Dravidian language families and is limited to investigating intra-family language relatedness. Future research will broaden this scope to include the Austroasiatic and Tibeto-Burman language groups, and investigating inter-family relationships. Furthermore, researchers can leverage data augmentation techniques to enhance training dataset and develop models capable of discerning hashtag usage patterns across the diverse cultural and linguistic landscape of the Indian subcontinent.
- [3] To recommend personalized hashtags for multimodal content, we interpret the task from both classification and generation perspectives. Future efforts will explore enhancing the model’s ability to recommend a wider range of relevant hashtags, including those with limited frequency, by incorporating keyword extraction techniques to generate supplementary keyword-based suggestions.
- [4] MISHON currently considers the static content of micro-videos when recommending hashtags. Nevertheless, micro-videos exhibit topicality and temporal sensitivity, implying that the relevance of hashtags associated with a given micro-video can evolve over time. In subsequent research, we intend to harness the temporal dimension inherent in micro-videos to yield recommendations that are more temporally aligned and pertinent. To further enhance MISHON, we envisage the incorporation of a broader spectrum of user-centric contextual information such as user profiles and geographical location data.
- [5] We predict popularity of multimodal content using visual demographic features, hashtag sentiment, and the interplay between texts, images, and hashtags. Future work will focus on enhancing the robustness of visual demographic features against challenging imaging conditions such as resolution, lighting, noise, occlusions through advanced feature enhancement techniques. Additionally, to improve performance across diverse contexts, we will incorporate culturally varied datasets and devise adaptive modeling approaches to address cultural disparities and platform-specific trends.

Bibliography

- [1] M. G. Rodriguez, K. Gummadi, and B. Schoelkopf, “Quantifying information overload in social media and its impact on social contagions,” in *Proc. International AAAI Conference on Web and Social Media (ICWSM)*, vol. 8, no. 1, 2014, pp. 170–179.
- [2] E. Salazar *et al.*, “Hashtags 2.0- An annotated history of the hashtag and a window to its future,” *Revista científica de Comunicación y Tecnologías emergentes (Revista ICONO 14)*, vol. 15, no. 2, pp. 16–54, 2017.
- [3] J. Li and H. Xu, “Suggest what to tag: Recommending more precise hashtags based on users’ dynamic interests and streaming tweet content,” *Knowledge-Based Systems (KBS)*, vol. 106, pp. 196–205, 2016.
- [4] D. Cao, L. Miao, H. Rong, Z. Qin, and L. Nie, “Hashtag our stories: Hashtag recommendation for micro-videos via harnessing multiple modalities,” *Knowledge-Based Systems (KBS)*, vol. 203, p. 106114, 2020.
- [5] S. Zhang, Y. Yao, F. Xu, H. Tong, X. Yan, and J. Lu, “Hashtag recommendation for photo sharing services,” in *Proc. AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 5805–5812.
- [6] M. Parkinson, “The power of visual communication,” *Billion Dollar Graphics*, vol. 660, 2012.
- [7] S. Bansal, K. Gowda, and N. Kumar, “Multilingual personalized hashtag recommendation for low resource Indic languages using graph-based deep neural network,” *Expert Systems with Applications (ESWA)*, vol. 236, p. 121188, 2024.

- [8] K. K. Pandey and S. Jha, “Exploring the interrelationship between culture and learning: the case of English as a second language in India,” *Asian Englishes*, pp. 1–17, 2021.
- [9] Y. Wang, J. Li, I. King, M. R. Lyu, and S. Shi, “Microblog hashtag generation via encoding conversation contexts,” in *Proc. North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers) (NAACL)*, 2019, pp. 1624–1633.
- [10] A. Javari, Z. He, Z. Huang, R. Jeetu, and K. Chen-Chuan Chang, “Weakly supervised attention for hashtag recommendation using graph data,” in *Proc. ACM Web Conference (WWW)*, 2020, pp. 1038–1048.
- [11] Q. Mao, X. Li, B. Liu, S. Guo, P. Hao, J. Li, and L. Wang, “Attend and select: A segment selective transformer for microblog hashtag generation,” *Knowledge-Based Systems (KBS)*, vol. 254, p. 109581, 2022.
- [12] Q. Zhang, J. Wang, H. Huang, X. Huang, and Y. Gong, “Hashtag recommendation for multimodal microblog using co-attention network,” in *Proc. International Joint Conference on Artificial Intelligence (IJCAI)*, 2017, pp. 3420–3426.
- [13] Y. Gong, Q. Zhang, and X. Huang, “Hashtag recommendation for multimodal microblog posts,” *Neurocomputing*, vol. 272, pp. 170–177, 2018.
- [14] M. Hasan, E. Agu, and E. Rundensteiner, “Using hashtags as labels for supervised learning of emotions in Twitter messages,” in *ACM SIGKDD Workshop on Health Informatics, New York, USA*, vol. 34, no. 74, 2014, p. 100.
- [15] T. Highfield and T. Leaver, “A methodology for mapping Instagram hashtags,” *First Monday*, vol. 20, no. 1, pp. 1–11, 2015.
- [16] K. W. Lim and W. Buntine, “Twitter opinion topic model: Extracting product opinions from tweets by leveraging hashtags and sentiment lexicon,” in *Proc. 23rd ACM International Conference on Information and Knowledge Management (CIKM)*, 2014, pp. 1319–1328.

- [17] C. H. Basch, J. Fera, A. Pellicane, and C. E. Basch, “Videos with the hashtag# vaping on TikTok and implications for informed decision-making by adolescents: descriptive study,” *JMIR Pediatrics and Parenting*, vol. 4, no. 4, p. e30681, 2021.
- [18] Y. Wei, Z. Cheng, X. Yu, Z. Zhao, L. Zhu, and L. Nie, “Personalized hashtag recommendation for micro-videos,” in *Proc. 27th ACM International Conference on Multimedia (MM)*, 2019, pp. 1446–1454.
- [19] A.-A. Liu, X. Wang, N. Xu, J. Guo, G. Jin, Q. Zhang, Y. Tang, and S. Zhang, “A review of feature fusion-based media popularity prediction methods,” *Visual Informatics*, vol. 6, no. 4, pp. 78–89, 2022.
- [20] N. J. G. Amala and K. Kumar, “Content popularity prediction methods-a survey,” in *Proc. 3rd International Conference on Communication and Electronics Systems (ICCES)*. IEEE, 2018, pp. 749–753.
- [21] M. A. Gonçalves, J. M. Almeida, L. G. dos Santos, A. H. Laender, and V. Almeida, “On popularity in the blogosphere,” *IEEE Internet Computing*, vol. 14, no. 3, pp. 42–49, 2010.
- [22] C. Li, Y. Lu, Q. Mei, D. Wang, and S. Pandey, “Click-through prediction for advertising in Twitter timeline,” in *Proc. 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2015, pp. 1959–1968.
- [23] J. Wang, B. Xu, and Y. Zu, “Deep learning for aspect-based sentiment analysis,” in *Proc. International Conference on Machine Learning and Intelligent Systems Engineering (MLISE)*. IEEE, 2021, pp. 267–271.
- [24] J. R. Saura, “Using data sciences in digital marketing: Framework, methods, and performance metrics,” *Journal of Innovation & Knowledge*, vol. 6, no. 2, pp. 92–102, 2021.
- [25] S. Bansal, M. Kumar, C. S. Raghaw, and N. Kumar, “Sentiment and hashtag-aware attentive deep neural network for multimodal post popularity prediction,” *Neural Computing and Applications (NCAA)*, vol. 37, no. 4, pp. 2799–2824, 2025.

- [26] M. Luca, “User-generated content and social media,” in *Handbook of media Economics*. Elsevier, 2015, vol. 1, pp. 563–592.
- [27] R. Zhu, D. Yang, and Y. Li, “Learning improved semantic representations with tree-structured LSTM for hashtag recommendation: An experimental study,” *Information*, vol. 10, no. 4, p. 127, 2019.
- [28] N. Kumar, E. Baskaran, A. Konjengbam, and M. Singh, “Hashtag recommendation for short social media texts using word-embeddings and external knowledge,” *Knowledge and Information Systems (KAIS)*, vol. 63, pp. 175–198, 2021.
- [29] P. Chakrabarti, E. Malvi, S. Bansal, and N. Kumar, “Hashtag recommendation for enhancing the popularity of social media posts,” *Social Network Analysis and Mining (SNAM)*, vol. 13, no. 1, p. 21, 2023.
- [30] Y.-C. Chen, K.-T. Lai, D. Liu, and M.-S. Chen, “TagNet: Triplet-attention graph networks for hashtag recommendation,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 3, pp. 1148–1159, 2021.
- [31] R. V. Lakshmi, D. Gerard, A. Santhanavijayan, and S. Radha, “A hybrid classifier-based ontology driven image tag recommendation framework for social image tagging,” *Procedia Computer Science*, vol. 218, pp. 67–73, 2023.
- [32] J. Sadr *et al.*, “Popular tag recommendation by neural network in social media,” *Computational Intelligence and Neuroscience*, vol. 2023, 2023.
- [33] S. Bansal, K. Gowda, and N. Kumar, “A hybrid deep neural network for multi-modal personalized hashtag recommendation,” *IEEE Transactions on Computational Social Systems (TCSS)*, vol. 10, no. 5, pp. 2439–2459, 2022.
- [34] H.-S. Won, S.-M. Roh, D. Kim, M.-J. Kim, H. Kim, and K.-M. Kim, “EXTRA: Integrating external knowledge into multimodal hashtag recommendation system,” in *Proc. IEEE International Conference on Web Services (ICWS)*. IEEE, 2023, pp. 719–721.

- [35] Z. Ding, X. Qiu, Q. Zhang, and X. Huang, “Learning topical translation model for microblog hashtag suggestion,” in *Proc. 23rd International Joint Conference on Artificial Intelligence (IJCAI)*, 2013.
- [36] S. Sedhai and A. Sun, “Hashtag recommendation for hyperlinked tweets,” in *Proc. 37th International ACM SIGIR Conference on Research & Development in Information Retrieval*, 2014, pp. 831–834.
- [37] L. Marujo, W. Ling, I. Trancoso, C. Dyer, A. W. Black, A. Gershman, D. M. de Matos, J. P. Neto, and J. G. Carbonell, “Automatic keyword extraction on Twitter,” in *Proc. 53rd Annual Meeting of the Association for Computational Linguistics (ACL) and the 7th International Joint Conference on Natural Language Processing (IJCNLP) (Volume 2: Short Papers)*, 2015, pp. 637–643.
- [38] Y. Gong, Q. Zhang, and X.-J. Huang, “Hashtag recommendation using Dirichlet process mixture models incorporating types of hashtags,” in *Proc. 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2015, pp. 401–410.
- [39] Y. Zhang, J. Li, Y. Song, and C. Zhang, “Encoding conversation context for neural keyphrase extraction from microblog posts,” in *Proc. 2018 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL): Human Language Technologies, Volume 1 (Long Papers)*, 2018, pp. 1676–1686.
- [40] Q. Zhang, Y. Wang, Y. Gong, and X.-J. Huang, “Keyphrase extraction using deep recurrent neural networks on Twitter,” in *Proc. 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2016, pp. 836–845.
- [41] J. Li, H. Xu, X. He, J. Deng, and X. Sun, “Tweet modeling with LSTM recurrent neural networks for hashtag recommendation,” in *Proc. International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2016, pp. 1570–1577.

- [42] X. Zheng, D. Mekala, A. Gupta, and J. Shang, “News meets microblog: Hashtag annotation via retriever-generator,” *arXiv preprint arXiv:2104.08723*, 2021.
- [43] Z. Li, X. Wang, W. Yang, J. Wu, Z. Zhang, Z. Liu, M. Sun, H. Zhang, and S. Liu, “A unified understanding of deep NLP models for text classification,” *IEEE Transactions on Visualization and Computer Graphics*, vol. 28, no. 12, pp. 4980–4994, 2022.
- [44] V. Dogra *et al.*, “A complete process of text classification system using state-of-the-art NLP models,” *Computational Intelligence and Neuroscience*, vol. 2022, 2022.
- [45] X. Li, X. Wu, Z. Luo, Z. Du, Z. Wang, and C. Gao, “Integration of global and local information for text classification,” *Neural Computing and Applications (NCAA)*, vol. 35, no. 3, pp. 2471–2486, 2023.
- [46] K. Lei, Q. Fu, M. Yang, and Y. Liang, “Tag recommendation by text classification with attention-based capsule network,” *Neurocomputing*, vol. 391, pp. 65–73, 2020.
- [47] M. Pathak and A. Jain, “ μ boost: An effective method for solving Indic multilingual text classification problem,” in *Proc. IEEE 8th International Conference on Multimedia Big Data (BigMM)*. IEEE, 2022, pp. 96–100.
- [48] D. Sanghvi, L. M. Fernandes, S. D’Souza, N. Vasaani, and K. Kavitha, “Fine-tuning of multilingual models for sentiment classification in code-mixed Indian language texts,” in *Proc. 19th International Conference on Distributed Computing and Intelligent Technology (ICDCIT)*. Springer, 2023, pp. 224–239.
- [49] M. Z. U. Rehman, S. Mehta, K. Singh, K. Kaushik, and N. Kumar, “User-aware multilingual abusive content detection in social media,” *Information Processing & Management (IPM)*, vol. 60, no. 5, p. 103450, 2023.
- [50] X. Zhang, Y. Malkov, O. Florez, S. Park, B. McWilliams, J. Han, and A. El-Kishky, “TwHIN-BERT: A socially-enriched pre-trained language model for mul-

- tilingual tweet representations at Twitter,” in *Proc. 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2023, pp. 5597–5607.
- [51] L. Besacier, E. Barnard, A. Karpov, and T. Schultz, “Automatic speech recognition for under-resourced languages: A survey,” *Speech communication*, vol. 56, pp. 85–100, 2014.
 - [52] Y. Khemchandani, S. Mehtani, V. Patil, and A. Awasthi, “Exploiting language relatedness for low web-resource language model adaptation: An Indic languages study,” in *Proc. 59th Annual Meeting of the Association for Computational Linguistics (ACL) and 11th International Joint Conference on Natural Language Processing (IJCNLP) (Volume 1: Long Papers)*, 2021.
 - [53] S. Aggarwal, S. Kumar, and R. Mamidi, “Efficient multilingual text classification for Indian languages,” in *Proc. International Conference on Recent Advances in Natural Language Processing (RANLP)*, 2021, pp. 19–25.
 - [54] J. Khatri, N. Saini, and P. Bhattacharyya, “Language relatedness and lexical closeness can help improve multilingual NMT: IITBombay@ MultiIndicNMT WAT2021,” in *Proc. 8th Workshop on Asian Translation (WAT2021)*, 2021, pp. 217–223.
 - [55] M. Marreddy, S. R. Oota, L. S. Vakada, V. C. Chinni, and R. Mamidi, “Multi-task text classification using graph convolutional networks for large-scale low resource language,” in *Proc. International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2022, pp. 1–8.
 - [56] M. Schlichtkrull, T. N. Kipf, P. Bloem, R. Van Den Berg, I. Titov, and M. Welling, “Modeling relational data with graph convolutional networks,” in *Proc. 15th International Conference on the Semantic Web (ESWC)*. Springer, 2018, pp. 593–607.

- [57] B. Sigurbjörnsson and R. Van Zwol, “Flickr tag recommendation based on collective knowledge,” in *Proc. 17th International Conference on World Wide Web (WWW)*, 2008, pp. 327–336.
- [58] Y. Gong, Y. Jia, T. Leung, A. Toshev, and S. Ioffe, “Deep convolutional ranking for multilabel image annotation,” *arXiv preprint arXiv:1312.4894*, 2013.
- [59] Y. Gong and Q. Zhang, “Hashtag recommendation using attention-based convolutional neural network,” in *Proc. International Joint Conference on Artificial Intelligence (IJCAI)*, vol. 16, 2016, pp. 2782–2788.
- [60] G. Wu, Y. Li, W. Yan, R. Li, X. Gu, and Q. Yang, “Hashtag recommendation with attention-based neural image hashtagging network,” in *Proc. International Conference on Neural Information Processing (ICONIP)*. Springer, 2018, pp. 52–63.
- [61] Y. S. Rawat and M. S. Kankanhalli, “ConTagNet: Exploiting user context for image tag recommendation,” in *Proc. 24th ACM International Conference on Multimedia (MM)*, 2016, pp. 1102–1106.
- [62] T. Yu, H. Yu, D. Liang, Y. Mao, S. Nie, P.-Y. Huang, M. Khabsa, P. Fung, and Y.-C. Wang, “Generating hashtags for short-form videos with guided signals,” in *Proc. 61st Annual Meeting of the Association for Computational Linguistics (ACL) (Volume 1: Long Papers)*, 2023, pp. 9482–9495.
- [63] C. Yang, X. Wang, and B. Jiang, “Sentiment enhanced multi-modal hashtag recommendation for micro-videos,” *IEEE Access*, vol. 8, pp. 78 252–78 264, 2020.
- [64] S. Mehta, S. Sarkhel, X. Chen, S. Mitra, V. Swaminathan, R. Rossi, A. Aminian, H. Guo, and K. Garg, “Open-domain trending hashtag recommendation for videos,” in *Proc. IEEE International Symposium on Multimedia (ISM)*. IEEE, 2021, pp. 174–181.

- [65] S. Tang, Y. Yao, S. Zhang, F. Xu, T. Gu, H. Tong, X. Yan, and J. Lu, “An integral tag recommendation model for textual content,” in *Proc. AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 5109–5116.
- [66] R. Ma, X. Qiu, Q. Zhang, X. Hu, Y.-G. Jiang, and X. Huang, “Co-attention memory network for multimodal microblog’s hashtag recommendation,” *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, vol. 33, no. 2, pp. 388–400, 2019.
- [67] M. Kaviani and H. Rahmani, “Emhash: Hashtag recommendation using neural network based on BERT embedding,” in *Proc. 6th International Conference on Web Research (ICWR)*. IEEE, 2020, pp. 113–118.
- [68] V. C. Tran, D. Hwang, and N. T. Nguyen, “Hashtag recommendation approach based on content and user characteristics,” *Cybernetics and Systems*, vol. 49, no. 5-6, pp. 368–383, 2018.
- [69] T. Durand, “Learning user representations for open vocabulary image hashtag prediction,” in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 9769–9778.
- [70] M. Peng, Y. Lin, L. Zeng, T. Gui, and Q. Zhang, “Modeling the long-term post history for personalized hashtag recommendation,” in *Proc. 18th China National Conference on Chinese Computational Linguistics (CCL)*. Springer, 2019, pp. 495–507.
- [71] D. Jeong, S. Oh, and E. Park, “DemoHash: Hashtag recommendation based on user demographic information,” *Expert Systems with Applications (ESWA)*, vol. 210, p. 118375, 2022.
- [72] U. Padungkiatwattana and S. Maneeroj, “Pac-man: Multi-relation network in social community for personalized hashtag recommendation,” *IEEE Access*, vol. 10, pp. 131 202–131 228, 2022.

- [73] M. Li, T. Gan, M. Liu, Z. Cheng, J. Yin, and L. Nie, “Long-tail hashtag recommendation for micro-videos with graph convolutional network,” in *Proc. 28th ACM International Conference on Information and Knowledge Management (CIKM)*, 2019, pp. 509–518.
- [74] S. Liu, J. Xie, C. Zou, and Z. Chen, “User conditional hashtag recommendation for micro-videos,” in *Proc. IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2020, pp. 1–6.
- [75] J. Shuai, L. Wu, K. Zhang, P. Sun, R. Hong, and M. Wang, “Topic-enhanced graph neural networks for extraction-based explainable recommendation,” in *Proc. 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2023, pp. 1188–1197.
- [76] N. Kumar, A. Yadandla, K. Suryamukhi, N. Ranabothu, S. Boya, and M. Singh, “Arousal prediction of news articles in social media,” in *Proc. 5th International Conference on Mining Intelligence and Knowledge Exploration (MIKE)*, vol. 10682 LNAI, 2017.
- [77] Z. Lin, F. Huang, Y. Li, Z. Yang, and W. Liu, “A layer-wise deep stacking model for social image popularity prediction,” *International World Wide Web Conference (WWW)*, vol. 22, 2019.
- [78] K. R. Purba, D. Asirvatham, and R. K. Murugesan, “Instagram post popularity trend analysis and prediction using hashtag, image assessment, and user history features,” *International Arab Journal of Information Technology*, vol. 18, 2021.
- [79] Q. Cao, H. Shen, J. Gao, B. Wei, and X. Cheng, “Popularity prediction on social platforms with coupled graph neural networks,” in *Proc. 13th International Conference on Web Search and Data Mining (WSDM)*, 2020.
- [80] K. Mannepalli, S. P. Singh, C. S. Kolli, S. Raj, G. R. Bojja, B. R. Rajakumar, and D. Binu, “Popularity prediction model with context, time and user sentiment

- information: An optimization assisted deep learning technique,” *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 31, 2023.
- [81] Y. Tan, F. Liu, B. Li, Z. Zhang, and B. Zhang, “An efficient multi-view multi-modal data processing framework for social media popularity prediction,” in *Proc. 30th ACM International Conference on Multimedia (MM)*, 2022, pp. 7200–7204.
 - [82] M. Zappavigna, “Searchable talk: The linguistic functions of hashtags,” *Social Semiotics*, vol. 25, no. 3, pp. 274–291, 2015.
 - [83] J. Liu, Z. He, and Y. Huang, “Hashtag2vec: Learning hashtag representation with relational hierarchical embedding model,” in *Proc. International Joint Conference on Artificial Intelligence (IJCAI)*, vol. 2018-July, 2018.
 - [84] Y. Y. Liao, “Leveraging hashtag networks for multimodal popularity prediction of Instagram posts,” in *Proc. 13th Language Resources and Evaluation Conference (LREC)*, N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk, and S. Piperidis, Eds. Marseille, France: European Language Resources Association, Jun. 2022, pp. 7191–7198.
 - [85] P. Chakrabarti, E. Malvi, S. Bansal, and N. Kumar, “Hashtag recommendation for enhancing the popularity of social media posts,” *Social Network Analysis and Mining (SNAM)*, vol. 13, 2023.
 - [86] K. Xu, Z. Lin, J. Zhao, P. Shi, W. Deng, and H. Wang, “Multimodal deep learning for social media popularity prediction with attention mechanism,” in *Proc. 28th ACM International Conference on Multimedia (MM)*, 2020.
 - [87] H. H. Lin, J. D. Lin, J. J. M. Ople, J. C. Chen, and K. L. Hua, “Social media popularity prediction based on multi-modal self-attention mechanisms,” *IEEE Access*, vol. 10, 2022.

- [88] S. Bansal, K. Gowda, and N. Kumar, “A hybrid deep neural network for multi-modal personalized hashtag recommendation,” *IEEE Transactions on Computational Social Systems (TCSS)*, vol. 10, no. 5, pp. 2439–2459, 2023.
- [89] J. Wang, S. Yang, H. Zhao, and Y. Yang, “Social media popularity prediction with multimodal hierarchical fusion model,” *Computer Speech and Language*, vol. 80, 2023.
- [90] M. Rhodan, “Please send help. Hurricane Harvey victims turn to Twitter and Facebook,” *Time, August*, vol. 30, 2017.
- [91] W. Frej, “Hurricane Florence flood victims turn to social media for rescue,” *HuffPost, 2018-09-14, Available at*, 2018.
- [92] L. Silverman, “Facebook, Twitter replace 911 calls for stranded in Houston,” *National Public Radio*, 2017.
- [93] J. R. Chowdhury, C. Caragea, and D. Caragea, “On identifying hashtags in disaster Twitter data,” in *Proc. AAAI Conference on Artificial Intelligence (AAAI)*, vol. 34, no. 01, 2020, pp. 498–506.
- [94] Q. Yang, G. Wu, Y. Li, R. Li, X. Gu, H. Deng, and J. Wu, “AMNN: Attention-based multimodal neural network model for hashtag recommendation,” *IEEE Transactions on Computational Social Systems (TCSS)*, vol. 7, no. 3, pp. 768–779, 2020.
- [95] S. Diao, S. S. Keh, L. Pan, Z. Tian, Y. Song, and T. Zhang, “Hashtag-guided low-resource tweet classification,” in *Proc. ACM Web Conference (WWW)*, 2023, pp. 1415–1426.
- [96] Z. Kong, W. Ping, J. Huang, K. Zhao, and B. Catanzaro, “Diffwave: A versatile diffusion model for audio synthesis,” in *Proc. International Conference on Learning Representations (ICLR)*, 2020.

- [97] J. Ho, T. Salimans, A. Gritsenko, W. Chan, M. Norouzi, and D. J. Fleet, “Video diffusion models,” *Advances in Neural Information Processing Systems*, vol. 35, pp. 8633–8646, 2022.
- [98] E. Austin, O. R. Zaïane, and C. Largeron, “Community topic: Topic model inference by consecutive word community discovery,” in *Proc. 29th International Conference on Computational Linguistics (COLING)*, 2022, pp. 971–983.
- [99] X. Li, J. Thickstun, I. Gulrajani, P. S. Liang, and T. B. Hashimoto, “Diffusion-LM improves controllable text generation,” *Advances in Neural Information Processing Systems*, vol. 35, pp. 4328–4343, 2022.
- [100] E. Zangerle, W. Gassler, and G. Specht, “Recommending #-tags in Twitter,” in *Proc. Workshop on Semantic Adaptive Social Web (SASWeb 2011). CEUR workshop proceedings*, vol. 730, 2011, pp. 67–78.
- [101] T. Chen, R. ZHANG, and G. Hinton, “Analog bits: Generating discrete data using diffusion models with self-conditioning,” in *Proc. 11th International Conference on Learning Representations (ICLR)*, 2022.
- [102] R. Strudel *et al.*, “Self-conditioned embedding diffusion for text generation,” in *Proc. European Chapter of the Association for Computational Linguistics (EACL) Conference*, 2022.
- [103] D. P. Kingma and M. Welling, “Auto-encoding variational Bayes,” in *Proc. International Conference on Learning Representations (ICLR)*, 2013.
- [104] M. Imran, P. Mitra, and C. Castillo, “Twitter as a lifeline: Human-annotated Twitter corpora for NLP of crisis-related messages,” in *Proc. 10th International Conference on Language Resources and Evaluation (LREC)*, 2016.
- [105] A. Olteanu, C. Castillo, F. Diaz, and S. Vieweg, “Crisislex: A lexicon for collecting and filtering microblogged communications in crises,” in *Proc. International AAAI Conference on Web and Social Media (ICWSM)*, vol. 8, no. 1, 2014, pp. 376–385.

- [106] F. Alam, F. Ofli, and M. Imran, “Crisismmd: Multimodal Twitter datasets from natural disasters,” in *Proc. International AAAI Conference on Web and Social Media (ICWSM)*, vol. 12, no. 1, 2018.
- [107] R.-Z. Fan, Y. Fan, J. Chen, J. Guo, R. Zhang, and X. Cheng, “RIGHT: Retrieval-augmented generation for mainstream hashtag recommendation,” in *Proc. European Conference on Information Retrieval (ECIR)*. Springer, 2024, pp. 39–55.
- [108] S. Gong, M. Li, J. Feng, Z. Wu, and L. Kong, “Diffuseq: Sequence to sequence text generation with diffusion models,” in *Proc. 11th International Conference on Learning Representations*, 2022.
- [109] H. Yuan, Z. Yuan, C. Tan, F. Huang, and S. Huang, “Text diffusion model with encoder-decoder transformers for sequence-to-sequence generation,” in *Proc. 2024 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL): Human Language Technologies (Volume 1: Long Papers)*, 2024, pp. 22–39.
- [110] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, “BERTScore: Evaluating text generation with BERT,” in *Proc. International Conference on Learning Representations (ICLR)*, 2019.
- [111] C.-Y. Lin, “ROUGE: A package for automatic evaluation of summaries,” in *Text summarization branches out*, 2004, pp. 74–81.
- [112] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, “Bleu: a method for automatic evaluation of machine translation,” in *Proc. 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2002, pp. 311–318.
- [113] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” in *Proc. 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL): Human Language Technologies, Volume 1 (Long and Short Papers)*, 2019, pp. 4171–4186.

- [114] T. Hachaj and J. Miazga, “Image hashtag recommendations using a voting deep neural network and associative rules mining approach,” *Entropy*, vol. 22, no. 12, p. 1351, 2020.
- [115] M. Park, H. Li, and J. Kim, “HARRISON: A benchmark on hashtag recommendation for real-world images in social networks,” *arXiv preprint arXiv:1605.05054*, 2016.
- [116] P. Kurunkar, O. Sawant, P. Mene, and N. Varghese, “An image-based hashtag recommendation system as a social media workflow tool,” in *Proc. International Conference on Smart Generation Computing, Communication and Networking (SMART GENCON)*. IEEE, 2022, pp. 1–5.
- [117] Y. Djenouri, A. Belhadi, G. Srivastava, and J. C.-W. Lin, “Deep learning based hashtag recommendation system for multimedia data,” *Information Sciences*, vol. 609, pp. 1506–1517, 2022.
- [118] P. Panchal and D. J. Prajapati, “The social hashtag recommendation for image and video using deep learning approach,” in *Proc. International Conference on Sentiment Analysis and Deep Learning (ICSADL)*. Springer, 2023, pp. 241–261.
- [119] Z. Yang and Z. Lin, “Interpretable video tag recommendation with multimedia deep learning framework,” *Internet Research*, vol. 32, no. 2, pp. 518–535, 2022.
- [120] V. Nama and G. Deepak, “DTagRecPLS: Diversification of tag recommendation for videos using preferential learning and differential semantics,” in *Proc. 14th International Conference on Soft Computing and Pattern Recognition (SoCPaR)*. Springer, 2023, pp. 887–898.
- [121] F.-F. Kou, J.-P. Du, C.-X. Yang, Y.-S. Shi, W.-Q. Cui, M.-Y. Liang, and Y. Geng, “Hashtag recommendation based on multi-features of microblogs,” *Journal of Computer Science and Technology*, vol. 33, pp. 711–726, 2018.

- [122] T. Pires, E. Schlinger, and D. Garrette, “How multilingual is multilingual BERT?” in *Proc. 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2019, pp. 4996–5001.
- [123] B. Perozzi, R. Al-Rfou, and S. Skiena, “Deepwalk: Online learning of social representations,” in *Proc. 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2014, pp. 701–710.
- [124] A. Grover and J. Leskovec, “node2vec: Scalable feature learning for networks,” in *Proc. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2016, pp. 855–864.
- [125] W. Hamilton, Z. Ying, and J. Leskovec, “Inductive representation learning on large graphs,” *Advances in Neural Information Processing Systems*, vol. 30, 2017.
- [126] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy, “Hierarchical attention networks for document classification,” in *Proc. 2016 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL): Human Language Technologies*, 2016, pp. 1480–1489.
- [127] D. Kakwani, A. Kunchukuttan, S. Golla, N. Gokul, A. Bhattacharyya, M. M. Khapra, and P. Kumar, “IndicNLP Suite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for Indian languages,” in *Findings of the Association for Computational Linguistics: EMNLP 2020*, 2020, pp. 4948–4961.
- [128] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, É. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, “Unsupervised cross-lingual representation learning at scale,” in *Proc. 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2020, pp. 8440–8451.
- [129] V. Sanh, “DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter.” in *Proc. 33rd Conference on Neural Information Processing Systems (NIPS2019)*, 2019.

- [130] A. Graves, “Supervised sequence labelling,” *Supervised Sequence Labelling with Recurrent Neural Networks*, pp. 5–13, 2012.
- [131] A. Andujar, “Analysing WhatsApp and Instagram as blended learning tools,” in *Recent Tools for Computer-and Mobile-Assisted Foreign Language Learning*. IGI Global, 2020, pp. 307–321.
- [132] K. Shively, “Simply measured q3 2014 Instagram study,” 2014.
- [133] P. Zhou, J. Liu, Z. Yang, and G. Zhou, “Scalable tag recommendation for software information sites,” in *Proc. IEEE 24th International Conference on Software Analysis, Evolution and Reengineering (SANER)*. IEEE, 2017, pp. 272–282.
- [134] N. Pervin, T. Q. Phan, A. Datta, H. Takeda, and F. Toriumi, “Hashtag popularity on Twitter: Analyzing co-occurrence of multiple hashtags,” in *Proc. International Conference on Social Computing and Social Media (SCSM)*. Springer, 2015, pp. 169–182.
- [135] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” in *Proc. International Conference on Learning Representations (ICLR)*, 2015.
- [136] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.
- [137] J. R. Anderson, *Cognitive psychology and its implications*. Macmillan, 2005.
- [138] D. Bahdanau, K. Cho, and Y. Bengio, “Neural machine translation by jointly learning to align and translate,” in *Proc. 3rd International Conference on Learning Representations (ICLR)*, 2015.
- [139] K. Xu, J. L. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. S. Zemel, and Y. Bengio, “Show, attend and tell: Neural image caption generation with

- visual attention,” in *Proc. 32nd International Conference on Machine Learning (ICML)*, vol. 3, 2015.
- [140] J. Lu, J. Yang, D. Batra, and D. Parikh, “Hierarchical question-image co-attention for visual question answering,” *Advances in Neural Information Processing Systems*, vol. 29, 2016.
- [141] P.-M. Caleffi, “The “hashtag”: A new word or a new rule?” *SKASE journal of Theoretical Linguistics*, vol. 12, no. 2, 2015.
- [142] Y. Wang, J. Qu, J. Liu, J. Chen, and Y. Huang, “What to tag your microblog: Hashtag recommendation based on topic analysis and collaborative filtering,” in *Asia-Pacific Web Conference (APWeb)*. Springer, 2014, pp. 610–618.
- [143] S. M. Kywe, T.-A. Hoang, E.-P. Lim, and F. Zhu, “On recommending hashtags in Twitter networks,” in *Proc. 4th International Conference on Social Informatics (SocInfo)*. Springer, 2012, pp. 337–350.
- [144] L. P. Torres and P. Valdiviezo-Diaz, “Hashtags recommendations in Twitter based on collaborative filtering,” in *Proc. International Conference of Digital Transformation and Innovation Technology (Incodtrin)*. IEEE, 2020, pp. 38–44.
- [145] A. Alsini, A. Datta, and D. Q. Huynh, “On utilizing communities detected from social networks in hashtag recommendation,” *IEEE Transactions on Computational Social Systems (TCSS)*, vol. 7, no. 4, pp. 971–982, 2020.
- [146] D. Kowald, S. C. Pujari, and E. Lex, “Temporal effects on hashtag reuse in Twitter: A cognitive-inspired hashtag recommendation approach,” in *Proc. 26th International Conference on World Wide Web (WWW)*, 2017, pp. 1401–1410.
- [147] X. Xing, W. Zhang, X. Zhang, and N. Xu, “Socitemrec: a framework for item recommendation in social networks,” *Journal of Theoretical and Applied Information Technology (JATIT)*, vol. 48, no. 3, 2013.

- [148] X. Wang, Y. Zhang, and T. Yamasaki, “User-aware folk popularity rank: User-popularity-based tag recommendation that can enhance social popularity,” in *Proc. 27th ACM International Conference on Multimedia (MM)*, 2019, pp. 1970–1978.
- [149] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, “An image is worth 16x16 words: Transformers for image recognition at scale,” in *Proc. International Conference on Learning Representations (ICLR)*, 2021.
- [150] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, “wav2vec 2.0: A framework for self-supervised learning of speech representations,” *Advances in Neural Information Processing Systems*, vol. 33, pp. 12 449–12 460, 2020.
- [151] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, “Librispeech: an ASR corpus based on public domain audio books,” in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.
- [152] B. M. Jafari, X. Luo, and A. Jafari, “Unsupervised keyword extraction for hashtag recommendation in social media,” in *Proc. International FLAIRS Conference*, vol. 36, 2023.
- [153] R. Cantini, F. Marozzo, G. Bruno, and P. Trunfio, “Learning sentence-to-hashtags semantic mapping for hashtag recommendation on microblogs,” *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 16, no. 2, pp. 1–26, 2021.
- [154] Y. Djenouri, A. Belhadi, G. Srivastava, and J. C.-W. Lin, “Toward a cognitive-inspired hashtag recommendation for Twitter data analysis,” *IEEE Transactions on Computational Social Systems (TCSS)*, vol. 9, no. 6, pp. 1748–1757, 2022.

- [155] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need,” *Advances in Neural Information Processing Systems*, vol. 30, 2017.
- [156] J. Chen, X. Song, L. Nie, X. Wang, H. Zhang, and T.-S. Chua, “Micro tells macro: Predicting the popularity of micro-videos via a transductive model,” in *Proc. 24th ACM International Conference on Multimedia (MM)*, 2016, pp. 898–907.
- [157] B. Thomee, D. A. Shamma, G. Friedland, B. Elizalde, K. Ni, D. Poland, D. Borth, and L.-J. Li, “Yfcc100m: The new data in multimedia research,” *Communications of the ACM*, vol. 59, no. 2, pp. 64–73, 2016.
- [158] S. Liu, H. Liu, Z. Chen, and X. Hu, “User-video co-attention network for personalized micro-video recommendation,” in *Proc. World Wide Web Conference (WWW)*, 2019, pp. 3020–3026.
- [159] Q. Wang, Y. Wei, J. Yin, J. Wu, X. Song, and L. Nie, “DualGNN: Dual graph neural network for multimedia recommendation,” *IEEE Transactions on Multimedia*, vol. 25, pp. 1074–1084, 2023.
- [160] F. Lei, Z. Cao, Y. Yang, Y. Ding, and C. Zhang, “Learning the user’s deeper preferences for multi-modal recommendation systems,” *ACM Transactions on Multimedia Computing, Communications and Applications*, vol. 19, no. 3s, pp. 1–18, 2023.
- [161] B. Wu, T. Mei, W.-H. Cheng, and Y. Zhang, “Unfolding temporal dynamics: Predicting social media popularity using multi-scale temporal decomposition,” in *Proc. AAAI Conference on Artificial Intelligence*, vol. 30, no. 1, 2016.
- [162] B. L. Aven, D. A. Burgess, J. F. Haynes, J. R. Merino, and P. C. Moore, “Using product and social network data to improve online advertising,” Sep. 23 2014, uS Patent 8,843,406.

- [163] A. Majid, L. Chen, G. Chen, H. T. Mirza, I. Hussain, and J. Woodward, “A context-aware personalized travel recommendation system based on geotagged social media data mining,” *International Journal of Geographical Information Science*, vol. 27, no. 4, pp. 662–684, 2013.
- [164] M.-T. Nguyen, D. H. Le, T. Nakajima, M. Yoshimi, and N. Thoai, “Attention-based neural network: A novel approach for predicting the popularity of on-line content,” in *Proc. IEEE 21st International Conference on High Performance Computing and Communications; IEEE 17th International Conference on Smart City; IEEE 5th International Conference on Data Science and Systems (HPCC/SmartCity/DSS)*. IEEE, 2019, pp. 329–336.
- [165] D. Liao, J. Xu, G. Li, W. Huang, W. Liu, and J. Li, “Popularity prediction on online articles with deep fusion of temporal process and content features,” in *Proc. AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 200–207.
- [166] J. Chen, D. Liang, Z. Zhu, X. Zhou, Z. Ye, and X. Mo, “Social media popularity prediction based on visual-textual features with XGBoost,” in *Proc. 27th ACM International Conference on Multimedia (MM)*, 2019, pp. 2692–2696.
- [167] Z. Zhang, T. Chen, Z. Zhou, J. Li, and J. Luo, “How to become Instagram famous: Post popularity prediction with dual-attention,” in *Proc. IEEE International Conference on Big Data (IEEE BigData)*. IEEE, 2018, pp. 2383–2392.
- [168] W. Zhang, W. Wang, J. Wang, and H. Zha, “User-guided hierarchical attention network for multi-modal social image popularity prediction,” in *Proc. World Wide Web Conference (WWW)*, 2018.
- [169] J. Wang, S. Yang, H. Zhao, and Y. Yang, “Social media popularity prediction with multimodal hierarchical fusion model,” *Computer Speech & Language*, vol. 80, p. 101490, 2023.

- [170] F. S. Abousaleh, W. H. Cheng, N. H. Yu, and Y. Tsao, “Multimodal deep learning framework for image popularity prediction on social media,” *IEEE Transactions on Cognitive and Developmental Systems*, vol. 13, 2021.
- [171] S. Bakhshi, D. A. Shamma, and E. Gilbert, “Faces engage us: Photos with faces attract more likes and comments on Instagram,” in *Proc. SIGCHI Conference on Human Factors in Computing Systems*, 2014.
- [172] F. Gelli, T. Uricchio, M. Bertini, A. D. Bimbo, and S. F. Chang, “Image popularity prediction in social media using sentiment and context features,” in *Proc. 23rd ACM International Conference on Multimedia (MM)*, 2015.
- [173] J. Li, Y. Gao, X. Gao, Y. Shi, and G. Chen, “SENTI2POP: Sentiment-aware topic popularity prediction on social media,” in *Proc. IEEE International Conference on Data Mining (ICDM)*, vol. 2019-November, 2019.
- [174] K. Mannepalli, S. P. Singh, C. S. Kolli, S. Raj, G. R. Bojja, B. Rajakumar, and D. Binu, “Popularity prediction model with context, time and user sentiment information: An optimization assisted deep learning technique,” *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 31, no. 02, pp. 283–302, 2023.
- [175] M. Arazzi, M. Cotogni, A. Nocera, and L. Virgili, “Predicting tweet engagement with graph neural networks,” in *Proc. ACM International Conference on Multimedia Retrieval (ICMR)*, 2023, pp. 172–180.
- [176] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, “Distributed representations of words and phrases and their compositionality,” *Advances in Neural Information Processing Systems*, 2013.
- [177] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “ImageNet: A large-scale hierarchical image database,” in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2009, pp. 248–255.

- [178] S. I. Serengil and A. Ozpinar, “Hyperextended lightface: A facial attribute analysis framework,” in *Proc. International Conference on Engineering and Emerging Technologies (ICEET)*, 2021.
- [179] M. Grootendorst, “BERTopic: Neural topic modeling with a class-based TF-IDF procedure,” <https://arxiv.org/abs/2203.05794>, 2022.
- [180] L. McInnes, J. Healy, and S. Astels, “hdbscan: Hierarchical density based clustering,” *The Journal of Open Source Software*, vol. 2, 2017.
- [181] L. McInnes, J. Healy, N. Saul, and L. Großberger, “UMAP: Uniform manifold approximation and projection,” *Journal of Open Source Software*, vol. 3, 2018.
- [182] C. D. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. J. Bethard, and D. McClosky, “The stanford coreNLP natural language processing toolkit,” in *Proc. 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations (ACL)*, vol. 2014-June, 2014.
- [183] C. J. Hutto and E. Gilbert, “Vader: A parsimonious rule-based model for sentiment analysis of social media text,” in *Proc. International AAAI Conference on Web and Social Media (ICWSM)*, 2014.
- [184] S. Aloufi, S. Zhu, and A. E. Saddik, “On the prediction of Flickr image popularity by analyzing heterogeneous social sensory data,” *Sensors (Switzerland)*, vol. 17, 2017.
- [185] I. T. Jolliffe and J. Cadima, “Principal component analysis: A review and recent developments,” 2016.
- [186] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, “Dropout: a simple way to prevent neural networks from overfitting. journal of machine learning research,” *Journal of Machine Learning Research (JMLR)*, vol. 15, 2014.

- [187] B. Wu, B. Liu, W. H. Cheng, Z. Zeng, P. Liu, and J. Luo, “SMP challenge: An overview of social media prediction challenge 2019,” in *Proc. 27th ACM International Conference on Multimedia (MM)*, 2019.
- [188] B. Wu, W. H. Cheng, Y. Zhang, Q. Huang, J. Li, and T. Mei, “Sequential prediction of social media popularity with deep temporal context networks,” in *Proc. International Joint Conference on Artificial Intelligence (IJCAI)*, vol. 0, 2017.
- [189] K. Ding, R. Wang, and S. Wang, “Social media popularity prediction: A multiple feature fusion approach with deep neural networks,” in *Proc. 27th ACM International Conference on Multimedia (MM)*, 2019.
- [190] C.-C. Hsu, C.-M. Lee, X.-Y. Hou, and C.-H. Tsai, “Gradient boost tree network based on extensive feature analysis for popularity prediction of social posts,” in *Proc. 31st ACM International Conference on Multimedia (MM)*, 2023, pp. 9451–9455.
- [191] S. Mao, W. Xi, L. Yu, G. Lü, X. Xing, X. Zhou, and W. Wan, “Enhanced catboost with stacking features for social media prediction,” in *Proc. 31st ACM International Conference on Multimedia (MM)*, 2023, pp. 9430–9435.