



# Leveraging the Subtle: Hidden Factors in Recommender Systems

DISSERTATION

submitted in partial fulfillment of the requirements for the degree of

**Doktor der Technischen Wissenschaften**

by

**Dipl.-Ing. Mete Sertkan, BSc**

Registration Number 0725297

to the Faculty of Informatics

at the TU Wien

Advisor: Univ.-Prof. i.R. Dipl.-Ing. Dr.techn. Hannes Werthner

The dissertation has been reviewed by:

---

Prof. Maria Soledad Pera, PhD  
TU Delft  
The Netherlands

---

Prof. Marko Tkalčič, PhD  
University of Primorska  
Slovenia

Vienna, 12<sup>th</sup> March, 2025

---

Mete Sertkan



# Declaration of Authorship

Dipl.-Ing. Mete Sertkan, BSc

I hereby declare that I have written this Doctoral Thesis independently, that I have completely specified the utilized sources and resources and that I have definitely marked all parts of the work - including tables, maps and figures - which belong to other works or to the internet, literally or extracted, by referencing the source as borrowed.

Vienna, 12<sup>th</sup> March, 2025

---

Mete Sertkan



# Abstract

Recommender systems are pivotal in various domains, aiding users in their decision-making. However, current systems often overlook subtle factors that significantly impact user preferences and choices. This work aims to bridge this gap by exploring the concept of implicit item characteristics – latent features that influence user decision-making in addition to explicit content. The investigation is divided into three key research areas.

Firstly, we explore how to systematically identify and expose implicit item characteristics to enhance recommender systems in two key domains: tourism and news. Using advanced analytics such as cluster analysis and multiple linear regression, we map tourist destinations to the established Seven-Factor Model in tourism. In the news, we employ natural language processing techniques to reveal hidden features essential for tailoring recommendations.

Secondly, we introduce a novel system called PicTouRe to elicit tourists' implicit preferences through pictures. Leveraging convolutional neural networks, we translate visual preferences into a Seven-Factor profile for each user, simplifying decision-making and capturing both immediate touristic desires and enduring personality traits.

Lastly, we enhance news recommender systems by leveraging sentiment and emotions of news articles. Two models, RobustSentiRec and EmoRec, were developed to capture these implicit characteristics, aligning recommendations more closely with user preferences but also raising ethical concerns around potential sentiment and emotional echo chambers.

Our findings offer a robust framework for more nuanced, user-sensitive recommendations, opening new avenues for future research and applications in recommender systems



# Contents

<b>Abstract</b>	<b>v</b>
<b>Contents</b>	<b>vii</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Research Questions . . . . .	4
1.2 Structure of This Work . . . . .	6
1.3 Significance of the Research . . . . .	13
<b>2 Background</b>	<b>15</b>
2.1 Recommender Systems . . . . .	15
2.2 Tourism & Recommenders . . . . .	18
2.3 News & Recommenders . . . . .	24
2.4 Recommender Systems & Users' Decision Making . . . . .	28
<b>3 Exploring &amp; Exposing Implicit Item Characteristics</b>	<b>33</b>
3.1 The "Personality" of Tourism Destinations . . . . .	34
3.2 A Multi-Level View on News Articles . . . . .	58
3.3 Summary . . . . .	72
<b>4 Unveiling Tourists' Implicit Preferences Through Pictures</b>	<b>75</b>
4.1 A Generic Profiler . . . . .	76
4.2 A User Study on Picture Collections . . . . .	85
4.3 Summary . . . . .	102
<b>5 Leveraging Sentiment &amp; Emotions for News Recommendation</b>	<b>105</b>
5.1 Exploiting & Diversifying Sentiments . . . . .	106
5.2 Exploring Expressed Emotions . . . . .	124
5.3 Summary . . . . .	142
<b>6 Conclusions</b>	<b>145</b>
6.1 Summary . . . . .	146
6.2 Practical Implications . . . . .	149
6.3 Future Directions . . . . .	150
	vii



# Introduction

In an age characterized by an insatiable human need for information, we are inundated by a deluge of data, articles, products, and services. This abundance presents a paradox: while we are surrounded by a wealth of potentially useful resources, discerning what is truly helpful becomes an increasingly complex task [89]. Traditional cognitive capacities and decision-making frameworks, grounded in theories of rationality [58, 22], are often stretched to their limits in this environment. Herbert Simon’s concept of “Bounded Rationality” [104, 22] aptly captures this dilemma, highlighting the constraints on human cognition and the subsequent need for efficient decision-making tools.

Emerging as a solution to this dilemma, recommender systems (RSs) have gained significant traction in recent years. Originating in the 1990s as an independent research area, these systems have evolved into some of the most pervasive and valuable applications of machine learning in industry today [89]. Operating on a wide array of platforms, they offer personalized suggestions tailored to individual preferences and needs [73]. In essence, RSs serve as digital counselors, mimicking the role traditionally played by human advisors [89]. They achieve this by leveraging algorithms capable of sifting through the vast expanse of the World Wide Web (WWW).

Ricci et al. [89] describe RSs as “*software tools and techniques that provide suggestions for items that are most likely of interest to a particular user*”. They can offer suggestions in either a personalized or a non-personalized manner. Non-personalized recommendations often leverage edited or curated lists and might also resort to presenting the top- $K$  popular items in a particular category. Although simple and straightforward, these approaches may not cater to individual tastes and can result in generic suggestions. In contrast, personalized recommendations consider the specific needs and preferences of the user, offering a tailored experience that generally results in higher user satisfaction [89].

The effectiveness of RSs hinges crucially on the quality of its underlying user model, which aims to capture the nuanced preferences and needs of individual users. A crucial,

and often overlooked, aspect of personalized RSs is the level of user interaction required to build this model. While some systems can operate in *cold-start* situations, relying primarily on content analysis or demographic data without any user interaction, most benefit from or even depend on some form of user engagement data. This interaction data can take various forms, broadly categorized as explicit and implicit feedback, impacting both the system's design and the user experience. Explicit feedback involves active user participation, where users deliberately provide ratings, reviews, or preferences, making their needs directly observable. Implicit feedback, on the other hand, relies on passive observation of user actions, such as browsing history, purchase patterns, or click-through rates. Both approaches have their merits and drawbacks. Explicit feedback, through active elicitation, generally provides more accurate information, but at the cost of user effort and potential biases. This approach can be justified in domains where the cost of consuming the recommended item is high, such as in tourism. Recommending a vacation package involves significant financial and time commitments; thus, users might be more willing to invest the effort. The relatively low frequency of such high-stakes decisions further justifies it. Implicit feedback, gathered through passive observation, is unobtrusive but might be less precise [127, 3]. However, passive observation is valuable when explicit feedback is scarce, for example, in dynamic domains like news recommendations, where new items are constantly introduced. Actively asking for feedback might be too burdensome; thus, implicitly modeling the user might lead to a better user experience. This thesis explores the spectrum of user interaction – from no interaction to passive observation (implicit feedback) and active preference elicitation (explicit feedback) – highlighting their strengths and weaknesses in different contexts, specifically within the tourism and news recommendation domains.

Among the array of recommendation techniques, collaborative filtering stands out for its ubiquity and effectiveness [93]. The foundation of collaborative filtering lies in the principle that users who have displayed similar preferences in the past are likely to do so in the future. This technique relies exclusively on a user-item interaction matrix, thereby eliminating the need for additional, domain-specific information such as demographics or content attributes [89]. This domain-independence is a key strength; it makes collaborative filtering versatile, easy to implement, and highly scalable. Additionally, by harnessing the “wisdom of the crowd,” collaborative filtering has demonstrated promising results across a multitude of domains, underlining its general applicability [61].

Despite the versatile nature of collaborative filtering, there are domains where it encounters limitations, particularly when faced with a sparse user-item interaction matrix or rapidly evolving content. News recommendation systems (NRS) are a prime example of such a domain. News articles often have a short life cycle, making it challenging for collaborative filtering techniques to accumulate sufficient interaction data for accurate recommendations. Additionally, new articles emerge constantly, exacerbating the *cold-start* problem where items lack sufficient interaction history. Content-based methods are better suited for these challenges, offering recommendations based on article content features and eliminating the need for a rich interaction history. These methods are more efficient in personalizing

---

recommendations based on a user’s specific reading history and can adapt quickly to the time-sensitive nature of news [61]. Similarly, in the tourism domain, content-based and/or knowledge-based approaches are more promising due to unique sector characteristics such as high risk, low churn, low heterogeneity, unstable preferences, and an explicit interaction style. Consumption costs in terms of time and money are higher, the value of items does not fluctuate rapidly, the range of needs that items can satisfy is narrow, past preferences may not be indicative of current choices, and interactions often require explicit user input [19].

Content-based methods learn to recommend items similar to those the user has liked, based on features or attributes such as genre or keywords. Semantic indexing techniques further enrich these features using external knowledge or large textual corpora [89]. However, traditional models assume that human decision-making is rational and preferences are static, which does not hold in practice. People operate under bounded rationality, where they trade off cognitive effort against decision quality [104, 22]. Preferences are dynamic and constructed during the decision-making process, often influenced by the way options are presented [58]. This extends to content-based approaches and RSs in general, where not only “what” is shown but also “how” it is presented can significantly affect consumer choices. Emotional and personality factors also play a crucial role, and ignoring these complexities can lead to suboptimal outcomes [58, 22, 109].

Uncontrolled decision biases, such as decoy and position effects, can lead to suboptimal decisions [58, 109]. To make effective recommendations, systems need to consider these biases and the dynamic, context-dependent nature of preferences. They must also account for the role of presentation and the emotional state of the user. This more nuanced view of human decision-making suggests that we should integrate both content-based methods and an understanding of bounded rationality and decision biases to better match human behavior and needs [58, 22, 109].

To address the aforementioned complexities and challenges, this thesis aims to elevate the quality of personalized recommendations in both the tourism and the news domains. Despite the distinct dynamics of these sectors, they share common issues such as sparse interaction matrices and a reliance on content-based methods. Existing research has predominantly focused on the overt attributes of recommended items. This work, however, delves into the value of implicit characteristics not immediately recognizable—such as the emotions expressed within news articles, or the connotation of tourism destinations with tourist roles and personalities. These subtle but crucial factors are essential for user decision-making and thus require detailed modeling.

We follow two strategies to explicitly incorporate implicit item characteristics into the user-item modeling and recommendation process. The first strategy uses a renowned domain model, such as the Seven-Factor Model [76, 77] in the tourism domain. This approach allows us to consider both overt content and subtle, implicit factors that influence decision-making. The second strategy involves deriving a domain model in a data-driven manner using deep learning, capturing both overt and covert factors. By accounting for both implicit and explicit characteristics of items and users, our overarching

goal is to create more robust user models that consider bounded rationality, decision biases, and human needs.

We will pursue this objective by exploring different levels of user interaction: i) No interaction, focusing on identifying implicit item characteristics through explorative analysis; ii) Active user involvement, eliciting user-specific preferences using implicit item features, following a picture-based approach grounded in the Seven-Factor Model in the tourism domain; and iii) Passive observation of users to build models based on their interaction history, integrating aspects like content preferences and emotional responses in the news domain. By addressing these aspects, this thesis holds the potential to significantly advance the state-of-the-art in user modeling and preference elicitation, thereby providing more personalized and satisfactory recommendations.

### 1.1 Research Questions

In bridging the gap between theoretical underpinnings and empirical investigation, this section translates the aforementioned aims, strategies, and approaches into distinct research questions. Central to our inquiry is the concept of implicit item characteristics, subtle yet impactful features that influence user decision-making in tandem with explicit content. To underscore the importance of these implicit factors, consider the emotions expressed within a news article, or how a tourist destination resonates with specific tourist roles—be it an adventurer, a culture seeker, or a relaxation-seeker. These latent variables often escape straightforward computational representation but play a decisive role in preference formation and decision biases.

The objective of unearthing these nuanced item-user relationships is dissected into three key research questions, each corresponding to the three-pronged approach laid out earlier. These questions serve as the backbone for dedicated chapters that delve into the intricacies of each research question, providing both a detailed conceptual framework and empirical answers.

**Exploring & Exposing Implicit Item Characteristics.** RSs in both tourism and news are ripe for improvement through data analytics. Our primary focus is to dig deeper into these areas, specifically targeting the less obvious, implicit characteristics that influence user preferences. In tourism, we automate the process of mapping destinations to the established Seven-Factor Model, a task previously performed manually. This not only simplifies decision-making for travelers but also grounds it in robust data. Additionally, the Seven-Factor Model captures both short-term touristic preferences and long-term personality traits, the latter being another subtle but influential factor in decision-making and bounded rationality choices. In the news realm, we use natural language processing to analyze articles from multiple angles—document, topic, and author levels—revealing hidden features that are essential for tailoring recommendations.

**RQ1** *How can we systematically identify and expose the implicit item characteristics embedded in structured and unstructured data to enhance recommender systems?*

In the tourism domain, we utilized the Seven-Factor Model as our guiding framework. Through advanced analytics like cluster analysis and multiple linear regression, we linked various destinations to this model effectively. This helps to make the recommendation process more data-driven and easier for users to navigate. Turning our attention to news, we tackled the ever-changing and short-lived nature of articles. We employed a variety of natural language processing techniques for a multi-level analysis, uncovering key hidden features. These insights allow us to deliver more personalized and varied news recommendations. Both these efforts lay the foundation for answering subsequent research questions in the following chapters.

**Unveiling Tourists’ Implicit Preferences Through Pictures.** Navigating the *cold-start* problem in the tourism industry, our second line of research actively involves users in the recommendation process through a novel system called PicTouRe. Here, we extend the utility of the Seven-Factor Model by mapping user-provided and ranked images onto it, effectively translating visual preferences into an aggregated Seven-Factor profile for each user. PicTouRe builds upon and extends previous approaches [77, 76] that were limited to a fixed set of pictures, by allowing the use of any picture set. This approach goes beyond simple personalization, encapsulating both immediate touristic desires and enduring personality traits that are crucial for decision-making.

**RQ2** *How can a generic picture-based user and item modeling – not limited to a fixed picture set – improve the efficacy and user satisfaction of tourism recommendation systems?*

We leverage convolutional neural networks (CNNs) to map pictures onto the Seven-Factor Model, advancing beyond traditional methods that rely solely on explicit preference statements or static picture collections [76, 77]. We engage users in a gamified manner by allowing them to provide and rank images, thereby reducing cognitive effort. By mapping these visual preferences onto the Seven-Factor Model, we simplify the decision-making process, making it more cognitively efficient. Moreover, the Seven-Factor Model captures psychological traits that are subtle but significant in shaping choices. An extensive user study affirmed the effectiveness of our approach, indicating a strong alignment between user self-perceptions and system predictions. This second line of research not only offers a unique, user-centric means of preference elicitation but also lays the groundwork for future advancements, complementing our initial focus on implicit item characteristics.

**Leveraging Sentiment & Emotions for News Recommendation.** Pivoting from active user involvement to passive observation, our third line of research focuses on enhancing the quality of news recommendations by tapping into underlying emotional and sentiment dimensions. We aim to capture not just the overt textual content of news articles but also the emotions and sentiments they express, categorizing these as implicit item characteristics. This enriches the user model, offering a more comprehensive understanding of user preferences and needs.

**RQ3** *How can sentiment and expressed emotions in news articles be effectively utilized in recommendation systems to improve recommendation performance, and what impact*

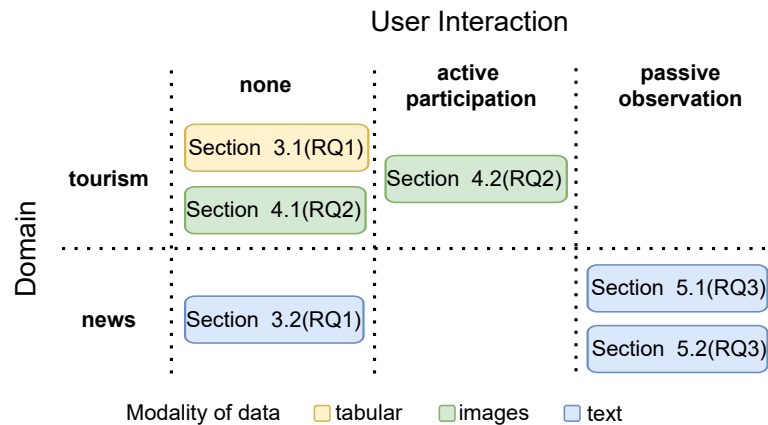


Figure 1.1: Overview of main sections of the thesis, categorized by domain, user interaction, and underlying modality of data.

*do these factors have on diversity?*

We embark on dual investigations: first, by reproducing and refining SentiRec [119], a neural news recommendation model, we highlight challenges in its generalizability and introduce an improved version called RobustSentiRec. Second, we develop EmoRec, which leverages emotions within news articles to enhance recommendations. Both models advance the state-of-the-art by aligning recommendations more closely with user preferences, yet they raise ethical concerns around potential sentiment and emotional filter bubbles. These efforts contribute to our ultimate goal of delivering more nuanced, user-sensitive recommendations.

In the subsequent chapters, we will unpack each of these questions, elaborating on the methods, datasets, and evaluation metrics employed to answer them. Through this rigorous inquiry, we aim to forge a deeper understanding of the subtle factors that contribute to personalized recommendations, offering a robust framework for future research and applications.

## 1.2 Structure of This Work

In Figure 1.1, we present an overview of the core sections of this thesis, categorizing them along several dimensions. The first categorization is based on the domain – either tourism or news – that each section focuses on. This domain dictates the strategy we use for feature representation. In the tourism domain, we align users and items with the widely-recognized Seven-Factor Model, capturing both overt and implicit characteristics. For the news domain, our approach models both emotion and sentiment – implicit item characteristics – as well as semantic attributes, which are more overt.

The second categorization is by the level of user interaction: no interaction, active involvement, or passive observation. The level of user involvement largely dictates our

strategic approach. When there is no user interaction, our focus is on exploring the domain to set the foundation for the more interactive strategies. With active user involvement, we opt for a less complex model, specifically employing the Seven-Factor Model, to minimize cognitive effort and facilitate better decision-making. For passive user observations, we model user content preference using latent semantic representations, while also explicitly accounting for emotion and sentiment dimensions while maintaining their interpretability.

The final classification is by the type of data considered in each section – tabular, images, or text – which largely influences the complexity of methods we employ, from simple multiple linear regression models to neural approaches.

Each of these main sections is grounded in peer-reviewed research, where Mete Sertkan has served as the lead investigator, responsible for planning, implementation, and writing. Following this, we offer a concise summary of each chapter and main section. Under each summary, we list the associated research paper(s), including publication year, venue or journal, title, authors, and a citation for easy reference.

**In Chapter 2**, we outline the background of the thesis and relate it to our work, and delve into four main areas:

**General Overview of Recommender Systems.** We explore the value and various techniques used in building RSs, setting the stage for our contributions in enhancing their efficiency and efficacy.

**Tourism & Recommenders.** Focusing on the travel and tourism domain, we discuss the existing RSs and introduce the Seven-Factor Model. This model allows for nuanced recommendations, blending both strong and weak signals to understand user preferences deeply.

**News & Recommenders.** We then move into the news sector, discussing how RSs function in this domain. Deep learning techniques are particularly highlighted for their capability to model user news consumption patterns. The role of sentiment in diversifying news recommendations is emphasized.

**Recommender Systems & Human Decision Making.** Finally, we examine how RSs interact with human decision-making processes. Cognitive theories such as the Effort-Accuracy framework and constructs like bounded rationality help us understand the complexities involved in decision-making, offering valuable insights for building more effective and empathetic RSs.

**Chapter 2** sets the foundation for our multi-disciplinary approach to improving RSs in both the tourism and news domains. We leverage advanced techniques like deep learning and integrate human cognitive theories to not just improve system performance, but also to enhance user satisfaction and emotional engagement. This is aligned with our ultimate aim to create a harmonious ecosystem where all stakeholders—users, providers, and system owners—derive value while maintaining transparency and integrity.

**Chapter 3** explores implicit item characteristics across two distinct domains: tourism and news. The chapter is partitioned into two key sections. The first section pioneers an innovative approach to mapping tourism destinations to the Seven-Factor Model, thereby offering a nuanced understanding of how touristic roles and personality can align with tourism destinations. The second section addresses the dynamic field of online news, introducing a multi-level analysis that tackles the challenges of content transience and *cold-start* problems. Across both sections, the unifying theme is uncovering less apparent but significant attributes—implicit item characteristics—that can be leveraged to enhance RSs, offering more personalized and effective recommendations to users.

**In Section 3.1**, we address the complexity and emotional resonance of tourism products by automating the mapping of tourism destinations onto the Seven-Factor Model—a comprehensive user model that captures the multifaceted preferences and personalities of travelers. We examine whether an underlying structure compatible with the Seven-Factor Model exists among tourism destinations and whether an automated mapping mechanism can be developed based on their attributes. Using a robust methodology involving exploratory data analysis, cluster analysis, and multiple linear regression models, we identify six conceptually meaningful clusters of tourism destinations and establish a statistically significant association between selected attributes of these destinations and the Seven-Factor Model. Our results demonstrate not only the feasibility of automated mapping but also its potential in improving the effectiveness of RSs within the tourism domain. Our contributions extend prior work by offering a nuanced understanding of the relationship between tourism destinations and user personalities, paving the way for more personalized and effective recommendations.

---

<b>2019</b>	<b>JITT</b>	What is the “Personality” of a Tourism Destination? <i>M. Sertkan, J. Neidhardt, H. Werthner</i>	[99]
-------------	-------------	---	------

---

**In Section 3.2**, we tackle the unique challenges posed by news content transience in the landscape of NRS. Recognizing that the transience of news items exacerbates the *cold-start* problem in NRS, we venture beyond traditional content-based approaches to explore latent characteristics within news articles. Our investigation is structured across three distinct yet interconnected dimensions: document-level, topic-level, and author-level. We illuminate how document-level analysis excels in extracting discriminative features and enables focused recommendations, while topic-level analysis reveals underlying themes and augments recommendation diversity. Author-level analysis provides a novel avenue for serendipitous recommendations based on stylistic congruence, achieving an accuracy rate of approximately 97%.

This multi-layered analysis lays the foundation for a future NRS that not only maximizes accuracy but also enriches user experience by promoting recommendation diversity and serendipity.

---

**2019 CBI** Documents, Topics, and Authors: Text-Mining of Online News [98]  
*M. Sertkan, J. Neidhardt, H. Werthner*

---

**Chapter 4** explores the interactive dimension of user preferences in the tourism industry, using image analysis as a medium for personalization. This chapter is divided into two focal sections. The first introduces a novel methodology—the Generic Profiler—that employs convolutional neural networks to analyze images within the framework of the Seven-Factor Model. This provides an automated method for profiling both users and tourism destinations. The second section shifts the focus to users, presenting a comprehensive study to evaluate the practical application of our *Picture-Based Tourism Recommender System*, or *PicTouRe*. Through this dual-pronged approach, the chapter aims to enhance tourism recommender systems (TRS) by incorporating a more user-centric model based on visual inputs. It scrutinizes the efficacy of this innovative methodology and offers actionable insights into its real-world applicability.

**In Section 4.1**, we introduce a novel methodology designed to characterize tourism-related entities, whether users or destinations, through an approach grounded in image analysis. The section revolves around the concept of a *Generic Profiler*, a composite of two core components: an *Image Classifier* and an *Aggregator*. The former employs convolutional neural networks to classify images based on the Seven-Factor Model, a representation we found effective in capturing various facets of tourism. Our model, initially trained on the ImageNet database, was further refined using a dataset of 300 expert-labeled images per factor and exhibited high validation accuracies ranging from 88% to 99%. The *Aggregator* compiles these individual classifications to create a cohesive profile. Our methodology was rigorously evaluated against expert-labeled datasets and showed promising results, although a few areas for improvement were identified. Contributions of this section include the development of a generic profiling methodology, the creation of an expert-labeled dataset, and the provision of our code and model weights for public access. This work represents a foundational step in utilizing image analysis for automating the profiling of tourism entities.

---

**2020 ENTER** From Pictures to Travel Characteristics: A Deep-Learning-Based Profiling of Tourists and Tourism Destinations [102]  
*M. Sertkan, J. Neidhardt, H. Werthner*

---

**In Section 4.2**, we delve into the user-centric evaluation of the *Generic Profiler* via a comprehensive user study integrated into *PicTouRe*, our picture-based TRS. Guided by four research questions, the study examines the efficacy of using user-provided pictures to generate touristic profiles. Specifically, we investigate the alignment between our predicted profiles and the users’ self-perceived profiles, and also explore whether the

sequence of pictures impacts these profiles. Among 81 participants, 65% agreed with their generated touristic profile, and 48% preferred our system’s recommendations over those based on their self-perceived profiles. Our results indicate that pictures can indeed serve as a powerful tool for extracting implicit preferences, thus substantiating the utility of *PicTouRe* as an early-stage travel planning aid. This section contributes by validating the system’s practical application, assessing the discrepancies between perceived and predicted touristic profiles, and scrutinizing the role of picture order in profile generation.

2020	RECSYS	PicTouRe - A Picture-Based Tourism Recommender <i>M. Sertkan, J. Neidhardt, H. Werthner</i>	[101]
2020	UMAP	Eliciting Touristic Profiles: A User Study on Picture Col- lections <i>M. Sertkan, J. Neidhardt, H. Werthner</i>	[100]

**Chapter 5** navigates the intricate landscape of sentiment and emotion in the realm of neural NRS. Comprising two essential sections, the chapter first scrutinizes the existing sentiment-aware recommendation model, SentiRec, and introduces an optimized variant—RobustSentiRec. The second section ventures into uncharted territory by incorporating emotional dimensions into news recommendations through a multi-level neural model called EmoRec. This chapter employs passive user data to construct comprehensive user models that incorporate the expressed emotions of news content users consume. It aims to enrich the relevance of recommendations by leveraging both textual and emotional elements, thereby offering a nuanced user experience that is both diverse and satisfying. Overall, this chapter contributes to the growing discourse on ethical and responsible machine learning by highlighting the implications of incorporating sentiment and emotion in NRS.

**In Section 5.1**, we explore the critical issue of sentiment diversity in neural NRS. Recognizing that contemporary methods often trap users in self-reinforcing cycles of sentiment, we undertake a rigorous reproduction and extension of SentiRec, a sentiment-aware recommendation model proposed by Wu et al. [119]. We employ the Microsoft MIND dataset for our experiments. Our analysis is multi-faceted, examining the performance of SentiRec against various baselines, its generalizability, and its effectiveness when substituting its rule-based sentiment analyzer with a neural one. Surprisingly, our results question the original findings by demonstrating that the baselines already perform competitively in both sentiment and topical diversity. Furthermore, we extend the scope of the original paper by examining topical and intra-list diversity. Notably, our version of SentiRec excels in intra-list sentiment diversity. To simplify the complexity and improve robustness, we introduce RobustSentiRec, an optimized variant of SentiRec, which offers comparable performance while reducing model intricacy.

These contributions not only deepen our understanding of sentiment diversity in NRS but also prompt critical questions about the replicability and generalizability of existing models.

---

<b>2022</b>	<b>PERSPECTIVES</b>  (RECSYS)	Diversifying Sentiments in News Recommendation  <i>M. Sertkan, S. Althammer, S. Hoftstätter, J. Neidhardt</i>	[103]
-------------	-------------------------------------	---	-------

---

**In Section 5.2**, we grapple with the under-explored dimension of incorporating expressed emotions in personalized NRS. Acknowledging the inherent emotional characteristics alongside semantic elements, we introduce a multi-level neural news recommendation model, *EmoRec*, that enriches user and item models with emotions extracted at various levels of an article, such as the title, abstract, category, and subcategory. We investigate the impact of emotion incorporation on recommendation performance, emotional diversity, and topical diversity. Extensive experiments reveal a nuanced picture: while *EmoRec* improves recommendation performance, it reduces both emotional and topical diversity, leading to the risk of a self-reinforcing “emotion chamber.” Our findings also indicate that coarse emotion taxonomies are more effective than fine-grained ones for this task. The section concludes by addressing ethical considerations and limitations, underscoring the need for a balanced, ethically grounded approach. This section serves as a critical addition to the existing literature, enriching our understanding of how emotions could be responsibly integrated into NRS.

---

<b>2022</b>	<b>UMAP</b>	Exploring Expressed Emotions for Neural News Recommendation  <i>M. Sertkan, J. Neidhardt</i>	[94]
<b>2023</b>	<b>INRA</b>  (RECSYS)	On the Effect of Incorporating Expressed Emotions in News Articles on Diversity within Recommendation Models  <i>M. Sertkan, J. Neidhardt</i>	[95]

---

**Chapter 6** concludes by summarizing the key contributions of our research and discussing their broader implications for the field of personalized recommendation systems. We move beyond a simple recapitulation of our methods and results, focusing instead on how our findings can inform the design of future systems in both the tourism and news domains, as well as other areas where nuanced user understanding is crucial.

Our research demonstrates that incorporating implicit item characteristics, such as emotional cues in news and personality-aligned attributes in tourism, can lead to more user-centric recommendations. This has significant implications for the design of systems that aim to go beyond simple content matching and truly understand user needs and preferences. For example, travel agencies and platforms can use our findings on destination clustering and the Seven-Factor Model to tailor their marketing and offerings, moving

beyond generic recommendations to provide experiences that resonate with individual traveler profiles. Similarly, news providers can leverage our multi-layered analysis of articles (document, topic, author) to create more diverse and engaging recommendation experiences, potentially mitigating the effects of filter bubbles.

The development and validation of *PicTouRe*, our picture-based preference elicitation technique, highlights the potential of innovative interaction methods in RSs. This approach offers a more engaging and less burdensome way to capture user preferences, particularly in domains like tourism where visual cues are highly relevant. This has implications beyond tourism, suggesting that other visually rich domains (e.g., fashion, real estate) could benefit from similar approaches.

Our work on sentiment- and emotion-aware news recommendation (RobustSentiRec and EmoRec) underscores the complex interplay between personalization and ethical considerations. While incorporating emotional cues can improve recommendation accuracy and user satisfaction, it also raises the risk of creating “echo chambers” that limit exposure to diverse viewpoints. This highlights a crucial area for future research: developing methods to balance personalization with diversity and social responsibility. This is not just a technical challenge but also a societal one, requiring interdisciplinary collaboration to ensure that RSs serve the broader public good.

We addressed the following leading research questions:

**RQ1** *How can we systematically identify and expose the implicit item characteristics embedded in structured and unstructured data to enhance recommender systems?*

We advanced the state-of-the-art in identifying and extracting implicit item characteristics, demonstrating practical methods for both structured (tourism) and unstructured (news) data. Our findings pave the way for richer, more nuanced representations of items in RSs, enabling a deeper understanding of user-item interactions.

**RQ2** *How can a generic picture-based user and item modeling – not limited to a fixed picture set – improve the efficacy and user satisfaction of tourism recommendation systems?*

We validated the effectiveness of picture-based preference elicitation as a user-friendly and accurate method for capturing implicit preferences, particularly within the framework of established domain models like the Seven-Factor Model. This approach offers a practical solution for reducing user effort while improving the quality of user profiles.

**RQ3** *How can sentiment and expressed emotions in news articles be effectively utilized in recommendation systems to improve recommendation performance, and what impact do these factors have on diversity?*

We explored the potential and pitfalls of incorporating sentiment and emotion into NRS. Our findings highlight the need for a careful balance between personalization and ethical considerations, paving the way for future research on responsible and diversity-aware recommendation algorithms.

Reflecting on this personal journey, what began as a rudimentary exploration into recommendation systems has now crystallized into a clear understanding of the paramount importance of centering the user in the recommendation process. To make *good* recommendations, systems must go beyond sophisticated algorithms and truly understand users, incorporating both their explicit preferences and the more subtle, implicit factors that drive their decisions. This includes understanding not only *what* users want but also *why* they want it and *how* their preferences are formed. This user-centric approach, encompassing both effectiveness and ethical considerations, will be essential in developing the next generation of recommendation systems.

### 1.3 Significance of the Research

In this thesis, we address challenges in personalized recommendation systems within the tourism and news sectors by focusing on implicit item characteristics such as emotions and personality factors, rather than just overt attributes. By exploring three research questions, we employ advanced analytics, image-based preference elicitation, and ethical considerations to create more robust and nuanced user models. Our approaches are empirically validated and demonstrate real-world scalability, contributing to the fields of personalized recommendations and user modeling.

The research has practical implications beyond its theoretical contributions. It addresses core challenges in the tourism and news sectors and has led to collaborations that validate its applicability. One collaboration with TU Munich resulted in a journal paper titled “A Comparative Study of Data-Driven Models for Travel Destination Characterization” [31], evaluating our approach of mapping destinations onto the Seven-Factor Model. Another joint venture with the Information Retrieval group at TU Wien explored the relevance of implicit characteristics like subjectivity and entertainment for podcast retrieval [50] at the TREC Deep-Learning Track 2021. These collaborations are part of a broader academic effort, resulting in over 30 scientific papers and more than 500 citations. Additionally, we have open-sourced our research materials on GitHub (<https://github.com/MeteSertkan>), allowing the wider community to analyze, reuse, and build upon our work. This research contributes to the domain of personalized recommendations and user modeling through its academic rigor and practical applications.



# Background

In this chapter, we provide an overview of RSs, focusing on their application in the tourism and news industries. We begin with a general exploration of recommendation techniques, before diving into their specific utility in the tourism sector, leveraging the Seven-Factor Model for nuanced user profiling. The chapter then transitions to an examination of NRS, highlighting the role of deep learning in capturing user preferences. We also incorporate human cognitive theories, such as the Effort-Accuracy model, to explore the intersection of RSs with human decision-making. The chapter sets the foundation for the thesis and upcoming chapters.

## 2.1 Recommender Systems

RSs serve as sophisticated software tools designed to guide users in making choices by offering personalized suggestions for items that might interest them (Figure 2.1). Originating in the 1990s, these systems have evolved to manage the modern problem of information overload on various platforms, including e-commerce websites like Amazon. While RSs are versatile and can be tailored to various types of items – from movies to news articles – their underlying goal remains constant: to assist users in navigating an overwhelming array of choices. By employing techniques like collaborative filtering, RSs generate ranked lists of items based on collected user data, such as explicit ratings or inferred preferences from user behavior. These systems not only alleviate the decision-making burden but also enhance the user experience by delivering more relevant and customized suggestions [89, 85].

In exploring the multifaceted nature of RSs, it is crucial to understand their objectives from the perspective of different stakeholders involved: Consumers, Providers, and System Owners. Essentially, RSs serve as a multi-role platform balancing various interests. For System Owners, the main goal is often to boost conversion rates—the percentage of users who actually go ahead and consume a recommended item. They aim to achieve

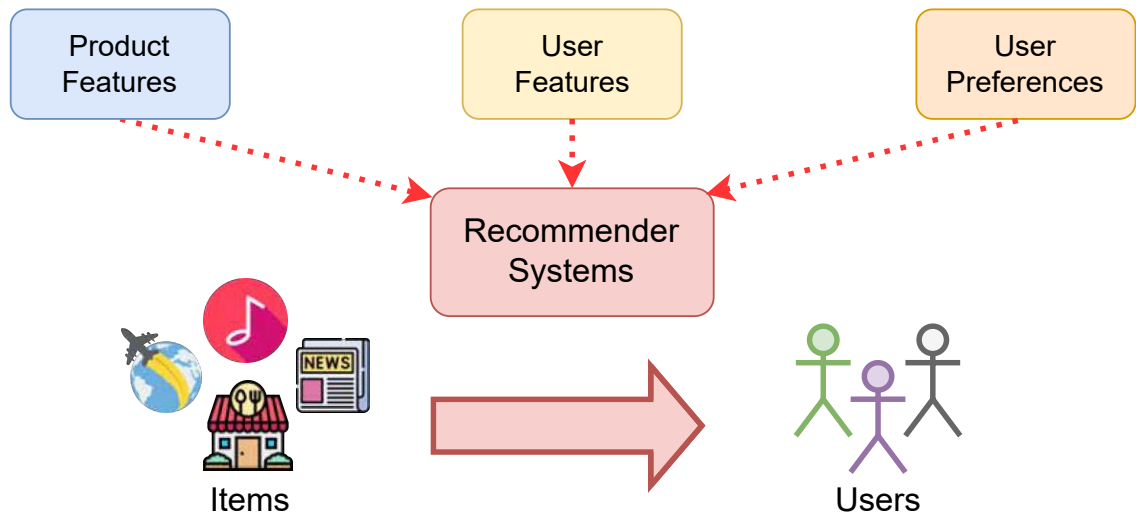


Figure 2.1: Recommender Systems overview.

this by offering personalized, accurate recommendations that meet users' needs, thus encouraging repeat visits and increased interaction with the system. Item Providers, such as manufacturers or suppliers, use RSs as a channel to increase sales and possibly diversify the range of items consumed. Consumers, or end-users, look for efficient, personalized suggestions that can help them make better choices, whether it is finding a new book to read, a tourist destination to explore, or a product to purchase. Therefore, a successful RSs is one that harmonizes these varied stakeholder objectives, offering value to each while maintaining user trust and system integrity. Advanced evaluation techniques are increasingly employed to measure this multifaceted impact and value generation across all stakeholder groups [87, 15, 89].

Interactions between users and items play a pivotal role in shaping the recommendations generated by RSs. These interactions are usually stored in logs, which can capture a myriad of details—ranging from what item a user purchased to the sequence of events leading to that purchase. Explicit user feedback, such as numerical ratings, is highly valuable but often hard to come by as users might not take the time to provide it. On the other hand, implicit feedback—like clicking or adding an item to the basket—is more abundant but provides a weaker signal of user preferences. The type of feedback, explicit or implicit, can greatly influence the recommendation algorithm employed by the system [93, 86, 89]. Furthermore, time also impacts the utility of interaction data; long-term data helps in building a robust user model, while short-term, session-based data is critical for catering to immediate user needs. Therefore, a well-designed RSs must skillfully navigate these factors to make effective recommendations [11, 12, 89].

Our thesis expands on this understanding of user-item interactions. In Chapter 3, we sidestep interactions entirely to first establish foundational knowledge on recommendation

items. However, we engage users more dynamically in Chapter 4, utilizing a gamified approach to elicit both explicit (pictures) and implicit (preferences mapped onto the Seven-Factor Model) feedback about tourism destinations. This method offers a balance between strong and weak signals, making it both engaging and insightful. Chapter 5 shifts gears to employ a more passive strategy, leveraging historical news-reading behaviors to model user preferences. Though the semantic representation is latent, we maintain the tonality dimension, capturing both short-term and long-term behavioral aspects. Overall, our work spans the spectrum from strong signals, as in Chapter 4’s gamified preference elicitation, to weak signals evidenced in Chapter 5’s retrospective analysis.

RSs serve to suggest items or choices tailored to individual users’ preferences or needs. They achieve this by employing various algorithms and techniques to predict the utility of items for a user. Broadly speaking, there are six major classes of recommendation techniques – Content-Based, Collaborative Filtering, Community-Based, Demographic, Knowledge-Based, and Hybrid – which are elaborated on next.

**Content-Based:** In content-based systems, items are matched with the user profile, where items liked by the user are positively correlated with the profile. The algorithm considers item features such as keywords, genre, or other meta-information. For example, if a user frequently watches comedy movies, the system will recommend other titles in the comedy genre. Advanced forms of content-based recommendations also employ semantic techniques, using ontologies or other forms of external knowledge to better understand the content [69, 89].

**Collaborative Filtering:** The core principle here is that users who agreed in the past tend to agree again in the future. Two types dominate: user-based and item-based. User-based methods find peers or users with similar past behavior to generate recommendations, while item-based methods recommend items similar to what the user has liked. Collaborative filtering is particularly powerful but can suffer from *cold-start* issues for new items or users and tends to focus on popular items, ignoring the long tail [44, 92, 89].

**Community-Based:** This approach expands upon collaborative filtering by utilizing social network metadata. The underlying premise is that friends or communities of users can provide more reliable recommendations. Recommendations are made by considering what a user’s social circle is interacting with. This approach is gaining importance with the prevalence of social networking sites [5, 9, 89].

**Demographic:** These systems categorize users based on demographic information like age, nationality, or gender and make recommendations based on these categories. For instance, a teenager might receive recommendations for high-school romance novels, while an adult might get recommendations for finance journals. While not always deeply personalized, these are effective for broad segmentation [14, 89].

**Knowledge-Based:** In scenarios where user-item interactions are rare or where specific expertise is required, knowledge-based systems come into play. They use explicit knowledge about users and items to make recommendations. For example, a knowledge-based

system might recommend a specialized medical journal to a healthcare professional but not to a general reader. They are excellent for niche markets but require a lot of domain-specific knowledge to be effective [17, 21, 89].

**Hybrid Recommender Systems:** Recognizing that no single approach is universally effective, hybrid systems combine features from multiple techniques. A hybrid system might use collaborative and content-based methods to recommend a new movie, for instance. Hybrid systems often yield more robust and accurate results but can be complex to implement [18, 80, 89].

By understanding the underlying techniques, one can better appreciate the strengths and limitations of various RSs. Each approach offers unique advantages depending on the specific needs of the users and the characteristics of the items to be recommended.

Building on this taxonomy of recommendation techniques, our thesis primarily focuses on *Content-Based* approaches, driven by the domains of News and Tourism which often have sparser interaction matrices compared to domains like music or movies. We aim to go beyond surface-level item features and model implicit characteristics to enhance efficacy, satisfaction, diversity, and explainability. However, our approach is not strictly confined to content-based methods. In Chapter 4, where tourism destinations and users are mapped onto the Seven-Factor Model, we also employ *Knowledge-Based* techniques. This model, developed in collaboration with domain experts, integrates expert knowledge into the utility function, lending a multidimensional perspective to recommendations. Finally, in Chapter 5, we incorporate both content-based and *Collaborative Filtering* techniques. We model users based on their past news consumption and develop the utility function based on the interactions of all users within the dataset. Thus, while our primary focus is on content-based recommendation, our work essentially adopts a *Hybrid* approach, reflecting the complexity and nuances inherent in catering to diverse user needs and domain specificities.

## 2.2 Tourism & Recommenders

### 2.2.1 The Travel & Tourism Domain

The Travel and Tourism domain is an expansive and diverse market that encompasses a variety of services such as package holidays, hotel stays, private vacation rentals, camping, and cruises<sup>1</sup>. The market is driven by a broad spectrum of users, namely travelers, who have the luxury of booking through an array of well-established providers like online travel agencies (OTAs) including Expedia and Opodo, and specialized providers for hotels and private accommodations like Hotels.com, Booking.com, and Airbnb.

Economically, the Travel and Tourism sector is an engine of growth with immense potential. For instance, the market is projected to generate an impressive revenue of approximately US\$855 billion in 2023, with hotels being the largest segment, expecting

---

<sup>1</sup><https://www.statista.com/study/40460/travel-tourism/>

to reach a volume of US\$410 billion the same year<sup>2</sup>. The market has consistently grown over the years, partly fueled by the millennial generation’s preference for experiences over material goods. However, it is important to note that the COVID-19 crisis did interrupt this growth, though recovery is ongoing.

Significant growth is also observed in the online aspects of this sector. With a surge in internet penetration globally, online booking of vacations has become the new norm. Digital services have made it convenient and efficient to plan trips, causing the online travel market to be valued at around US\$475 billion in 2022, with projections of it exceeding US\$521 billion in 2023<sup>3</sup>. Indeed, online channels are expected to account for 69% of the total market revenue in 2023<sup>4</sup>.

Advancements in technology have played a crucial role in shaping the market. Artificial Intelligence (AI) technologies, such as chatbots and virtual assistants, are being integrated into travel booking websites and apps to aid customers throughout the booking process. This integration not only enhances user experience but also paves the way for more personalized services, where travel companies can analyze customers’ previous trips and online behavior to offer tailored travel recommendations.

Furthermore, the rise in remote and hybrid working models has led to an interesting intersection of work and vacation, a trend often termed as “digital nomadism”. This offers an opportunity for travel and accommodation providers to cater to a clientele that is not just on vacation but also potentially working part of the time, thus extending the traditional travel seasons.

To sum up, the Travel and Tourism domain is a dynamic and ever-evolving field with numerous opportunities and challenges. It is influenced by various factors, from technological advancements to generational preferences and global events, making it a fascinating area of study and business.

### 2.2.2 Recommender Systems in Tourism

Travel and tourism serve as a significant application area for RSs, offering a unique set of challenges and opportunities. While travelers often face a multitude of decisions ranging from choosing a destination to planning activities, accommodations, and transportation, RSs aim to simplify this complex process. However, the task is not straightforward due to various factors. For instance, unlike domains like movies or music, where user data is abundant, tourism suffers from data sparsity owing to the high costs and time commitments involved in traveling [77, 113, 91].

Additionally, tourism products are complex bundles of various components like activities, accommodations, and transportations, which are not only intangible but also emotionally charged [113]. This emotional aspect further necessitates a highly personalized and

<sup>2</sup><https://www.statista.com/study/40460/travel-tourism/>

<sup>3</sup><https://www.statista.com/study/15218/online-travel-market-statista-dossier/>

<sup>4</sup><https://www.statista.com/study/87036/digitalization-of-the-travel-industry/>

context-aware approach in RSs. New users often face the *cold-start* problem, where the system has limited data to generate personalized recommendations [69]. Technological barriers also exist, particularly in adopting complex machine learning and AI techniques [15]. Finally, accurately capturing user preferences continues to be a persistent challenge [76, 77].

This thesis aims to innovate within this complex landscape by addressing some of these challenges. One of the key contributions is the deep characterization of tourism destinations using the Seven-Factor Model. Introduced in Chapter 3, this model enables more nuanced and accurate matching between user profiles and destinations, improving the quality of recommendations [30]. To tackle the *cold-start* problem, Chapter 4 introduces a picture-based approach for user profiling. Users can upload their pictures, which are mapped to the Seven-Factor Model to determine preferences – similar to [76, 77] where a static set of 63 pictures is utilized. This creative solution not only gamifies the process of preference elicitation but also captures implicit, emotional preferences that are difficult to articulate.

The focus of this work is not just technological sophistication but also user-centricity and emotional engagement. While existing systems may focus either on destination selection or on activities within a destination, this thesis prioritizes the former, making it broadly applicable and foundational for further research [40]. Moreover, the emphasis is on developing a system easily integrable into web-based and mobile platforms, making the technology accessible and user-friendly [15].

In conclusion, by leveraging the Seven-Factor Model and a novel picture-based approach for user profiling, this thesis not only advances the state of research but also has the potential to significantly improve the end-user experience in TRS. It injects fresh perspectives and solutions into a field ripe for innovation, particularly in user profiling and destination characterization.

### 2.2.3 Tourist Roles & The Seven-Factor Model

In the realm of tourism, understanding the nuanced behavior, motivations, and preferences of tourists is critical. Traditionally, various frameworks have been developed to categorize tourists into roles or types, aimed at capturing these subtleties. Cohen’s seminal work laid down a typology consisting of four tourist roles—organized mass tourist, individual mass tourist, explorer, and drifter [23]. Pearce later extended this by identifying 15 roles, including Businessman, Migrant, and Conservationist, among others [81]. Yiannakis and Gibson further fine-tuned the focus by restricting it to leisure travelers and introduced 17 refined roles [125, 42].

These categorizations have proved valuable for understanding travel behavior and psychology. For instance, Gretzel et al. [47] demonstrated that these tourist roles could be effectively used in RSs to suggest suitable travel activities and destinations. However, practical applications like Airbnb’s Trip Matcher<sup>5</sup> or BuzzFeed’s travel personality

---

<sup>5</sup><https://press.atairbnb.com/tripmatcher/>

quiz<sup>6</sup> have not always employed a scientific basis for their typologies, leaving room for improvement.

This leads us to the Seven-Factor Model introduced by Neidhardt et al. [76, 77], which forms the cornerstone of this thesis. This model is an evolution of previous frameworks, reducing the 17 roles of Gibson and Yiannakis into seven comprehensible factors by incorporating “The Big Five” personality traits [70]. The resulting Seven-Factor Model captures both short-term travel behaviors (such as activities preferred during a trip) and long-term personality traits (like openness or conscientiousness). This makes the model a robust and efficient tool for understanding and predicting tourist behavior over varying time frames.

In this thesis, we take the Seven-Factor Model to the next level. Chapter 3 focuses on representing tourism destinations automatically within the Seven-Factor framework based on their attributes, providing a structured way to assess and compare different locations. In Chapter 4, we explore the potential of employing picture collections to map both tourists and destinations onto the Seven-Factor spectrum, thereby facilitating a more personalized and effective recommendation process.

The inclusion of “The Big Five” personality traits in the Seven-Factor Model lends it long-term relevance and robustness [115]. The Seven-Factor Model is not just easier to process computationally and cognitively, but it also enhances our understanding of tourism behavior. This is demonstrated through empirical data collected from questionnaires and factor analyses, revealing significant insights into travel behavioral patterns. The Seven-Factors are summarized in Table 2.1.

Therefore, the Seven-Factor Model offers a well-rounded, efficient, and scientifically sound approach to understanding tourist roles. When contrasted with the largely anecdotal or non-empirical methods currently employed in the industry, the Seven-Factor Model presents a more structured and insightful framework. This is especially crucial for the development of intelligent RSs, which is the ultimate goal of this thesis.

By building on these foundations, this thesis aims to contribute significantly to the field, both in terms of academic research and practical applications, pushing the boundaries of how we understand and cater to the dynamic roles and preferences of tourists.

#### 2.2.4 A Picture-Based Approach

In Chapter 4, we introduce and delve into our model, PicTouRe – a picture based TRS, which builds upon the Seven-Factor Model for travel behavior patterns. This model serves as the cornerstone for creating both user profiles and item descriptions in RSs, as previously established by Neidhardt et al. [77, 76].

The Seven-Factor Model offers a seven-dimensional vector space, where each dimension represents a different aspect of travel behavior. Within this space, user profiles and

<sup>6</sup><https://www.buzzfeed.com/jadayoungatchett/what-type-of-traveler-are-you>

Table 2.1: Seven-Factor Model [76, 77]

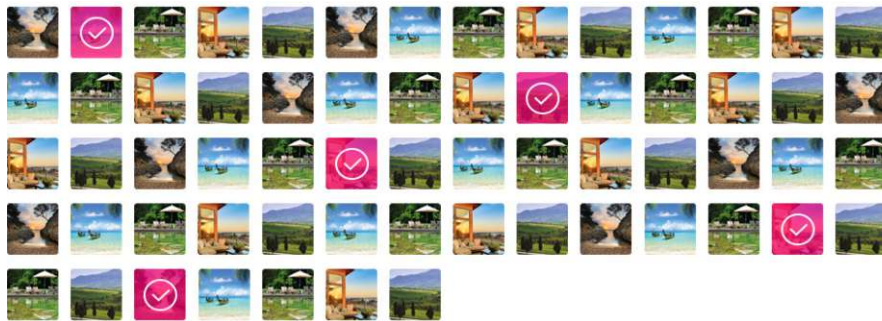
Factor	Description
<i>Sun &amp; Chill-Out</i>	a neurotic sun lover, who likes warm weather and sun bathing and does not like cold, rainy or crowded places;
<i>Knowledge &amp; Travel</i>	an open minded, educational and well-organized mass tourist, who likes travelling in groups and gaining knowledge, rather than being lazy;
<i>Independence &amp; History</i>	an independent mass tourist, who is searching for the meaning of life, is interested in history and tradition, and likes to travel independently, rather than organized tours and travels;
<i>Culture &amp; Indulgence</i>	an extroverted, culture and history loving high-class tourist, who is also a connoisseur of good food and wine;
<i>Social &amp; Sports</i>	an open minded sportive traveller, who loves to socialize with locals and does not like areas of intense tourism;
<i>Action &amp; Fun</i>	a jet setting thrill seeker, who loves action, party, and exclusiveness and avoids quiet and peaceful places;
<i>Nature &amp; Recreation</i>	a nature and silence lover, who wants to escape from everyday life and avoids crowded places and large cities.

Points of Interest (POIs) are treated as individual points. Recommendations are made by calculating the distance between these points, specifically using Euclidean distance, as done by Neidhardt et al. [77, 76].

One of the novel aspects of our approach is its dynamic nature. Unlike Neidhardt et al. [77, 76], who utilized a fixed set of 63 travel-related pictures to capture the Seven-Factors, we allow users to upload their own pictures. This feature adds a layer of flexibility and personalization to the user profiling process.

In the traditional framework, users would identify their travel preferences, needs, and personality by selecting a subset of pictures from a given set (see Figure 2.2). This non-verbal method of preference elicitation addresses the challenge many people face when explicitly stating their needs and desires, especially when it comes to emotional and impulsive decision-making, like travel planning. Our model maintains this advantage while extending it to include a broader and more personalized range of images.

Neidhardt et al. [77, 76] conducted multiple studies to validate their approach. They first used workshop participants to associate a collection of 102 travel pictures with the Seven-Factors. After this, they narrowed down the picture set to a more concise 63 images that held the most information, based on user selection and ranking. Finally, travel experts were used to link these pictures to over 10,000 POIs, each rated on the Seven-Factors on a scale from 0 to 1. Neidhardt et al. [77, 76] employed statistical methods like multiple regression analysis to quantify the relationship between the selected



(a) Preference elicitation - using a fixed set of 63 pictures



(b) User profile - Depicting the user within the Seven-Factor Model



(c) Recommendations - Ranking based on the Euclidian distance between POIs and the user profile

Figure 2.2: PixMeAway - A picture-based recommender based on [77, 76].

pictures and the Seven-Factors.

In contrast, our PicTouRe model adopts a different technique for this quantification. Instead of using regression models, we utilize and fine-tune pre-trained CNNs to map the uploaded images to the Seven-Factors, as detailed in Chapter 4. This provides a more dynamic and technically advanced way to relate images to travel behaviors.

Our approach also allows pictures of the recommendation items, such as tourist destinations, to be represented in terms of the Seven-Factors. This creates a more dynamic and comprehensive system.

In summary, PicTouRe enhances the foundational work of Neidhardt et al. [77, 76] by adding a dynamic and flexible layer to the picture-based approach for TRS. Our method not only captures the complexity and nuance of individual user preferences but also enriches the description of recommendation items, offering a more complete and personalized travel planning experience.

## 2.3 News & Recommenders

### 2.3.1 The News Domain

Over the past two decades, the realm of news consumption has undergone transformative changes influenced significantly by the advent of digital technology. Traditional print newspapers, once the cornerstone of media engagement, have been pushed to the periphery as digital platforms ascend to prominence. This shift has been universal but not uniform, revealing nuanced patterns across various geographical, demographic, and psychological landscapes<sup>7</sup>.

In the United States, for example, well-known publications like *The New York Times* have weathered the storm by adopting digital-first strategies, accumulating over five million subscribers for their digital-only products. However, they represent an exception rather than the rule, as many other outlets struggle to maintain profitability in both print and digital formats<sup>8</sup>.

Similarly, in Europe, the decline in the influence of print newspapers is palpable. Advertising spending on newspapers plummeted by -23.3% in 2020, further strained by the COVID-19 pandemic. Even as some recovery was noted in 2021, long-term forecasts continue to signal a bleak future for the traditional written press, not just in Western Europe but also in Central and Eastern parts of the continent<sup>9</sup>.

Global patterns of news consumption reveal further layers of complexity. For instance, social networks are the primary news source for Greeks and Bulgarians, while audiences in the UK, Germany, and France display less reliance on such platforms. Trust levels in

---

<sup>7</sup><https://www.statista.com/study/112006/global-news-consumption>

<sup>8</sup><https://www.statista.com/topics/994/newspapers>

<sup>9</sup><https://www.statista.com/topics/3965/newspaper-market-in-europe>

traditional media sources vary significantly across nations, and the willingness to pay for news also follows different patterns<sup>10</sup>.

The rapidly evolving landscape presents a multitude of research opportunities, particularly in the area of personalization and recommendation in news. Exploring the impact of sentiment, emotion, and diversity of topics is not merely an academic exercise but a necessity for the industry. Addressing the inherent negativity bias in news and its implications for societal well-being becomes particularly crucial in this context.

In conclusion, the news domain is a fertile ground for research, teeming with challenges and opportunities. As we tread deeper into the digital age, it becomes imperative to focus on innovative models that prioritize credible reporting while catering to a diverse audience. For researchers like myself, interested in personalization and recommendation, there has never been a more pressing need to explore avenues for exploration, personalization, and diversification, especially given the critical role news plays in shaping public opinion and emotional well-being.

### 2.3.2 Recommender Systems in News

The newspaper industry has undergone a significant transformation over the last two decades. With the proliferation of digital platforms, finding news has never been easier. However, the information overload presents a new set of challenges for readers, making it difficult to sift through the deluge of articles and identify the ones that are most relevant. RSs have emerged as crucial tools for mitigating this issue, guiding users through the ever-changing news landscape [59].

NRS exhibit unique characteristics that distinguish them from other recommendation systems, such as those for movies or products. For instance, the high volatility of news relevance and “item churn” necessitate constant model updates [2, 25, 79]. Furthermore, the dynamic nature of users’ interests, influenced by various contextual factors like time, device, and location, adds another layer of complexity [20, 65].

Research has explored various recommendation paradigms, including collaborative filtering, content-based filtering, and hybrid approaches [59]. Though collaborative filtering is popular in academia, news recommendation primarily employs content-based and hybrid methods [79, 61]. The main challenge lies in accurately gauging the effectiveness of these algorithms. Offline metrics such as precision and recall may not always predict online performance, and the issue of balancing accuracy with other quality factors like novelty or diversity persists [61].

This thesis contributes to the field by enhancing the granularity of feature exploration and introducing emotion and sentiment dimensions into news recommendation. Chapter 3 of this thesis dives deep into the intricate features of news articles at various levels—title, abstract, and document. This enables a nuanced understanding of how different features can be leveraged to increase the effectiveness and diversity of news recommendations.

<sup>10</sup><https://www.statista.com/topics/9584/news-consumption-worldwide>

Chapter 5 takes this a step further by introducing a hybrid approach that integrates semantic, sentiment, and emotional elements into the news recommendation process. The emotion and sentiment dimensions are explicitly modeled, thereby providing greater interpretability and facilitating an informed consumption decision. This multi-faceted approach aims not only to improve efficacy but also to increase awareness of diversity in the sentiment and emotional landscape of news consumption.

User modeling in news recommendation is another critical aspect, generally constructed based on explicit or implicit feedback [25, 2, 61]. Various systems have employed different strategies ranging from content-based profiles to community behavior to demographic information. However, capturing both short-term and long-term interests of the user is often cited as important but not well-addressed [2, 25, 61].

In our approach presented in Chapter 5, we use a neural NRS that models both semantic and emotional aspects of users. Utilizing attention mechanisms, the system learns from the sequence of consumed news articles, thereby achieving a fine-grained understanding of user behavior over time.

Diversity is an essential criterion for any recommender system. In the context of news, it is often a tug of war between accuracy and diversity [29, 34, 84, 61]. The ability of a system to present a balanced list of articles, catering to multiple topics and perspectives, can significantly influence user satisfaction.

Chapter 5 specifically addresses this by incorporating sentiment and emotions, allowing for an analysis of the trade-off between efficiency and diversity. One unique focus is on the well-known ‘negativity bias,’ which predisposes users to consume more negative news. Our approach counterbalances this by incorporating sentiment diversification, thereby promoting a more balanced and informed news consumption.

### 2.3.3 Deep-Learning for News Recommendation

Deep learning has gained significant attention in the field of NRS due to its ability to capture intricate patterns in data, eliminating the need for manual feature engineering. Researchers have explored various deep learning architectures and techniques for both news modeling and user modeling, which serve as the backbone of any NRS. For example, the two tower architecture learns representations of users and items separately and models their interaction at a late stage (see Figure 2.3).

The cornerstone of personalized news recommendation is understanding the content of news articles. Traditional methods, limited by the short lifecycle and dynamic nature of news articles, often struggle to build effective news representations. Deep learning has shown promise in addressing these challenges. CNNs are frequently employed for modeling the text of news articles, particularly their titles, due to their effectiveness in capturing local contexts [116, 120, 123]. Attention mechanisms and Transformer architectures have also been incorporated to selectively focus on important features in news content [118, 121]. To enrich the semantic representations, some methods

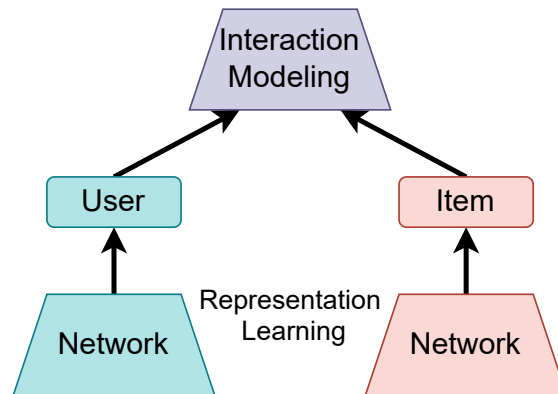


Figure 2.3: Deep learning based interaction modeling for recommendation [126].

integrate entities from knowledge graphs or use additional features like topic categories and sentiments [117, 119, 123].

However, existing methods often overlook the dynamic, timely nature of news and the commonsense knowledge embedded within. Techniques have been introduced to incorporate high-order information from user-news bipartite graphs, but they are static and may struggle with newly published articles [41, 52, 123].

The modeling of user interests has also seen advances with the adoption of deep learning techniques. Historical click behaviors serve as the primary data source for user interest inference [49, 123]. Attention-based mechanisms are prevalent for aggregating historical clicked news to form user embeddings [116, 123]. Some methods extend this by using candidate-aware attention, which computes attention weights based on the relevance to candidate news. Recurrent Neural Networks (RNN), particularly with Gated Recurrent Units (GRU) or Long Short-Term Memory (LSTM) layers, are used to capture the sequential dependencies among different clicked news [123]. However, traditional RNNs have limitations in capturing global interest information and may not suit the diverse preferences in news consumption [122].

To overcome these challenges, self-attention and co-attention mechanisms have been employed, capable of capturing complex relationships between different user behaviors [118]. Some methods have also considered auxiliary user features like device type and location to improve the recommendation [123].

In summary, deep learning techniques have brought significant advancements in both news and user modeling for news recommendation. However, there are challenges such as capturing commonsense knowledge, handling dynamic news content, and the noisy nature of click data that are yet to be fully addressed. In Chapter 5, we explore neural news recommendation methods including Neural News Recommendation with Multi-Head Self-Attention (*NRMS*) [118], Neural News Recommendation with Attentive

Multi-View Learning (*NAML*) [116], and Neural News Recommendation with Long- and Short-term User Representations (*LSTUR*) [4] as baselines and propose a novel method that incorporates sentiment and emotions into the recommendation process.

## 2.4 Recommender Systems & Users' Decision Making

RSs serve as digital counterparts to human sales experts, guiding users through the complex landscape of online product and service choices. While these systems have become integral tools for online retailers, their development has mainly focused on modeling user preferences and crafting sophisticated algorithms for item suggestion. Less emphasis has been placed on comprehensively understanding the nuanced decision-making processes that users undergo, an aspect that remains critical for enhancing user experience and engagement [89, 58, 22].

Traditional economic models of decision-making presuppose that consumers are rational actors with stable, pre-defined preferences. However, research in both behavioral economics and psychology has increasingly contested these assumptions, revealing that decision-making often entails a dynamic construction of preferences. For instance, a consumer might initially set an upper price limit for a product but later adjust it upon learning more about the item's features [58, 22].

To capture this complexity, alternate frameworks such as the Effort-Accuracy model have been proposed. This model posits that decision-making is a trade-off between cognitive effort and decision accuracy, acknowledging that consumers employ a variety of heuristics depending on the context. Furthermore, constructs like bounded rationality illustrate that decision-making is susceptible to various biases and constraints, including cognitive limitations and emotional factors [58, 22].

In line with this, our work aims to extend the capabilities of RSs by integrating deeper insights into human decision-making processes. We do not only aim to improve recommendation accuracy but also aim to analyze and increase diversity by addressing users non-verbally and on an emotional level, moving beyond surface-level attributes.

Incorporating theories from cognitive and decision psychology in the development of RSs can greatly enhance their performance and user experience. We provide a brief overview of five key theories that bear significant relevance to the construction and evaluation of RSs as highlighted by Jannach et al. [58]:

**Context Effects.** Context effects posit that the presentation context of item alternatives has a crucial impact on decision-making. Notably, even the addition of inferior choices can alter choice behavior, contradicting traditional economic models that assume rational, optimal decisions.

**Primacy/Recency Effects.** Primacy and recency effects are cognitive phenomena indicating that items at the beginning and end of a list are more likely to be remembered

and evaluated. Such effects can influence the selection behavior of consumers and must be accounted for in recommendation sequences.

**Framing Effects.** Framing effects, which form the core of our thesis, describe how the presentation manner of decision alternatives impacts user behavior. Whether it is price framing or attribute framing, the way information is presented can considerably influence user choice.

**Priming.** Priming makes certain attributes of an item more accessible in memory, thereby affecting consumer evaluations. Background and attribute priming can even influence the questions a recommender system might ask and the choices it might present.

**Defaults.** Defaults influence consumer choices by capitalizing on the status quo bias. Defaults can reduce interaction effort and even subtly guide user choices towards options that are beneficial to the service provider.

The aforementioned cognitive theories offer invaluable insights into user behavior and decision-making, which are directly applicable to the construction of effective RSs. Our focus primarily lies on the framing effect – the subtle often lies in the framing of the items and/or system – as it not only serves as the backbone of this research but also introduces a nuanced layer to the systems we develop. Specifically, in Chapter 3, we delve into multi-layered aspects of news articles, as well as the intricate interplay between tourist destinations and personality-aware tourist roles. We go beyond surface-level features to expose underlying factors, including employing the Seven-Factor Model, which encapsulates utility based on tourist roles and personality. This focus on framing effects resonates deeply with the need for more human-centric, psychologically attuned systems.

Our approach aims not just at understanding *what* is being presented but also *how* it is framed. This multidimensional perspective guides us through different facets of items, from news articles and tourist destination mappings in Chapter 3 to sentiment-aware news recommendations in Chapter 5. In Chapter 5, we consider the tonality of news articles, exploring how the sentiment and emotions expressed within the text affect the efficacy and diversity of recommendations. This comprehensive examination underscores the thesis' central argument: that recognizing and implementing cognitive theories, particularly the framing effect, can significantly augment the capabilities and efficiency of modern RSs.

Moreover, Jannach et al. [58] highlight following eight highly relevant phenomena from personality and social psychology theory which play a key in the design and development of RSs applications:

**Five Factor Model (FFM).** FFM originates from the lexical hypothesis and identifies five core personality traits: Openness to Experience, Conscientiousness, Extraversion, Agreeableness, and Neuroticism. This model is essential for understanding individual preferences and behaviors, making it particularly valuable for personalizing RSs.

**Locus of Control (LOC).** The LOC theory addresses the extent to which individuals believe they can control events affecting them. In RSs, this theory helps design interfaces that match users' desire for control over the recommendation process.

**Need for Closure (NFC).** NFC reflects an individual's urge to reach a quick decision and minimize cognitive effort. RSs can incorporate NFC by providing feedback mechanisms, like progress bars, that tell users how close they are to completing a task.

**Maximizer and Satisficer (MaxSat).** The MaxSat model identifies two basic behavioral patterns: Maximizers seek the best possible outcome, while Satisficers look for "good enough" solutions. Understanding these patterns can help RSs adapt to user needs.

**Conformity.** Conformity is the adjustment of one's thoughts or actions to align with those of others. RSs can influence user opinions and behaviors, sometimes leading users to conform to the views expressed by the system or its user base.

**Trust.** Trust plays a pivotal role in influencing consumer decisions. For online RSs, establishing trust involves ensuring transaction security, preserving privacy, and enhancing the reputation of the online platform.

**Emotions.** Emotions are states triggered by events important to the individual, impacting both cognitive and physical aspects of behavior. While most RSs overlook this, incorporating emotional aspects can enhance the adaptability of the system.

**Persuasion.** Persuasion theories suggest that human decision-making is not just about preference elicitation but also about preference construction. RSs can act as tools for both, thereby influencing consumer decisions more broadly.

In this thesis, particularly in Chapters 3 and 4, we focus on the Seven-Factor Model. This model, based on both the FFM and 17 tourist roles, provides a nuanced understanding of users' personalities and preferences through a unique methodology. Unlike traditional methods that rely on questionnaires, we use fine-tuned CNNs to analyze images provided by users and map them to the Seven-Factors. This automated mapping creates a gamified way of understanding user attributes and personalities without subjecting them to tedious questionnaires.

Emotion is another core aspect explored in this thesis, specifically in Chapter 5, where we include sentiments and emotions expressed in news articles to make the recommendation process more nuanced. Not only do we build a semantic representation of articles consumed by users, but we also make the emotional and sentiment dimensions interpretable. This offers the ability to raise user awareness about emotional diversity and intervene, if necessary. For example, in cases of evident negativity bias, we can either confront the user with their emotional consumption patterns or diversify the emotional tone of their recommendations.

Lastly, the notion of Trust is implicitly addressed in our methodology. By allowing users to validate or correct their Seven-Factor model representations, and by retaining the interpretability of the emotional and sentiment dimensions, we enhance the system's trustworthiness. The transparency of this approach empowers users to understand how recommendations are generated, thus fostering a sense of trust and control.



# Exploring & Exposing Implicit Item Characteristics

In this chapter we investigate **RQ1** *How can we systematically identify and expose the implicit item characteristics embedded in structured and unstructured data to enhance recommender systems?* We focus on uncovering and understanding implicit item characteristics without involving user interactions. This analysis aims to go beyond the obvious item characteristics and delve into the less apparent, yet significant attributes that contribute to user preferences.

In the context of tourism, this involves trying to map structured data, or the visible features of tourism destinations, to a well-established domain model, i.e., the Seven-Factor-Model, which captures implicit item and user characteristics. Essentially, this means drawing connections between what is readily observable about a destination and deeper aspects related to it, such as cultural nuances or potential experiences that are not immediately visible.

In the news domain, this research conducts a multi-level analysis at the article, author, and topic levels. The aim is to expose implicit item characteristics hidden within these elements and discuss their role in enhancing RSs and diversity. This includes detecting patterns in articles that are not overtly evident or understanding an author's subtle biases that may influence a reader's preferences.

Overall, this exploratory data analysis serves as a foundation for understanding the implicit item characteristics in both domains, setting the stage for user involvement in the subsequent stages of research.

### 3.1 The “Personality” of Tourism Destinations

The intertwined relationship between Information and Communication Technologies (ICT) and the tourism sector has dramatically reshaped the industry, especially with the rise of the WWW [113]. Contemporary consumers enjoy an unparalleled access to vast information, enabling them to exchange experiences and compare travel offerings effortlessly. However, this information abundance often induces cognitive overload, highlighting the need for sophisticated tools to efficiently manage, categorize, and personalize information for consumers.

The pervasive influence of online platforms is evident in how 65% of leisure travellers initiate their travel research online [72]. In these preliminary decision-making phases, many struggle to articulate specific preferences and desires [128]. RSs offer a solution by suggesting products that align with user preferences, especially vital in tourism where offerings amalgamate diverse components like accommodation, transport, and experiences, which are complex and emotional by nature [114].

Neidhardt et al. [76, 77] pioneered a picture-based methodology to gauge user preferences, proposing a Seven-Factor Model that merges the renowned “Big Five” personality traits [45] with 17 tourist roles. This model orchestrates a seven-dimensional vector space, categorizing travel behaviours into distinct patterns. To offer personalized recommendations, both users and items, like POIs, need representation within this space. But manual mapping, as initially done for over 10,000 POIs, is infeasible for scalability.

The challenge herein is twofold, which forms the central research inquiries of this section:

**RQ1.1.1** *Is there a latent underlying structure that aligns with the Seven-Factor Model?* Exploratory data analysis and cluster analysis identified six meaningful clusters of tourism destinations. However, representing destinations with diverse attributes is challenging due to the mutual exclusivity of the clusters.

**RQ1.1.2** *Can we devise a mechanism to automatically map tourism destinations to the Seven-Factor Model considering their attributes?*

Multiple linear regression (MLR) models significantly linked destination attributes to the Seven-Factor Model, outperforming their non-linear counterparts K-Nearest-Neighbour regression (KNN) and Random-Forest regression (RF) in accuracy and interpretability, despite some data and model limitations.

While prior work focused on mapping POIs and hotels, our research emphasizes tourism destinations. Unlike previous attempts that classified hotels under singular factors, we aim for comprehensive scoring across all factors. Our objective is to discern conceptual groups among destinations, offering profound insights into their resemblances and variances. This deepened understanding aids RSs to function optimally. Moreover, our approach critically evaluates the connection between the Seven-Factors and destination attributes, extending the pioneering analysis in Sertkan, Neidhardt, and Werthner [96]. Our methodology integrates advanced data preprocessing, extensive cluster analyses, and employs diverse regression methods, concluding with a thorough evaluation and discussion.

Our contributions can be summarized as follows:

- We conduct an unsupervised cluster analysis to discover latent structures in the data, identifying six conceptually meaningful clusters of destinations based on their attributes, without the need for prior expert knowledge.
- We establish seven MLR models, each corresponding to a factor in the Seven-Factor Model, demonstrating a significant relationship between selected destination attributes and the Seven-Factors.
- We compare the performance of three regression models (MLR, RF, and KNN) on the data, finding that MLR offers a balance of good performance and interpretability.

### 3.1.1 Related Work

The role of ICT in reshaping the tourism industry is profound [113]. With the advent and growth of the WWW, the surge in available information has led to the challenge of information overload. RSs emerged as a solution to this issue, aiming to match consumers with appropriate products or services, improving their overall experience, and simplifying decision-making processes. In the context of tourism, these decisions can span from choosing travel destinations, modes of transportation, accommodations, to activities and beyond.

Over the years, several recommendation techniques have been developed, including content-based, collaborative-filtering, knowledge-based, demographic, community-based, and hybrid. Notably, many of these methods depend on user ratings. Such techniques have been widely applied in domains like movies, music, and books. However, the unique nature of tourism — often being expensive, time-consuming, and emotionally charged — means fewer rating data is available, making accurate personalization a challenge [77].

In this backdrop, the challenge of creating effective RSs for the complex, intangible, and emotion-driven tourism products becomes evident [113]. To tackle this, RSs often rely heavily on content and knowledge, as evidenced by Neidhardt et al. [77]. Supporting this, Burke and Ramezani [19] suggest that the most fitting recommendation techniques for tourism are knowledge-based or content-based.

The current work is positioned to automate the determination of the Seven-Factor representation of tourism destinations, aiming to align them with user profiles. The innovative picture-based approach to RSs [76, 77] provides a user-friendly method to elicit the Seven-Factors of a user. Through this approach, users’ preferences are captured intuitively and emotionally without explicit declarations. This strategy, being both content and knowledge-driven, addresses the often-encountered challenge of users struggling to articulate their preferences, and it offers a potential solution to the *cold-start* problem in RSs. Our approach can be viewed as a natural continuation or complementation of the picture-based method. The automated mapping from destination features to the Seven-Factor representation enables a straightforward way to populate the recommendation base.

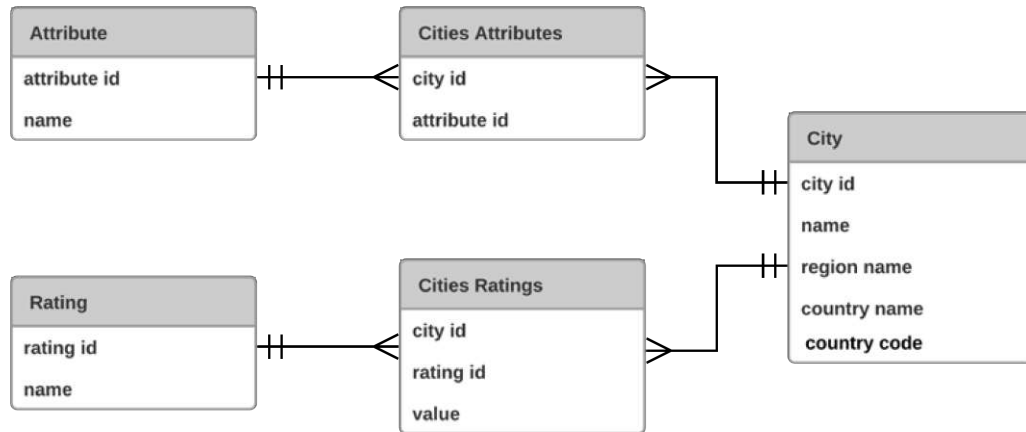


Figure 3.1: Entity-relationship model of the provided destinations SQL-dump.

Essentially, while the picture-based approach maps users onto the Seven-Factors, our method maps destinations to the same Seven-Factors. With both users and destinations mapped to the same vector space, recommendations can easily be generated using simple distance measures.

Distinguishing between different types of TRS, Garcia, Sebastia, and Onaindia [40] categorizes them into those focusing on destination selection and those highlighting activities within a chosen destination. This work primarily aligns with the former, emphasizing tourism destinations as the core items for recommendation. This is a shift from the approach of Neidhardt et al. [76, 77], who prioritize POIs. Although research on destination RSs is extensive [37, 15], many studies target specific regions or individual POIs; fewer integrate user personality traits and motivations [16].

### 3.1.2 The Data

Beirman et al. [7] refer to a tourism destination as “a country, state, region, city or town which is marketed or markets itself as a place for tourists to visit”. In this work destinations are defined in a similar way, except that the range is wider, i.e., from a hamlet with a population smaller than 100 to a metropolis with a population larger than one million. The data is provided as a SQL-dump by a German e-Tourism company and consists of more than 30,000 destinations all around the world.

Fig. 3.1 shows the structure of the tables in the SQL-dump and the relations among them. Destinations are described through 22 geographical attributes and 27 motivational ratings.

**Motivational ratings** lie in the interval  $[0,1]$  and describe the degree of suitability for a particular motif. The following 27 motifs are listed: *nightlife, wellness, shopping, nature & landscape, image & flair, culture, sightseeing, entertainment, mobility,*

*price level, accommodations, gastronomy, beach & swimming, golf, scuba diving, kite & windsurfing, hiking, cycling, horseback riding, winter sports, sports, family, peacefulness, surfing, sailing, gays, mountain biking.* The motivational ratings are determined by the e-Tourism company by considering factors such as infrastructure, climate, user opinions, number of services, image, and marketing. However, not all details are disclosed and thus it is not known how exactly the scores are determined.

**Geographical attributes** are given in binary format and describe the presence or absence of a particular geographical attribute. The following 22 attributes are listed: *sea, mountain, lake, island, sandy beach, metropolis, forest, river, desert, old town, pebble beach, sand & pebble beach, hill, swamp, volcano, fjord, flat decaying sandy beach, beach promenade, wine-growing, heath, health resort, winter sports resort.*

All possible attributes and ratings are persisted in the tables *Attribute* and *Rating*. As previously mentioned, there are over 30K tourism destinations. They are persisted in the table *City*. This table contains an identifier for each destination and textual descriptions to capture destination name, region name, country name, and country code in ISO 3166-1 alpha-2 format, for example AT for Austria. In the table *Cities Attributes* tuples of geographical attributes and tourism destinations are recorded, e.g. (Vienna, *old town*). Similarly, the table *Cities Ratings* persists the motivational ratings of a tourism destination with corresponding (rating) value, e.g. (Vienna, *culture*, 0.99). A major drawback of such a structure is that a tourism destination does not necessarily have an entry for each rating or attribute. Thus, in many cases it is not clear if a destination does in fact not have such attribute or rating, or the data is missing. This ambiguity leads to many “missing values” and in turn to a sparse data set, which are treated as follows.

First, destination attributes, which are denoted by the data provider as experimental, are deleted. Afterwards, destination attributes with similar meaning are combined, e.g. *sandy beach, flat decaying sandy beach, pebble beach, sand and pebble beach* are combined to the attribute *beach*. Then, destinations with many missing values are discarded. Since there are seven destination features (i.e., *price level, gastronomy, sports, accommodations, shopping, nightlife, and entertainment*) which are mostly non-missing (i.e., in about 90% of the cases), only destinations with minimum ten non-missing features are kept in the data set in order to have at least three more aspects of a tourism destination.

This leads to a more concise data set with 16950 destinations and 38 attributes (i.e., 26 motivational ratings and 12 geographical attributes). Finally, a model-based procedure is defined (i.e., naive imputation for geographical attributes and SOFT-IMPUTE [71] for motivational ratings) in order to intelligently replace the remaining missing values (i.e., 62% of the overall cells).

Almost all countries are represented in the database, but the majority (65%) of destinations are located in the USA, Germany, France, Italy, Spain, Great Britain, Austria,

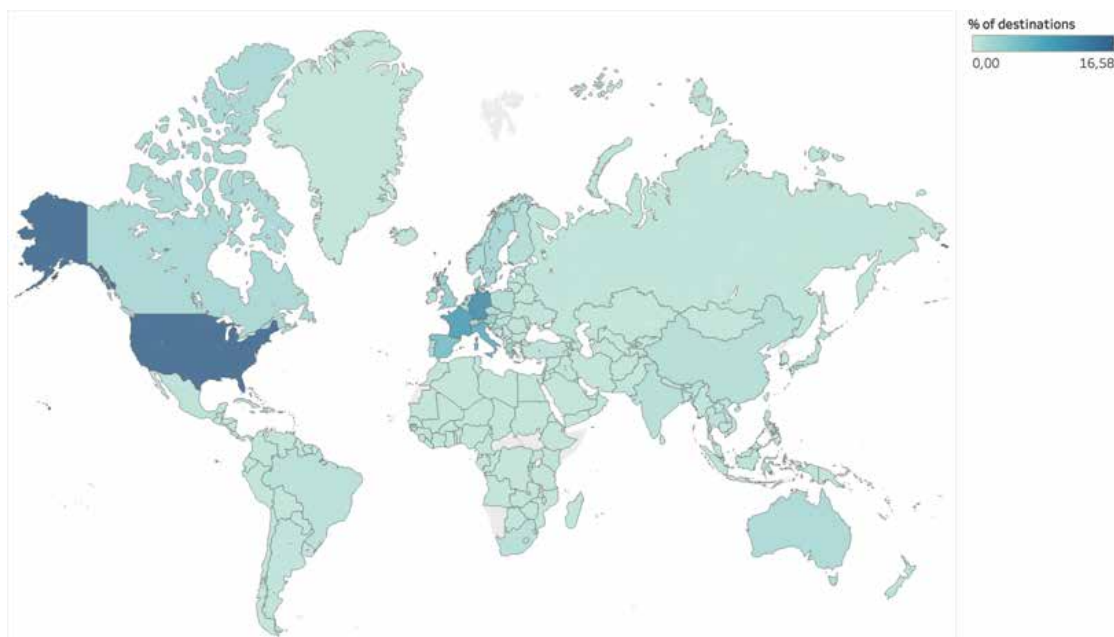


Figure 3.2: Distribution of tourism destinations over countries.

Greece, Switzerland, and Sweden. This can also be observed in Fig. 3.2, where the distribution of tourism destinations over countries is presented as a heat map.

Summary statistics of the distributions of the different motivational ratings for the tourism destinations are listed in Table 3.1. The motivational ratings *nature & landscape*, *peacefulness*, *hiking*, *cycling*, and *mountain biking* have similar and high average values of 0.61-0.71. Those ratings are not only similar, but they can also be considered as nature and recreation related. On the other end, some specific water sports related ratings (i.e., *sailing*, *diving*, *kite- and windsurfing*, and *surfing*) have low average values (0.21-0.29).

In Table 3.2 the relative frequencies of the different geographical attributes are listed. It can be observed that 21-25% of destinations are either on an island and/or at the seaside and/or near a beach. However, it is noteworthy that in the data set 43% of destinations do not have any geographical attribute listed at all.

In order to analyse similarities among all destination attributes (i.e., motivational ratings and geographical attributes) a correlation matrix comprising all pairwise Pearson correlation coefficients is calculated. To get a better understanding and overview, the correlation matrix is visualized as a clustered heat map (see Fig. 3.3).

Following meaningful groups are identified (marked by red rectangular):

- Group one comprises attributes with a relationship to recreational travelling. *Peacefulness*, *nature & landscape*, *mountain biking*, *hiking*, *cycling* and *winter sports* are forming this group.

Table 3.1: Summary statistics of the motivational ratings.

	mean	std	min	median	max
nightlife	0.51	0.15	0.05	0.50	0.99
wellness	0.37	0.16	0.05	0.32	1.00
shopping	0.52	0.14	0.02	0.51	1.00
nature_landscape	0.71	0.16	0.10	0.75	1.00
image_flair	0.64	0.14	0.09	0.62	1.00
culture	0.51	0.15	0.03	0.46	1.00
sightseeing	0.40	0.17	0.05	0.35	1.00
entertainment	0.31	0.21	0.01	0.26	0.98
mobility	0.45	0.15	0.13	0.41	1.00
pricelevel	0.57	0.15	0.10	0.56	1.00
accommodations	0.41	0.19	0.05	0.38	0.97
gastronomy	0.40	0.20	0.05	0.36	0.98
beach_swimming	0.45	0.21	0.01	0.36	0.98
golf	0.39	0.13	0.01	0.36	1.00
diving	0.28	0.16	0.01	0.22	1.00
kite_windsurfing	0.26	0.15	0.01	0.19	1.00
hiking	0.68	0.17	0.12	0.69	0.98
cycling	0.63	0.15	0.09	0.66	0.96
horsebackriding	0.31	0.11	0.01	0.28	1.00
wintersports	0.27	0.12	0.03	0.26	0.93
sports	0.39	0.19	0.05	0.36	0.98
family	0.44	0.14	0.04	0.41	1.00
peacefulness	0.70	0.18	0.06	0.74	1.03
surfing	0.21	0.14	0.01	0.15	1.00
sailing	0.29	0.15	0.01	0.23	1.00
mountainbiking	0.62	0.18	0.09	0.59	0.99

- Group two consist of *island and volcano*, which might be a sign for nature and adventure.
- In group three there are typical attributes of destinations at the countryside, namely *river, hill, mountains, forest, lake and health resort*.
- Group four comprises attributes related to city trips and metropolitan areas. Those attributes are *shopping, price level, metropolis, and old town*.
- Group five contains attributes related to mass tourism, such as *image & flair, entertainment, gastronomy, accommodations, sports, wellness, mobility, nightlife, culture, and sightseeing*. The attributes within this group are strongly positively correlated.

Table 3.2: Frequencies of geographical attributes.

	relative frequency
island	0.25
sea	0.22
beach	0.21
mountains	0.09
forest	0.07
old_town	0.07
hill	0.07
volcano	0.06
lake	0.04
health_resort	0.04
metropolis	0.03
river	0.03

- Group six mainly comprises attributes related to water sports and beach vacations. Attributes within this group are *kite & windsurfing, sailing, sea, beach, surfing, beach & swimming, diving, family, golf, and horseback riding*. The last three attributes are a bit detached from the other attributes within this group, which is also reflected by the low correlation coefficients.

Overall, one can observe a contrast between attributes related to mass tourism (green rectangular) and attributes related to recreational destinations (first group), especially in the case of motivational rating *peacefulness*. However, the correlation analysis delivers first insights of a latent structure in the given data set and furthermore the results can be used in order to substitute or merge highly correlated attributes.

Out of all destinations, 561 destinations were chosen randomly and mapped manually to the Seven-Factors by experts. These experts were members of an Austrian e-Tourism company using an implementation of the picture based approach. Thus, they were familiar with both characteristics of tourism destinations and the Seven-Factor Model. For the 561 destinations, three experts assigned first individually a score for each factor using the scale 0 - 0.25 - 0.50 - 0.75 - 1. The higher the score the more suitable, in the expert's opinion, the destination for that specific factor. After the individual mappings, a final mapping was determined in a joint discussion. The majority of destinations in the sample are located in Germany, USA, France, Greece, Great Britain, Italy, Denmark, Spain, Austria, and Netherlands (62%), which is similar to the distribution in the whole data set. However, after the essential missing value treatment, also the expert sample got smaller in size, namely from N=561 to N=350 (62%).

In Fig. 3.4 the rating behaviour of the experts can be examined, i.e., the score distribution for each of the Seven-Factors are shown. For example, in case of the factor *Sun & Chill-Out* 30% of the destinations scored with 0, 17.1% with 0.25, 15.4% with 0.5, 10.6% with

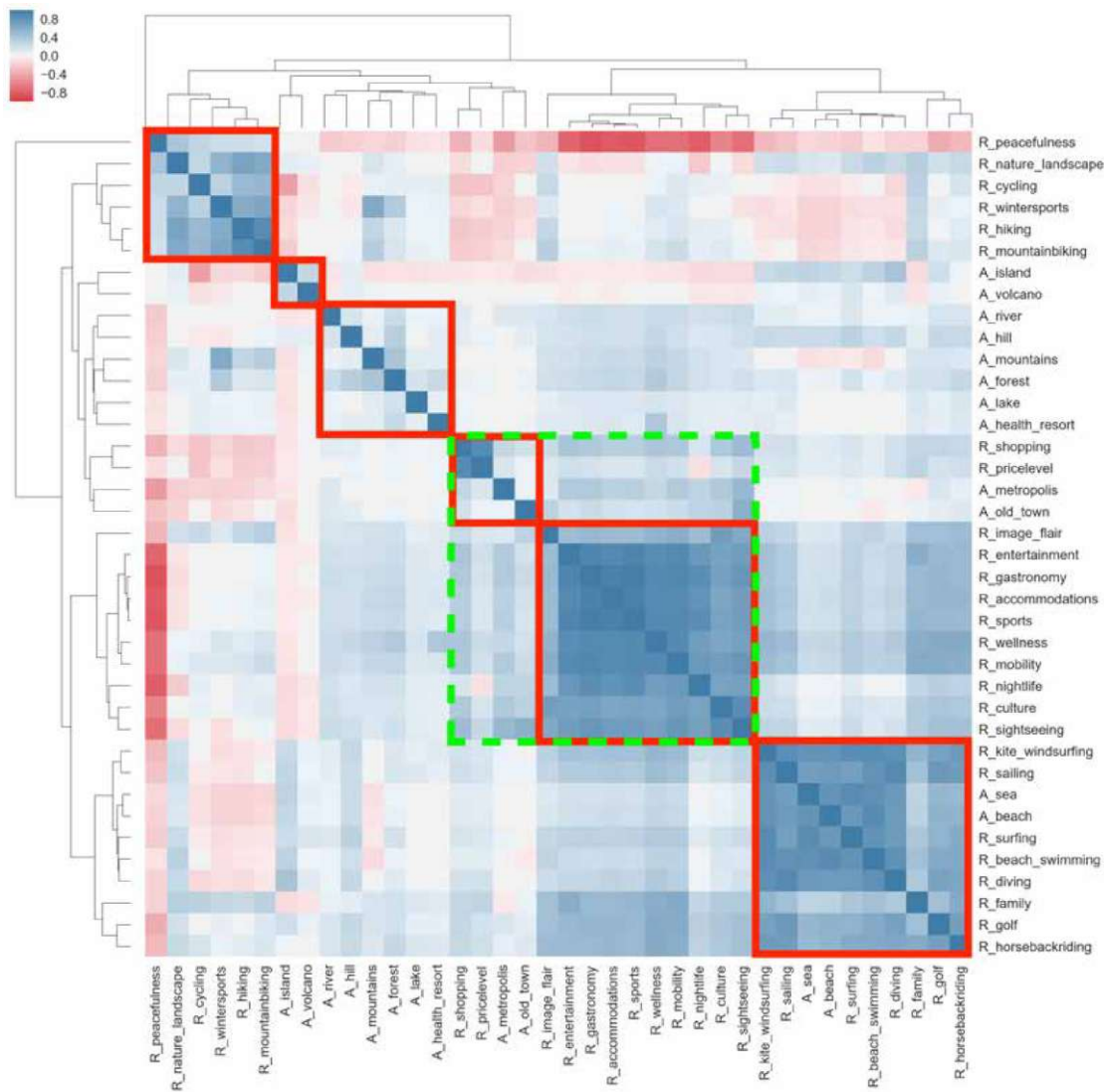


Figure 3.3: Clustered correlation heatmap of all destination attributes.

0.75, and 26.9% with 1. Thus, the majority of destinations (56.9%) scored with 0 or 1 in the factor *Sun & Chill-Out*. Whereas, the majority of destinations (55.7%) in case of *Knowledge & Travel* scored with either 0 or 0.25 and a few with 1 (10.6%), similar to the distribution in the factor *Action & Fun*. Almost half of the destinations have a score in the “lower middles,” i.e., 0.5 (28.3%) and 0.25 (21.4%), in the factor *Culture & Indulgence*. This is similar to the distribution in the factor *Nature & Recreation*, where the majority of destinations have scores in the “upper middles,” i.e., 0.5 (24%) and 0.75 (27.4%). An extreme case of this “upper middles” can be seen in factor *Social & Sports*, where almost all destinations (87.1%) scored with either 0.5 (53.4%) or with 0.75 (33.7%). Only *Independence & History* shows (approximately) a normal distribution.

### 3. EXPLORING & EXPOSING IMPLICIT ITEM CHARACTERISTICS

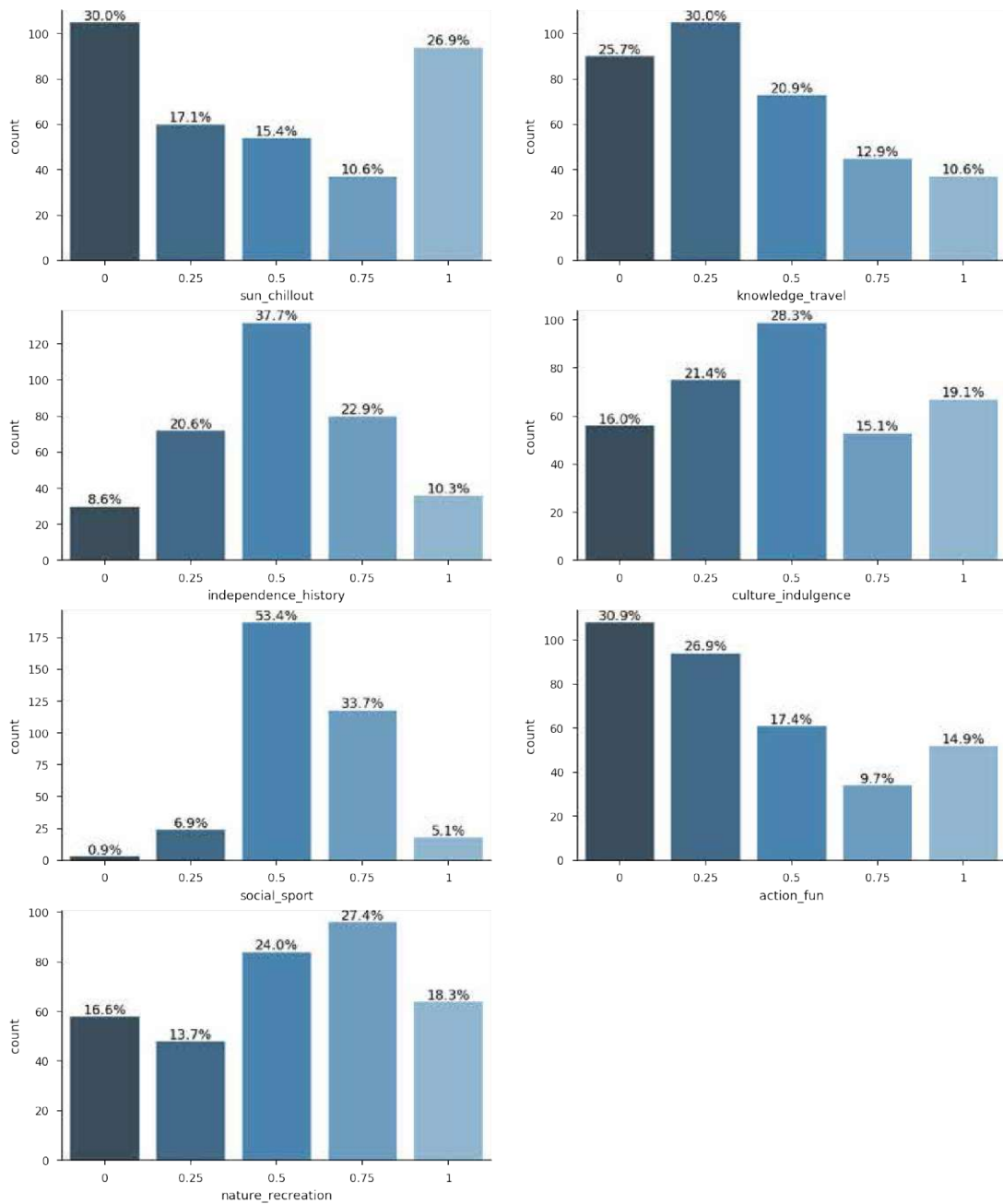


Figure 3.4: Distribution of Seven-Factor scores in the expert sample.

#### 3.1.3 Cluster Analysis

Identifying conceptually meaningful groups of destinations with shared common characteristics will help to further understand the data and its structure, which may contribute

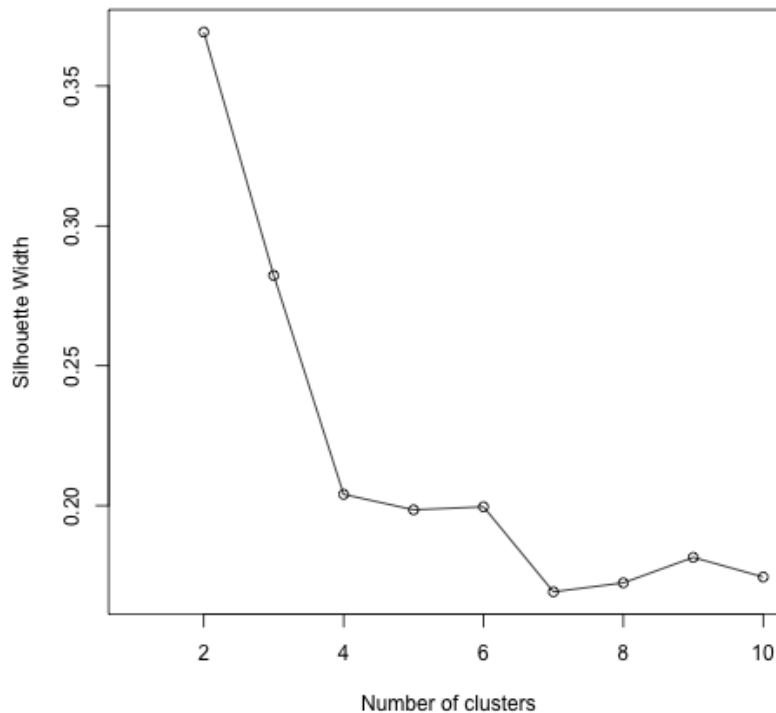


Figure 3.5: Average silhouette width in different  $k$  numbers of clusters. Higher scores indicate better separation and consistency of clusters.

to a more generalizable solution. Furthermore, clustering is considered as an unsupervised learning method. Hence, it does not rely on a priori labelled data sets and thus on resource intensive (e.g., time, costs, etc.) expert knowledge. Therefore, it can be used as an initial approach for recommending destinations if no further knowledge about the destinations is available. The cluster analysis comprises 16950 destinations (i.e., the data set after pre-processing). Partitional clustering techniques are considered, where most prominent ones are K-means and K-medoids. Since the data comprises binary attributes, using the Euclidean distance and thus centroids (both are essentials of the K-means algorithm) are not meaningful. Therefore, K-medoids is applied. A medoid corresponds per definition to an actual data point, which is considered as the most representative point for the cluster [107]. Specifically, Partitioning Around Medoids (PAM) [64], the most common K-medoids algorithm, is used. Since the data consists of two different data types, i.e., binary (geographical attributes) and continuous (motivational ratings), the Gower distance (appropriate for mixed datatypes) [46] is used as distance metric.

In order to find an appropriate number of clusters, the internal evaluation metric silhouette width (i.e., silhouette coefficient) [90] is used for assessment. The silhouette

width measures how similar an object is to its own cluster compared to other clusters, with higher values indicating better-defined clusters. Fig. 3.5 shows the average silhouette width within different cluster sizes. Based on the average silhouette width two, three, four and six cluster solutions are considered, but for the sake of interpretability and diversity a six-cluster solution is chosen. For example, the four-cluster solution distinguishes between beach resorts and non-beach-resorts. On the other hand, the six-cluster solution enables to distinguish between mass touristic beach resorts, recreational beach resorts, and non-beach-resorts. This adds diversity to the cluster solution and facilitates finding associations with the Seven-Factors. Next, the resulting clusters are examined in detail. The number of destinations in each cluster is provided at the beginning of each paragraph.

**C1 (N = 1940).** The medoid of cluster C1 is Paralia, a small city in Greece. Paralia means in Greek beach and as the name already suggests, the city is located directly at the beach. It is a popular and vibrant seaside resort with many nightlife and shopping opportunities. Interestingly, 93% of the destinations in C1 are located at the sea and 92% directly at the beach, whereas globally only about 20% of destinations are located at the sea or beach. Also, the rating *beach & swimming* has a high mean value of 0.81. Additionally, ratings *gastronomy*, *nightlife*, *sports*, *accommodations*, and *culture* are showing an increased average value (0.62 - 0.66). To conclude, destinations in C1 are mainly located on the beach, vibrant and lively, and also attractive for various sports.

**C2 (N = 2177).** The medoid of cluster C2 is Gubbio, a city located on the lowest slope of Mt. Ingino in Italy. Its origins are ancient and reach back to the Bronze Age. Thus, many cultural and sightseeing activities are provided. Features *image & flair*, *hiking*, *culture*, *gastronomy*, *nightlife*, *mobility*, *accommodations*, *sports*, *sightseeing*, and *entertainment* are showing increased mean values in C2 (0.61 - 0.77). Interestingly, 17% of the destinations in C2 are metropolises, which is about six times more considering the whole data set. Plus, only 1% of the destinations in C2 are located at the sea or beach. Hence, destinations in C2 can be considered as mainly vibrant cities or metropolises not located at the beach, offering many nightlife, cultural, sightseeing, gastronomy, and entertainment opportunities.

**C3 (N = 1774).** The medoid of cluster C3 is Aghios Markos, a small, peaceful village in the nature on the island of Corfu, Greece. In C3, 90% of the destination are located at the sea, 88% at the beach, and 70% on an island. Whereas, in the whole data set only 20% of destinations are located at the sea or beach and 25% on an island. Ratings *beach & swimming*, *nature & landscape*, and *peacefulness* have an increased mean value of 0.77-0.79. Furthermore, there is only one metropolis in C3. Therefore, destinations in C3 can be seen as small and peaceful towns at the seaside, probably on an island, with a few sports opportunities and not many tourists.

**C4 (N = 5576).** The medoid of cluster C4 is Montbrió del Camp, a small, peaceful village in Catalonia, Spain. The average value of motivational rating *peacefulness* in C4

is 0.81. Also, ratings *nature & landscape*, *hiking*, *cycling* and *mountain biking* have an increased mean value of 0.62-0.67. Interestingly, none of the 5576 destinations is located on an island or is a metropolis. Furthermore, the mean values of all other attributes are relatively low for destinations in C4. Hence, destinations of C4 can be considered as small and peaceful villages, probably in the nature, and more or less good for hiking, cycling, and mountain biking.

**C5 (N = 1877).** The medoid of cluster C5 is Reynoldston, a small, peaceful village in Wales, Great Britain. Interestingly, all destinations within this cluster are located on an island, only 2% are at the beach, and there is only one metropolis. Further, only ratings *peacefulness* (0.76), *nature & landscape* (0.71), and *hiking* (0.64) are showing an increased mean, all other destination attributes have a relatively low average value. C5 is quite similar to C4, except that destinations of C5 are only located on islands, where destinations of C4 are not. Thus, destinations of C5 can be considered as mainly small, peaceful villages, located on an island and in the nature, with some recreational sports offers.

**C6 (N = 3606).** The medoid of cluster C6 is Irun, a city in Spain at the border to France and on the Atlantic coast. It offers some cultural and sightseeing activities, but also some sports and recreational activities in the nature. Following ratings have an increased average values (0.63 - 0.71) within this cluster: *nature & landscape*, *peacefulness*, *image & flair*, *mountain biking*, *cycling*, *nightlife*, and *culture*. In C6, 12% of the destinations are located near a mountain, which is about three times more compared to the whole data set. Only 1% of the destinations are considered as metropolises and only 1% are located at the beach or sea. Thus, destinations within C6 can be considered as small cities, probably in the nature, with recreational, cultural, and entertainment offers, but none of them dominating.

In summary, it can be said that there is an underlying structure of the data leading to six conceptually meaningful groups of destinations. For a better understanding, these groups or clusters can be simplified and summarized as follows: C1 - *vibrant beach resorts*, C2 - *energetic cities*, C3 - *tranquil seaside resorts*, C4 - *peaceful towns*, C5 - *idyllic island villages*, C6 - *ordinary towns*. Furthermore, the identified clustering is showing a clear contrast between tranquil and vibrant, land and island, and seaside and inland.

The silhouette plot in Fig. 3.6 displays the silhouette coefficients of each destination in a cluster in an ordered way. The red dashed line shows the average silhouette width of 0.2, which assisted to find the right cluster size. The silhouette plot enables a visual assessment of the relative quality of the developed clustering. A negative silhouette coefficient indicates an incorrect assignment of a destination to a cluster and a very low silhouette coefficient points out that a destination is located in-between two clusters. Hence, almost none of the destinations in C4 and C5 are incorrectly assigned, but some

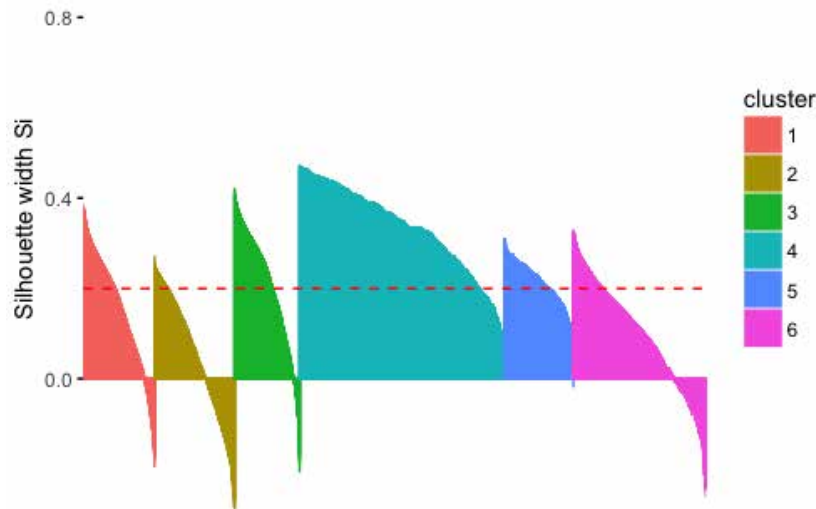


Figure 3.6: Silhouette plot of the six-cluster solution.

might be located between two clusters. Whereas, in all other clusters there are falsely assigned destinations, especially in C2 and C6.

Furthermore, the proposed clustering can be analysed in the light of the provided expert mapping (N=350). Table 3.3 lists the mean Seven-Factor scores and corresponding standard deviations (sd) of the expert sample as a whole and in different clusters. Note, cluster means, which are significantly different to the average factor scores of the rest of the sample, are in bold. Significance is tested with the Mann-Whitney-U-test [51] at a significance level of 0.05.

The factor *Sun & Chill-Out* shows, as expected, an increased mean value in cluster C1 - *vibrant beach resorts* and C3 - *tranquil seaside resorts*. The factors *Knowledge & Travel*, *Independence & History*, and *Culture & Indulgence* have an increased average score in clusters C1 - *vibrant beach resorts* and C2 - *energetic cities*, where in contrast to the other clusters more sightseeing, cultural, entertainment, and nightlife activities are offered. Furthermore, the mentioned factors have a decreased mean value (in comparison to the sample mean) in the clusters with more tranquil and peaceful destinations (C3, C4, C5), where the probability of activities fitting the characteristic aspects of the three mentioned factors is low. In comparison to the sample mean, the factor *Social & Sports* shows an increased score in cluster C1 - *vibrant beach resorts* and a decreased score in C3 - *tranquil seaside resorts*, but overall it does not show much diversity. In other words, the cluster means are similar (i.e. 0.50 - 0.64) throughout the clusters and in comparison to the sample mean. The factor *Action & Fun* has an increased mean value in destinations, which are considered as vibrant and energetic (i.e., destinations in clusters C1 and C2), and a decreased mean value in destinations, which are considered as peaceful, idyllic, or ordinary (i.e., destinations in clusters C4, C5, and C6). Such average Seven-Factor score

Table 3.3: Average factor scores and standard deviations (sd) of the expert sample (N=350) as a whole and in different clusters.

		sample	C1	C2	C3	C4	C5	C6
<i>Sun &amp; Chill-Out</i>	mean	0.37	<b>0.71</b>	<b>0.18</b>	<b>0.96</b>	<b>0.22</b>	0.39	<b>0.30</b>
	sd	0.38	0.31	0.26	0.11	0.25	0.43	0.33
<i>Knowledge &amp; Travel</i>	mean	0.30	<b>0.46</b>	<b>0.65</b>	<b>0.24</b>	<b>0.14</b>	<b>0.18</b>	0.30
	sd	0.30	0.31	0.30	0.21	0.21	0.23	0.24
<i>Independence &amp; History</i>	mean	0.45	<b>0.58</b>	<b>0.73</b>	<b>0.42</b>	<b>0.31</b>	<b>0.32</b>	0.45
	sd	0.28	0.23	0.22	0.21	0.24	0.28	0.24
<i>Culture &amp; Indulgence</i>	mean	0.40	<b>0.58</b>	<b>0.76</b>	0.41	<b>0.23</b>	<b>0.24</b>	0.44
	sd	0.34	0.30	0.26	0.22	0.26	0.30	0.30
<i>Social &amp; Sports</i>	mean	0.57	<b>0.64</b>	0.57	<b>0.50</b>	0.59	0.58	0.58
	sd	0.19	0.16	0.19	0.16	0.18	0.14	0.20
<i>Action &amp; Fun</i>	mean	0.26	<b>0.59</b>	<b>0.56</b>	0.30	<b>0.05</b>	<b>0.12</b>	<b>0.19</b>
	sd	0.33	0.34	0.34	0.19	0.15	0.17	0.20
<i>Nature &amp; Recreation</i>	mean	0.54	<b>0.35</b>	<b>0.31</b>	<b>0.76</b>	<b>0.82</b>	<b>0.79</b>	<b>0.72</b>
	sd	0.34	0.28	0.30	0.18	0.17	0.22	0.25

Note Cluster means, which are significantly different to the average factor scores of the rest of the sample, are in bold

distribution is reasonable and in line with the characteristics of the factor. Finally and obviously, the factor *Nature & Recreation* has an increased score in tranquil, peaceful, ordinary, and idyllic places (i.e., destinations in C3, C4, C5, and C6), rather than in destinations with mass tourism characteristics (i.e., destinations in C1 and C2).

### 3.1.4 Regression Analysis

The main goal is to automatically map destinations onto the seven-dimensional vector space of travel behavioural patterns based on attributes of the destinations. However, additionally the relationships between the Seven-Factors and these attributes should be understood. James, Witten, Hastie, and Tibshirani [56] suggest to choose linear models over more complex ones if inference and interpretability is the goal. Taking this into account, a multiple linear regression model [55] with step-wise variable selection [54] is applied. All Seven-Factors are considered as independent from each other, since they are obtained from factor analysis. Therefore, they can be treated separately by fitting a model for each travel behavioural pattern, which takes the attributes of a destination as input and returns the factor score in the interval  $[0, 1]$  as output. The expert sample is split into a training and test set in a ratio of 80/20. Model performance is assessed by  $R^2$ , the proportion of variance explained, and root mean square error (RMSE), the standard deviation of the residuals / prediction errors [48, 56]. Furthermore, a performance evaluation is conducted, in order to compare the performance of the linear model against following two more complex models: KNN regression [55] and RF

### 3. EXPLORING & EXPOSING IMPLICIT ITEM CHARACTERISTICS

regression [57]. Finally, the outcomes are evaluated by assessing the performance against a baseline and by examining the distributions of the predicted factors.

The resulting multiple linear regression models comprise both motivational ratings and geographical attributes. After the variable selection 15 out of 26 motivational ratings and seven out of twelve geographical attributes in total are used. Table 3.4 summarizes the outcomes of the regression analysis. Motivational ratings *sightseeing*, *peacefulness*, *nightlife*, *culture*, *nature & landscape*, and *shopping* appear in more than one model. Also, geographical attributes *health resort* and *sea* are used in several models. In the following, the models are discussed in more detail.

Table 3.4: Results of the regression analysis conducted on the expert sample (N=350).

	F1	F2	F3	F4	F5	F6	F7
<i>(Intercept)</i>	0.41***	-0.18***	0.08	-0.09	0.28**	0.02	0.61***
<i>sightseeing</i>	-	1.02***	0.39***	0.30*	-0.29***	-	-0.27***
<i>nightlife</i>	-0.76***	-	-	-	-	0.41***	-0.57***
<i>peacefulness</i>	-	-	-	-	0.27**	-0.53***	0.51***
<i>health resort</i>	0.27***	-	-	-	-	-0.11***	0.09**
<i>sea</i>	0.23***	-0.12***	-	-	-	0.17***	-
<i>culture</i>	-	-	0.61***	0.47***	-	-	-
<i>nature &amp; landscape</i>	-	-	-0.07*	-0.27**	-	-	-
<i>shopping</i>	-	-	-	-	-	0.42***	-0.22**
<i>winter sports</i>	-	-0.24*	-	-	-	0.54***	-
<i>beach &amp; swimming</i>	0.73***	-	-	-	-	-	-
<i>family</i>	-	-	-	-	-	-0.33***	-
<i>hiking</i>	-	-	-	-	-	-	0.35***
<i>image &amp; flair</i>	-	-	-	0.48***	-	-	-
<i>kite &amp; windsurfing</i>	-	-	-	-	-	0.10***	-
<i>mobility</i>	-	0.26**	-	-	-	-	-
<i>sports</i>	-	-	-	-	0.85***	-	-
<i>wellness</i>	-	-	-	-	-0.31**	-	-
<i>metropolis</i>	-	-	-	-	-	0.19***	-
<i>mountains</i>	-	-	-	-	0.09***	-	-
<i>old town</i>	-	-	-	0.17***	-	-	-
<i>beach</i>	-	-	-	-	-	-	-0.05*

Note Significance level is coded as follows: \*\*\* $p < 0.001$ , \*\* $p < 0.01$ , \* $p < 0.05$

**F1 - Sun & Chill-Out.** The geographical attributes *sea*, *health resort*, and especially the motivational rating *beach & swim* have a significant positive impact on the factor *Sun & Chill-Out*. Those attributes can be interpreted as indicators for sun and relaxation. On the other side, the motivational rating *nightlife* has a significantly strong, negative impact, which can be associated with crowded places and mass tourism.

**F2 - Knowledge & Travel.** The motivational rating *sightseeing* has a significant, strongly positive relation with the factor *Knowledge & Travel*. Hence, it is capturing the knowledge part of the factor. Whereas, the motivational rating *mobility* is also positively related to the factor and one can say it captures the travel part. On the other hand, the geographical attributes *sea* and *winter sports resort* are significantly negatively related with the factor. This is reasonable, since in such areas usually the tourism focus does not lie on gaining knowledge.

**F3 - Independence & History.** The motivational ratings *culture* and *sightseeing* are significantly, positively related to the factor *Independence & History*. Those attributes can be seen as the main motivation of travellers with interests in history and tradition. Whereas, the motivational rating *nature & landscape* has a significant negative impact on the factor. Since cultural and historical interests are short coming in nature and recreation related destinations, such negative association is reasonable.

**F4 - Culture & Indulgence.** The motivational ratings *sightseeing*, *culture*, *image & flair*, and geographical attribute *old town* are significantly, positively related to the factor *Culture & Indulgence*. Those ratings can be interpreted as the main motivation of a culture and history interested high class tourist. On the other side, the motivational rating *nature & landscape* has a significant, negative impact on the factor. Again, this might show that destinations branded with a nature and landscape motif have shortcomings in cultural tourism.

**F5 - Social & Sports.** The motivational rating *sports* has a strong, significant, positive impact on the factor *Social & Sports*, which is obvious. Also, the motivational rating *peacefulness* and the geographical attribute *mountains* have a significant positive relation to the factor. Since the factor *Social & Sports* factor avoids crowded areas and locations of mass tourism, and prefers more tranquil places, positive associations of both attributes with the factor are reasonable. On the other hand, the motivational rating *sightseeing* has a significant, negative impact on the factor. It can be seen as an indicator of crowded areas and mass tourism. Surprisingly, the motivational rating *wellness* is significantly, negatively associated with the factor *Social & Sports*. This is caused by an unsound sample, as 55% of the destinations in the expert sample have a larger wellness rating (>0.5) and are located at the beach and 25% are metropolises, which is far less in the whole data set.

**F6 - Action & Fun.** The motivational ratings *peacefulness*, *family*, and the geographical attribute *health resort* have a significant, negative impact on the factor. This fits perfectly to the character traits of the factor *Action & Fun*. Whereas, the motivational ratings *nightlife*, *winter sports*, *shopping*, and the geographical attributes *sea* and *metropolis* are significantly positively related to the factor. Those can be interpreted as attributes of energetic, vibrant and action loaded places, which are main aspects of destinations for thrill seeking and action loving travellers.

**F7 - Nature & Recreation.** The motivational rating *peacefulness*, *hiking*, and the geographical attribute *health resort* are significantly positively related to the factor *Nature & Recreation*, which is obvious and does not need further explanation. On the other side, the motivational ratings *nightlife*, *sightseeing*, *shopping*, and the geographical attribute *beach* have a significant, negative impact on the factor. Those attributes can be interpreted as signs of mass tourism and crowded areas. Hence, a negative association on a recreational and escapist traveller is reasonable.

The resulting models are evaluated by assessing both in-sample and out-of-sample performance. In other words, the performance measures are determined using both training set (in-sample) and test set (out-of-sample). Obviously, out of sample performance plays a bigger role, because it delivers an approximation to the question, how the model will perform using unseen data. Still, in sample performance also provides some crucial insights. For example, it might give some hint, whether the developed models are overfitting. Furthermore, resulting linear regression models are compared to an appropriate baseline function  $f_0$  in order to show, whether the resulting models actually did learn something. Since the target variables, i.e. the Seven-Factors, are continuous, a simple mean function is chosen as baseline. Additionally, two more complex and non-linear models, namely KNN and RF, are fitted, in order to challenge the performance of the simple linear model.

In Table 3.5 the training and test performance of  $f_0$ , MLR, KNN, and RF are listed. Note that  $f_0$  is always a constant function. Thus, it does not explain any variance in the factor scores. Therefore,  $R_{train}^2$  and  $R_{test}^2$  of  $f_0$  are always zero. Training and test performance of the MLR models is close together, which shows that this model is not much overfitting. Whereas the RF models and especially the KNN models are overfitting the training set, i.e. the training performance is much better than the test performance. For example, an extreme case is the KNN model for factor *Action & Fun*, where  $R_{train}^2$  is 1.00 (100% of the variance in the factor is explained) and  $R_{test}^2$  is 0.63 and also  $RMSE_{train}$  is 0.01 (almost perfect) and  $RMSE_{test}$  is 0.21. Although both models are well tuned, the overfitting can be a sign of too few training data, but it also shows a potential for enhancement if more data is used.

Overall, the out-of-sample performance of all three models MLR, KNN, and RF are pretty close. Hence, one can expect that they will perform similar if confronted with unseen data. The overall performance of all three models (MLR, KNN, RF) are always better than the simple mean function  $f_0$ , which indicates that the models must have learned something out of the data. The difference is in most cases clear to observe, except in factor *Social & Sports*. Here, the  $RMSE_{test}$  of  $f_0$  is 0.19 and MLR, KNN, and RF have a  $RMSE_{test}$  of 0.17-0.18. This is caused by an uneven distribution of the expert mapping, where 87% of the destinations have scored with 0.5 or 0.75. Hence, a constant prediction of 0.58, like  $f_0$  does, is performing pretty well, but it also means that there is less information to learn from. On the other hand, the models are performing the best in factor *Nature & Recreation*, where  $RMSE_{test}$  is 50% smaller than the baseline. The

Table 3.5: Comparison of performance measures of baseline function ( $f_0$ ), MLR, KNN regression, and RF regression.

		$f_0$	MLR	KNN	RF
<i>Sun &amp; Chill-Out</i>	$R_{train}^2$	0.00	0.68	0.78	0.94
	$R_{test}^2$	0.00	0.62	0.61	0.64
	$RMSE_{train}$	0.40	0.23	0.19	0.10
	$RMSE_{test}$	0.40	0.25	0.25	0.24
<i>Knowledge &amp; Travel</i>	$R_{train}^2$	0.00	0.72	0.62	0.85
	$R_{test}^2$	0.00	0.71	0.64	0.70
	$RMSE_{train}$	0.32	0.17	0.20	0.12
	$RMSE_{test}$	0.33	0.18	0.20	0.18
<i>Independence &amp; History</i>	$R_{train}^2$	0.00	0.65	0.46	0.71
	$R_{test}^2$	0.00	0.59	0.58	0.62
	$RMSE_{train}$	0.27	0.17	0.20	0.14
	$RMSE_{test}$	0.28	0.17	0.18	0.17
<i>Culture &amp; Indulgence</i>	$R_{train}^2$	0.00	0.69	0.99	0.79
	$R_{test}^2$	0.00	0.61	0.58	0.67
	$RMSE_{train}$	0.33	0.20	0.03	0.15
	$RMSE_{test}$	0.35	0.21	0.22	0.20
<i>Social &amp; Sports</i>	$R_{train}^2$	0.00	0.28	0.22	0.54
	$R_{test}^2$	0.00	0.22	0.06	0.16
	$RMSE_{train}$	0.18	0.15	0.16	0.12
	$RMSE_{test}$	0.19	0.17	0.18	0.17
<i>Action &amp; Fun</i>	$R_{train}^2$	0.00	0.73	1.00	0.88
	$R_{test}^2$	0.00	0.68	0.63	0.70
	$RMSE_{train}$	0.35	0.18	0.01	0.12
	$RMSE_{test}$	0.36	0.20	0.21	0.19
<i>Nature &amp; Recreation</i>	$R_{train}^2$	0.00	0.80	1.00	0.92
	$R_{test}^2$	0.00	0.77	0.69	0.75
	$RMSE_{train}$	0.33	0.15	0.02	0.10
	$RMSE_{test}$	0.34	0.17	0.19	0.17

*Note* All models have been applied on the expert sample (80% for training and 20% for testing)

out of sample performance of the KNN model is always a tick worse than the MLR and RF model, whereas there is almost no difference in the performance of RF and MLR. Thus, discarding the KNN model and choosing the MLR model over RF is reasonable since they are performing similar but the MLR model is much simpler to fit and easier to interpret.

MLR delivers promising results by explaining 59-77% of the variance of factor scores in the test set. Except the model for *Social & Sports*, where only 22% of the variance is explained. Such poor performance is caused by an uneven distribution of the factor scores in the expert mapping.

In contrast to the previous analysis, where the focus is predictive performance, now the distribution of the predicted factor scores is analysed. In detail, the factor score distribution of the expert mapping is compared to the distribution behaviour of the predicted factor scores. In order to do so, the MLR model of each factor is fed with the complete data set as input. Then the resulting distribution in factor scores is compared to the one in the expert mapping, which is shown in Fig. 3.7. This comparison will foster a better understanding of the generalization power of the developed models.

**Sun & Chill-Out.** Here, 49% of the destinations in the complete set are scoring with 0.25. This is not observable in the expert sample. The expert sample shows an increased amount of destinations with score 0 (30%) and 1 (27%), whereas 43% of destinations score either with 0.25, 0.5 or 0.75. A similar but damped behaviour can be observed in the predicted factor scores of the complete set (setting aside the peak at score 0.25).

**Knowledge & Travel.** Taking into account the expert sample, the majority of destinations score either with 0 or 0.25 and with increasing factor score the amount of destinations decays. A similar behaviour can be observed in the predicted factor scores of the complete set.

**Independence & History.** Considering the predicted factors of the complete set, once again one can see a peak at score 0.25 like previously in *Sun & Chill-Out*. Besides that, the distribution has more or less a normal shape (bell), similar to the factor score distribution of the expert mapping.

**Culture & Indulgence.** Looking at the predicted factors of the complete set, there is again a peak at score 0.25 (57%), which is not observable in the expert mapping. At score 0.5 and 0.75 the percentage of destinations in the expert sample (28% and 15%) are relatively close to the ones in the complete set (23% and 11%). On the other hand, this is not the case for scores 0 or 1, where the difference is much higher.

**Social & Sports.** The vast majority of destinations score either with 0.5 or 0.75 in both, whereas only few destinations score with 0, 0.25 or 1.

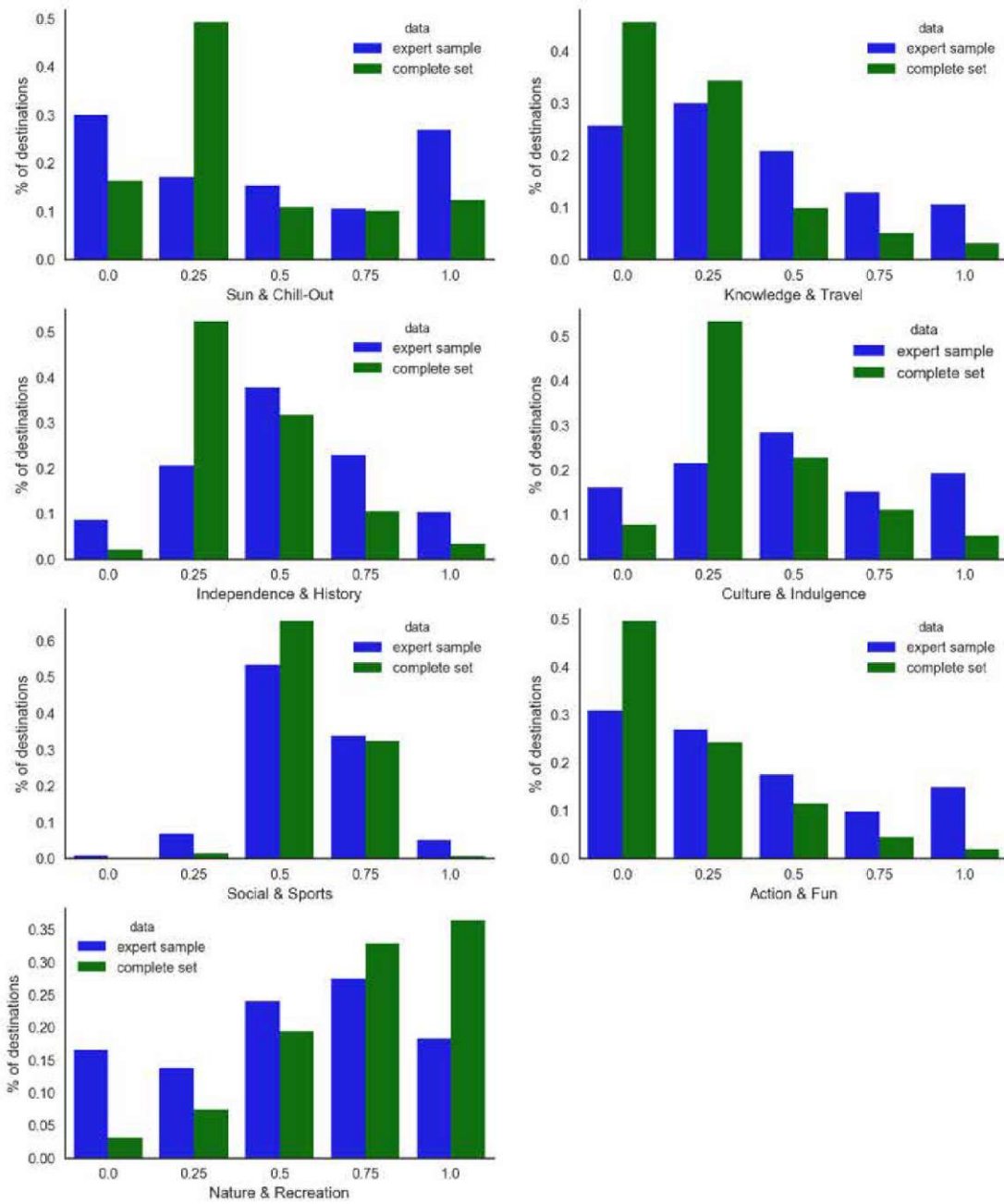


Figure 3.7: A comparison of factor score distributions in the expert sample versus the distributions of predicted factor scores of the complete data set.

**Action & Fun.** Considering the expert mapping, one can see that the majority of destination score either with 0 or 0.25 and that this amount is decaying the higher the score gets. A similar behaviour can be observed by looking at the predicted factors of the complete set.

**Nature & Recreation.** In the predicted scores of the complete set the amount of destinations is increasing with increasing factor scores. This cannot be observed in the expert mapping. Still, in both, the expert sample and the complete set, most of the destinations are scoring with 0.5 or more.

To sum up, there are some differences in the distributions of factor scores between the manually labelled expert sample and the predicted scores of the complete set. But overall, both show similar trends in the distributions. This shows that the build multiple linear regression models are mimicking the experts quite good. Hence, one can expect a sufficient generalization.

#### 3.1.5 Discussion

In general, one can distinguish between knowledge-poor recommendation techniques (i.e., only basic data like user ratings for items in use) and knowledge-dependent recommendation techniques (i.e., ontological description of users and items, constraints, or social relations and activities of users in use) [88]. The picture based approach [76, 77], where both users and items are described through the Seven-Factor Model, belongs to the latter group. Primarily, this work's aim is to identify and explain associations between destination attributes and the Seven-Factor Model to enable an automated mapping of destinations onto the Seven-Factors. Ultimately, such a mapping should enable a match-making between users and destinations. To reach the stated goals an exploratory data analysis, a cluster analysis, and a regression analysis were conducted.

Since the overall aim is not only to project destinations into the seven-dimensional vector space of travel behavioural patterns, but also to explain which attributes of destinations are more important for this purpose, a MLR analysis with step wise variable selection was conducted. Seven models were established, one for each factor of the Seven-Factor Model. The resulting models are providing strong evidence that there is a significant relation between selected destination attributes and the Seven-Factors. Table 3.6 lists all independent variables (destination attributes) of the fitted MLR models of the Seven-Factors. Note that a minus sign indicates a negative impact on the corresponding factor. For example, the model of the factor *Sun & Chill-Out* consists out of indicators of sun and beach as expected, but there are also indicators of crowdedness, which have a negative impact on the factor. Such model structure is in line with the characteristics of *Sun & Chill-Out*, where crowdedness and mass tourism are negatively associated with the factor.

Furthermore, the developed linear models were challenged by two conceptually different non-linear models, namely RF regression and KNN regression. Additionally, the predictive performances of all three models MLR, RF, and KNN were compared to a baseline function

Table 3.6: Used destination attributes in the resulting MLR models.

Factor	MLR coefficients
<i>Sun &amp; Chill-Out</i>	- <i>nightlife, beach &amp; swimming, health resort, sea</i>
<i>Knowledge &amp; Travel</i>	<i>sightseeing, mobility, - winter sports resort, - sea</i>
<i>Independence &amp; History</i>	<i>culture, sightseeing, - nature &amp; landscape</i>
<i>Culture &amp; Indulgence</i>	<i>image &amp; flair, culture, sightseeing, - nature &amp; landscape, old town</i>
<i>Social &amp; Sports</i>	<i>sports, - wellness, - sightseeing, peacefulness, mountains</i>
<i>Action &amp; Fun</i>	<i>winter sports, - peacefulness, shopping, nightlife, - family, metropolis, sea, - health resort, kite &amp; windsurfing</i>
<i>Nature &amp; Recreation</i>	- <i>nightlife, peacefulness, hiking, - sightseeing, - shopping, health resort, - beach</i>

*Note* A minus sign indicates a negative impact on the corresponding factor

(in this case simple mean of each factor). The evaluation showed that all three models were always better than the baseline function, which indicated that they had learned something out of the data. Although the RF models were excelling in the training phase of the models their performance in the test phase dropped drastically. Thus, the RF models were clearly overfitting the training data. Also, fine-tuning and regularization did not bring any performance gain. Thus, the big difference in training and test performance might be caused by too small samples. Consequently, bigger samples will be aimed in the follow up studies. However, it has been demonstrated that the performance of the MLR model is similar to the performance of the RF model and both are outperforming the KNN model. In the end the MLR model was chosen over the RF model since MLR is simpler to fit and easier to interpret than RF.

Overall, all travel behavioural patterns are well described (59-77% of the variance) by the resulting models, except for the factor *Social & Sports*, where only 22% of the variance can be explained. This is caused by an uneven distribution of scores of the factor *Social & Sports* in the expert sample. Thus, statistically sounder samples will be targeted in future work. Furthermore, the fitted MLR models show a RMSE between 0.17 and 0.25 (in a scale of 0 to 1). The interpretation of a RMSE value is totally domain dependent. Unfortunately, there is no reference to look at or a rule of thumb to follow, which indicates whether such RMSE is sufficient enough or not. For example, in critical decisions like in medicine a RMSE of 0.01 (in a scale of 0 to 1) might still be too much. But, for simple recommendation items like for example movies in the Netflix prize case [8] a RMSE of about 0.9 (in a scale of 1 to 5) might be sufficient. Tourism products are usually more complex and more expensive than movies and thus one might need a lower RMSE than about 20% of the scale. Note that the models here are trained and tested on a small dataset evaluated by experts, which is substantially different to the setting of the Netflix

prize. Ultimately, the decision whether the fitted models are applicable or not has to be evaluated within a RSs and will be addressed in follow up studies.

As already mentioned, the MLR models were obtained by using just a small subset of the given data set (i.e., 2%). In order to get a better understanding of the generalization performance of the fitted models with respect to the whole data set the Seven-Factor score distribution of the expert sample was compared to the predicted Seven-Factor score distribution of the whole data set. There was some discrepancy between both distributions but overall, they showed similar trends and thus one can expect a sufficient generalization within the given data set. Most differences were experienced at a score of 0 and 0.25, which might indicate the difficulty in differentiating between “bad” and “very bad” (i.e., between 0.25 and 0). A finer scale than the currently used one (i.e., 0-0.25-0.50-0.75-1) might counter this issue.

Anyway, one should keep in mind that the fitted models are based on a proprietary data set. Although the introduced approach can easily be replicated with other datasets, it brings the disadvantage that for each new data set new models have to be trained and tested (since commonly data sets in the tourism landscape are heterogeneous and not standardized). To counter this generalization issue, a comprehensive tourism product model (partly based on existing ontological representations) will be developed in order to harmonize heterogeneous data sources. Afterwards, analyses conducted in this study will be applied on the harmonized data to obtain a more general solution. First steps in this direction have already been taken.

In addition to the regression analysis (i.e., a supervised learning method, where pre-labelled data is needed) a cluster analysis (i.e., an unsupervised learning method) was conducted. The goal was to find latent, conceptually meaningful structures within the given data set without the need of prior expert knowledge. Six conceptually meaningful clusters were identified, namely *vibrant beach resorts*, *energetic cities*, *tranquil seaside resorts*, *peaceful towns*, *idyllic island villages*, and *ordinary towns*.

The resulting cluster solution is supported by the exploratory data analysis, where destination attributes could be group based on their pairwise correlation (see Fig. 3.3). The identified groups of attributes are covering following aspects of tourism destinations: *recreational*, *island*, *countryside*, *urban area*, *mass tourism*, and *seaside*. Considering both the six clusters and the groups of destination attributes one can observe a clear contrast of *vibrant* to *tranquil*, *land* to *island*, *seaside* to *inland (urban area)*. Also, the expert sample supports to some extent the six cluster solution, where a reasonable average factor score distribution over the clusters (see Table 3.3) could be observed. Only in case of the factor *Social & Sports* no clear associations with the clusters could be drawn. The ambiguity of *Social & Sports* is due to an uneven distribution of factor scores in the expert mapping, where 87% of the destinations scored either with 0.5 or 0.75.

The benefit of the cluster solution is that it works totally unsupervised. In other words, clusters can be obtained without the need of a previous manual mapping of experts and only by considering the given data set. Further, the resulting clusters

could be addressed directly by RSs, where they can be used as an initial approach for recommending destinations if no further data is available. But the biggest limitation of the clustering approach is mutual exclusivity, i.e. destinations are members of only one cluster. For example, Rio De Janeiro is a huge city offering many cultural, historical, nightlife and entertainment activities, but it is also located at the seaside with famous beaches like Copacabana and Ipanema. Therefore, assigning Rio to just one cluster for example *energetic cities* would not consider customers interested in beach and swimming adequately. Furthermore, people tend to have a combination of travel preferences rather than just one interest [47, 76, 77], where an explicit assignment would have shortcomings.

Finally, the exploratory data analysis and also the developed models showed that the used data set is not able to cover all characteristic aspects of the factors of the Seven-Factor Model. Especially, there were no features indicating independence, the passion for knowledge gain, indulgence, or socialization with locals. This is also pointed out by Glatzer et al. [43], where they argue that this shortcoming might be the cause of some performance loss in their models. Other data sources might be able to cover the characteristic aspects of the factors of the Seven-Factor Model better and will be used to enhance the models.

### 3.1.6 Conclusions

This work embarked on a journey to discern and quantify the associations between tourism destination attributes and the Seven-Factor Model, with the ultimate aim of automated mapping for enhanced match-making between users and destinations. Through extensive analysis, we addressed two main research questions:

**RQ1.1.1** *Is there a latent underlying structure that aligns with the Seven-Factor Model?*

The results of the exploratory data analysis and cluster analysis are affirmative. The data revealed six conceptually meaningful clusters encompassing varied aspects of tourism destinations. However, the mutual exclusivity of clusters presents challenges in adequately representing destinations with diverse attributes.

**RQ1.1.2** *Can we devise a mechanism to automatically map tourism destinations to the Seven-Factor Model considering their attributes?*

MLR models, built for each of the seven factors, demonstrated a significant association between selected destination attributes and the Seven-Factor Model. A testament to their efficacy, these models outperformed non-linear counterparts (KNN and RF) in predictive accuracy and interpretability, although certain limitations with the data and the chosen model approach were evident.

The primary contributions of the work lie in its empirical insights. The findings, as tabulated in Table 3.6, underline the strengths and directions of the relationship between destination attributes and the Seven-Factor Model. The introduced models, albeit built on a limited dataset, have demonstrated potential for scalability and applicability in a real-world RSs setting.

The current research, while pivotal, is not without its challenges. The data's inherent limitations and the expert sample's constraints underpin most of the observed shortcomings. Looking ahead, future endeavors will pivot towards harnessing more comprehensive and heterogeneous data sources to enhance model robustness and applicability. Evaluating the proposed models within an operational RSs and expanding the scope to aggregate tourism products for better recommendation outcomes are also on the horizon.

In essence, this study has laid down a promising foundation, pushing the boundaries of knowledge-dependent recommendation techniques in the realm of tourism. However, there is still room for further improvement in the development of a more comprehensive destination RSs.

#### 3.1.7 Limitations

While this research offers valuable insights, several limitations stemming from the data and scope must be acknowledged. The original dataset, provided by a German e-Tourism company, contained a substantial number of missing values, necessitating imputation techniques that, while necessary, may have introduced some uncertainty. Furthermore, the dataset's inherent focus on certain geographical regions and its reliance on a specific set of attributes (some with subjective motivational ratings) limit the universal applicability of our findings. The expert sample used for model training and validation, though comprised of experienced professionals, was relatively small and geographically skewed, potentially impacting the generalizability of the models, particularly for the *Social & Sports* factor, which showed uneven score distribution. The choice of multiple linear regression, while beneficial for interpretability, assumed linear relationships between attributes and factors, potentially overlooking more complex interactions that non-linear models might have captured, had they not been prone to overfitting given the dataset's size. Finally, the cluster analysis, while revealing meaningful groupings, imposed a mutual exclusivity on destinations that does not fully represent the multifaceted nature of many tourism locations.

Addressing these limitations will be crucial for future research. This includes exploring alternative data sources, expanding the expert sample, employing more sophisticated machine learning techniques, and developing more nuanced approaches to clustering and recommendation.

## 3.2 A Multi-Level View on News Articles

The digital revolution, fueled by the WWW, has significantly transformed news consumption, shifting from traditional print media to online platforms, as evidenced by rising statistics of digital news access [78]. This shift offers readers a vast array of choices and tailored experiences but also leads to information overload and consumer disorientation. To address this, NRS have emerged, aiming to connect readers with content that matches their preferences and interests [62]. However, the transient nature of news, with its

constantly evolving items, presents a *cold-start* problem for NRS, highlighting the importance of content-based techniques that build user profiles based on their engagement. In response to the need for effective NRS, this chapter explores the latent characteristics of news articles, leading to two pivotal research questions.

**RQ1.2.1** *What latent characteristics, which are not immediately evident, can be exposed within news articles when analyzed at different levels (document, topic, and author levels)?*

Document-level analysis excels in extracting objective knowledge and identifying discriminative terms, influenced by the document type. Topic-level analysis, done in collaboration with *der FALTER*, identifies latent themes aligning with document-level insights. Author-level analysis uncovers unique writing styles of authors with approximately 97% accuracy.

**RQ1.2.2** *How can these identified latent characteristics be harnessed in a news recommender system setting, and what implications might they have on the diversity of recommendations?*

Document-level analysis enables focused recommendations based on cosine similarity between *text-frequency inverse-document-frequency* (TF-IDF) vectors. Topic-level analysis adds thematic diversity by offering content-unified items with varied content. Author-level analysis allows for serendipitous recommendations based on writing style, unbounded by document content.

The overarching goal is to establish a foundation for a comprehensive NRS. Such a system, rather than being solely accuracy-driven, would emphasize novelty, diversity, and serendipity, ensuring a more holistic and enriching user experience.

Our contributions can be summarized as follows:

- We conduct a detailed multi-level analysis on news articles, illuminating latent characteristics across document, topic, and author levels.
- We demonstrate the potential of these latent features in the realm of NRS, elucidating their implications on recommendation diversity.
- We employ an integrated approach, synergizing insights from all three levels to proffer comprehensive guidelines for the development of future RSs.

### 3.2.1 Related Work

RSs have been studied extensively, with the main goal of assisting users to discover items they might be interested in from a vast sea of choices. Among the several methods of recommendations, collaborative filtering has seen extensive use due to its domain-independent nature and vast availability of datasets [60, 62]. Nonetheless, NRS are more inclined towards content-based approaches or hybrids of these with other methods [62].

The distinction in the choice of techniques for NRS arises primarily due to the inherent nature of news. The permanent *cold-start* problem, a consequence of the ever-fresh content in the news domain, poses challenges for systems relying on historical user-item

interactions. The textual nature of news articles also steers the inclination towards techniques from data mining, especially text mining and information retrieval [62]. These techniques, as described by [108], allow both the extraction of useful patterns and predictive modeling from vast repositories of data.

Moreover, mining textual data is closely aligned with the realms of text mining and information retrieval. While information retrieval is essential for connecting users with the right content, text mining provides tools for in-depth analysis, helping users digest and make decisions based on the content [1]. The specificity of news articles, authored predominantly by professionals adhering to journalistic standards and societal consensus, imposes distinct challenges in text mining [10]. These challenges span various facets, from dealing with thematic diversity to recognizing story redundancies.

Although well-established text mining and information retrieval techniques have been applied to NRS, there remains room for a more comprehensive approach that considers multiple aspects of a document simultaneously. The unique focus of this work is on the concurrent analysis of document-level, topic-level, and author-level in the context of NRS. While numerous studies have focused on the accuracy of recommendations, there has been a noticeable lack in research emphasizing diversity and serendipity in NRS, both of which are addressed in this work [62].

#### 3.2.2 Methods

##### Source of Data

The primary source of data for this research is *der FALTER*, a weekly newspaper founded in 1977, headquartered in Vienna. Over the years, it has earned a reputation for providing comprehensive coverage on various subjects, ranging from politics and media to culture and urban life. Additionally, their digital offerings, which span from restaurant reviews to event listings, are a treasure trove of information.

Our study zeroes in on two specific text corpora, each presented in CSV file format. The first corpus, referred to as the *Thurnher corpus*, is a collection of comments written by Armin Thurnher – one of the founders of *der FALTER* – on contemporary issues. The second, named the *Klenk corpus*, is composed of articles by Florian Klenk that delve into intricate narratives and story arcs. Both individuals currently serve as Editors-in-Chief for *der FALTER*.

From the *Klenk corpus*, a total of 360 articles were extracted, while the *Thurnher corpus* contributed 216 comments. All these pieces were published between June 2013 and June 2018. The subsequent distribution of these articles on a year-by-year basis is illustrated in Figure 3.8. For context, the average word count stands at 1055 for the *Klenk corpus* and 975 for the *Thurnher corpus*.

##### Analysis Framework

For methodological clarity, our analysis approach is divided into three primary dimensions:

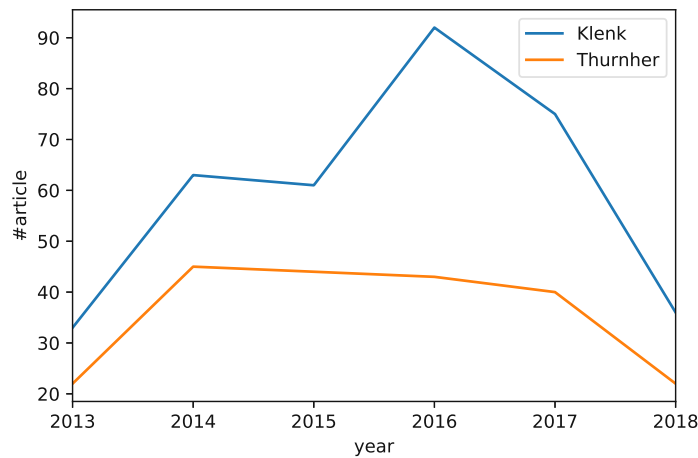


Figure 3.8: Yearly article distribution from June 2013 to June 2018.

- **Document-level Analysis:** This level is centered on identifying crucial terms within each document. By converting these documents into vectors, we facilitate comparisons between them. The mechanisms underpinning this procedure, involving *text-frequency inverse-document-frequency* (TF-IDF), are elaborated in Section 3.2.2.
- **Topic-level Analysis:** Here, our focus shifts to ascertaining latent themes dispersed across the content. By recognizing these themes, we can categorize documents based on topical relevance. Techniques like topic modeling and co-occurrence networks, which are integral to this dimension, are expanded upon in Sections 3.2.2 and 3.2.2 respectively.
- **Author-level Analysis:** This dimension seeks to distinguish between authors based on their unique writing styles—a method known as stylometry. The nuances of this technique are discussed further in Section 3.2.2.

### Data Preparation

Before diving into the main analyses, it is imperative to preprocess both corpora. This stage is essential to transform the raw text data into a structured format conducive for subsequent operations. Our preprocessing routine involves the following steps:

1. **Lowercase Transformation:** All letters are converted to lowercase.
2. **Cleaning:** Punctuation, special characters, and numbers are removed.
3. **Tokenization:** Each document is split into words (i.e., tokens). This step is fundamental for document vector generation (i.e., quantified document), as the frequency of token occurrences forms the basis of such a vector [112].

4. **Lemmatization:** Each word (i.e., token) is transformed to its lexical form, e.g., *transformed* to *transform*, *corpora* to *corpus*, etc.
5. **Stopword Filtering:** Words that almost never have any predictive capability, such as articles (e.g., *a* and *the*) and pronouns (e.g., *it* and *they*) [112], are removed.

Apart from processing individual words, this research also gives importance to bi-grams and tri-grams—combinations of two and three words, respectively. This ensures that terms like *Natural Language Processing*, which lose their meaning when split, are treated as unique entities.

Further, the datasets are enriched with meta-information. This includes implementing Named Entity Recognition (NER) to identify and tag entities such as people, places, and organizations present within each document. Moreover, the structure of sentences is retained, which proves beneficial for constructing co-occurrence networks and performing topic modeling. Using Part of Speech (POS) tagging, we also categorize and store the grammatical role of every token.

#### Document Vectors and Similarity Measures

To conduct document-level analysis, tokens (including bi-grams and tri-grams) are utilized to create a document-term matrix. In this matrix, every row represents a document vector. The individual elements of this vector are the *TF-IDF* scores of terms within the document. By emphasizing terms that are pertinent to a particular document and devaluing those that are omnipresent across the corpus, *TF-IDF* effectively highlights a term's importance. Using the cosine similarity measure, we can then quantify the degree of likeness between documents.

#### Uncovering Latent Topics

Topic modeling is a technique grounded in probabilistic mathematics, aiming to unveil hidden thematic structures in a collection of documents. It assumes that each topic is a distribution over terms, while each document is a mix of topics. Our research employs the Latent Dirichlet Allocation (LDA) model for this purpose. By primarily focusing on nouns and proper nouns, we can achieve more insightful results. The number of topics is optimized using the  $C_V$  coherence measure, and the *mallet* library is integrated with the *gensim* LDA-wrapper for enhanced performance.

#### Entity Relations: Co-Occurrence Networks of Named Entities

These networks are instrumental in deciphering the underlying relationships between various entities present within a text. We identify entities by employing NER [106], and in our case, named entities are Organizations, Locations, or Persons. If two entities appear within the same context (typically a sentence), an edge is formed between them. The strength or weight of this edge mirrors the frequency of their co-occurrence, providing



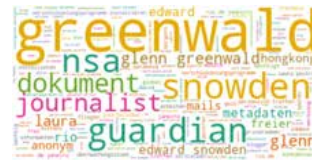
(a) Klenk Corpus



(b) Turnher Corpus



(c) Das Überwachte Wien



(d) Sind Sie ein freier Mann, Mister Greenwald?



(e) Klenk Corpus 2015



(f) Klenk Corpus 2018

Figure 3.9: Word clouds representing significant words based on their  $TF-IDF$  values.

insights into how strongly the entities are interrelated. By deploying network metrics, such as degreeG

### Stylometry: Detecting Authorial Signature

Stylometry is the analytical study of literary style, aiming to discern patterns unique to individual authors. By evaluating various textual attributes like the average sentence length, choice of words, and even the distribution of punctuation marks, it is possible to derive an “authorial signature”. For this research, we consider the 1000 most frequently used words and other indicative markers to define this signature. Subsequently, machine learning models are trained on these markers, offering insights into the distinct stylistic nuances of different authors.

### 3.2.3 Analysis Outcomes

#### TF-IDF Word Clouds: Document Analysis

Figure 3.9 displays word clouds derived from significant  $TF-IDF$  scores in various documents. The font size indicates the prominence of a word’s score: the larger the size, the higher its  $TF-IDF$  value.

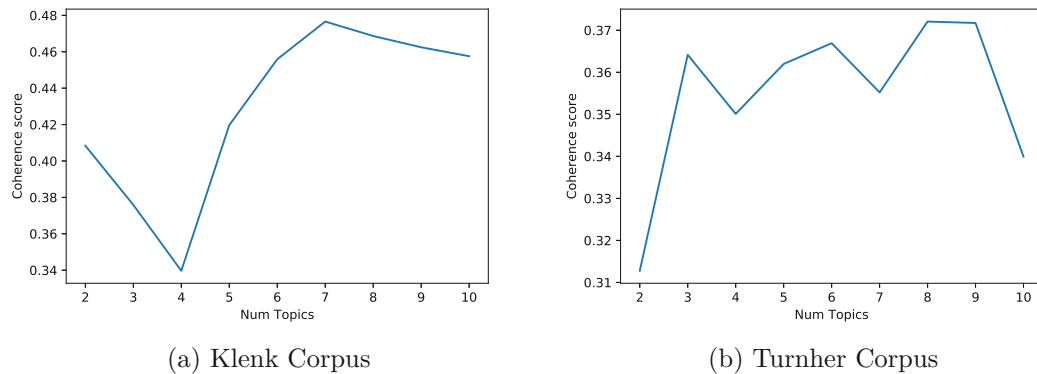


Figure 3.10: Coherence scores of LDA models over varying topic counts.

The *Klenk corpus*, represented in Figure 3.9a, emphasizes individual protagonists, while the *Thurnher corpus* in Figure 3.9b spotlights organizations and broader concepts. This distinction reflects the individual story-focused nature of the *Klenk corpus* and the event commentary orientation of the *Thurnher corpus*.

For a focused look, Figure 3.9c highlights an article from the *Klenk corpus* about surveillance in Vienna. A similar article is shown in Figure 3.9d, displaying significant overlap in keywords, affirming its relevance despite a low cosine-similarity score.

Moreover, year-wise aggregated *TF-IDF* scores show evolving themes in the *Klenk corpus*, shifting from the refugee crisis in 2015 to a domestic intelligence controversy in 2018 (see Figures 3.9e and 3.9f).

### Topic Analysis using LDA

The ideal topic number for LDA modeling must be pre-specified. Figure 3.10 gauges the coherence scores (indicating how well topics resonate with the corpus content) for different topic counts. Best fits were found with seven topics for the *Klenk corpus* and eight for the *Thurnher corpus*.

Tables 3.7 and 3.8 encapsulate the derived topics. These were further fine-tuned during a workshop with *der FALTER*.

Both corpora share topics like *FPÖ*, *Austrian Media*, and *Refugee Crisis*. Yet, their interpretation varies: the *Klenk corpus* takes a localized stance, while the *Thurnher corpus* adopts a broader, more international perspective.

### Exploring Named Entity Co-occurrence Networks

Co-occurrence networks in this research pivot around Named Entities—persons, locations, or organizations—that appear together in sentences. Defined during the data preparation phase using NER, nodes represent these entities, while edges signify their

Table 3.7: 7 Topic LDA Solution of the *Klenk Corpus*

Topic	Top 10 terms	Label
1	fpö, wahl, strache, kickl, bvt, leute, partei, innenminister, grund, sache	FPÖ
2	journalist, medium, falter, geschichte, pilz, politiker, datum, frau, orf	Austrian Media
3	österreich, flüchtling, spö, kern, land, eu, grüne, övp, partei, pesendorfer	Refugee Crisis
4	kind, stadt, frau, eltern, leute, lorenz, schulen, problem, vater, aslan	Family & Education
5	polizei, polizist, frau, fall, video, mann, hand, beamte, strasser, josef	Police
6	euro, grasser, fall, staatsanwaltschaft, falter, million_euro, meischberger, ermittler, pröll, konto	BUWOG Affair
7	beamte, fall, insasse, falter, häftling, gefängnis, brandstetter, haft, arzt, opfer	Police Violence & Prison Conditions

concurrent appearance. The node size mirrors its betweenness centrality [39], indicating its prominence in the network.

Observing Figure 3.11a, one notes the prominence of nodes like *Vienna* alongside major Austrian political entities such as *ÖVP*, *FPÖ*, *SPÖ*. Furthermore, the representation empowers an in-depth examination of entity relationships. An illustrative instance would be the blue-circled nodes, all municipalities in Lower Austria, and their connection to the red-circled node, *Erwin Pröll*, the preceding governor of Lower Austria.

Contrastingly, Figure 3.11b displays the Thurnher corpus co-occurrence network. Here, *Austria* and *Europa* claim dominance. Other notable entities like *Germany*, *USA*, *Donald Trump* grace the upper section, while domestic political players populate the lower half. These patterns accentuate the varying perspectives captured within each corpus—encompassing broad international commentaries and narrower domestic narratives.

Further deep-dives explored networks solely centered on individuals, offering insights into protagonists' interrelations. To encapsulate the evolution of these relations, separate networks were constructed for varying years. Figure 3.12 showcases such networks derived from *Klenk articles* in 2015 and 2018. A notable observation is the shifting prominence from *Heinz Christian Strache* in 2015 to *Sebastian Kurz* in 2018. Additionally, the network alludes to topical events. The 2018 network, for instance, highlights main players of the *BVT Affair* like *Herbert Kickl* and *Peter Gridling*.

Table 3.8: 8 Topic LDA Solution of the *Thurnher Corpus*

Topic	Top 10 terms	Label
1	fpö, wahl, kandidat, amt, republik, öffentlichkeit, profil, falter, fall, kickl	FPÖ
2	eu, europa, linke, politik, krise, regierung, idee, deutschland, grenze, land	Refugee Crisis
3	österreich, land, krone, bild, seite, krone_zeitung, million, spielen, jeannée, art	Austrian Media
4	medium, orf, öffentlichkeit, trump, social_media, facebook, demokratie, art, debatte, interesse	(Social)Media & Democracy
5	staat, stadt, welt, markt, kapitalismus, land, wort, banken, finanzminister, folge	Economy
6	politik, partei, övp, spö, kanzler, sebastian, grüne, strache, politker, österreich	Politics
7	demokratie, usa, staat, präsident, freiheit, krieg, kraf, türkei, weise, menschenrechte	Democracy & Human Rights
8	falter, zeitung, satz, journalist, gesellschaft, seite, fall, aufklärung, autor, frau	Falter

### Exploring Authorial Style Through Stylometry

Stylometry, often referred to as the study of linguistic style, seeks to discern the distinctive writing styles of authors. In this analysis, our goal is to distinguish between the writings of Klenk and Thurnher, even when their articles are merged into a single corpus (combining both *Klenk corpus* and *Thurnher corpus*).

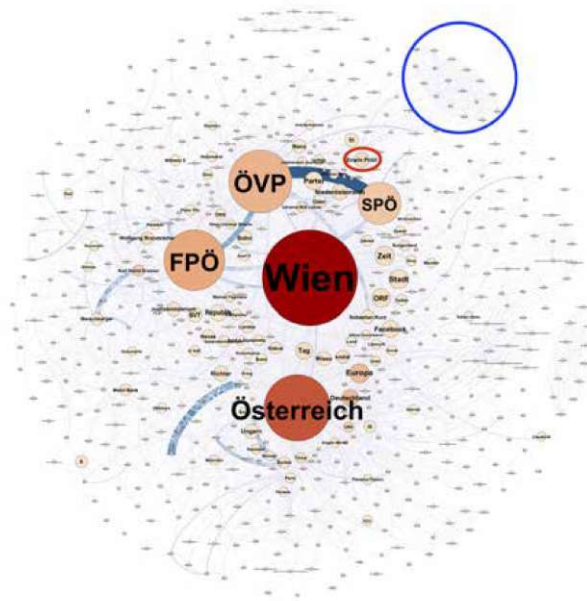
To achieve this distinction, we assess the Most Frequent Words (MFW) in the corpus. The premise is that authors have unique patterns of word usage, serving as a “fingerprint” of their individual writing style. By examining these patterns, we can compute the distance between articles using Burrows’s Delta metric, a measure of stylistic difference.

As illustrated in Figure 3.13, clustering was performed based on the top 200 MFW without any form of regularization, known as culling. The resulting clusters were impressively accurate, with a mere 2.95% (or 17 out of 576) of the documents being misclassified.

#### 3.2.4 Discussion and Implications

In their seminal work, Daelemans et al. [24] categorize the types of knowledge that can be gleaned from textual data into three main classes:

- *Objective knowledge*, which involves straightforward factual information answering

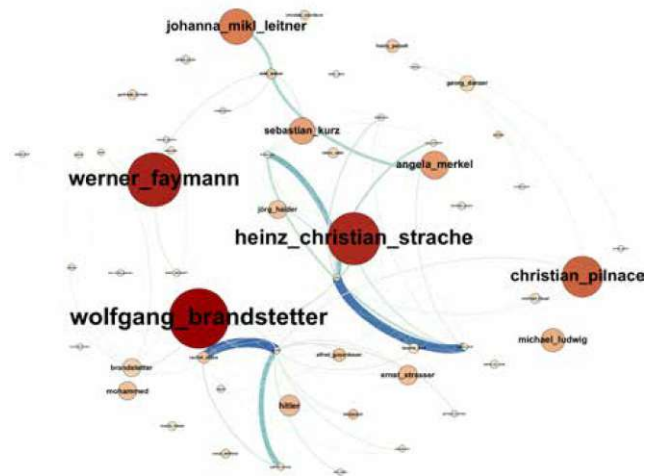


(a) Klenk Corpus

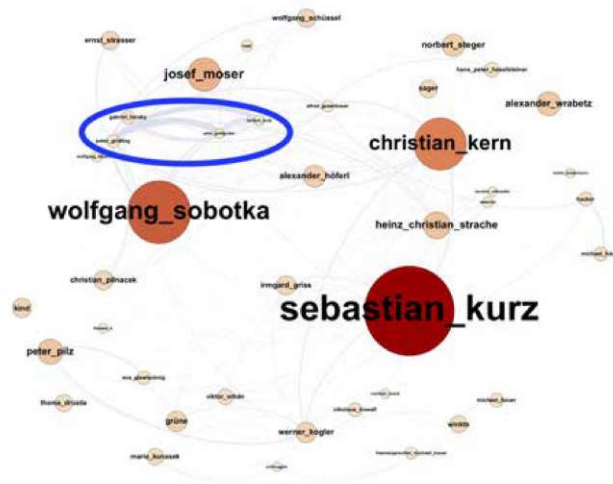


(b) Turnher Corpus

Figure 3.11: Co-occurrence networks centered on Named Entities across different corpora.



(a) 2015



(b) 2018

Figure 3.12: Temporal person-centric co-occurrence networks from Klenk articles in 2015 and 2018.

questions such as who did what, when, and where;

- *Subjective knowledge*, which addresses the sentiment or opinions in the text, encapsulated in the question “Who holds what opinion about a particular subject?”;
- *Metaknowledge*, a meta-level understanding that is not tied to the specific content of the text but could include factors like authorship style.

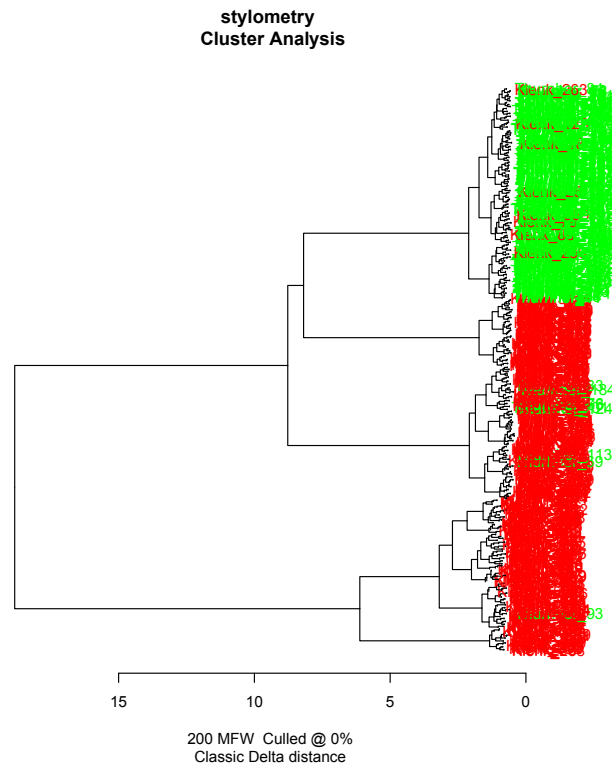


Figure 3.13: Document clustering based on the 200 most frequently used words without culling. Klenk’s documents are marked in red, while Thurnher’s are in green.

Our study primarily focuses on extracting objective knowledge during both the document-level and topic-level analyses. However, the author-level analysis is an exception; it aims at understanding metaknowledge, particularly focusing on discerning the unique writing styles of authors rather than the content of their documents.

In the document-level analysis, each document was transformed into a sparse vector representation based on the frequency of terms it contains. We then calculated *TF-IDF* (Term Frequency-Inverse Document Frequency) scores to highlight terms with the most discriminative power. This allowed us to use cosine similarity for document comparison. When used in a NRS, this approach might yield recommendations that are closely related or similar, as conceptually represented in Figure 3.14.

The document-level analysis also revealed that the nature of the document significantly impacts the keywords that get extracted. For example, narrative articles—those with a storyline—usually contained keywords associated with specific individuals. In contrast, opinion pieces or commentaries focused more on broader topics like EU, politics, and media. We also captured temporal changes by aggregating the *TF-IDF* scores of documents published in the same year.

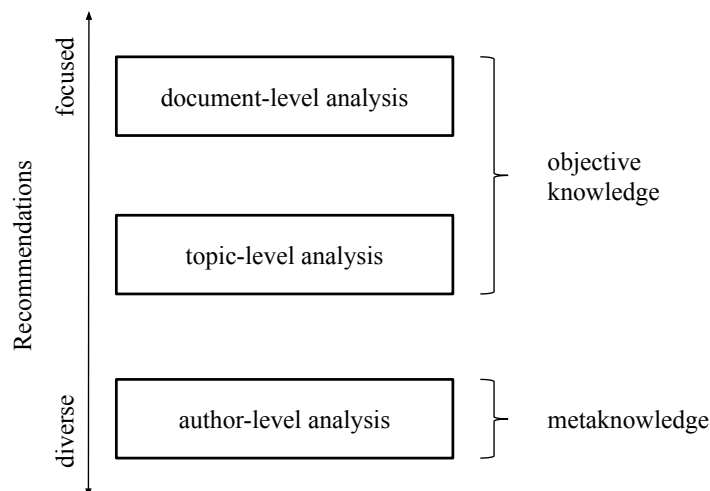


Figure 3.14: Conceptual overview of the types of analysis conducted and their potential applications in recommender systems, mapped onto a spectrum of diversity.

For topic-level analysis, we used a combination of LDA and exploratory techniques like co-occurrence networks to identify latent topics within the text corpora. We collaborated with our data provider, *der FALTER*, in a workshop to label these latent topics based on the most frequently occurring words. Though some topics led to debates, we usually reached a general consensus. The analysis also highlighted how the nature of journalistic documents (i.e., articles versus comments) influenced the topics that emerged.

Temporal trends were analyzed using person-based co-occurrence networks, which aligned well with our other findings. In the context of an NRS, we expect that topic-level analysis would recommend items that are related in theme but could potentially be more diverse in content compared to recommendations from document-level analysis, as depicted in a conceptual manner in Figure 3.14.

In our author-level analysis, the objective was to differentiate between documents from different authors in a merged corpus. We achieved a high level of accuracy, correctly attributing approximately 97% of the documents to their respective authors. When applied to an NRS, this level of analysis has the potential to deliver a diverse range of recommendations, as it is not constrained by the actual content of the documents. For instance, a reader might receive a surprisingly diverse recommendation based on stylistic similarities between the currently read and recommended items.

### 3.2.5 Conclusion

This study embarked on a multifaceted exploration of textual analysis in the domain of news articles. By diving deep into the layers of document, topic, and author, we aspired to answer two overarching research questions.

**RQ1.2.1** *What latent characteristics, which are not immediately evident, can be exposed within news articles when analyzed at different levels (document, topic, and author levels)?*

We found that document-level analysis excels in extracting objective knowledge, focusing mainly on identifying discriminative terms and keywords. This level of analysis also illustrated that the type of document—be it a narrative or opinion piece—significantly influences the extracted keywords. Topic-level analysis, in collaboration with *der FALTER*, helped identify latent themes within the text, which also align well with document-level insights. Finally, author-level analysis allowed us to delve into metaknowledge, specifically the unique writing styles of authors, with an impressive accuracy rate of approximately 97%.

**RQ1.2.2** *How can these identified latent characteristics be harnessed in a news recommender system setting, and what implications might they have on the diversity of recommendations?*

We found that document-level analysis can facilitate highly focused recommendations based on cosine similarity between TF-IDF vectors. For those seeking more thematic diversity, topic-level analysis comes into play. It can offer items that share thematic unity but vary in content, thus adding an extra layer of richness to the recommendations. Most intriguingly, author-level analysis opens the door for serendipitous recommendations, entirely unbounded by the content of the documents, focusing instead on stylistic congruence.

While our study has significantly contributed to our understanding of news article analysis and its application in NRS, limitations persist. Our dataset is relatively small and specialized, raising potential concerns about the generalizability of our findings. Furthermore, the absence of user interaction data in our analysis calls for future work to examine how these different layers of analysis resonate with user behavior in a real-world NRS setting.

In sum, our work lays a robust foundation for the development of a comprehensive NRS. It not only considers the palpable characteristics of news articles but also leverages their latent facets to offer a rich spectrum of recommendations—ranging from focused, to diverse, to surprisingly delightful.

### 3.2.6 Limitations

This study’s multi-layered approach to news article analysis, while insightful, is subject to certain limitations that warrant careful consideration. Primarily, our reliance on *der FALTER*, a Vienna-based weekly newspaper, as the sole data source introduces potential regional and cultural biases. The findings, derived from articles published between June 2013 and June 2018, may not fully capture the dynamic nature of news, long-term trends, or variations in journalistic style across different outlets or geographic regions. The relatively modest size of the corpora (360 articles from the *Klenk corpus* and 216 from the *Thurnher corpus*), and the fact that the text is in German, also restricts the

generalizability of our results, especially regarding topic modeling and cross-linguistic applicability.

Crucially, while subsequent chapters incorporate user interaction data alongside content and subtle cues, this Section ( 3.2) focuses solely on content analysis *without* considering those interactions. Therefore, the conclusions drawn here about the *potential* effectiveness of the identified latent characteristics in a NRS are, at this stage, theoretical and based purely on the textual content itself. The static nature of the content analysis in this section, the inherent subjectivity in topic labeling (despite collaboration with *der FALTER*), and potential limitations of the author-level analysis in generalizing to a larger and more diverse author set, are additional factors to acknowledge. Finally, it must be recognized that the simplified relationships modeled with the co-occurrence networks do not capture the full, complex nature of such relationships (e.g., multi-entity, sentiment). Future research, and indeed later chapters, will build upon this foundation by integrating user interaction data, providing a more comprehensive evaluation of the proposed methods.

### 3.3 Summary

This chapter examined the challenges and opportunities presented by RSs in two distinct yet increasingly digitalized domains—tourism and news. The overarching narrative revolved around the need to improve the user experience in these areas, both of which have been significantly impacted by the digital revolution.

In the tourism sector, we sought to personalize recommendations by aligning them with the multifaceted Seven-Factor Model. The research addressed two core questions: whether a latent structure that fits this model exists within tourism destinations and whether an automatic mapping mechanism could be devised to categorize these destinations. The findings were promising; our cluster analysis yielded six meaningful conceptual groups of tourism destinations, and multiple linear regression models successfully mapped these destinations to the Seven-Factor Model.

Switching gears, the chapter then navigated the ever-changing currents of digital news consumption. Here, the transient nature of news—its ephemeral and constantly evolving quality—poses a unique challenge for RSs. While most items in other domains have a longer shelf life, news items are often relevant only for a short period, causing a persistent *cold-start* problem. This makes it essential for RSs to adapt quickly and efficiently, focusing on content-based techniques to build timely user profiles. We asked: what hidden traits can be uncovered in news articles when analyzed through multiple lenses, and how can these be harnessed to improve news recommendations? Through multi-level analyses, we unearthed insightful details at the document, topic, and author levels of news articles. These latent characteristics were conceptually discussed for their potential to offer a more enriching and diversified news recommendation experience.

Both segments of the chapter make key contributions to their respective fields. In tourism, the research clears a path toward making recommendations more personalized

and data-driven, with real-world scalability in sight. In digital news, the chapter revealed new dimensions that could enrich the quality and diversity of recommended articles, especially in addressing the challenges presented by the transient nature of news.

While the progress has been significant, this chapter also serves as a cornerstone for the immediate topics it covers and lays the foundation for further, more nuanced analyses. It is important to acknowledge the scope of the approaches adopted, recognizing opportunities for future refinement and expansion. For instance, in the tourism domain, the current dataset has certain geographical and attribute-based constraints, and the expert sample, while valuable, was relatively small. Future work could benefit from a broader, more diverse dataset. Additionally, exploring non-linear relationships could potentially uncover more complex interactions than those captured by the regression models. In the news domain, the current analysis focused on a single, regionally specific news source (*der FALTER*) within a limited time frame. Expanding the scope to include multiple sources and a longer time horizon would enhance the generalizability of the findings. Furthermore, incorporating user interaction data, which was not considered in Section 3.2, would provide valuable empirical validation of the identified latent characteristics' effectiveness in an NRS. These considerations point towards exciting avenues for future research that can build upon the foundations established in this chapter.



# Unveiling Tourists' Implicit Preferences Through Pictures

In this chapter we explore **RQ2** *How can a generic picture-based user and item modeling – not limited to a fixed picture set – improve the efficacy and user satisfaction of tourism recommendation systems?* . We actively involve users in the preference elicitation process to provide a more personalized RSs.

We aim to map any given image to a domain model, i.e., Seven-Factor-Model, that encapsulates high-level implicit user/item characteristics – different from previous work, which used a fixed picture set loaded with the Seven-Factors. This involves identifying deeper traits associated with an image beyond just its surface-level details and how these align with the Seven-Factor-Model.

The users are then asked to provide their images and rank them, serving as an interactive mechanism to express their preferences. By analysing these ranked images, we are able to construct a model of the users and their needs, effectively placing them within the Seven-Factor-Model.

This process, our picture-based tourism recommender - PicTouRe, essentially maps users' preferences expressed through their image selections to corresponding tourism recommendations. The effectiveness of this approach is evaluated by conducting a user study, which provides insights into how well the system caters to users' implicit preferences and needs. This approach aims to deliver a more nuanced and user-centric RSs in the tourism industry.

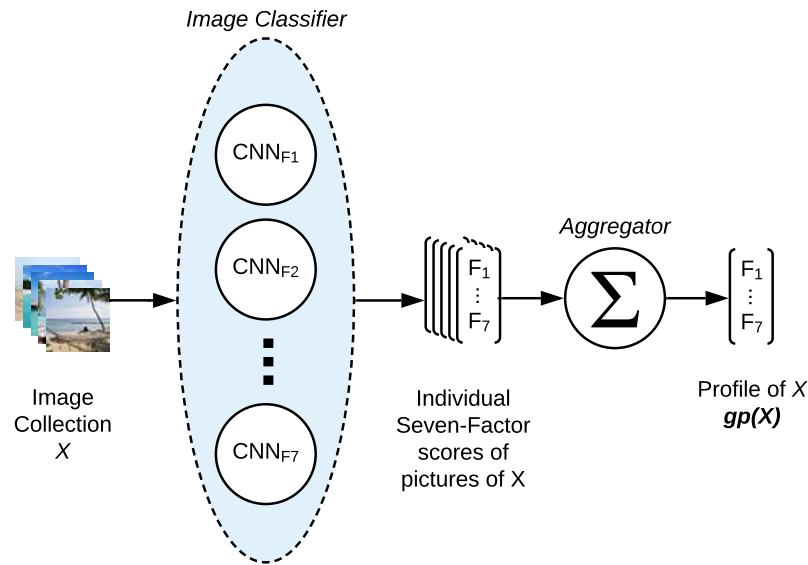


Figure 4.1: Generic Profiler - an overview of the framework. Note, F1 is Sun & Chill-Out, F2 is Knowledge & Travel, F3 is Independence & History, F4 is Culture & Indulgence, F5 is Social & Sports, F6 is Action & Fun, and F7 is Nature & Recreation

## 4.1 A Generic Profiler

The focal point of our research, as illustrated in Figure 4.1, revolves around the development of a method to establish a vector representation—referred to as a profile—of a given image collection. This collection could either belong to a user or represent an item, such as a tourism destination. For a user-provided collection, this profile is conceptualized as a model reflecting the user’s preferences and needs. Alternatively, when the collection comprises images of an item, the profile is treated as a detailed descriptor or set of characteristics pertaining to that item.

Central to our approach is a *Generic Profiler*, composed of two primary elements: an *Image Classifier*, responsible for evaluating individual images, and an *Aggregator*, that compiles these evaluations into a cohesive profile. This section provides an introduction to these key components and delves into their respective functionalities and implementations.

Our investigation is guided by following research questions:

**RQ2.1.1** *To what extent can we determine the Seven-Factor representation of a given image?*

To address this question, we commissioned experts to collect a set of 300 images per factor of the Seven-Factor Model, yielding a balanced dataset with 150 positive and 150 negative examples for each factor. We allocated 200 of these images (per factor) for training our model and reserved the remaining 100 (per factor) for evaluation purposes. We then employed a pre-trained image classifier, initially trained on ImageNet, and

further fine-tuned this classifier using our expert-labeled dataset. The evaluation results are promising: our fine-tuned classifier demonstrates an accuracy range between 0.88 and 0.99 on the validation set, with this range varying depending on the specific Factor under prediction.

**RQ2.1.2** *To what extent can we determine the Seven-Factor representation of a collection of images, specifically images of tourism destinations downloaded from the internet?*

For this aspect of our study, we utilized the fine-tuned classifier from the first research question. We amassed a collection of images for 422 different destinations, obtaining 25 images for each destination from Flickr, Google Search, and Google Travel by searching for the destination name appended with “travel”. For each of these destinations, we already had expert-labeled Seven-Factor scores, scaled between 0 and 1 in increments of 0.25. The outputs generated by our Generic Profiler exhibited a strong positive correlation with the expert labels for almost all factors. However, an exception was observed for Factor F6, labeled “Action & Fun”, where no correlation or a negative correlation was noted. Furthermore, the Mean Absolute Error (MAE) of the predicted factors, when compared to the expert labels, ranged between 0.18 and 0.28.

Our contributions can be summarized as follows:

- We introduce a generic touristic profiler that is capable of determining the touristic profile of either an item (in our case, a destination) or users, given a corresponding collection of pictures.
- We have created an expert-labeled dataset to train and evaluate our models, which have shown promising results.
- We are making our code and the model weights publicly available under: <https://github.com/MeteSertkan/PicTouRe>.

#### 4.1.1 *Image Classifier*

Our approach leverages the power of CNNs without the need for extensive resources. Instead of training CNNs from scratch, which demands a large dataset and significant computation, we adopt a Transfer Learning strategy with pre-trained models.

We specifically utilize a pre-trained ResNet50 model from PyTorch<sup>1</sup>. Originally trained on the extensive ImageNet dataset [27], ResNet50 is a 50-layer CNN that can classify images into 1000 different classes.

**Dataset and Training:** Our dataset, generously provided by an Austrian Tourism company, consists of 300 images per factor of the Seven-Factor Model – 150 positive and 150 negative examples for each. These images, which range from everyday life scenes to nature and urban areas, align well with the variety found in the ImageNet dataset. Each model is trained with 200 images (100 positive and 100 negative) and validated with 100

<sup>1</sup>[https://pytorch.org/hub/pytorch\\_vision\\_resnet/](https://pytorch.org/hub/pytorch_vision_resnet/)

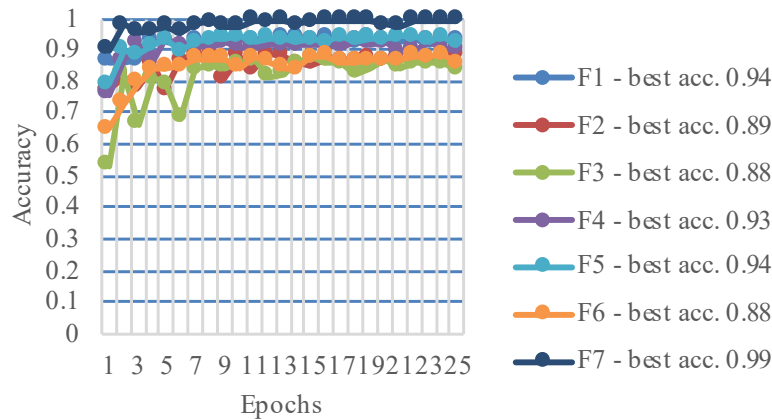


Figure 4.2: Test Accuracy, where F1 is Sun & Chill-Out, F2 is Knowledge & Travel, F3 is Independence & History, F4 is Culture & Indulgence, F5 is Social & Sports, F6 is Action & Fun, and F7 is Nature & Recreation.

images (50 positive and 50 negative). Data augmentation techniques, such as random cropping and horizontal flipping, are used to enrich the training data.

**Adaptation Strategy:** We use ResNet50 as a feature extractor, replacing only its final layer [63]. This custom layer is a simple linear node that outputs a score indicating how likely an image belongs to a specific tourism factor. We apply a SoftMax function to these scores to convert them into probabilities between 0 and 1.

**Multiple Models for Factors:** Since the factors of the Seven-Factor Model are independent, we train a separate CNN for each factor. Thus, for any given image, our complete system – depicted as the *Image Classifier* in Figure 4.1 – effectively assembles the outputs from these seven distinct CNNs into a coherent Seven-Factor score vector.

**Training Strategy:** For training our models, we employ Stochastic Gradient Descent (SGD) as the optimizer and use cross-entropy loss as the loss function [63].

Figure 4.2 illustrates the validation performance of each model. It is evident that our strategy of using pre-trained models gives us a strong starting point, with initial validation accuracies ranging from 53% to 90% after just one training epoch. Ultimately, all seven models perform commendably, with the best models achieving validation accuracies of 88% or higher, notably for factor  $F7$ , where 99% of the images in the validation set were correctly classified.

#### 4.1.2 Aggregator

Once we have processed the images through the *Image Classifier*, each image in a collection is represented by a Seven-Factor score vector, denoted as  $f^p$ . The primary role of the *Aggregator* is to combine these individual Seven-Factor representations  $f_i^p$  of all

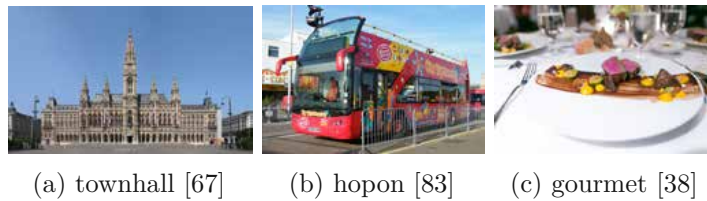


Figure 4.3: Example pictures for profile development. Pictures are taken from *Wikimedia Commons*<sup>1</sup> and *Flickr*<sup>2</sup>.

images in a collection  $X$ , where  $i$  ranges from 1 to  $N$  (the total number of images), into a single representation that characterizes the entire collection.

In this work, we have chosen a straightforward yet effective approach for aggregation: calculating the mean of the Seven-Factor scores across all images in the collection. Thus, the *Generic Profile* of a collection  $X$ , denoted as  $gp(X)$ , is defined as follows:

$$gp(X) = \frac{1}{N} \sum_{i=1}^N f_i^p \quad (4.1)$$

where  $gp(X)$  represents the Generic Profile of collection  $X$ ,  $N$  is the total number of images in the collection, and  $f_i^p$  is the Seven-Factor score vector of the  $i$ -th image.

Despite its simplicity, we designed the *Aggregator* as a separate component. This decision allows for potential flexibility in future work. For instance, future iterations could consider the order of images or incorporate additional sources of information when generating the profile of a collection.

### 4.1.3 Profile Development and Validation

To better understand how profiles are generated using images, let's walk through an example. We will use the images shown in Figure 4.3 to develop a profile. The *Generic Profiler* takes these images as input to generate a profile. For instance, when using the image in Figure 4.3a alone, we obtain the profile shown in Figure 4.4a. In this example, cultural, historical, and knowledge-related factors ( $F_2$ ,  $F_3$ , and  $F_4$ ) score highly, while other factors, like sports and nature, have almost zero scores since these elements are not present in the image.

Adding more images to the collection adjusts the profile. For example, combining the images in Figure 4.3a and Figure 4.3b produces the profile in Figure 4.4b. Compared to the initial profile,  $F_3$  and  $F_4$  factors decrease because tour buses are often associated with mass tourism, not with independent or high-class travelers.

<sup>1</sup><https://commons.wikimedia.org>

<sup>2</sup><https://www.flickr.com>

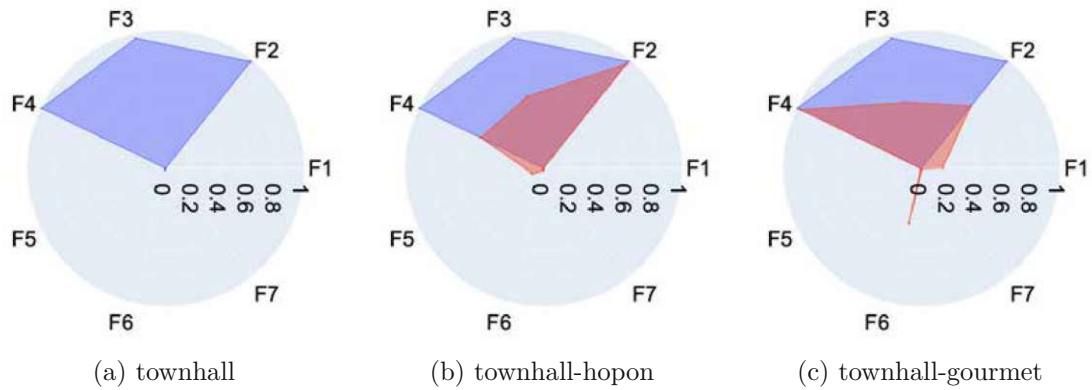


Figure 4.4: Profile development. Note, F1 is Sun & Chill-Out, F2 is Knowledge & Travel, F3 is Independence & History, F4 is Culture & Indulgence, F5 is Social & Sports, F6 is Action & Fun, and F7 is Nature & Recreation.



Figure 4.5: Pictures of the example collection *Action*.

Next, we will compare our *Generic Profiler* approach with a static-picture-based approach, as described in previous work [76, 77]. In the static-picture-based approach, a user selects 3 to 7 pictures from a set of 63, and a profile is determined based on this selection.

For example, Figure 4.6a compares the profiles generated by both approaches using the “Action” collection (Figure 4.5). In both profiles, the *F6* factor scores highly, consistent with the action theme of the images.

The profiles generated using the “Sea” collection (Figure 4.7) are depicted in Figure 4.6b. Interestingly, the CNN-based approach gives a higher score to the *F7* (nature) factor than the static-picture-based approach does, which seems reasonable as all pictures are related to nature.

We also compare our CNN-based approach with expert assessments for two destinations: Vienna and Las Vegas, using data from [96, 99].

For instance, in the case of Vienna (Figure 4.8a), the experts’ opinion offers a broader view of the city. The limitations of our CNN approach seem to stem more from the image collection used as input, which focuses mostly on cultural and historical sights.

The examples provided earlier are designed to showcase the capabilities of the *Generic Profiler* and to facilitate understanding of how it works. We will now delve deeper into the evaluation of the profiler using a comprehensive dataset of labeled destinations.

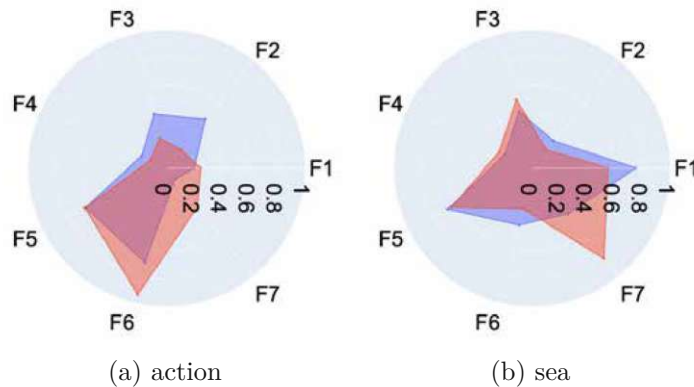


Figure 4.6: Profiles of the example collections, where red is based on the *Generic Profiler* and blue on the fixed picture-set approach. Note F1 is Sun & Chill-Out, F2 is Knowledge & Travel, F3 is Independence & History, F4 is Culture & Indulgence, F5 is Social & Sports, F6 is Action & Fun, and F7 is Nature & Recreation.



Figure 4.7: Pictures of the example collection *Sea*.

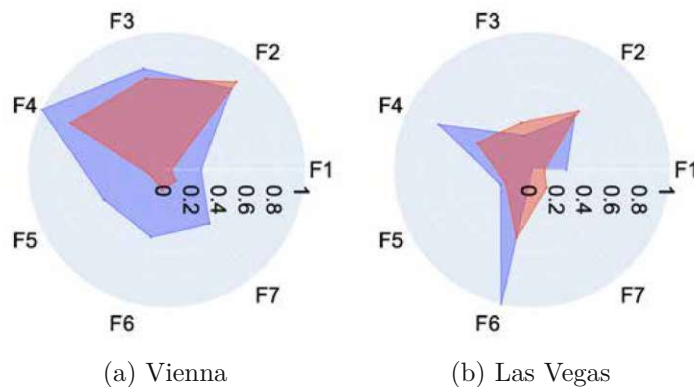


Figure 4.8: Profiles of the example destinations, where red is based on the *Generic Profiler* and blue on the experts. Note F1 is Sun & Chill-Out, F2 is Knowledge & Travel, F3 is Independence & History, F4 is Culture & Indulgence, F5 is Social & Sports, F6 is Action & Fun, and F7 is Nature & Recreation.

Table 4.1: Correlation between the destination profiles generated by Generic Profiler and the expert labels.

Factor	Image Source		
	Google Travel	Google Search	Flickr
F1 - Sun & Chill-Out	.69	.72	.60
F2 - Knowledge & Trave	.66	.46	.46
F3 - Independence & History	.49	.36	.28
F4 - Culture & Indulgence	.59	.32	.53
F5 - Social & Sports	.50	.38	.36
F6 - Action & Fun	.15	-.17	.04
F7 - Nature & Recreation	.67	.53	.58

Our evaluation utilizes a labeled dataset of 422 destinations from our previous work [96, 99]. For each of these destinations, experts assigned a score between 0 and 1, in steps of 0.25, for each factor of the Seven-Factor Model. This expert scoring serves as our reference or gold standard, resulting in a seven-factor representation for each of the 422 destinations.

For each of the 422 destinations, we curate a collection of 75 images. This collection is assembled by downloading and scrutinizing 25 images from each of three data sources: Flickr, Google Search, and Google Travel<sup>2</sup>. We use the destination name combined with the keyword “travel” to fetch these images.

For each destination, we obtain three different Seven-Factor representations by employing the Generic Profiler on the corresponding 25-image collections from the three data sources. We then compare these representations with the gold standard (expert labeling). Our comparison metrics include the correlation with the gold standard and MAE relative to the gold standard.

Our results indicate that, for all factors and utilizing all data sources, there is a correlation with the gold standard, with one notable exception. For factor F6 - “Action & Fun,” we observed either no correlation or a negative correlation, depending on the image source used (see Table 4.1).

The MAE between the profiler’s output and the expert labels ranges from 0.18 to 0.28. This is noteworthy because it corresponds to approximately one point in the labeling scale used by the experts (see Table 4.2).

In summary, the Generic Profiler performs reasonably well in terms of correlation with expert opinions, suggesting its potential utility in practical applications. However, the correlation varies by factor and data source, with Factor F6 being a notable area for improvement.

<sup>2</sup><https://www.google.com/travel>

Table 4.2: MAE between the destination profiles generated by Generic Profiler and the expert labels.

Factor	Image Source		
	Google Travel	Google Search	Flickr
F1 - Sun & Chill-Out	.21	.20	.23
F2 - Knowledge & Trave	.22	.22	.22
F3 - Independence & History	.18	.21	.19
F4 - Culture & Indulgence	.20	.25	.23
F5 - Social & Sports	.22	.28	.18
F6 - Action & Fun	.21	.24	.22
F7 - Nature & Recreation	.19	.23	.21

#### 4.1.4 Conclusions

In this Section 3.1, we aimed to address two main research questions regarding the ability of our Image Classifier, based on CNNs, to determine the Seven-Factor representation of a given image and a collection of images. Below, we summarize our findings in relation to each of these research questions:

**RQ2.1.1** *To what extent can we determine the Seven-Factor representation of a given image?*

Our Image Classifier demonstrates a significant capability to determine the Seven-Factor representation of a single image. The trained CNN models have shown promising results, with validation accuracies between 88% to 99% depending on the respective factor of the Seven-Factor Model. This high level of accuracy in image classification suggests that the Image Classifier can effectively map individual images to their respective Seven-Factor representations. These results are particularly encouraging because they indicate the potential for accurate, automated assessment of how individual images contribute to each of the seven factors. This capability is crucial for building a nuanced understanding of visual content, moving beyond simple object recognition to capturing the more subtle connotations and associations that images evoke.

**RQ2.1.2** *To what extent can we determine the Seven-Factor representation of a collection of images, specifically images of tourism destinations downloaded from the internet?*

Our evaluation utilized a labeled dataset of 422 destinations, with each destination represented by a collection of 75 images curated from three different internet sources. Our approach, employing the Generic Profiler on these image collections, was rigorously evaluated against a gold standard set by expert labels. For most of the factors, our approach correlated with the expert gold standard, indicating a significant capacity to determine the Seven-Factor representation of a collection of images. Notably, the MAE between the profiler’s output and the expert labels was within a range of 0.18 to 0.28, indicating room for improvement. A notable exception was observed for factor F6 - “Action & Fun”, for which the correlation was inconsistent across different image sources.

While the correlations for most factors are strong, the lower performance on F6 (“Action & Fun”) highlights a key area for future development. This discrepancy suggests that the visual cues associated with “Action & Fun” may be more complex, context-dependent, or less consistently represented in online imagery than those for other factors. It also underscores the importance of diverse and representative image sets. The MAE values, while relatively low, indicate that the system’s estimations can deviate from expert judgments. This deviation, while acceptable for some applications, might require refinement for high-precision scenarios. The varying performance across different image sources (Google Travel, Google Search, Flickr) further emphasizes the impact of data source bias on the profiler’s output. A key implication is the need for careful consideration of image sourcing and potentially source-specific model adjustments.

Our work reveals the challenges in curating a proper collection of pictures that comprehensively characterizes a destination. The data acquisition process, as revealed by our research, should be conducted systematically to avoid biases and ensure comprehensive capture of the Seven-Factors. A future avenue could involve developing a taxonomy of tourism-related pictures or products to guide the data acquisition process. This taxonomy would ideally capture a wide range of visual elements and themes associated with each factor, providing a more structured approach to image selection.

We have identified the need for a more sophisticated Aggregator in future work, which could consider the order and/or the probability distribution of the pictures in a collection. The data acquisition process is planned to be conducted more systematically in future work, possibly utilizing a larger and more diverse dataset. Specifically, future aggregators could incorporate attention mechanisms or weighted averaging based on image quality, relevance scores, or even temporal information (e.g., recency of the image).

In summary, this chapter demonstrates a pioneering step towards automated profiling of tourist destinations based on images. The Image Classifier and Generic Profiler introduced in this work show a promising ability to determine the Seven-Factor representation of both individual images and collections of images, thus providing substantial evidence in addressing our research questions RQ2.1.1 and RQ2.1.2. However, the findings also emphasize the importance of data quality, source bias, and the need for more sophisticated aggregation techniques. The success in most factors, coupled with the challenges in F6, provides a roadmap for future improvements, focusing on both model refinement and data curation strategies. The ultimate implication is the potential to create highly personalized and nuanced recommendations in the tourism domain, moving beyond simple keyword matching to a deeper understanding of visual preferences.

### 4.1.5 Limitations

The research presented in Section 4.1, while promising, is subject to several limitations. These limitations are primarily related to the data used, the model’s architecture, and the evaluation methodology.

Firstly, the dataset used for training and evaluation, although extensive, is not exhaustive. The 300 images per factor for the Image Classifier training, while providing a reasonable starting point, may not fully capture the vast diversity of visual representations for each factor, especially for the more abstract or subjective factors like “Action & Fun”. Similarly, the 422 destinations used for evaluating the Generic Profiler, while covering a range of locations, do not represent the entirety of global travel destinations. The reliance on images sourced from Flickr, Google Search, and Google Travel introduces a potential bias towards destinations and imagery that are popular or readily available online. This may not accurately reflect the full spectrum of experiences or less-known destinations.

Secondly, the Generic Profiler’s reliance on a pre-trained ResNet50 model, while efficient, might limit its ability to capture highly domain-specific visual features. Although fine-tuning was performed, the underlying architecture was originally trained on ImageNet, which, despite its breadth, may not perfectly align with the nuances of tourism-related imagery. The simple averaging approach used by the Aggregator is another limitation. It treats all images equally, regardless of their quality, relevance, or the confidence of the Image Classifier’s predictions.

Thirdly, the evaluation relies heavily on expert labels as the gold standard. While expert opinion is valuable, it is inherently subjective and may not fully capture the diverse preferences of individual users. The correlation and MAE metrics, while providing quantitative assessments, do not fully capture the user experience or the perceived accuracy of the generated profiles. The evaluation also focuses primarily on destination profiling, and the applicability of the Generic Profiler to user profiling, while conceptually similar, has not been directly validated with user studies in this section.

Finally, the research does not fully address the temporal aspect of destination preferences. Images and destination characteristics can change over time, and the model does not currently incorporate mechanisms to account for these temporal dynamics.

## 4.2 A User Study on Picture Collections

In the preceding section, we presented the *Generic Profiler*, a tool capable of deriving the *Seven-Factor* representation of a given collection of pictures. This representation takes one of two forms: a *user profile*, constructed from user-provided pictures, or an *item profile*, created when the pictures are sourced from a specific destination, for example. We illustrated the effectiveness of this approach through both exemplar cases and quantitative analysis. Specifically, we applied the Generic Profiler to destination images, obtained from various sources, and compared the resulting profiles with expert-generated labels. This analysis, however, was conducted from the perspective of the item, not the user.

In this section, we shift our focus to evaluate the *Generic Profiler* from the *user’s perspective*. We integrate and deploy our model within an operational system named *PicTouRe*. To assess its utility, we conducted a user study with 81 participants. Furthermore, we

introduced an *aggregation method* that considers the ordering of images in the analysis, and we compared this approach to our initial method, which we refer to as the *averaging method*.

Our research in this section was guided by the following questions, each of which we answer based on our study:

**RQ2.2.1** *To what extent can we determine users' touristic profile using the pictures they provide?*

65% of the participants were satisfied with their predicted touristic profile, while 18% disagreed with it. In particular, the participants disagreed most with the Sun & Chill-Out factor, with a disagreement rate of 37%.

**RQ2.2.2** *How does the touristic profile predicted by our system compare with the users' self-perceived touristic profile?*

Our study showed that the presented touristic profile closely aligns with the adjusted (i.e., perceived) touristic profile, with an average difference between 0.09 and 0.16 across the corresponding factors.

**RQ2.2.3** *Is the order in which user-provided pictures are arranged significant in this context?*

Our analysis found no significant differences attributable to the aggregation strategy, suggesting that the order of user-provided pictures does not significantly affect the results.

**RQ2.2.4** *Which set of recommendations is more appealing to users—those based on the predicted profile, or those based on the user's self-perceived profile?*

Most participants (48%) preferred recommendations generated based on the touristic profile determined by our approach. In contrast, 34% of participants preferred recommendations based on their perceived profile.

Our contributions in this section are summarized as follows:

- We introduced *PicTouRe*, a picture-based TRS that incorporates the *Generic Profiler*.
- We designed and evaluated a user study aimed at establishing users' touristic profiles based on the pictures they provide.
- We assessed the discrepancies between the perceived and predicted touristic profiles of users.
- We scrutinized the impact of picture order in our analysis, comparing the performance of our approach with and without this consideration.
- Our findings indicate that users generally prefer recommendations based on the predicted touristic profile over those derived from their self-perceived profile.
- We make our code - ready to deploy *PicTouRe* - publicly available at: <https://github.com/MeteSertkan/PicTouRe>

### 4.2.1 Methods

#### Accounting for the Picture Order

Insights of a study conducted in [77] show that most people tend to select three to seven pictures out of a given set of pictures. Furthermore, the order of the pictures might carry valuable information, since in the same study people often re-ranked their initial selection. In order to consider the order (i.e., rank) of the pictures in the user's selection they experimented with different strategies. The best strategy not only considered the rank of the pictures in the user's selection, but also the number of pictures in the user's selection.

We adapt those insights and also follow the best performing strategy by 1) Limiting the size of the input collection to minimum three and maximum seven pictures; and 2) Aggregating the individual Seven-Factor scores  $f_i^p$  (i.e., output of the *Classification* step) of an input collection  $X$  with  $n = 3, \dots, 7$  pictures through weighted averaging. Thus, the profile of  $X$ , i.e.,  $gp(X)$ , is defined as follows:

$$gp(X) = \frac{\sum_{i=1}^n \omega_i f_i^p}{\sum_{i=1}^n \omega_i} \quad (4.2)$$

where  $gp(X)$  represents the aggregated profile of the input collection  $X$ ,  $\omega_i$  is the weight assigned to the  $i$ -th picture, and  $f_i^p$  is the Seven-Factor score of the  $i$ -th picture in the collection.

$$\omega_i = 7 \frac{-r + n + 1}{\sum_{k=1}^n k} \quad (4.3)$$

where  $\omega_i$  is the weight of each picture,  $n$  is the total number of pictures in the input collection  $X$  (ranging from 3 to 7), and  $r$  denotes the rank of the picture within the selection (e.g.,  $r = 1$  for the first picture,  $r = 2$  for the second, and so on). For instance,  $\omega_i$  for the first ranked picture in a collection of three pictures equals to  $\frac{21}{6}$ ,  $\frac{14}{6}$  for the second ranked picture, and finally  $\frac{7}{6}$  for the third ranked picture. The sum of all weights always equals seven.

#### Experimental Design

Figure 4.9 shows the five main steps of our study. Here is a straightforward breakdown:

**Starting Point:** Upon arrival, participants see detailed information about the study and are asked to agree to participate.

**Step 1: Picture Selection.** Participants imagine their next vacation and choose 3 to 7 photos that fit this vision. These can be their own photos or images from the internet. We assure participants that their photos will not be saved or shown anywhere, respecting their privacy.

**Step 2: Picture Ranking.** Participants arrange their selected photos in order of importance.

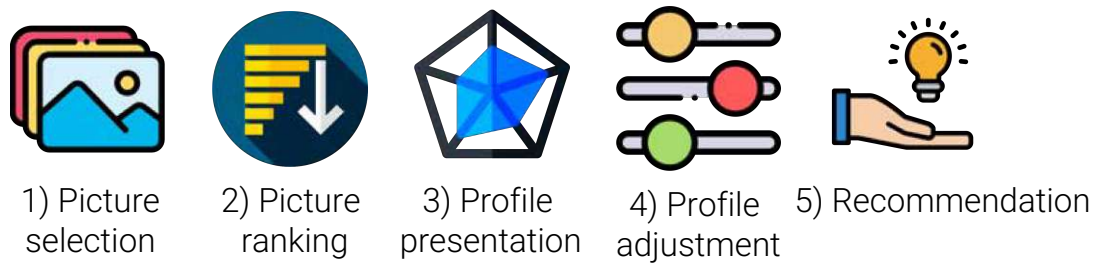


Figure 4.9: Study Procedure: 1) Picture Selection; 2) Picture Ranking; 3) Profile Presentation; 4) Profile Feedback & Questions; 5) Destination Recommendation.

**Step 3: Profile Presentation.** From the chosen photos, we generate a touristic profile (i.e., Seven-Factor representation) using the Generic Profiler. This profile could be-with equal chance-calculated using either an average method (referred to as *AVG*) or a rank-weighted average method (referred to as *RWA*). Participants also receive a brief overview of the Seven-Factor model and a link to a more detailed explanation.

**Step 4: Profile Adjustment & Questions.** We gather participant feedback through a series of questions, ranging from the ease of finding photos to their thoughts on the generated profile and some basic personal details. For clarity, these questions are (\* marks mandatory questions):

Q01 - \* It was easy to find 3 to 7 pictures.

Q02 - \* I mainly used pictures downloaded from the internet (e.g., Google, Flickr, etc.).

Q03 - \* I mainly used my own pictures.

Q04 - \* I understood the explanations of the Seven-Factors.

Q05 - \* The resulting profile matches my preferences.

Q06 - Which factor in the resulting profile does not match well? (multiple answers allowed)

Q07 - How would you adjust the resulting profile? (multiple adjustments allowed)

Q08 - \* What is your age?

Q09 - \* What is your gender?

Q10 - \* What is your highest degree of education?

Q11 - \* How often do you travel for pleasure (leisure/tourism)?

Q12 - Comments/Suggestions.

For questions *Q01-Q05* we provide a five-point Likert scale ranging from *strongly disagree* to *strongly agree*. For question *Q06* we provide seven checkboxes, each for one factor of the Seven-Factor Model. In case of *Q07*, we provide seven sliders (again each for one factor of the Seven-Factor Model), where the values are pre-set to the Seven-Factor scores of the predicted touristic profile. Questions *Q08-Q11* can be answered via Radio buttons, where we always provide the option “prefer not to say”. Question *Q12* is an open question, which can be answered via text field.

We categorize questions into three main themes: 1) Picture Choices (*Q01-Q03*), 2) Feedback on the Touristic Profile (*Q04-Q07*), and 3) Participant Background (*Q08-Q011*).

Additionally, we track:

- How long they take to select and rank pictures.
- Time spent understanding their profile and answering questions.
- How often they rearrange the order of pictures.

**Step 5: Destination Recommendation.** Lastly, we suggest travel destinations to participants. These suggestions are based on their perceived profile, the initially generated profile, and also some random selections. The recommended tourism destinations stem from a previous study [99] and were mapped manually onto the Seven-Factor Model by experts. Recommendations are presented in a random order to mitigate position bias. If two top choices are the same, we count it for both methods used. Participants can choose their favorite recommendation or opt not to choose at all.

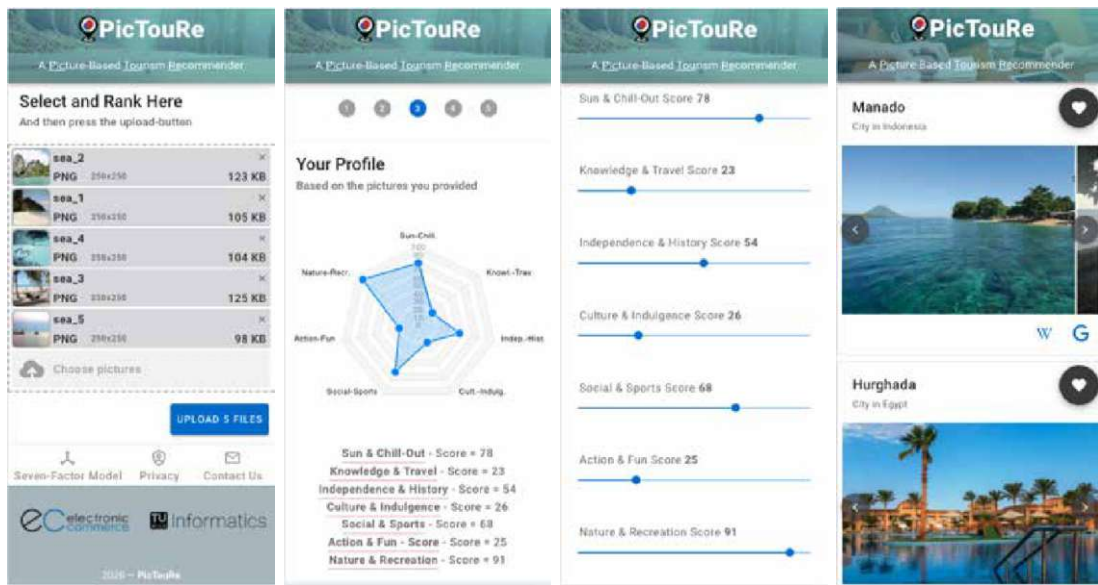
### System Overview

The system’s core is built using Python, with Flask serving the web pages and PyTorch handling our pre-set image models. The visual part of our website, what users interact with, is crafted with Vue.js and styled using Vuetify for a sleek design. PicTouRe is designed to work smoothly on both mobile devices and computers.

Here is how users interact with PicTouRe:

- **Introduction:** Upon entering the site, users learn about PicTouRe. We assure them of their anonymity and confirm we will not store their uploaded pictures.
- **Selecting Pictures:** Users choose 3-7 pictures from their device. They can either drag and drop these photos or use a file picker. Once uploaded, they can rearrange them in their preferred order as shown in Fig. 4.10a.
- **Profile Creation:** From these pictures, we create a user’s travel profile, showcasing it as a radar-chart in Fig. 4.10b. Note: The profile can be calculated using various methods for a balanced comparison.

#### 4. UNVEILING TOURISTS' IMPLICIT PREFERENCES THROUGH PICTURES



(a) Picture Selection (b) Profile (c) Profile Adjustment (d) Recommendation

Figure 4.10: Screenshots of PicTouRe's mobile interface, showing the journey from choosing pictures to receiving a travel recommendation.

- **Profile Adjustment:** If users want, they can fine-tune their profile using adjustable sliders, as seen in Fig. 4.10c.
- **Receiving Recommendations:** Considering the predicted profile, the perceived (adjusted) profile, and a random profile (control), we suggest travel destinations. These are shown in Fig. 4.10d. We mix up the order of these suggestions to keep it fair (position bias). If two top choices are the same, both methods used to make that suggestion get recognition. These destinations stem from [99], where experts matched them to the Seven-Factor Model.
- **Learning More:** Curious users can dig deeper into each recommendation, getting details from Google Travel or Wikipedia.
- **Feedback:** Lastly, users have the option to share their thoughts by picking one recommendation.

#### Evaluation

The purpose of questions *Q01-Q04* and *Q08-Q11* is to get more insights about the participants and their behaviour, and to find support for generalizability statements. Also, tracking interactions and time might give insights about the behaviour and hints about difficulties participants face. Altogether, those insight can be used to further improve the introduced concept and moreover its presentation (i.e., user interface).

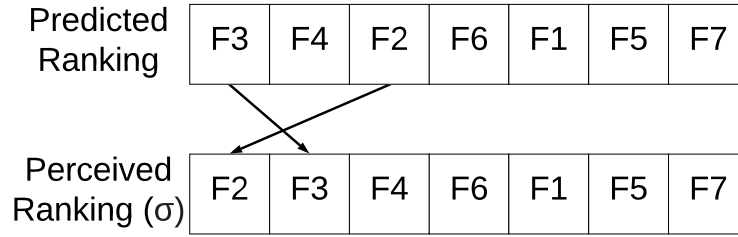


Figure 4.11: Kendall's Tau distance - Total number of inversion in  $\sigma$ . Note, in this example Kendall's Tau distance is 2.

Questions *Q05-Q07* are used to assess the overall performance and also the difference in performance with respect to the aggregation strategies (i.e., *AVG* and *RWA*). We use the *MAE* in order to assess the difference in each factor of the Seven-Factor Model between predicted touristic profile and the user's perception (i.e., user's adjustment to the presented profile). Besides considering the predictive performance in each factor, we also treat the user's touristic profile as such by considering its distance to the user's perception. Therefore, we use Kendall's Tau distance ( $DIST_\tau$ ), Spearman's Footrule ( $DIST_{SPEAR}$ ), and the Euclidean distance ( $DIST_{EUCL}$ ).

Comparing the predicted ranking (i.e., rank of the factors based on scores in the predicted profile) and perceived ranking ( $\sigma$ ) (i.e., rank of the factors based on scores in the adjusted profile), the Kendall's Tau distance can be interpreted as the total number of inversions in  $\sigma$  (see Figure 4.11) [32]. Here, a pair of elements  $F_i$  and  $F_j$  is considered as inverted if  $R_{F_i} > R_{F_j}$  and  $R_{\sigma(F_i)} < R_{\sigma(F_j)}$ , where  $R_{F_i}$  stands for the ranking of an element and  $R_{\sigma(F_i)}$  the ranking of an element in  $\sigma$ . Thus the Kendall's Tau distance is defined as following:

$$DIST_\tau = \sum_{R_{F_i} < R_{F_j}} 1_{R_{\sigma(F_i)} > R_{\sigma(F_j)}} \quad (4.4)$$

Where:

- $DIST_\tau$  represents the Kendall's Tau distance.
- $R_{F_i}$  is the rank of factor  $F_i$  in the predicted ranking.
- $R_{F_j}$  is the rank of factor  $F_j$  in the predicted ranking.
- $R_{\sigma(F_i)}$  is the rank of factor  $F_i$  in the perceived ranking ( $\sigma$ ).
- $R_{\sigma(F_j)}$  is the rank of factor  $F_j$  in the perceived ranking ( $\sigma$ ).
- $1_{R_{\sigma(F_i)} > R_{\sigma(F_j)}}$  is an indicator function, which equals 1 if  $R_{\sigma(F_i)} > R_{\sigma(F_j)}$  (indicating an inversion) and 0 otherwise.

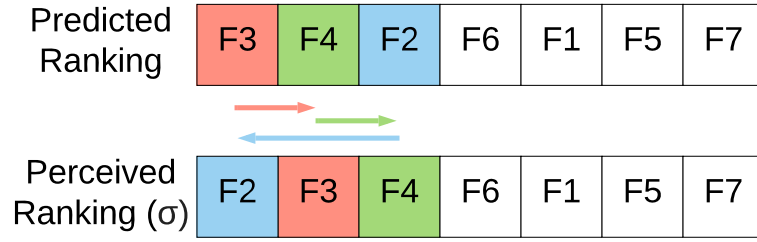


Figure 4.12: Spearman's Footrule distance - Total displacements of elements in  $\sigma$ . Note, in this example Spearman's Footrule distance is 4.

- The summation iterates over all pairs of factors where  $R_{F_i} < R_{F_j}$ .

On the other hand, the Spearman's Footrule distance (see Figure 4.12) can be interpreted as the total number of displacement of all elements [32]. Here, a displacement is considered as the distance an element  $F_i$  has to be moved to match  $\sigma(F_i)$ , which can also be written as  $|R_{F_i} - R_{\sigma(F_i)}|$ . Since the Spearman's Footrule is defined as the total number of displacements, it can be written as following:

$$DIST_{SPEAR} = \sum_i |R_{F_i} - R_{\sigma(F_i)}| \quad (4.5)$$

Where:

- $DIST_{SPEAR}$  represents Spearman's Footrule distance.
- $R_{F_i}$  is the rank of factor  $F_i$  in the predicted ranking.
- $R_{\sigma(F_i)}$  is the rank of factor  $F_i$  in the perceived ranking ( $\sigma$ ).
- The summation iterates over all factors (indexed by  $i$ ).
- $|R_{F_i} - R_{\sigma(F_i)}|$  calculates the absolute difference in ranks, representing the displacement of a factor.

By comparing the ranking, we account for the change in relevance of each factor of the Seven-Factor Model, for instance, the factor *Sun & Chill-Out* might be more relevant (i.e., ranked higher) in the user's perception than in the predicted touristic profile. Besides that, we also consider the distance of the presented and the perceived touristic profile based on the actual difference in Seven-Factor scores by using the Euclidian distance, which is defined as following:

$$DIST_{EUCL} = \sqrt{\sum_{i=1}^7 (predicted\_F_i - perceived\_F_i)^2} \quad (4.6)$$

Where:

- $DIST_{EUCL}$  represents the Euclidean distance.
- $predicted\_F_i$  is the predicted score for the  $i$ -th factor of the Seven-Factor Model.
- $perceived\_F_i$  is the perceived (adjusted by the user) score for the  $i$ -th factor.
- The summation iterates over all seven factors ( $i$  from 1 to 7).

Finally, in order to identify significant distributional differences between the predicted Seven-Factor scores and the perceived Seven-Factor scores, we use the paired Student's t-test or the Wilcoxon signed-rank test depending on the outcome of the Shapiro-Wilk normality test. Furthermore, to compare differences based on the two aggregation strategies (i.e., *AVG* and *RWA*) and the nature of the considered variable we either use Mann-Whitney U test or Fisher's exact test.

#### 4.2.2 Results

In this section, we delve into the results of our user study. We describe the demographics of the participants, their engagement with the picture selection and ranking process, and conclude with an evaluation of our model's performance. This includes comparisons with user perceptions and contrasts between the two aggregation strategies.

##### Participants

The study attracted 81 participants. Their recruitment took place in January 2020 through i) an announcement at the ENTER2020 international eTourism conference and targeted emails to tourism professionals, and ii) dissemination on social media and personal networks. Notably, the recruitment channel did not significantly impact the outcomes.

Demographically:

- Gender: 62% identified as male, and 38% as female.
- Age distribution:
  - 60%: 25-34 years
  - 20%: 35-44 years
  - 7%: 45-55 years
  - 7%: Above 55 years
  - 3%: 18-24 years
  - 3%: Below 18 years

Table 4.3: Agreement statistics for Q05, differentiated by aggregation strategy and collectively. Note that a score of 0 represents strong disagreement, while 4 signifies strong agreement.

	mean	sd	min	median	max
Overall (N=81)	2.69	0.81	0	3	4
AVG (N=41)	2.68	0.81	0	3	4
RWA (N=40)	2.70	0.81	1	3	4

- Education: 87% held a bachelor's, master's, or PhD degree, 9% had a high school diploma, 2% had less than a high school education, and 2% categorized their education as "other."
- Travel frequency for leisure:
  - 62%: 1-3 times a year
  - 15%: 3-4 times a year
  - 4%: 4-5 times a year
  - 12%: More than 5 times a year
  - 7%: Less than once a year
- Device usage: 42% used a mobile device.
- Understanding of the Seven-Factors: 90% either agreed or strongly agreed that they grasped the concept, as indicated by their response to Q04.

Each participant's touristic profile was based on either the *AVG* or the *RWA* aggregation strategy, assigned randomly. Of the total participants, 51% were in the *AVG* group and 49% in the *RWA* group. Subsequent sections provide detailed outcomes for the entire group (N=81) and for the individual *AVG* (N=41) and *RWA* (N=40) subgroups.

### Overall Performance

To evaluate our approach's effectiveness, we assessed participant satisfaction by inquiring how well the provided touristic profile aligned with their preferences (Q05). Furthermore, we identified which factors in the profile were perceived as misaligned (Q06).

A majority, 65%, expressed satisfaction with the touristic profiles generated, indicating agreement or strong agreement with Q05. This sentiment was consistent regardless of the aggregation strategy—either *AVG* or *RWA*, both randomly assigned.

The summarized response data for Q05 across both strategies and collectively are presented in Table 4.3. Notably, no discernible differences between the strategies regarding satisfaction were observed.

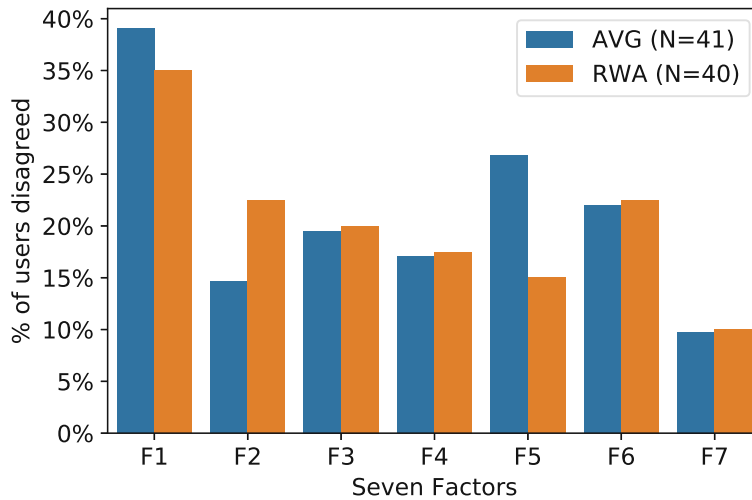


Figure 4.13: Factors' disagreement rates as per the presented touristic profile (*Q06*), disaggregated by strategy. Note, F1 is Sun & Chill-Out, F2 is Knowledge & Travel, F3 is Independence & History, F4 is Culture & Indulgence, F5 is Social & Sports, F6 is Action & Fun, and F7 is Nature & Recreation.

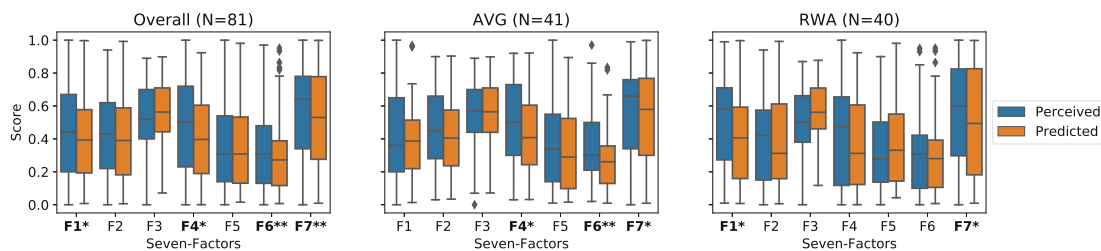


Figure 4.14: Comparison of predicted and perceived Seven-Factor scores, split by aggregation strategy. Noteworthy variances are emphasized with significance levels: \*  $p < 0.05$  and \*\*  $p < 0.01$ . Note, F1 is Sun & Chill-Out, F2 is Knowledge & Travel, F3 is Independence & History, F4 is Culture & Indulgence, F5 is Social & Sports, F6 is Action & Fun, and F7 is Nature & Recreation.

Regarding the factors, *Sun & Chill-Out* had the highest disagreement, with 37% of participants marking it as unaligned, while *Nature & Indulgence* fared better with only a 10% disagreement rate. The disagreement percentages for other factors ranged between 17-22%. Disagreements related to both aggregation strategies are visualized in Figure 4.13. Again, no significant differences between strategies were found in this context.

The next figure, Figure 4.14, juxtaposes the predicted profile against the perceived Seven-Factor scores, offering insights across aggregation strategies. Significant variances, where present, are highlighted.

Lastly, utilizing the predicted, perceived, and a random (control) profile, we proposed

travel destinations, as showcased in Fig. 4.10d. We randomized the presentation order of these recommendations to neutralize position bias. When the top picks matched, credit was attributed to both methods. These destinations originate from [99], where they were mapped to the Seven-Factor Model.

Almost half of the participants (48%) showed a preference for destinations recommended by our approach's touristic profile, while 34% favored those based on their perceived profile.

### **Predicted vs. Perceived Touristic Profile**

Besides the binary feedback of whether or not a factor of the predicted touristic profile fits well (i.e., *Q06*), we also provided the possibility to adjust the respective profile via seven sliders, each for one factor of the Seven-Factor Model (i.e., *Q07*). We consider the resulting profile after *Q07* as the user's perceived touristic profile. About 90% of the participants adjusted one or more factors of the predicted touristic profile, but on average three factors. Similar observations are made, if the initial sample is split with respect to both aggregation strategies and analysed separately. No statistically significant difference was observed with respect to both strategies and the number of taken adjustments.

We conducted statistical significance tests in order to capture if the Seven-Factor scores of the predicted touristic profiles differ significantly from the Seven-Factor scores of the perceived touristic profiles. In particular, based on the outcome of the Shapiro Wilk normality test we either used Wilcoxon signed-rank test or paired Student's t-test. Figure 4.14 summarizes the outcome of this comparison.

Overall (N=81), the distribution of scores between predicted touristic profile and perceived touristic profiles were significantly different in factors: *Sun & Chill-Out (F1)* with  $p < 0.05$ , *Culture & Indulgence (F4)* with  $p < 0.05$ , *Social & Sports (F6)*  $p < 0.01$ , and *Nature & Recreation (F7)* with  $p < 0.01$ . On average, the participants corrected those factors by plus 0.05-0.06.

Focusing only on the responses to the *AVG* aggregation strategy (N=41), the distributions of Seven-Factor scores of the predicted touristic profiles compared to the distributions of the perceived touristic profiles were significantly different in factors: *Culture & Indulgence (F4)* with  $p < 0.05$ , *Social & Sports (F6)*  $p < 0.01$ , and *Nature & Recreation (F7)* with  $p < 0.05$ . On average, the participants corrected those factors by plus 0.05-0.09.

On the other hand, by only considering the responses of participants, who were assigned the *RWA* aggregation strategy (N=40), following factors showed significant differences in Seven-Factor scores distributions when the predicted and the perceived profiles were compared: *Sun & Chill-Out (F1)*  $p < 0.05$  and *Nature & Recreation (F7)* with  $p < 0.05$ . On average, the participants corrected those factors by plus 0.07-0.10.

Besides identifying differences in Seven-Factor scores distributions, we also calculated the MAE, i.e., mean of the absolute differences, between predicted and perceived Seven-Factor scores, in order to capture the predictive performance of our models. Table 4.4 lists the

Table 4.4: MAE, where F1 is Sun & Chill-Out, F2 is Knowledge & Travel, F3 is Independence & History, F4 is Culture & Indulgence, F5 is Social & Sports, F6 is Action & Fun, and F7 is Nature & Recreation.

	F1	F2	F3	F4	F5	F6	F7
MAE-Overall	0.16	0.10	0.09	0.10	0.10	0.10	0.9
MAE-AVG	<b>0.16</b>	<b>0.08</b>	<b>0.07</b>	0.11	0.12	0.12	<b>0.07</b>
MAE-RWA	0.17	0.13	0.10	<b>0.09</b>	<b>0.08</b>	<b>0.09</b>	0.10

resulting overall MAEs for each factor of the Seven-Factor model and for both strategies. Overall, our approach showed promising performance with MAEs between 0.09 and 0.16 on a scale from 0 to 1. Furthermore, the largest deviation from the perceived Seven-Factor scores was the one for factor *Sun & Chill-Out (F1)* with a MAE of 0.16. For all other factors our approach showed similar performance.

Similar MAEs were observed if the predicted and perceived Seven-Factor scores were considered for both strategies separately. Moreover, a comparison of MAEs between both strategies showed that *AVG* results in a slight better predictive performance in factors *Sun & Chill-Out (F1)*, *Knowledge & Travel (F2)*, *Independence & History (F3)*, and *Nature & Recreation (F7)*. On the other hand, *RWA* slightly performed better in factors *Culture & Indulgence (F4)*, *Social & Sports (F5)*, and *Action & Fun (F6)*. But, the differences in performance were overall not significant.

Analysing mean absolute differences or distributional differences between Seven-Factor scores of predicted and perceived touristic profiles does not consider the user representation as a such, but rather the factors of the Seven-Factor model. Therefore, we also took into account the distance between the predicted and perceived user representations (i.e., touristic profiles) into account. We analysed how far apart both representations in Euclidean space are by using  $DIST_{EUCL}$ . Furthermore, we used  $DIST_{\tau}$  and  $DIST_{SPEAR}$  to capture whether or not changes in Seven-Factor scores lead to changes in factor relevance (i.e., ranking). For instance, after the user’s predicted profile adjustment *Sun & Chill-Out (F1)* might score better and thus get more relevant (i.e., ranked higher) than *Nature & Recreation (F7)*.

We determined all three distances between predicted and perceived touristic profile for all participants and then averaged them (i.e.,  $\overline{DIST}_{\tau}$ ,  $\overline{DIST}_{SPEAR}$ , and  $\overline{DIST}_{EUCL}$ ) in order to draw conclusions about our approach with respected to the distances. The results are listed in Table 4.5. *RWA* performed on average relatively better than *AVG* with respect to the ranking of the factors (i.e., lower  $\overline{DIST}_{\tau}$  and  $\overline{DIST}_{SPEAR}$ ). On the other hand, *AVG* performed relatively better with respect to prediction accuracy (i.e., lower  $DIST_{EUCL}$ ). However, the Mann-Whitney-U test showed that the differences are not significant.

Table 4.5: Differences in predicted and perceived touristic profile with respect to changes in ranking (i.e., relevance) of the factors and point-wise difference. Note,  $\overline{DIST}_\tau$  is the average  $DIST_\tau$  between predicted and perceived touristic profiles; similarly  $\overline{DIST}_{SPEAR}$  is the average  $DIST_{SPEAR}$ , and  $\overline{DIST}_{EUCL}$  the average  $DIST_{EUCL}$ .

	Overall	AVG	RWA
$\overline{DIST}_\tau$	3.58	3.90	<b>3.25</b>
$\overline{DIST}_{SPEAR}$	6.68	7.20	<b>6.15</b>
$\overline{DIST}_{EUCL}$	0.44	<b>0.41</b>	0.47

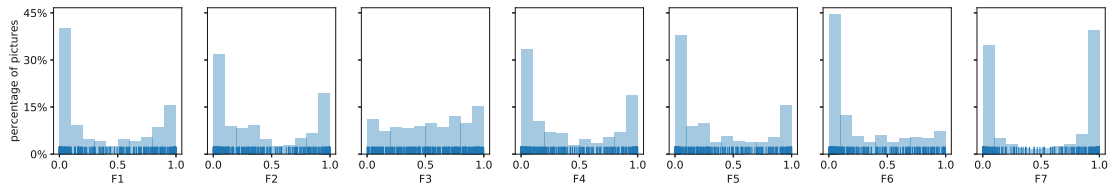


Figure 4.15: Distribution of Seven-Factor scores of the uploaded pictures. Note, F1 is Sun & Chill-Out, F2 is Knowledge & Travel, F3 is Independence & History, F4 is Culture & Indulgence, F5 is Social & Sports, F6 is Action & Fun, and F7 is Nature & Recreation.

### Picture Selection & Ranking

As already mentioned, we gave the participants the option to select between three and seven pictures. The majority of the participants (i.e., 52%) selected only three pictures, 16% selected four pictures, 11% six pictures, another 11% seven pictures, and finally 10% five pictures. We also asked the participants to re-consider the initial ranking of the selected pictures, where only 21% did actually a re-ranking. Those, who considered a re-ranking, changed the initial ranking between one and four times. Half of the participants finished the selection and ranking task within 2.7 minutes, 75% of the participants completed after 5.8 minutes, and after 10.7 minutes already 90% were finished. The majority of the participants (72%) agreed or strongly agreed with Q01 (i.e., “It was easy to find 3 to 7 pictures”), which is in line with the reported timing above.

The distributions of Seven-Factor scores of the uploaded pictures (see Figure 4.15) have overall a similar shape, where either there are strong signals in the pictures for the considered factor (i.e., high score) or there are no signals (i.e., low score). Except in case of factor *Independence & History* (F3), where the scores are more evenly distributed. Furthermore, there only were very few signals for the factor *Action & Fun* (F6) in the user provided pictures. Both observations might indicate that some factors are harder to capture, leaving the room for further improvements.

In contrast to analysing the distribution of the Seven-Factor scores of all uploaded pictures, where we treated the Seven-Factors individually and all pictures at once, we also analysed the diversity of the provided pictures per user selection. In other words,

Table 4.6: Diversity in users’ picture selection of different sizes. Note, “#Pics” is selection size; “Kendall’s” is the mean of the average pairwise  $DIST_{\tau}$  of the users’ picture selection; “Spearman’s” is the mean of the average pairwise  $DIST_{SPEAR}$ ; and “Euclidean” is the mean of the average pairwise  $DIST_{EUCL}$ .

#Pics	Kendall’s	Spearman’s	Euclidean
3	8.41	13.08	1.11
4	8.07	12.94	1.06
5	10.82	16.34	1.35
6	7.73	12.77	<b>1.00</b>
7	<b>7.30</b>	<b>11.47</b>	1.03

we investigated whether the uploaded images are homogeneous (e.g., only images of nature) or more diverse (e.g., images of nature, sports, and beach). Here, diversity of a users picture selection is defined as the average of the pairwise distances of the pictures in the respective selection. We used  $DIST_{\tau}$ ,  $DIST_{SPEAR}$ , and  $DIST_{EUCL}$  as distance measure.

The results are listed in Table 4.6, for instance, in case of the selection size of three one has to swap on average 8 times adjacent (based on ranking) factors in order to match the ranking of factors with another picture’s ranking in the same selection. Similarly, one has to move the factors (based on ranking) 13 times in order to match the ranking of factors with another picture’s ranking in the same selection. Also, the point-wise difference, i.e., diversity based on  $DIST_{EUCL}$ , is relatively high. Thus, in case of picture selection size three, the participants selected relatively diverse pictures, which is also true in case of all other picture sizes. However, the diversity in pictures in selections of sizes six or seven is relatively lower compared to diversity of pictures in selections of all other sizes, which is not expected (since there are more pictures to compare with).

### 4.2.3 Discussion

Based on the responses to the demographic questions (i.e.,  $Q08-Q11$ ) the majority of participants were male, between 25 and 44 years old, and hold at least a bachelor’s degree. Thus, generalizing the outcomes and implications might only be possible in a limited way. Future work will consider a more systematic approach of survey distribution and sampling, and eventually further distribution channels to increase and diversify the pool of participants.

We are aware that by asking the users for their next hypothetical trip, where they have nothing to win or lose, might influence the users’ behaviour and thus introduce some bias. Future work will consider user incentives. Ultimately, we aim at pre-trip and post-trip studies. Furthermore, we plan to enrich the questionnaire (e.g., by following [84] and [34]) in order to obtain even more valuable information.

Following the approach in [77], we gave participants the option to select three to seven pictures. Approximately half of them only selected three pictures. This might have happened because of convenience or they may already have had a very focused idea about their next tourism destination and thus uploaded few most important pictures. Difficulty in finding pictures might not be the reason, since the majority of the participants reported that it was easy to find pictures and also the timing indicates that they were relatively quick in selecting and ranking pictures. Furthermore, only few participants took the chance to re-order their initial pictures selection by relevance. Deciding, which of the selected pictures is more important than the others, might have been a difficult task. Especially, in case of only three pictures, it might have felt unnecessary. Another possible explanation, why users select only few pictures and do not re-consider their ordering, is a lack of involvement, which can be addressed with user incentives in future.

Overall, our approach got positive feedback. Most of the participants were satisfied with the predicted touristic profile capturing their preferences. The participants mostly disagreed with the predicted score for factor *Sun & Chill-Out* in comparison to all other factors. Furthermore, no significant differences in performance could be shown between both aggregation strategies, i.e. *AVG* and *RWA*. In the future, we will further improve our models, e.g., by training the CNNs systematically with more pictures. Moreover, we plan to adapt other aggregation strategies, for instance, variations of ordered weighted averaging.

In addition to the binary feedback, whether or not a factor fits well, we also provided the option to directly adjust the predicted factors via slider inputs. The vast majority of participants used this opportunity and adjusted at least one of the factors. The participants adjusted the factor *Sun & Chill-Out* more often than the other factors of the Seven-Factor Model. This is in line with the outcomes of the binary feedback (i.e., people disagreed the most with factor *Sun & Chill-Out*). Similar observations were made by considering the MAE between predicted and perceived touristic profiles, where the biggest difference was observed in factor *Sun & Chill-Out*. Thus, the participants not only reported that they disagree with *Sun & Chill-Out* more often than the other factors, but they also adjusted this factor the most in comparison to the others. Overall, our approach has a tendency to underrate the predicted factor scores, i.e., the participants were usually correcting the predicted factors upwards. However, based on the resulting MAEs in each factor our approach showed promising performance.

Finally, we analysed the differences in predicted relevancy (i.e., ranking) and perceived relevancy of the factors of the Seven-Factor representation (i.e., touristic profile). Therefore, we captured to what extent the ranking of the factors (based on the predicted Seven-Factor scores) did change after the user's adjustment. Our results indicates that the predicted ranking is relatively close to the perceived ranking. However, we did not consider the relative position of the rankings. Discrepancies in top ranked (i.e., highly relevant) factors might have a higher impact than in low ranked factors.

#### 4.2.4 Conclusions

In this research, we sought to address a prevalent challenge encountered by travelers: the articulation of their unique preferences and needs. Drawing inspiration from the adage “a picture is worth a thousand words,” we innovatively employed images as a means of communication. With the aim of surmounting extant communication barriers, we developed PicTouRe - a picture-based TRS.

To gauge the efficacy and utility of PicTouRe, we carried out an in-depth user study. In doing so, we posed several research questions:

**RQ2.2.1** *To what extent can we determine users’ touristic profile using the pictures they provide?*

Building on previous research concepts [97, 102], participants, with their upcoming hypothetical journeys in mind, uploaded between three to seven images, arranging them by importance. Our findings revealed that 65% of participants felt their predicted touristic profile mirrored their tastes, while a minority (18%) disagreed. The *Sun & Chill-Out* factor elicited disagreements more frequently, with 37% of participants expressing divergent views, than the other factors which garnered disagreements in the range of 10-22%. This discrepancy may stem from the inherent subjectivity in interpreting the *Sun & Chill-Out* factor. While some may associate it solely with beach vacations, others might envision a broader spectrum of activities, such as exploring vibrant cities with ample outdoor cafes and relaxation spots. This highlights the challenge of capturing the multifaceted nature of individual preferences, even within well-defined categories. A manual error analysis showed that the training data consisted mostly of beach vacation pictures, whereas destinations like Marrakesh or Valencia, which are also associated with Sun & Chill-Out, were underrepresented. This highlights the limitation and challenge of curating a representative collection of images for accurate predictions.

**RQ2.2.2** *How does the touristic profile predicted by our system compare with the users’ self-perceived touristic profile?*

After creating an initial touristic profile based on their image selections, participants had the chance to adjust this profile. The accuracy of the system’s predictions compared to the users’ self-perceptions was demonstrated by the MAEs for the Seven Factors, which ranged from 0.09 to 0.16 on a scale of 0 to 1. This suggests that while our model captures the general trend of user preferences, there is room for improvement in fine-tuning the accuracy of individual factor predictions. The relatively low MAEs, however, underscore the potential of our image-based approach in effectively approximating user preferences.

**RQ2.2.3** *Is the order in which user-provided pictures are arranged significant in this context?*

We introduced an improved aggregation technique that took into account the order in which participants arranged their images, recognizing the possible nuances in their sequencing. However, the results indicated that considering the image order did not significantly improve our model’s accuracy. This suggests that the initial hypothesis about the importance of image order may not hold true. It is possible that the act of

selecting images itself carries more weight in revealing preferences than the specific order in which they are arranged. Alternatively, the current method of incorporating image order may not be sensitive enough to capture its subtle influence. Further investigation into alternative methods of incorporating image order, such as attention mechanisms or sequential modeling, could shed more light on this aspect.

**RQ2.2.4** *Which set of recommendations is more appealing to users—those based on the predicted profile, or those based on the user's self-perceived profile?*

As a final step, we recommended tourism destinations, considering both the predicted and self-perceived profiles. Of the participants, 48% exhibited a preference for destinations proposed by our model, whereas 34% opted for destinations based on their self-perceived profiles. This result indicates that our model's recommendations are indeed competitive with, and often preferred to, those based on self-assessment. This highlights the potential of our approach in not only capturing user preferences but also in suggesting destinations that align with these preferences, even if they were not explicitly stated. It also suggests that users may not always be fully aware of their own preferences, and that our model can help them discover new and exciting destinations that they might not have considered otherwise.

In conclusion, our findings suggest that image-based preference elicitation can be an effective tool for understanding traveler needs and preferences, leading to more personalized and satisfying travel recommendations. The relatively high level of user satisfaction with the predicted profiles and the preference for our model's recommendations underscore the potential of our approach in revolutionizing the way travel planning is conducted.

### 4.2.5 Limitations

While our research demonstrates the potential of image-based TRS, it is essential to acknowledge certain limitations. First, the sample size of our user study was relatively small and may not fully represent the diversity of travelers. Second, the study relied on hypothetical travel scenarios, which may not fully capture the complexities of real-world travel planning. Third, our model's accuracy was limited by the quality and representativeness of the training data, particularly for the *Sun & Chill-Out* factor. Fourth, the current implementation of PicTouRe does not incorporate contextual factors such as budget, travel dates, or companion preferences, which could significantly influence travel decisions. Finally, the interpretability of our model's predictions remains a challenge, as it is difficult to pinpoint the specific image features that drive the prediction of a particular factor. Addressing these limitations in future research will further strengthen the validity and applicability of our findings.

## 4.3 Summary

This chapter delves into the realm of RSs, focusing on an innovative technique for the tourism industry: *PicTouRe*. Recognizing the inherent challenge of the *cold-start*

problem—wherein user data is often unavailable—the chapter champions the use of images as a communication bridge between users and RSs, transcending the need for explicit preference statements.

*PicTouRe* leverages images to create personalized tourism recommendations. Users provide and rank images that serve as interactive reflections of their preferences. These images are then mapped to a domain model, specifically the *Seven-Factor Model*, capturing high-level implicit user/item characteristics. By doing so, *PicTouRe* translates visual preferences into tailored tourism suggestions.

Historical techniques in RSs have largely depended on collecting user data or explicit preference input. Recent works, however, including those by Ferwerda et al. [36], Ferwerda and Tkalcic [35], demonstrate the viability of images, especially social media photos, as indicators of personal traits and preferences. The research presented in this chapter evolves from past approaches by allowing users to interact with a dynamic set of pictures rather than being restricted to a fixed set [77]. Furthermore, it leverages CNNs to directly discern the *Seven-Factor* scores, bypassing the potential loss of information from intermediate classification.

The effectiveness of *PicTouRe* is underscored by two sets of conclusions. First, the *Image Classifier*, based on CNNs, manifests a robust capability to map individual images, and even collections of images, to their respective *Seven-Factor* representations. The classifier achieved high validation accuracies, suggesting that this approach could comprehensively characterize tourist destinations. However, performance varied across factors, with “Action & Fun” showing less consistent results, highlighting the influence of data source bias and the need for more sophisticated aggregation. Furthermore, challenges were highlighted in curating a holistic collection of images, prompting suggestions for more systematic data acquisition in future iterations.

Second, an exhaustive user study evaluated *PicTouRe*’s practical application. The study revealed that a substantial majority of participants (65%) agreed with their generated touristic profiles, and the MAEs between system predictions and users’ self-perceptions were relatively low. Interestingly, while predicted and self-perceived profiles aligned, image order during aggregation did not significantly impact prediction accuracy. Notably, a considerable portion of users preferred recommendations based on the system’s predicted profile, demonstrating *PicTouRe*’s potential for real-world application. Nevertheless, there remains room for growth, notably in expanding the training dataset, refining aggregation strategies, and broadening the applicability of findings across diverse user groups.

While the results are promising, several limitations should be acknowledged. Section 4.1.5 discusses the reliance on a pre-trained ResNet50 model, the simple averaging approach of the Aggregator, and the subjectivity of the expert labels used as the gold standard. The dataset’s limitations, including potential biases from online image sources and the limited representation of the full spectrum of destinations, were also considered. Similarly, the user study (Section 4.2.5) faced constraints related to sample size, the

#### 4. UNVEILING TOURISTS' IMPLICIT PREFERENCES THROUGH PICTURES

---

use of hypothetical scenarios, the underrepresentation of certain visual concepts in the training data (particularly for the “Sun & Chill-Out” factor), and the absence of contextual factors in the current *PicTouRe* implementation. Future research should focus on addressing these limitations to enhance the system’s robustness and generalizability.

In essence, this chapter highlights the promising potential of using images to capture and convey user preferences in tourism recommendations. As the research progresses, we envision multiple enhancements to further refine, validate, and build upon the foundation established by *PicTouRe*.

# Leveraging Sentiment & Emotions for News Recommendation

In this chapter we investigate **RQ3** *How can sentiment and expressed emotions in news articles be effectively utilized in recommendation systems to improve recommendation performance, and what impact do these factors have on diversity?* In contrast to the previous chapter, we switch from active user involvement to passive observation, aiming to model users- both their content-wise and emotional/sentiment preferences -within a latent vector space implicitly. This involves observing users' online interactions and behaviors, particularly in the context of news reading. However, we retain the interpretability of the emotional/sentiment dimensions.

We consider not only the textual content of news articles, which represent the obvious item characteristics, but we also pay attention to sentiment and emotions expressed within these articles, which we categorize as implicit item characteristics. This multi-faceted approach allows us to build a comprehensive user model, capturing a wider spectrum of their preferences and needs.

Our goal here is to improve the relevance of recommendations by considering these underlying emotional dimensions. By focusing on sentiment and emotion, we better understand the user's response to content, going beyond simple interaction data to explore more subtle user preferences.

We also explore diversity and diversification approaches. By considering the sentiment/emotion dimensions, we aim for balanced and diverse recommendations to provide a more satisfying and enriching user experience. This line of research essentially leverages passive user data to build dynamic, robust, and user-sensitive RSs.

## 5.1 Exploiting & Diversifying Sentiments

Content-based RSs traditionally suggest items to users based on their past preferences [88]. In the realm of neural news recommendation, recent advancements still adhere to this principle, drawing insights from users' past interactions with news articles to rank potential news items for them [124]. One significant limitation of these methods, however, is their susceptibility to offering a monotonous array of recommendations, often lacking in diversity [119].

It is worth noting that news articles with negative sentiment tend to attract more user clicks than their positive counterparts. As a result, ensuring a diverse mix of sentiments in the recommendations is paramount in the field of news recommendation [119].

Recognizing the importance of sentiment diversity, Wu et al. [119] proposed *SentiRec*, an innovative neural news recommendation algorithm that emphasizes sentiment diversity. SentiRec's unique approach lies in its sentiment-aware news representations which it derives from the actual content of the news. This is achieved by synchronously training the recommendation model while also incorporating an auxiliary sentiment prediction task. The user modeling is based on both their clicked and unclicked (i.e., seen but not clicked) news articles. To bolster sentiment diversity, SentiRec systematically penalizes news articles that align closely with the user's prevailing sentiment orientation. Central to the tasks of sentiment regularization and prediction in SentiRec is the application of VADER [53], a rule-oriented sentiment analyzer, to assign sentiment polarity scores.

In this section, we first reproduce SentiRec without the availability of the original source code and dataset. We attempted to contact the authors of SentiRec but were unsuccessful. Following this replication, we build upon the foundational work and delve deeper into an analytical exploration of SentiRec's framework. We also introduce adaptations to enhance its efficacy and robustness. For our experimental analysis, we utilized the Microsoft MIND dataset [124], which, while coming from the same origin as the SentiRec study, comprises a distinct subset of data (i.e., different time frame). During our replication efforts, we identified gaps in the details regarding the data preprocessing, the design of both the news and user encoders, and the hyperparameters. To bridge these gaps, we turned to the related *NRMS* model [118] and its open-source version<sup>1</sup>. Furthermore, we refined the hyperparameters by tuning on a validation set.

To align our replication with the original findings, we first compare the outcomes of our SentiRec implementation with existing baselines on the MIND dataset:

**RQ3.1.1** *How does our reproduced SentiRec implementation compare to baselines on the MIND dataset?*

While our results confirm the effectiveness of SentiRec in increasing sentiment diversity without compromising performance, the sentiment diversity margins are not as broad as the ones reported in the original paper. A potential reason could be dataset variances, but this underscores potential challenges in the generalizability of SentiRec.

<sup>1</sup><https://github.com/microsoft/recommenders>

Our exploration further segues into examining the impact of integrating a more efficient neural sentiment analyzer as opposed to a rule-based one, thus we study:

**RQ3.1.2** *What downstream effect does a more effective neural sentiment analyzer create in comparison to a rule-based one?*

We consider the potential advantages of a pre-trained neural sentiment analyzer (BERT-SA<sup>2</sup>) over a rule-based model (VADER-SA). Interestingly, our observations indicate no significant improvements in recommendation performance or sentiment diversity when transitioning to BERT-SA.

To highlight the trade-off between effectiveness and sentiment diversity we study:

**RQ3.1.3** *What influence do sentiment prediction and sentiment diversity loss regularization hyperparameters have on the resulting sentiment diversity and recommendation performance?*

In contrast to the original paper, we observe a decrease in sentiment diversity if the sentiment prediction task's influence increases. In line with the original paper, a too strong influence of sentiment prediction as well as sentiment regularization leads to a drop in recommendation performance.

Beyond the scope of the original study, which only explored user-centric sentiment diversity, we further investigate:

**RQ3.1.4** *How does our reproduced SentiRec implementation compare to the MIND baselines concerning topical diversity?*

As in the user-centric sentiment-diversity case, the baselines, especially the *NAML* baseline, already yield significantly better topical diversity than our SentiRec model while maintaining reasonable recommendation performance – demonstrating the competitiveness of the baselines again.

We also extend our research horizon to inspect intra-list sentiment and topical diversity, focusing on the nuances within a single recommendation list:

**RQ3.1.5** *How does our reproduced SentiRec implementation compare to the MIND baselines concerning intra-list sentiment-diversity and intra-list topical-diversity?*

In contrast to the user-centric evaluation, our *SentiRec* reproduction significantly outperforms most baselines, including the strong *NAML* baseline if intra-list sentiment diversity is considered – demonstrating the capacities of the *SentiRec* model.

Moreover, we introduce *RobustSentiRec*, an enhanced version of SentiRec optimized for sentiment diversity robustness. In this model, instead of relying on the auxiliary sentiment prediction task, we leverage sentiment labels directly. This leads us to:

**RQ3.1.6** *To what extent can the sentiment prediction task be omitted and instead the sentiment-labels be used as an addition signal?*

RobustSentiRec offers improved recommendation performance while maintaining sentiment diversity levels and simplifying the overall model.

<sup>2</sup><https://huggingface.co/distilbert-base-uncased-finetuned-sst-2-english>

Our contributions can be summarized as follows:

- We re-implemented SentiRec [119] from scratch using the MIND dataset [124], as the original source code and dataset were unavailable. Our implementation shows similar trends but does not reproduce the original findings, likely due to dataset differences. Notably, our baselines already demonstrate good recommendation and sentiment diversity performance.
- We experimented with a pre-trained neural sentiment analyzer instead of a rule-based one but observed no improvement in effectiveness or sentiment diversity.
- We extended the experiment to include topical diversity and intra-list topical and sentiment diversity. While the baselines performed better in topical diversity, our approach significantly outperformed most baselines in intra-list sentiment diversity.
- We introduced RobustSentiRec, a simplified version of SentiRec, achieving comparable sentiment diversity results while improving recommendation performance.
- We published the first open-source implementation of SentiRec, along with RobustSentiRec and all baselines, available at: <https://github.com/MeteSertkan/newsrec>.

### 5.1.1 Methods

#### Diversifying Sentiments in News Recommendation

Our primary objective is to enhance both the accuracy of our recommendations and the diversity of sentiments within those recommendations. Naturally, striving for both can introduce a trade-off between the two criteria.

Here is the task broken down:

**Objective.** Rank a list of potential news articles based on a user’s past interactions.

**User Profile.** For a given user,  $u$ , we have a historical dataset,  $H$ , consisting of  $n$  news articles they have previously viewed, represented as  $[D_1, D_2, \dots, D_n]$ . Each of these articles has an associated sentiment polarity score:  $[s_1, s_2, \dots, s_n]$ .

**Candidate Articles.** We have a set,  $C$ , of  $p$  potential news articles to recommend, represented as  $[D_1^C, D_2^C, \dots, D_p^C]$ . Each of these candidates also has a sentiment polarity score:  $[s_1^C, s_2^C, \dots, s_p^C]$ .

**Ranking.** Our task is to assign each candidate article in set  $C$  a score, resulting in a list:  $[\hat{y}_1, \hat{y}_2, \dots, \hat{y}_p]$ .

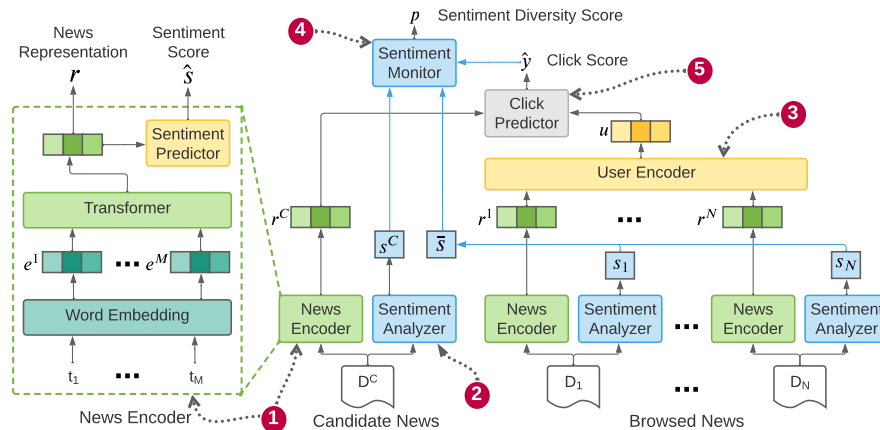


Figure 5.1: Overview of SentiRec [119] comprising following major components: ① *News Encoder*, which learns to encode news by their content and simultaneously to predict a sentiment score based on the learned encoding; ② *Sentiment Analyzer*, which assigns a sentiment score to each news article based on its content; ③ *User Encoder*, which models users based on their previous news interactions; ④ *Click Predictor*, which determines a score for a given user and candidate news pair; and ⑤ *Sentiment Monitor*, which monitors and regularizes the sentiment diversity.

**Diversity Goal.** The ultimate aim is to offer a diverse sentiment palette in our recommendations. This means that the top-ranking articles in our recommendation should have sentiment polarity scores that differ from the average sentiment,  $\bar{s}$ , of the user’s historical articles. The average sentiment is calculated as:  $\bar{s} = \text{mean}([s_1, s_2, \dots, s_N])$ .

In simpler terms, we want our recommendations to not only be accurate based on the user’s history but also offer a diverse range of sentiments, different from what the user typically views.

### SentiRec

SentiRec [119] creates sentiment-aware embeddings of news using their content and user embeddings by considering the users’ history track of interacted news. Candidate news is then ranked by the dot-product between the user and item embedding and penalized if the news sentiment gets closer to the users’ overall sentiment orientation. In the following we describe the different SentiRec components as shown in Figure 5.1.

① *News Encoder.* The task of the news encoder is to find a representation  $r^c$  of candidate news  $D^c$  as well as representations  $[r_1, \dots, r_N]$  of browsed news  $[D_1, \dots, D_N]$  by taking their title as input. It consists of an embedding layer followed by a transformer layer to obtain a representation  $r$  out of a sequence of terms. Since no details about the transformer layer were given, we follow the architecture of the closely related *NRMS* [118] model.

Thus, we use multi-head self-attention for contextualization and additive-attention to obtain a unified embedding out of the contextualized word embeddings. The news encoder is jointly trained with an auxiliary sentiment prediction task in order to infuse sentiment awareness to the news representation. The sentiment score  $\hat{s}$  is predicted using a linear layer, i.e.,  $\hat{s} = V_s \times r + v_s$ , where  $V_s$  and  $v_s$  are learnable parameters and  $r$  is the news representation. As loss function the MAE between predicted sentiment scores  $\hat{s}_i$  and the sentiment scores determined by the sentiment analyzer  $s_i$  is used as follows :

$$\mathcal{L}_{senti} = \frac{1}{S} \sum_{i=1}^S |\hat{s}_i - s_i| \quad (5.1)$$

where  $S$  denotes the set of training samples,  $\hat{s}_i$  represents the predicted sentiment score for the  $i$ -th news article, and  $s_i$  is the corresponding sentiment score obtained from the sentiment analyzer.

② *Sentiment-Analyzer.* Given the title of a news article, the sentiment analyzer determines the sentiment polarity score ranging in  $[-1, 1]$ , which is considered as the sentiment label of the respective news article. The original paper uses VADER [53] (a rule-based method) as sentiment analyzer (VADER-SA). In addition, we also study a pre-trained neural sentiment analyzer<sup>3</sup> (BERT-SA).

③ *User Encoder.* The user encoder gets the sentiment-aware representations of the previously browsed news, i.e.,  $[r_1, \dots, r_N]$ , as input and uses a transformer layer (i.e., multi-head self-attention followed by additive attention according to *NRMS* [118]) to obtain a representation  $u$  of the user.

④ *Click Predictor.* The click predictor uses the dot-product between user and candidate embedding, i.e.,  $u^\top r^c$ , to determine a click score  $\hat{y}$ .

⑤ *Sentiment Monitor.* The sentiment monitor observes to what extent the sentiment polarity score (obtained by the sentiment analyzer)  $s^c$  of a candidate news article diverges from the users' overall sentiment orientation  $\bar{s} = \text{mean}([s_1, \dots, s_N])$  (i.e., the mean sentiment polarity score of the users browsing history). This diversity in sentiment is measured by  $p = \max(0, \bar{s}s^c\hat{y})$ , where  $\bar{s}$  is the user's overall sentiment orientation,  $s^c$  is the sentiment score of the candidate news, and  $\hat{y}$  is the predicted click score, where larger values of  $p$  indicate less sentiment diversity. The sentiment diversity score  $p$  is further used to regularize and steer the model into a more sentiment diverse direction. Following loss function is used for this purpose:

$$\mathcal{L}_{div} = \frac{1}{|S|} \sum_{i \in S} p_i \quad (5.2)$$

where  $S$  is the training set and  $p_i$  the sentiment diversity score of the  $i$ -th sample.

Negative sampling is used in order to create a labeled dataset for the recommendation task. For each clicked news in a user impression,  $K$  non-clicked samples from the same

<sup>3</sup><https://huggingface.co/distilbert-base-uncased-finetuned-sst-2-english>

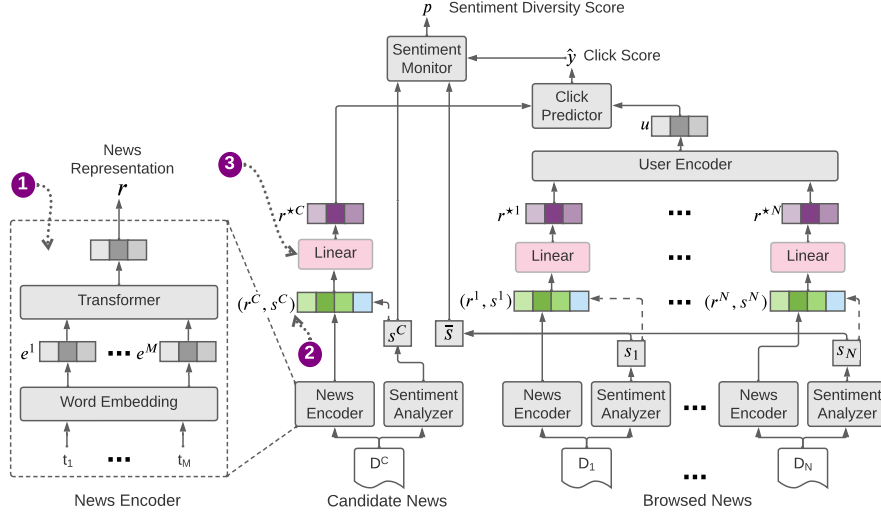


Figure 5.2: Overview of RobustSentiRec. In contrast to SentiRec we: ① Skip the sentiment prediction; ② Concatenate the news representation and the corresponding sentiment labels; and ③ Use a linear layer to obtain a sentiment-aware news representation.

impression are randomly selected. The recommendation loss is the negative log-likelihood of the clicked samples and is defined as follows:

$$\mathcal{L}_{rec} = - \sum_{i \in S} \log \left( \frac{\exp(\hat{y}_i^+)}{\exp(\hat{y}_i^+) + \sum_{j=1}^K \exp(\hat{y}_{i,j}^-)} \right) \quad (5.3)$$

where  $\hat{y}_i^+$  is the click score of  $i$ -th clicked news and  $\hat{y}_{i,j}^-$  the click score of the  $j$ -th sample of the corresponding  $K$  negative samples, and  $S$  is the training set. The final loss function brings all three losses, i.e., recommendation loss, sentiment prediction loss, and sentiment diversity loss, together as follows:

$$\mathcal{L} = \mathcal{L}_{rec} + \lambda \mathcal{L}_{senti} + \mu \mathcal{L}_{div} \quad (5.4)$$

where  $\mathcal{L}_{rec}$  represents the recommendation loss,  $\mathcal{L}_{senti}$  denotes the sentiment prediction loss,  $\mathcal{L}_{div}$  is the sentiment diversity loss, and  $\lambda$  and  $\mu$  are hyperparameters controlling the influence of the sentiment prediction loss and sentiment diversity loss respectively.

### RobustSentiRec

In addition to our SentiRec reproduction, we introduce RobustSentiRec, where we adapt SentiRec by omitting the auxiliary sentiment prediction task (compare ① in Figure 5.2 with ① in Figure 5.1). Instead, we directly incorporate the sentiment label by concatenating it with the news embedding (compare ② in Figure 5.2 with the output

of ❶ in Figure 5.1) and applying a linear transformation (see ❸ in Figure 5.2). The adapted sentiment-aware news representation  $r^*$  is computed as Equation 5.5:

$$r^* = V_{r^*} \times (r, s) + v_{r^*}, \quad (5.5)$$

where  $r^*$  is the adapted sentiment-aware news representation,  $(r, s)$  is the concatenation of the news representation  $r$  and the corresponding sentiment label  $s$ , and  $V_{r^*}$  and  $v_{r^*}$  are learnable parameters of the linear layer. Finally, we keep the sentiment regularization task and thus the resulting loss function is defined as Equation 5.6:

$$\mathcal{L} = \mathcal{L}_{rec} + \mu \mathcal{L}_{div} \quad (5.6)$$

where  $\mathcal{L}$  represents the total loss,  $\mathcal{L}_{rec}$  is the recommendation loss,  $\mathcal{L}_{div}$  the sentiment-diversity loss, and  $\mu$  is a hyperparameter regularizing the sentiment diversity loss.

### Evaluation Perspectives

We evaluate our reproduction from five different perspectives: effectiveness, user-centric sentiment diversity, intra-list sentiment diversity, user-centric topical diversity, and intra-list topical diversity. Note, in contrast to the intra-list diversity measures, the user-centric measures assess diversity in relation to the user’s previous news consumption. We compare the results of our reproduction against all baselines and our extensions, using paired t-test with Bonferroni correction [110, 105].

**Effectiveness.** We evaluate effectiveness using  $AUC$ ,  $MRR$ ,  $nDCG@5$ , and  $nDCG@10$ .

**User-Centric Sentiment Diversity.** We evaluate user-centric sentiment diversity using the sentiment alignment metrics  $S_{MRR}$  and  $S@K$ , introduced by WU et al. [119], which is defined as follows:

$$S_{MRR} = \max(0, \bar{s} \sum_{i=1}^C \frac{s_i^c}{i}), \quad S@K = \max(0, \bar{s} \sum_{i=1}^K s_i^c) \quad (5.7)$$

where  $C$  is the length of the recommendation list (i.e., number of candidate items),  $s_i^c$  is the sentiment polarity score of the news article ranked at position  $i$  in this list, and  $\bar{s}$  is the overall sentiment orientation of the corresponding user. Hence, the closer top-ranked candidates’ sentiment to the users’ overall sentiment orientation, the higher the sentiment alignment metrics. Ergo, lower sentiment alignment scores indicate more sentiment-diverse recommendations.

**Intra-List Sentiment Diversity.** As the sentiment polarity score  $s_i$  of a news article is only one scalar, we compute the intra-list sentiment diversity by averaging the absolute difference of sentiment polarity scores  $s_i$  and  $s_j$  between each news pair in the Top-K list of recommended candidate articles:

$$ILS_S@K = \frac{2}{K(K-1)} \sum_{s_i, s_j \in C@K} |s_i - s_j| \quad (5.8)$$

where  $K$  is the number of recommended candidate articles considered (i.e., Top-K),  $s_i$  and  $s_j$  are the sentiment polarity scores of two recommended news articles, and  $C@K$  represents the set of Top-K recommended candidate articles.

The intra-list sentiment diversity score lies between 0 and 1, with 0 being maximal divers.

**User-Centric Topical Diversity.** We consider the news articles' categories (e.g., sports) and subcategories (e.g., soccer) to compute topical diversity. We represent a (sub)category of a news article with a 1-hot-encoding. We compute the user's category representation  $c_u$  by summing up all browsed news category representations. Similarly, we compute the recommendations list's category representation  $c_{C@K}$  by summing up the category representations of the recommended top-K candidate news articles. We then measure diversity  $T@K$  by taking cosine similarity between  $c_u$  and  $c_{C@K}$ . This leads to a measure between 0 and 1, with 0 being maximal divers. Similarly, we measure  $T_{MRR}$  with the difference being computing a weighted average of all candidates' category representations to obtain a representation  $c_{MRR}$  of the recommendation list, where the weight is the rank of corresponding news articles.

$$T_{MRR} = \text{cos}_{sim}(c_{MRR}, c_u), \quad T@K = \text{cos}_{sim}(c_{C@K}, c_u) \quad (5.9)$$

where  $\text{cos}_{sim}$  denotes the cosine similarity,  $c_{MRR}$  represents the weighted average of all candidates' category representations in the recommendation list,  $c_u$  is the user's category representation,  $c_{C@K}$  represents the sum of the category representations of the recommended top-K candidate news, and  $K$  denotes the number of considered items.

**Intra-List Topical Diversity.** We again represent a (sub)category of a news article with a 1-hot-encoding. We measure the intra-list topical diversity of the recommendation list by computing the average pairwise cosine similarity between the 1-hot-encoded category representations  $c$  of the recommended top-k news articles. This leads to a measure between 0 and 1, with 0 being maximal divers.

$$ILS_T@K = \frac{2}{K(K-1)} \sum_{c_i, c_j \in C@K} \text{cos}_{sim}(c_i, c_j) \quad (5.10)$$

where  $K$  is the number of recommended candidate articles considered (i.e., Top-K),  $c_i$  and  $c_j$  are the 1-hot-encoded category representations of two recommended news articles,  $C@K$  represents the set of Top-K recommended candidate articles, and  $\text{cos}_{sim}$  denotes the cosine similarity.

### 5.1.2 Experimental Setting

**Dataset.** The dataset of the original paper is constructed from MSN News<sup>4</sup> logs collected from October 31, 2018, to January 29, 2019, but has not been open-sourced, and our access request has not been answered yet. Thus, we use the MIND dataset - specifically the MIND-small<sup>5</sup> version - in our experiments, as it stems from the same source. It was randomly sampled from 50K users (with at least five clicks) during six weeks, from October 12 to November 22, 2019, where the first five weeks are for training and the last week for testing. One sample is composed of a timestamp, the user-id, a list of chronologically ordered news-ids representing the user’s click history, and a list of shuffled candidate news-ids with corresponding labels (i.e., 1 for clicked and 0 for seen but not clicked). Detailed statistics of the datasets are summarized in Table 5.1. MIND-small has five times more users with about two times fewer impressions and on average about seven times fewer positive interactions per user (seven clicks vs. 49) than the SentiRec dataset.

Table 5.1: SentiRec dataset (as reported) and MIND-small dataset statistics.

Dataset	#Users	#News	#Impression	#Clicks	#Non-Clicks
SentiRec	10,000	42,255	445,230	489,644	6,651,940
MIND-small	50,000	65,238	230,117	347,727	8,236,715

**Training.** All models are trained on 90% of the training data. The remaining 10% is used to tune the hyperparameters by optimizing AUC. We use early-stopping with a minimum delta of 0.0001 AUC and patience of 5. Note that we use 300-dimensional Glove embeddings [82] in all models to initialize the word embedding layer and NLTK [13] word tokenizer for tokenization. Further, we limit the number of browsed news in each impression to 50 and the title length to 20 terms (smaller sequences are zero-padded).

**Parameter Settings.** We apply 20% dropout to the word embeddings. The negative sampling ratio  $K$  is set to 4. We use multi-head self-attention with 15 attention heads followed by an additive-attention layer with a 200-dimensional query vector. We use the ADAM [66] optimizer with a learning rate of 0.0001 and a batch size of 128. For the VADER-SA based model ( $SentiRec_V$ ) we set  $\lambda = 0.4$  and  $\mu = 10$  and for the BERT-SA based model ( $SentiRec_B$ ) we set  $\lambda = 0.4$  and  $\mu = 1$ . Likewise, we obtain the same parameters for  $RobustSentiRec_V$  (based on VADER-SA) and  $RobustSentiRec_B$  (based on BERT-SA) following the same approach. Except we omit  $\lambda$  since the sentiment prediction is skipped.

**Baselines.** We compare the reproduced and adapted models against following baselines suggested by the dataset providers [124]:

<sup>4</sup><https://www.msn.com/en-us/news>

<sup>5</sup><https://msnews.github.io/index.html>

*LSTUR* [4] - Neural NRS capturing users' long- and short-term interests. We initialize the GRU network with the user embedding. We set the masking probability of the users' long-term interests to 50%. We apply 20% dropout to the word embeddings. The negative sampling ratio  $K$  is set to 4. For the CNN, we set the number of filters to 300 and the window size to 3. We use a 200-dimensional query vector for the additive-attention layer. We use the ADAM [66] optimizer with a learning rate of 0.0001 and a batch size of 256.

*NAML* [116] & *NAML<sub>T</sub>* - Neural NRS incorporating multiple views (i.e., title, category, and abstract) into the news representation. We limit the abstract length to 50 terms. We apply 20% dropout to the word embeddings. We set the category embeddings dimension to 100. The number of CNN filters is set to 400 and the window size to 3. We use 200-dimensional query vectors in the additive-attention layers. The negative sampling ratio  $K$  is set to 4. We use the ADAM [66] optimizer with a learning rate of 0.0001 and a batch size of 256. We also trained *NAML<sub>T</sub>* - a "title only" version - as used in the original paper [119]. We obtained the same parameters as *NAML* without the need for category dimensions.

*NRMS* [118] - Neural NRS which utilizes multi-head self-attention within both the news encoder and the user encoder. We use multi-head self-attention with 15 attention heads followed by an additive-attention layer with a 200-dimensional query vector. We apply 20% dropout to the word embeddings. We set the negative sampling ratio  $K$  to 4. We use the ADAM [66] optimizer with a learning rate of 0.0001 and a batch size of 128.

### 5.1.3 Results

In this section, we present and analyze our results and answer our previously stated research questions. We investigate whether the reproduced models perform as described in the original paper and study:

**RQ3.1.1** *How does our reproduced SentiRec implementation compare to baselines on the MIND dataset?*

We compare the recommendation performance (i.e., *AUC*, *MRR*, *nDCG@5*, and *nDCG@10*) of the reproduced model (i.e., *SentiRec<sub>V</sub>*) against the baselines (i.e., *LSTUR* [4], *NAML* & *NAML<sub>T</sub>* [116], *NRMS* [118], and *Random*), which is summarized in rows 1-6 of Table 5.2. Opposing the original work, our sentiment reproduction does not significantly outperform all baselines concerning recommendation effectiveness. Moreover, it performs similarly to the closely related *NRMS* baseline.

Furthermore, we investigate sentiment diversity by comparing the sentiment alignment scores (i.e., *S<sub>MRR</sub>*, *S@5*, and *S@10* – lower scores indicate higher sentiment diversity) of our reproduced model, i.e., *SentiRec<sub>V</sub>*, and the baselines (see rows 1-6 in Table 5.3). In the original work [119], *SentiRec* outperforms all baselines in sentiment diversity - even the *Random* model - while maintaining the highest recommendation performance scores. We can not confirm these findings. Moreover, our results suggest that the baselines already perform well in all aspects, i.e., recommendation performance and sentiment diversity. In particular, we do not observe large margins in sentiment diversity as in the original

Table 5.2: Comparing effectiveness (i.e., AUC, MRR, nDCG@5, and nDCG@10). Higher effectiveness scores indicate better performance. Subscripts V (VADER-SA) and B (BERT-SA) indicate the used sentiment analyzer. Note, † indicates a statistically significant difference to *SentiRec<sub>V</sub>* at alpha 0.05.

Model		AUC	MRR	nDCG	
				@5	@10
1	Random	.4994 <sup>†</sup>	.2190 <sup>†</sup>	.2236 <sup>†</sup>	.2863 <sup>†</sup>
2	<i>NAML<sub>T</sub></i>	.6194	.2982	.3190	.3804
3	<i>NAML</i>	.6206	.2913 <sup>†</sup>	.3185	.3782 <sup>†</sup>
4	<i>LSTUR</i>	.6210 <sup>†</sup>	.2840 <sup>†</sup>	.3101 <sup>†</sup>	.3721 <sup>†</sup>
5	<i>NRMS</i>	<u>.6228</u>	.2946	.3191	.3817
6	<i>SentiRec<sub>V</sub></i>	.6224	<u>.2952</u>	<u>.3211</u>	.3818
7	<i>SentiRec<sub>B</sub></i>	.6219	.2942	.3203	<u>.3820</u>
8	<i>RobustSentiRec<sub>V</sub></i>	<b>.6243<sup>†</sup></b>	<b>.2979<sup>†</sup></b>	<b>.3238<sup>†</sup></b>	<b>.3847<sup>†</sup></b>
9	<i>RobustSentiRec<sub>B</sub></i>	.6211	.2930	.3193	.3815

paper. We argue that these discrepancies are due to dataset differences highlighting the shortcomings of SentiRec concerning generalizability. Our dataset contains five times more users and about 23K more news than the original paper; however, it contains relatively few positive feedback (i.e., clicks) and spans only over six weeks (compared to nine weeks). Thus, inherently more diverse behavior is contained in the used dataset than in the original paper.

Setting aside the random model, the *NAML* [116] model outperforms all other models regarding sentiment diversity while maintaining comparable recommendation performance to our SentiRec reproductions. Besides the title of a news article, it also considers category, subcategory, and abstract. Thus, we reason that considering different modalities supports the diversification task. Note, in the original paper *NAML* is fed with only one modality (i.e., title) - in this work denoted as *NAML<sub>T</sub>*. To further strengthen our understanding of the problem, we also investigate:

**RQ3.1.2** *What downstream effect does a more effective neural sentiment analyzer create in comparison to a rule-based one?*

In addition to the rule-based sentiment analyzer (VADER-SA), as utilized in [119], we conduct our experiments using a pre-trained neural sentiment analyzer (BERT-SA). Figure 5.3a shows the sentiment polarity score distribution of all news articles determined by VADER-SA and BERT-SA, respectively. VADER-SA assigns mostly neutral scores and positive and negative scores are almost balanced (mean  $-0.03$ , standard deviation 0.39). On the other hand, BERT-SA shows more classification characteristics by either assigning relatively positive or relatively negative scores and few neutral (mean  $-0.15$ , standard deviation 0.91). Figure 5.3b shows the overall sentiment orientation ( $\bar{s}$ ) of

Table 5.3: Comparing user-centric sentiment and topic alignment (i.e.,  $S_{MRR}$ ,  $S@5$ ,  $S@10$ ,  $T_{MRR}$ ,  $T@5$ ,  $T@10$ ). Lower alignment scores indicate better diversity. Subscripts V (VADER-SA) and B (BERT-SA) indicate the used sentiment analyzer. Note,  $\dagger$  indicates a statistically significant difference to  $SentiRec_V$  at alpha 0.05.

Model	VADER-SA Labels			BERT-SA Labels			$T_{MRR}$	$T@5$	$T@10$
	$S_{MRR}$	$S@5$	$S@10$	$S_{MRR}$	$S@5$	$S@10$			
1 <i>Random</i>	<b>.0086<math>\dagger</math></b>	<b>.0150<math>\dagger</math></b>	<b>.0188<math>\dagger</math></b>	<b>.1095<math>\dagger</math></b>	<b>.1748<math>\dagger</math></b>	<b>.2638<math>\dagger</math></b>	<b>.4315<math>\dagger</math></b>	<b>.3680<math>\dagger</math></b>	<b>.4428<math>\dagger</math></b>
2 <i>NAMLT</i>	.0157 $\dagger$	.0276 $\dagger$	.0382	.1741 $\dagger$	.2623 $\dagger$	.3933 $\dagger$	.5091 $\dagger$	.4570 $\dagger$	.5047 $\dagger$
3 <i>NAML</i>	<u>.0131<math>\dagger</math></u>	<u>.0210<math>\dagger</math></u>	<u>.0248<math>\dagger</math></u>	<u>.1132<math>\dagger</math></u>	<u>.1749<math>\dagger</math></u>	<u>.2936<math>\dagger</math></u>	<u>.4504<math>\dagger</math></u>	<u>.3744<math>\dagger</math></u>	<b>.4270<math>\dagger</math></b>
4 <i>LSTUR</i>	.0158 $\dagger$	.0281 $\dagger$	.0412 $\dagger$	.1655 $\dagger$	.2637 $\dagger$	.4297 $\dagger$	.4735 $\dagger$	.4220 $\dagger$	.4867 $\dagger$
5 <i>NRMS</i>	.0149 $\dagger$	.0282	.0390	.1317 $\dagger$	.2317 $\dagger$	.3869 $\dagger$	.4883	.4353	.4926 $\dagger$
6 <i>SentiRec_V</i>	.0161	.0284	.0386	.1300	.2153	.3651	.4872	.4328	.4891
7 <i>SentiRec_B</i>	.0174 $\dagger$	.0325 $\dagger$	.0449 $\dagger$	.1560 $\dagger$	.2675 $\dagger$	.4330 $\dagger$	.4905 $\dagger$	.4414 $\dagger$	.4942 $\dagger$
8 <i>RobustSentiRec_V</i>	.0169	.0292	.0440 $\dagger$	.1536 $\dagger$	.2678 $\dagger$	.4325 $\dagger$	.4909 $\dagger$	.4364 $\dagger$	.4939 $\dagger$
9 <i>RobustSentiRec_B</i>	.0163	.0288	.0437 $\dagger$	.1537 $\dagger$	.2679 $\dagger$	.4410 $\dagger$	.4936 $\dagger$	.4414 $\dagger$	.4962 $\dagger$

users' based on their history track. The users' overall sentiment orientation is primarily negative, and this is more prevalent using BERT-SA (mean  $-0.27$ , standard deviation  $0.33$ ) than VADER-SA (mean  $-0.08$ , standard deviation  $0.15$ ). We show that replacing the rule-based sentiment analyzer as used in the original paper with a more effective neural one does not yield performance gains (compare *SentiRec\_V* and *SentiRec\_B*, i.e., rows 6 and 7, in Table 5.2) but a significant drop in sentiment diversity (compare *SentiRec\_V* and *SentiRec\_B*, i.e., rows 6 and 7, in Table 5.3). To validate the loss regularization ablation of the original paper and to demonstrate the trade-off between effectiveness and sentiment diversity we study:

**RQ3.1.3** *What influence do sentiment prediction and sentiment diversity loss regularization hyperparameters have on the resulting sentiment diversity and recommendation performance?*

While acknowledging this addresses hyperparameter optimization, an engineering concern, we frame it as a research question for narrative clarity. Following [119] we first omit the sentiment regularization task and tune the sentiment prediction loss regularization hyperparameter  $\lambda$ . Figure 5.4 shows the influence of  $\lambda$  in both settings, *SentiRec\_V* and *SentiRec\_B*. Opposing the original work's findings (in both settings), increasing the influence of the sentiment prediction task yields overall to a higher alignment of the recommended news sentiment and the users' sentiment orientation, and thus, decreases sentiment diversity. We argue that by infusing more sentiment awareness into the news and consequently into the user representation, we move news and users with similar sentiment orientation to closer proximity in their shared vector space. In line with the original paper, a too strong influence of the sentiment prediction task leads to a drop in recommendation performance. Therefore, we set  $\lambda = 0.4$  for both models.

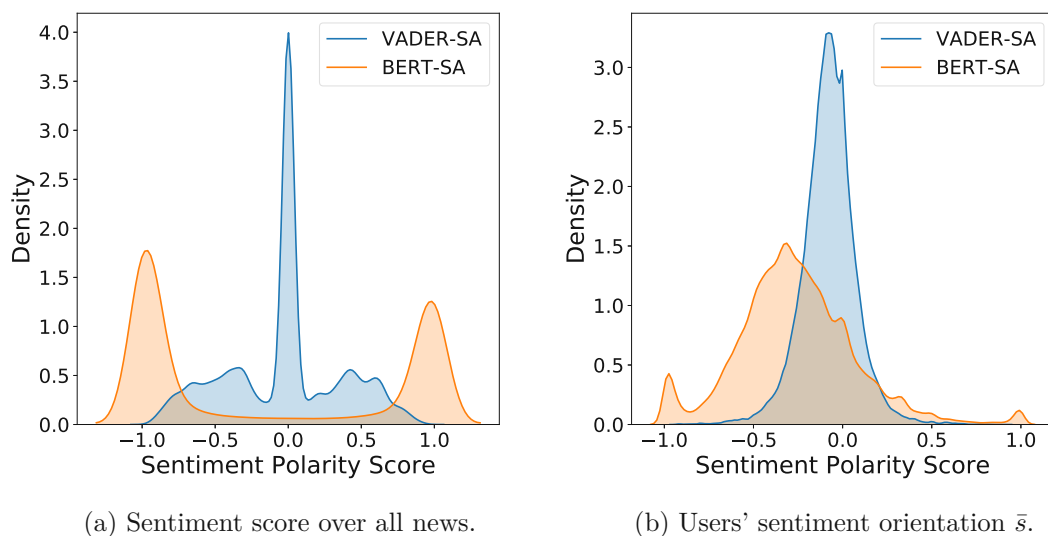
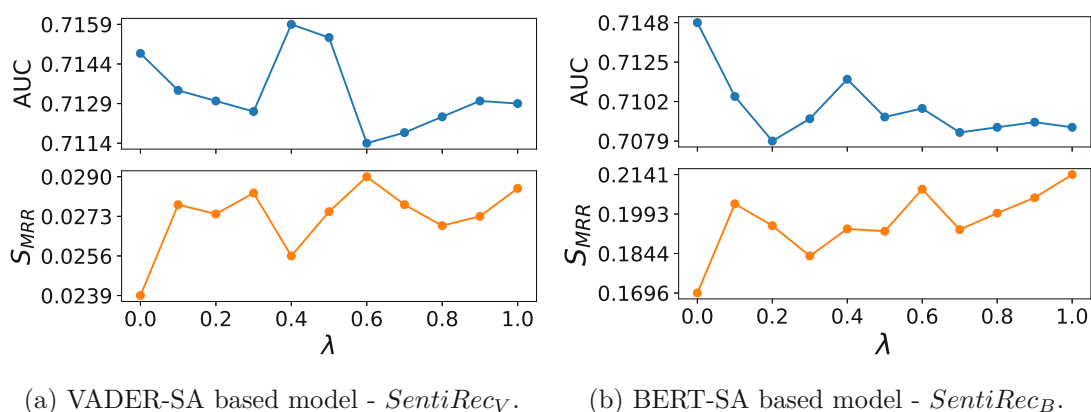
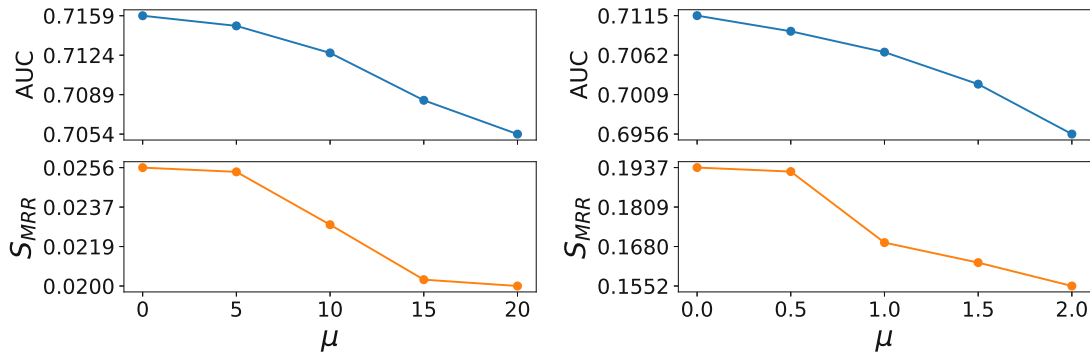


Figure 5.3: Sentiment polarity score distribution in MIND-small.

Figure 5.4: Influence of sentiment prediction loss hyperparameter  $\lambda$ .

In a second step, we search for the sentiment diversity loss regularization hyperparameter  $\mu$  by fixing  $\lambda = 0.4$ . Figure 5.5 shows the influence of  $\mu$  for both models,  $SentiRec_V$  and  $SentiRec_B$ . In both settings, increasing  $\mu$  leads to a steady improvement of sentiment diversity, i.e., drop of  $S_{MRR}$ , as anticipated and in line with the original work's findings. However, Figure 5.5 demonstrates the trade-off between sentiment diversity and recommendation performance. A too stark regularization of sentiment diversity might ultimately lead to a drop in user satisfaction. Thus, we moderately regularize sentiment diversity by setting  $\mu = 10$  for  $SentiRec_V$  and  $\mu = 1$  for  $SentiRec_B$ . In addition to the original work, which only investigates diversity based on sentiment, we study:

**RQ3.1.4** How does our reproduced *SentiRec* implementation compare to the *MIND*

(a) VADER-SA based model - *SentiRec<sub>V</sub>*.(b) BERT-SA based model - *SentiRec<sub>B</sub>*.Figure 5.5: Influence of sentiment regularization loss hyperparameter  $\mu$  under  $\lambda = 0.4$ .

### baselines concerning topical diversity?

We adapt the sentiment alignment metrics and introduce topical alignment metrics, i.e.,  $T_{MRR}$  and  $T@K$ , by considering the categorical membership of the news articles. Lower topical alignment metrics indicate higher topical diversity. The last three columns of Table 5.3 summarize our analysis. The *Random* model recommends the most topically diverse news articles to the users' previously browsed news articles, except if the top 10 recommendations are considered, where the *NAML* model excels. The *NAML* and the *LSTUR* baselines significantly reach better topical diversity than our reproduced *Sentirec* models while maintaining reasonable recommendation performance – demonstrating the competitiveness of the baseline models. As previously, we argue that additional modalities may foster the diversification of news.

The sentiment alignment metrics and topical alignment metrics are user-centric; in other words, they highlight the sentiment and topical differences of the users' previous interaction history to the recommendation list. However, they do not yield any information about sentiment diversity or topical diversity within the list; thus, we investigate:

### RQ3.1.5 How does our reproduced *SentiRec* implementation compare to the *MIND* baselines concerning intra-list sentiment-diversity and intra-list topical-diversity?

We compute the intra-list sentiment similarity at cutoff  $K$ , i.e.,  $ILS_S@K$ , and intra-list topical similarity at cutoff  $K$ , i.e.,  $ILS_T@K$ , by considering the pairwise differences of news articles within a top  $K$  recommendation list. Table 5.4 (rows 1-7) summarizes our outcomes. A lower intra-list similarity score indicates better diversity. In contrast to our user-centric diversity findings, where the baselines already exhibit decent performance, we observe that our reproduced model, i.e., *Sentirec<sub>V</sub>*, significantly outperforms most baselines concerning intra-list sentiment diversity. In comparison, the *NAML* baseline shows the worst performance. Suggesting that additional modalities might foster user-centric diversity (see Table 5.3) but hurt intra-list diversity by recommending top  $K$  news articles with a rather higher topic and sentiment similarity.

Finally, we introduce RobustSentiRec, a modified version of SentiRec which aims for a

Table 5.4: Comparing sentiment- and topic-based intra-list similarity (i.e.,  $ILS_S@5$ ,  $ILS_S@10$ ,  $ILS_T@5$ ,  $ILS_T@10$ ). Lower intra-list similarity scores indicate better diversity. Subscripts V (VADER-SA) and B (BERT-SA) indicate the used sentiment analyzer. Note,  $\dagger$  indicates a statistically significant difference to  $SentiRec_V$  at alpha 0.05.

Model	VADER-SA Labels		BERT-SA Labels		$ILS_T@5$	$ILS_T@10$
	$ILS_S@5$	$ILS_S@10$	$ILS_S@5$	$ILS_S@10$		
1 <i>Random</i>	.2393 $\dagger$	.2394 $\dagger$	.5047 $\dagger$	.5045 $\dagger$	<b>.0774<math>\dagger</math></b>	<b>.0775<math>\dagger</math></b>
2 <i>NAMLT</i>	.2336 $\dagger$	.2377 $\dagger$	.4770 $\dagger$	.4863 $\dagger$	.1396 $\dagger$	.1089 $\dagger$
3 <i>NAML</i>	.2600 $\dagger$	.2480 $\dagger$	.5221 $\dagger$	.5049 $\dagger$	.3377 $\dagger$	.1886 $\dagger$
4 <i>LSTUR</i>	.2313	.2347	.4826 $\dagger$	.4826	.1223 $\dagger$	.1026
5 <i>NRMS</i>	.2376 $\dagger$	.2393 $\dagger$	.4700	.4819	.1290	.1016
6 <i>SentiRec_V</i>	<u>.2310</u>	<b>.2337</b>	.4682	<u>.4812</u>	.1289	<u>.1013</u>
7 <i>SentiRec_B</i>	.2423 $\dagger$	.2404 $\dagger$	<u>.4444<math>\dagger</math></u>	<u>.4648<math>\dagger</math></u>	.1429 $\dagger$	.1063 $\dagger$
8 <i>RobustSentiRec_V</i>	<b>.2286<math>\dagger</math></b>	<u>.2339</u>	.4536 $\dagger$	.4697 $\dagger$	.1344 $\dagger$	.1036
9 <i>RobustSentiRec_B</i>	.2329	.2369 $\dagger$	<b>.4349<math>\dagger</math></b>	<b>.4641<math>\dagger</math></b>	.1316 $\dagger$	.1023

more robust sentiment diversity, and investigate:

**RQ3.1.6** *To what extent can the sentiment prediction task be omitted and instead the sentiment-labels be used as an addition signal?*

We introduce RobustSentiRec, a modified version of SentiRec, where we omit the auxiliary sentiment prediction task. Instead, we use the sentiment labels determined by the sentiment analyzer as additional input (see Section 5.1.1). We compare the effectiveness of our reproduced model, i.e.,  $SentiRec_V$ , against  $RobustSentiRec$  in both settings, i.e.,  $RobustSentiRec_V$  and  $RobustSentiRec_B$ , in Table 5.2 (rows 6 vs. 8-9), the user-centric sentiment and topical alignment in Table 5.3 (rows 6 vs. 8-9), and the intra-list similarity in Table 5.4 (rows 6 vs. 8-9).  $RobustSentiRec_V$  significantly outperforms  $SentiRec_V$  in effectiveness while reaching comparable user-centric sentiment diversity results if the VADER-SA labels are considered. Moreover,  $RobustSentiRec$  in both configurations yields significantly better intra-list sentiment-diversity than  $SentiRec_V$ . Like  $SentiRec$ , also  $RobustSentiRec$  achieves better performance if the rule-based sentiment analyzer is utilized. However, we also observe that  $SentiRec_V$  outperforms  $RobustSentiRec$  concerning intra-list topical diversity if top-5 recommendations are considered. Overall, we conclude that by directly incorporating the sentiment labels, we omit the ambiguity of the sentiment prediction and consequently the related information loss, which leads to the performance gains. However, we acknowledge the end-to-end encoding capabilities of  $SentiRec$ , once trained.

### 5.1.4 Discussion

Overall, we cannot confirm the findings of the original work, where they outperformed all baselines in effectiveness and user-centric sentiment diversity. We argue that the effectiveness and diversity discrepancies between the original SentiRec and our reproduction are due to dataset differences highlighting the shortcomings of *SentiRec* concerning generalizability. Our dataset contains five times more users and about 23K more news than the original paper; however, it contains relatively few positive feedback (i.e., clicks) and spans only over six weeks (compared to nine weeks). Thus, inherently more diverse behavior is contained in the used dataset than in the original paper. One might argue that the sentiment diversity issue in our sample is not as prevalent as in the sample of the original work. However, we demonstrate that the *NAML* baseline significantly outperforms our reproduction and gets close to the *Random* model’s performance. This highlights that there is room for improvement, which is not utilized by the *SentiRec*’s diversification approach.

As mentioned, the *NAML* [116] model outperforms all other models (except the *Random* model) regarding user-centric sentiment diversity while maintaining comparable recommendation performance to our *SentiRec* reproductions. Besides the title of a news article, it also considers category, subcategory, and abstract. Thus, we reason that considering different modalities supports the diversification task. Note, in the original paper *NAML* is fed with only one modality (i.e., title) - in this work denoted as *NAML<sub>T</sub>*.

Besides the user-centric view of sentiment diversity, we also analyze a more generic perspective, i.e., intra-list sentiment diversity. We demonstrate that our reproduction achieves an outstanding intra-list sentiment diversity, although optimized for user-centric sentiment diversity. Setting both perspectives alongside opens the room for the following question, which we will tackle in future work: Which view of sentiment diversity should we optimize while maintaining user satisfaction? Optimizing for the user-centric perspective is more conservative. This will rank news articles with an orthogonal sentiment to the overall sentiment of the user’s news consumption higher. Such an approach has a strong nudging power but might drop user satisfaction by recommending more the “unusual”. On the other hand, optimizing for the intra-list perspective is more relaxed by suggesting news articles with different sentiments. However, it bears the risk that users might still follow their previous behavior and consume, for example, only negative news articles.

Our last evaluation perspective, which the original work does not consider, is topical diversity. In particular, we consider categorical differences between recommended news articles and the users’ browsed news, i.e., user-centric topical diversity and categorical differences within the news articles in the recommendation list, i.e., intra-list topical diversity. In both measures, the *Random* model achieves the most topically diverse recommendations. Setting aside the *Random* model, while in the user-centric perspective, our reproduction *Sentirec<sub>V</sub>* is outperformed by most baselines, in the intra-list perspective, it is on par or better than the baselines. With different sentiment distributions within news categories, we plan to analyze whether topical diversification already yields

sentiment diversification and higher user satisfaction in future work.

Moreover, we introduced RobustSentiRec. Instead of relying on the auxiliary sentiment prediction task, it integrates sentiment labels directly from a sentiment analyzer. RobustSentiRec shows promising results in efficacy and diversity compared to SentiRec. The key distinction here is the strategic approach to sentiment handling: RobustSentiRec sidesteps potential ambiguities in sentiment prediction, resulting in a more streamlined performance. Yet, this distinction comes with implications for practical implementation. Choosing between SentiRec and RobustSentiRec hinges on specific needs. Opting for SentiRec has the advantage of a standalone, integrated system, where only one model is required for recommendations. On the other hand, the allure of RobustSentiRec lies in its simplified architecture, which, while eliminating the sentiment prediction phase during training, necessitates an additional sentiment analyzer during the recommendation process. Such considerations are crucial when deciding on a model based on efficiency, simplicity, and deployment constraints.

### 5.1.5 Conclusion

This research embarked on the reproduction of SentiRec [119], a sentiment diversity-aware neural news recommendation model, absent the privilege of accessing the original source code and dataset. We re-implement SentiRec from scratch and make it publicly available. We use the MIND [124] dataset, which has the same source as the original paper, albeit a different time period. Overall, we can not confirm the significant findings of the SentiRec paper. The reproduced model does not outperform the random model in (user-centric) sentiment diversity while maintaining the best recommendation performance compared to the baselines as in the original work. Moreover, our results suggest that the baselines already perform well. In particular, the *NAML* [116] model delivers the most sentiment-diverse recommendations (w.r.t. to the users' overall consumption behavior) apart from the random model while holding a comparable recommendation performance to all other baselines. We conclude that these discrepancies are due to dataset differences high-lighting the shortcomings of SentiRec concerning generalizability.

In addition to the original paper, we also consider the topical diversity of the recommended list compared to the users' previous user history. Similar to previously, we show that the baselines, particularly *NAML*, significantly yield better topical diversity than our reproduced *Sentirec* model.

In contrast to the original paper, we observe a decrease in sentiment diversity (user-centric) if the influence of the sentiment prediction task increases. We argue that infusing more sentiment awareness into the news and user representations moves users and news with similar sentiment orientation closer in the shared vector space. In line with the original paper, we demonstrate that an overly strong setting of the influence of the sentiment prediction and diversity regularization loss leads to a drop in recommendation performance.

In addition to a rule-based sentiment analyzer, as used by Wu et al. [119], we conducted our experiments with a pre-trained neural sentiment analyzer to study whether a neural model leads to better sentiment labels and thus to improved overall training performance. However, we do not observe improvements in recommendation performance or sentiment diversity.

While the original paper only focuses on sentiment diversity by comparing the users' overall user history with the recommendation list (i.e., user-centric diversity), we also investigate the sentiment and the topical diversity between news articles within the recommendation list (intra-list diversity). In contrast to the user-centric evaluation, the intra-list evaluation shows that our *SentiRec* reproduction significantly outperforms most baselines, while the strong *NAML* baseline performs poorly.

Finally, we propose RobustSentiRec, a modified version of SentiRec, intending to increase sentiment diversity robustness. Our RobusSentiRec model significantly outperforms SentiRec in recommendation performance and intra-list sentiment diversity while achieving comparable user-centric sentiment diversity and overall benefits in reduced model complexity.

Looking ahead, our vision is to integrate diverse auxiliary information into end-to-end recommendation models, like personality or emotion awareness/diversity. Our ultimate aspiration is to craft recommendation models that harmonize a spectrum of goals, aiming to enrich society with more thoughtful and responsible recommendations.

### 5.1.6 Limitations

While this study offers several advancements and insights into sentiment-aware news recommendation, certain limitations warrant discussion. Firstly, our reproduction of SentiRec was based on a different dataset (MIND) than the one used in the original study, due to the unavailability of the original dataset. While both datasets stem from the same source, differences in time frame and data characteristics might be the reason of the observed discrepancies in results. Secondly, the reliance on sentiment analysis, whether rule-based (VADER-SA) or neural (BERT-SA), introduces a potential source of error. The accuracy and nuances of sentiment detection can significantly impact the model's performance, and inherent limitations in these sentiment analysis tools could affect the outcomes. Especially, the sentiment might vary among different user groups. Thirdly, our study primarily focuses on sentiment and topical diversity, but other diversity aspects, such as novelty or serendipity, are not explicitly addressed. Moreover, this study did not address long-term user satisfaction which might be influenced by getting more or less sentiment diverse news. Finally, while RobustSentiRec offers a simplified architecture, it still depends on an external sentiment analyzer, introducing a dependency that might not be present in a purely end-to-end model.

## 5.2 Exploring Expressed Emotions

Personalized NRS help manage the vast number of daily news items by creating user models and generating personalized suggestions, addressing challenges like rapid turnover and *cold-start* problems. These systems must consider both the semantic content and the emotional tones of news articles to effectively engage readers

In this section, we focus on the emotions conveyed in news articles, defined as the emotions the article’s creator expressed during its creation [68]. We examine how incorporating these emotional elements, along with semantic content, affects both recommendation performance and the emotional diversity of suggested articles. We introduce a multi-level emotion-aware news recommendation framework that considers emotions at the title, abstract, category, and subcategory levels. For example, for a news article under the “sports” category and the “soccer” subcategory, our framework aggregates the emotions extracted from the titles and abstracts of all articles in these categories and subcategories to determine the overall emotions at the category and subcategory levels. We did not use the body of the articles because the title, abstract, category, and subcategory are the first elements users see, which form the basis for their implicit feedback (i.e., whether they click or not) in the dataset used.

Despite being subtle, expressed emotions are distinguishing features of news articles. We hypothesize that integrating these emotions in recommendations will lead to better alignment with users’ preferences, potentially resulting in reduced emotional diversity in recommendations. Thus, we aim to scrutinize the extent of such alignments and additionally, observes topical diversity, acknowledging that different topics possess unique emotional distributions.

Existing research supports the idea that incorporating emotions can offer a more in-depth understanding of users, thereby enhancing the quality of recommendations [68]. While the role of emotion has been explored in various RSs contexts, its application in NRS has received relatively less attention. Previous works, such as those by Mizgajski and Morzy [74], have focused on users’ self-reported emotions and their reading choices. In contrast, our approach exploits the emotions conveyed through the news content itself and models users based on their implicit feedback.

We also expand on previous studies Babanejad et al. [6], Wu et al. [119] by employing three distinct emotion taxonomies of varying granularity, including the seven-dimensional Ekman taxonomy [33], basic sentiments, and the more nuanced GoEmotion taxonomy [26]. Unlike the Ekman taxonomy, which features a single positive emotion (joy), the GoEmotion taxonomy allows for the distinction between various positive, negative, neutral, and ambiguous emotions. This granularity aims to capture and utilize the nuances of emotional variation more effectively.

Specifically, for our proposed multi-level emotion-aware news recommendation framework, we study the following research questions:

**RQ3.2.1** *To what extent does the incorporation of emotions and the usage of different*

*emotion taxonomies affect recommendation performance?*

Our findings indicate notable performance enhancements when emotions are integrated into the recommendation process. The proposed model consistently delivers superior results compared to existing state-of-the-art neural news recommendation models. Our ablation study reveals that both text-based (title and abstract level) emotions and category-based (category and subcategory level) emotions contribute to this improvement, with text-based emotions having the most substantial impact. Moreover, we observe that a coarse emotion taxonomy is more effective than its finer-grained counterparts.

Besides recommendation performance, we are also interested in emotional diversity since incorporating emotions risks decreasing the recommendations' emotional diversity. Thus, we investigate:

**RQ3.2.2** *To what extent does the incorporation of emotions and the usage of different emotion taxonomies affect emotional diversity?*

To evaluate emotional diversity, we introduce two distinct metrics. Our results indicate that the integration of emotions tends to reduce the emotional diversity of recommendations. According to our ablation study, the inclusion of text-based (title and abstract level) emotions results in the most significant reduction in emotional diversity. Notably, our model utilizing the Ekman taxonomy produces more emotionally diverse recommendations than other variations of our model, although it does not surpass the non-emotion-aware baseline models in this regard.

As topics inherently possess different emotional distributions, it is essential to explore the potential consequences of emotion integration on the variety of topics recommended to users. We seek to understand how emotion-aware recommendations might shift the topical landscape presented to readers. With this in mind, we investigate:

**RQ3.2.3** *To what extent does the incorporation of emotions and the usage of different emotion taxonomies affect topical diversity?*

Incorporating emotions into the recommendation process generally leads to a decrease in topical diversity. Regardless of whether these emotions are derived from text or categories, the inclusion of emotions tends to narrow the range of topics that are recommended to users. This decrease in topical diversity might be due to the matchmaking, where the user's emotional profile matches (to some extent) with the overall emotional profile of a topic (category/subcategory). Interestingly, when both text-based and category-based emotions are integrated, some configurations of our model are able to counterbalance this reduction to a degree, resulting in recommendations that are more topically diverse than when emotions are excluded entirely. Nonetheless, a random recommendation model consistently outperforms our emotion-aware models in terms of delivering the most topically diverse recommendations.

Our contributions can be summarized as follows:

- We propose a multi-level emotion-aware neural news recommendation framework, which disentangles semantic and emotional user and item modeling; and incorporates

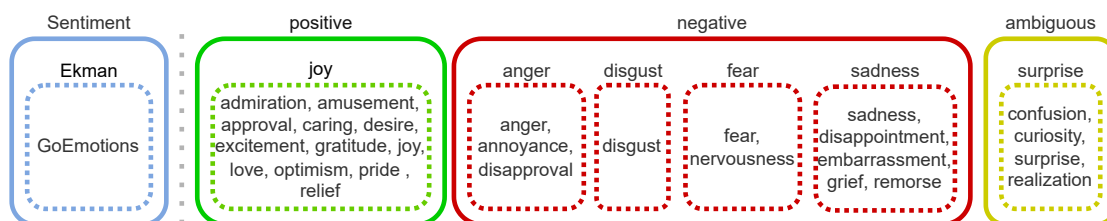


Figure 5.6: Hierarchical structure of the emotion taxonomies used, as described by Demszyk et al. [26]. Sentiments (positive, negative, ambiguous) are broken down into the Ekman taxonomy [33] (joy, anger, disgust, fear, sadness, surprise), which are further detailed in the GoEmotions taxonomy [26]. Note that the neutral dimension is consistent across all taxonomies and is not depicted here.

emotions extracted from different levels (i.e., title, abstract, category, subcategory) into the recommendation process.

- We introduce user-centric and intra-list emotional diversity measures to assess whether incorporating emotions yields emotional diversity deficits.
- We introduce user-centric and intra-list topical diversity measures to assess whether incorporating emotions yields topical diversity deficits.
- Extensive experiments on a real-world dataset demonstrate that our approach significantly outperforms state-of-the-art neural recommendation models. However, we also observe that incorporating emotions does yield emotional and topical diversity decreases.
- We investigate the impact of three distinct emotion taxonomies of different granularity and show that the approach considering the coarse taxonomy significantly outperforms all other models.
- We publish our code (including all baselines, configs, pre-processing steps, and documentation) under: <https://github.com/MeteSertkan/EmoRec>

### 5.2.1 Methods

#### Incorporating Emotions

**A Multi-Level Perspective on Expressed Emotions in News Articles.** News articles are structured into various components, including the title, abstract, category (e.g., *sports*”), and subcategory (e.g., *soccer*”). In our approach, we utilize this inherent hierarchical structure of news articles to analyze emotions at different levels. For each level—namely, the title, abstract, category, and subcategory—we compute and index emotion scores, denoted as  $e_T$ ,  $e_A$ ,  $e_C$ , and  $e_S$ . These scores are computed with varying levels of granularity, represented as  $\mathbb{R}^4$ ,  $\mathbb{R}^7$ , and  $\mathbb{R}^{28}$  (refer to Figure 5.6 for details). Our BERT-based emotion classifiers are employed to directly extract emotions present in

the title ( $e_T$ ) and abstract ( $e_A$ ) of each article. To calculate the emotion scores at the category and subcategory levels, we aggregate the emotion scores from the titles and abstracts of all articles belonging to a specific category or subcategory. Formally, for each category or subcategory  $X$ , we aggregate the title and abstract emotions— $e_T^D$  and  $e_A^D$ , respectively—of news articles  $D$  that are part of  $X$ . This aggregation is mathematically represented as:

$$e_X = \frac{1}{2|X|} \sum_{D \in X} e_T^D + e_A^D, \quad (5.11)$$

where  $|X|$  denotes the number of articles in the category or subcategory  $X$ ,  $e_X$  represents the aggregated emotion score for category or subcategory  $X$ ,  $e_T^D$  represents the emotion score derived from the title of news article  $D$ , and  $e_A^D$  represents the emotion score derived from the abstract of news article  $D$ . The summation iterates over all news articles  $D$  within the category or subcategory  $X$ . The result is then normalized by the total number of articles in  $X$  multiplied by two (to account for both title and abstract contributions).

**Task** The overall task is to rank a list of candidate news articles that a user might be interested in, based on that user’s past interactions with news articles. Let us consider a user  $u$  who has a browsing history  $H$ , consisting of  $M$  previously read news articles denoted as  $[D_1^H, \dots, D_M^H]$ . Our objective is to assign a score to each article in a set  $C$  of  $P$  candidate news articles, denoted as  $[D_1^C, \dots, D_P^C]$ . These scores, represented as  $[\hat{y}_1, \dots, \hat{y}_P]$ , are used to rank the candidate articles in order of relevance to the user.

To make our recommendation model more insightful and personalized, we integrate emotion awareness into the process. This is achieved by incorporating emotion scores at various levels—namely, title-level  $e_T$ , abstract-level  $e_A$ , category-level  $e_C$ , and subcategory-level  $e_S$ —for both the previously browsed articles  $[(e_T, e_A, e_C, e_S)_1^H, \dots, (e_T, e_A, e_C, e_S)_M^H]$  and the candidate articles  $[(e_T, e_A, e_C, e_S)_1^C, \dots, (e_T, e_A, e_C, e_S)_P^C]$ .

As already mentioned, for each article, we pre-compute and index:

- Title-level emotions, denoted as  $e_T \in \mathbb{R}^{d_e}$ ,
- Abstract-level emotions, denoted as  $e_A \in \mathbb{R}^{d_e}$ ,
- Category-level emotions, denoted as  $e_C \in \mathbb{R}^{d_e}$ , and
- Subcategory-level emotions, denoted as  $e_S \in \mathbb{R}^{d_e}$ ,

where  $d_e$  represents the dimensionality of the emotion taxonomy used in our model.

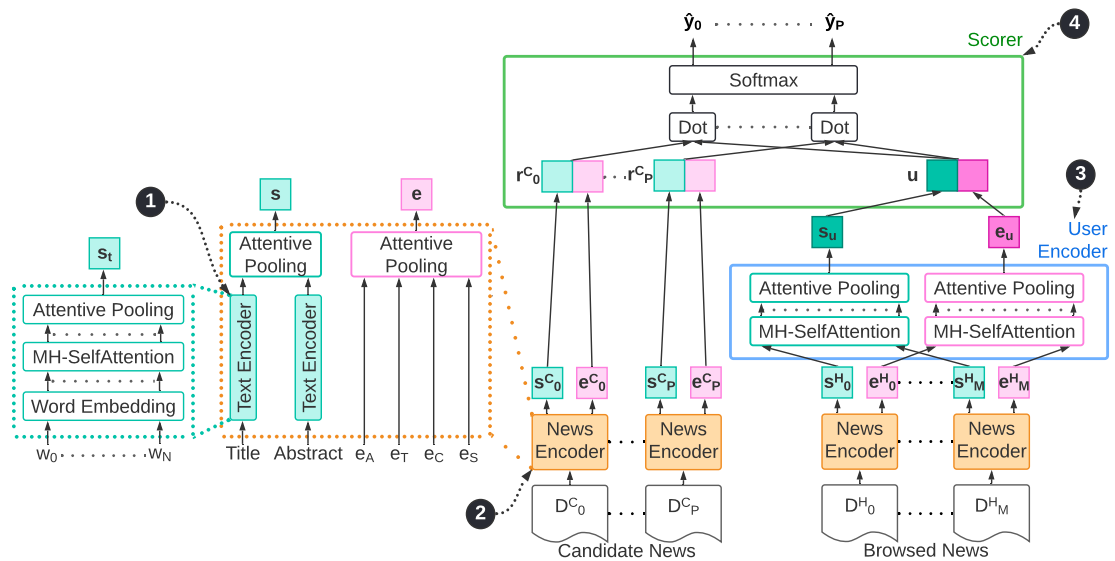


Figure 5.7: Overview of our recommendation framework comprising following major components: **1** *Text Encoder*, which learns a semantic representation  $s_t$  of any given sequence of words; **2** *News Encoder*, which utilizes the Text Encoder to obtain a semantic representation  $s$  of a news article by its title and abstract; and which combines the pre-computed emotional representations  $e_T$ ,  $e_A$ ,  $e_C$ , and  $e_S$  (i.e., title-, abstract-, category-, and subcategory-emotions) of a news article, to one representation  $e$ ; **3** *User Encoder*, which separately models a semantic representation  $s_u$  and an emotional representation  $e_u$  of users based on their previous news interactions; **4** *Scorer*, which determines a score for a given user and candidate news pair. Note that the final representation of candidate news  $r_i^C$  and users  $u$  are simply the concatenation of their corresponding semantic and emotional representations.

**Recommendation Framework** Our recommendation framework constructs distinct semantic and emotional profiles for news articles based on their titles, abstracts, and multi-level emotion scores. Similarly, for user modeling, we independently handle the semantic and emotional representations of the articles with which the user has previously interacted. In both cases, this separation is designed to retain the interpretability of the emotional dimensions.

For the final representation of each news article, we concatenate its semantic and emotional profiles. Likewise, for users, we combine the separate semantic and emotional profiles of their previously interacted articles to create a comprehensive user representation.

To suggest personalized articles for a user, we compute the score for each candidate news article. This is done by calculating the dot product between the user’s representation and the representations of the candidate news articles. Based on these scores, we then rank the candidate articles.

Next, we delve into the key components of our recommendation framework, which are visually depicted in Figure 5.7.

① *Text Encoder.* The text encoder is used to learn a semantic representation of a given word sequence  $[w_1, \dots, w_N]$ . The word embedding layer maps the given sequence into a sequence of low-dimensional embeddings  $[\omega_1, \dots, \omega_N]$ . We use multi-head self-attention [111] to put each word into the context of each other in the sequence and thus obtain contextualized word embeddings  $[\omega_1^*, \dots, \omega_N^*]$ . In order to obtain a unified semantic representation  $s_t$  of the given word sequence, we pool the contextualized word embeddings using an attention mechanism. First, we compute for each word an attention weight  $\alpha_i$  as follows

$$a_i = q_t^T \tanh(V_t \omega_i^* + v_t), \quad \alpha_i = \frac{\exp(a_i^w)}{\sum_{j=1}^N \exp(a_j^w)}, \quad (5.12)$$

where  $a_i$  represents the unnormalized attention score for the  $i$ -th word,  $q_t$ ,  $V_t$ , and  $v_t$  are learnable parameters of the *text encoder*,  $\omega_i^*$  is the contextualized word embedding for the  $i$ -th word, and  $\alpha_i$  represents the normalized attention weight for the  $i$ -th word, obtained by applying the softmax function to the unnormalized attention scores.

Then, we take the weighted sum of the contextualized word embeddings to obtain the semantic representation  $s_t$  of the given word sequence as follows:

$$s_t = \sum_{i=1}^N \alpha_i \omega_i^* \quad (5.13)$$

where  $s_t$  is the final semantic representation of the text,  $\alpha_i$  represents the normalized attention weight for the  $i$ -th word (as defined in Equation 5.12), and  $\omega_i^*$  is the contextualized word embedding for the  $i$ -th word.

② *News Encoder.* The news encoder utilizes the *text encoder* to compute a semantic representation  $s_T$  given a news article's title and a semantic representation  $s_A$  given its abstract. Both are then combined using an attentive pooling mechanism – similarly as described in Equations 5.12 and 5.13 – to obtain the semantic representation  $s$  of the news article. In addition to a news article's semantic representation, the news encoder also learns and produces its emotional representation  $e$ . As already mentioned, we pre-compute emotion vectors of multiple levels for each news article, i.e., title-level  $e_T$ , abstract-level  $e_A$ , category-level  $e_C$ , and subcategory-level  $e_S$ . We combine those multi-level emotional representations of a news article using attentive pooling (like in Equation 5.12 and 5.13) to obtain its unified emotional representation  $e$ .

③ *User Encoder.* The user encoder learns a semantic representation  $s_u$  and an emotional representation  $e_u$  of users based on their interaction history  $H$ . We use multi-head self-attention to put each news' semantic representation  $s_i^H$  into the context of each others' in the interaction history  $H$ . We then pool the contextualized semantic representations  $s_i^{*H}$  using an attention mechanism, following a similar approach as previously, to obtain

the user’s semantic representation  $s_u$ . Similarly, we use multi-head self-attention followed by attentive pooling to obtain the user’s emotional representation  $e_u$  by utilizing each news’ emotional representation  $e_i^H$ .

④ *Scorer Encoder.* We obtain the final candidate news representations  $r_i^C$  and the final user representation  $u$  by concatenating their corresponding semantic and emotional representations, i.e.,  $r_i^C = \text{concat}(s_i^C, e_i^C)$  and  $u = \text{concat}(s_u, e_u)$ . In this way, we retain the interpretability of the emotion scores. We compute the ranking score (click probability)  $\hat{y}_i$  of each candidate news article  $D_i^C$  by utilizing the dot-product between user and candidate news representations and a softmax function:

$$\hat{y}'_i = u^T r_i^C, \quad \hat{y}_i = \frac{\exp(\hat{y}'_i)}{\sum_{j=0}^P \exp(\hat{y}'_j)} \quad (5.14)$$

where  $\hat{y}'_i$  is the unnormalized ranking score,  $u$  represents the final user representation,  $r_i^C$  represents the final representation of the  $i$ -th candidate news article, and  $\hat{y}_i$  is the normalized ranking score (probability) for the  $i$ -th candidate news, obtained through a softmax function.  $P$  is the number of candidate news articles.

**Model Training** Exploiting negative feedback has been shown to improve recommendation performance [119]. Following [119], we use negative sampling to train our models. Thus, for each clicked news article, i.e., positive sample, we randomly sample  $K$  news articles from the same session seen but not clicked, i.e., negative samples. To train our models, we minimize the negative log-likelihood of the clicked news articles (i.e., positive samples):

$$\mathcal{L} = - \sum_{y_j \in S} \log(\hat{y}_j), \quad (5.15)$$

where  $\mathcal{L}$  denotes the loss function,  $y_j$  is the actual label (1 for clicked news, 0 otherwise) of the  $j$ -th sample,  $S$  is the set of all positive training samples (i.e., clicked news articles), and  $\hat{y}_j$  represents the predicted probability that the  $j$ -th sample is a positive sample (i.e., will be clicked). We employ the ADAM [66] optimizer to minimize our loss. Furthermore, we train three different models based on the three distinct emotion taxonomies, i.e., **Sentiment**, **Ekman**, and **GoEmotion** (see Section 5.2.1). We call our (multi-level **emotion**-aware news **re**commendation) models  $EmoRec_S$ ,  $EmoRec_E$ , and  $EmoRec_G$ , where the subscripts correspond to the first letter of the used taxonomy.

### Diversity Metrics

We employ following diversity metrics: user-centric emotional diversity  $E_{UCD}$ , intra-list emotional diversity  $E_{ILD}$ , user-centric topical diversity  $T_{UCD}$ , and intra-list topical diversity  $T_{ILD}$ . While the intra-list diversity metrics compare news articles within the

recommended list, the user-centric metrics put them in contrast to the users' previous consumption behavior. We take the cosine distance as the basis for our diversity metrics:

$$\text{dist}(v_{source}, v_{target}) = 1 - \frac{v_{source} \cdot v_{target}}{\|v_{source}\| \|v_{target}\|} \quad (5.16)$$

where  $\text{dist}(v_{source}, v_{target})$  represents the cosine distance between two vectors,  $v_{source}$  and  $v_{target}$ .  $v_{source} \cdot v_{target}$  denotes the dot product of the two vectors, and  $\|v_{source}\|$  and  $\|v_{target}\|$  represent the magnitudes (Euclidean norms) of the respective vectors. Depending on the computed metric,  $v_{source}$  and  $v_{target}$  are either emotion vectors of dimension 4, 7, or 28 (depending on the considered emotion taxonomy) or category embeddings of dimension 100.

**Emotional Diversity.** In comparing emotion-aware and non-emotion-aware recommendation models, we exclude the learned user emotion representation  $e_u$  and the weights used to combine various views such as title, abstract, category, and subcategory. We extract the emotion representation  $e_{TA}^D$  of a news article  $D$  using BERT-based classifiers, taking the title and abstract as input. This approach is consistent with the majority of baselines that rely solely on text. We then average emotion representations of all news articles in a user's history  $H$  to form the user's overall emotion representation, denoted as  $\bar{e}_u$ . Taking all this into account and given a ranked recommendation list  $L$  with  $R$  articles  $[D_0, \dots, D_R]$ , we define the intra-list emotional diversity  $E_{ILD}$  as the average pairwise distance at cutoff  $K$ :

$$E_{ILD@K} = \frac{2 \sum_{D_i \in L@K} \sum_{D_j \in L@K \setminus \{D_i\}} \text{dist}(e_{TA}^{D_i}, e_{TA}^{D_j})}{K(K-1)}. \quad (5.17)$$

where  $E_{ILD@K}$  is the intra-list emotional diversity at cutoff  $K$ ,  $L@K$  represents the list of recommended items at cutoff  $K$ ,  $D_i$  and  $D_j$  are individual news articles within the recommendation list  $L$ ,  $e_{TA}^{D_i}$  and  $e_{TA}^{D_j}$  are the emotion representations of articles  $D_i$  and  $D_j$  (extracted from their title and abstract), and  $\text{dist}$  is the cosine distance function as defined in Equation 5.16.

It provides insight into the emotional diversity of the ranked lists; greater diversity results in a higher  $E_{ILD@K}$ . Similarly, we define user-centric emotional diversity  $E_{UCD}$  as the average distance between the emotional representations of all news articles in the ranked recommendation list at cutoff  $K$  and the user's overall emotion orientation  $\bar{e}_u$ :

$$E_{UCD@K} = \frac{1}{K} \sum_{D_i \in L@K} \text{dist}(e_{TA}^{D_i}, \bar{e}_u). \quad (5.18)$$

where  $E_{UCD@K}$  represents the user-centric emotional diversity at cutoff  $K$ ,  $D_i$  is a news article in the recommendation list  $L$  at cutoff  $K$ ,  $e_{TA}^{D_i}$  is the emotion representation of article  $D_i$  (extracted from its title and abstract),  $\bar{e}_u$  is the user's overall emotion

representation (averaged across their history), and  $dist$  is the cosine distance function (Equation 5.16). It reflects how the ranked lists differ emotionally from the user’s overall orientation. A greater difference in the top-K ranks results in higher values of  $E_{UCD@K}$

**Topical Diversity.** We create 100 dimensional embeddings for categories  $c_C$  (e.g., for sports) and subcategories  $c_S$  (e.g., for soccer) of news articles. We average the (sub)category embeddings of all browsed news articles of users’ to obtain their categorical representation  $c_u$ . Similarly, we average the (sub)category embeddings of top-K recommended news articles to obtain the recommendations category representation  $c_{L@K}$ . Having both, user-centric topical diversity is defined as:

$$T_{UCD@K} = dist(c_{L@K}, c_u), \quad (5.19)$$

where  $T_{UCD@K}$  denotes the user-centric topical diversity at cutoff  $K$ ,  $c_{L@K}$  is the category representation of the recommendation list at cutoff  $K$ ,  $c_u$  represents the user’s categorical representation, and  $dist$  is the cosine distance function (Equation 5.16). It indicates to what extent the consumed news articles differ from the recommended ones categorically (the higher the more diverse). For any given article  $D$  we compute its categorical representation  $c_D$  by averaging the article’s category  $c_C$  and subcategory  $c_S$  embeddings. Therefore, we define the intra-list topical diversity  $T_{ILD}$  as the average pairwise distance at cutoff  $K$ :

$$T_{ILD@K} = \frac{2 \sum_{D_i \in L@K} \sum_{D_j \in L@K \setminus \{D_i\}} dist(c_{D_i}, c_{D_j})}{K(K-1)}, \quad (5.20)$$

where  $T_{ILD@K}$  is the intra-list topical diversity at cutoff  $K$ ,  $L@K$  is the recommendation list at cutoff  $K$ ,  $D_i$  and  $D_j$  are individual news articles within the list,  $c_{D_i}$  and  $c_{D_j}$  represent the categorical representations of articles  $D_i$  and  $D_j$ , respectively, and  $dist$  is the cosine distance function (Equation 5.16). It provides an intuition how the top-K ranked news articles diverge topically.

### 5.2.2 Experimental Setting

**Dataset** We use the MIND [124] dataset, in particular the MIND-small<sup>6</sup> version. The dataset is constructed from MSN News<sup>7</sup> logs collected from October 12 to November 22, 2019, where the first five weeks are for training and the last week for testing. It contains 50K randomly sampled users (with at least five clicks), 65K news articles, 230K impressions with 350K clicks and 8M non-clicks. One sample, i.e., impression log, is composed of a timestamp, a user-id, a list of chronologically ordered news-ids representing the user’s interaction history, and a list of shuffled candidate news-ids with corresponding labels (i.e., “*clicked*” and “*seen but not clicked*”).

<sup>6</sup><https://msnews.github.io/index.html>

<sup>7</sup><https://www.msn.com/en-us/news>

**Training** We train all models on 90% of the training data. We use the remaining 10% for tuning the hyperparameters by optimizing AUC. We employ early-stopping with a minimum delta of 0.0001 AUC and patience of 5. Note that we use 300-dimensional Glove embeddings [82] in all models to initialize the word embedding layer and NLTK [13] word tokenizer for tokenization. Further, we limit the number of browsed news in each impression to 50, the title length to 20 and the abstract length 50 terms (smaller sequences are zero-padded).

**Baselines** We compare our approach to following baselines suggested by the dataset providers [124]: *LSTUR* [4] - a neural NRS capturing users' long- and short-term interests; *NAML* [116] - a neural NRS incorporating multiple views (i.e., title, abstract, category, and subcategory) into the news representation; *NRMS* [118] - a neural NRS which utilizes multi-head self-attention within both the news encoder and the user encoder. Additionally, we compare our approach to *SentiRec* [119] - a sentiment diversity-aware neural NRS.

**Evaluation.** We evaluate model effectiveness using *AUC*, *MRR*, *nDCG@5*, and *nDCG@10* as suggested by the dataset providers [124] and adopted in related work [119, 118, 4, 116]. We evaluate emotional diversity and topical diversity using *EUCD*, *EILD*, *TUCD*, and *TILD* - introduced previously. We compare our results using paired *t*-test with Bonferroni correction [110, 105]. Please, refer to our NewsRec<sup>8</sup> and EmoRec<sup>9</sup> repositories for reproducibility.

### 5.2.3 Results and Analysis

We propose a multi-level emotion-aware recommendation framework to incorporate emotions into the recommendation process and trained three models - *EmoRec<sub>S</sub>*, *EmoRec<sub>E</sub>*, and *EmoRec<sub>G</sub>* - based on three distinct emotion taxonomies of different granularity, i.e., *Sentiment*, *Ekman*, and *GoEmotion* (see Section 5.2.1). First, we investigate:

**RQ3.2.1** *To what extent does the incorporation of emotions and the usage of different emotion taxonomies affect recommendation performance?*

We compare our models to strong baselines on the MIND dataset [124], to SentiRec [119], a sentiment/sentiment-diversity aware NRS, and additionally to a random model. Table 5.5 summarizes our effectiveness comparison. Overall, our models consistently outperform all baselines. In particular, *EmoRec<sub>S</sub>* significantly ( $p < 0.001$ ) outperforms the most effective baseline, *NRMS*, in all positions. Also, *EmoRec<sub>E</sub>* and *EmoRec<sub>G</sub>* significantly ( $p < 0.001$ ) outperform *NRMS* in almost all positions except when considering the *MRR* of the fully ranked lists; both models indicate an improvement, but the differences are not significant. Thus, we conclude that incorporating emotions improves news recommendation effectiveness. We demonstrate that utilizing a coarse taxonomy yields the most improvements.

<sup>8</sup><https://github.com/MeteSertkan/newsrec>

<sup>9</sup><https://github.com/MeteSertkan/EmoRec>

Table 5.5: Comparing effectiveness (i.e., AUC, MRR, nDCG@5, and nDCG@10) of  $EmoRec_S$ ,  $EmoRec_E$ ,  $EmoRec_G$ , and the baselines. Subscripts S, E, and G indicate the used taxonomy for model training. Higher effectiveness scores indicate better performance. Note, \* indicates a statistically significant difference to (our most effective model)  $EmoRec_S$  and † indicates a statistically significant difference to (the most effective baseline)  $NRMS$ , both at alpha 0.001.

Model	AUC	MRR	nDCG	
			@5	@10
Random	.4994*†	.2190*†	.2236*†	.2863*†
$NAML$	.6206*	.2913*	.3185*	.3782*
$LSTUR$	.6210*†	.2840*†	.3101*†	.3721*†
$SentiRec$	.6224*	.2952*	.3211*	.3818*
$NRMS$	.6228*	.2946*	.3191*	.3817*
$EmoRec_S$	<b>.6384†</b>	<b>.2980†</b>	<b>.3276†</b>	<b>.3890†</b>
$EmoRec_E$	<u>.6374*†</u>	.2968*	<u>.3261*†</u>	<u>.3878*†</u>
$EmoRec_G$	.6345*†	<u>.2970</u>	.3233*†	.3865*†

Additionally, we conduct an ablation study to highlight the contributions of different views (i.e., text vs. category) on the effectiveness. Therefore, we train our models once without any incorporated emotion, once with only incorporating text-based emotions (i.e., title- and abstract-based), once with only incorporating category-based emotions (i.e., category- and subcategory-based), and once at total capacity. Figure 5.8 illustrates the impact of each configuration on the effectiveness. The heavy lifting is done by semantic matching (i.e., if no emotions are used). Utilizing text-based or category-based emotions yield performance improvements, where text-based emotions bring in more gains. However, combining both views, i.e., our model at full capacity, leads to the best effectiveness.

Alongside effectiveness, we are also interested in beyond accuracy capabilities, particularly emotional diversity. Incorporating emotions improves effectiveness but might bear the risk of decreasing the recommendations’ emotional diversity. Thus, we investigate:

**RQ3.2.2** *To what extent does the incorporation of emotions and the usage of different emotion taxonomies affect emotional diversity?*

In our analysis of emotional diversity among various models, we employ two specific evaluation metrics:  $E_{UCD}$  and  $E_{ILD}$  (details in Section 5.2.1). These diversity measures are influenced by the chosen emotion taxonomy (vector space), and therefore, we calculate three distinct sets of emotional diversity metrics for each model. A summary of the results is presented in Table 5.6. Among the models, the  $NAML$  baseline demonstrates superior emotional diversity, surpassing all competitors, including the random model.

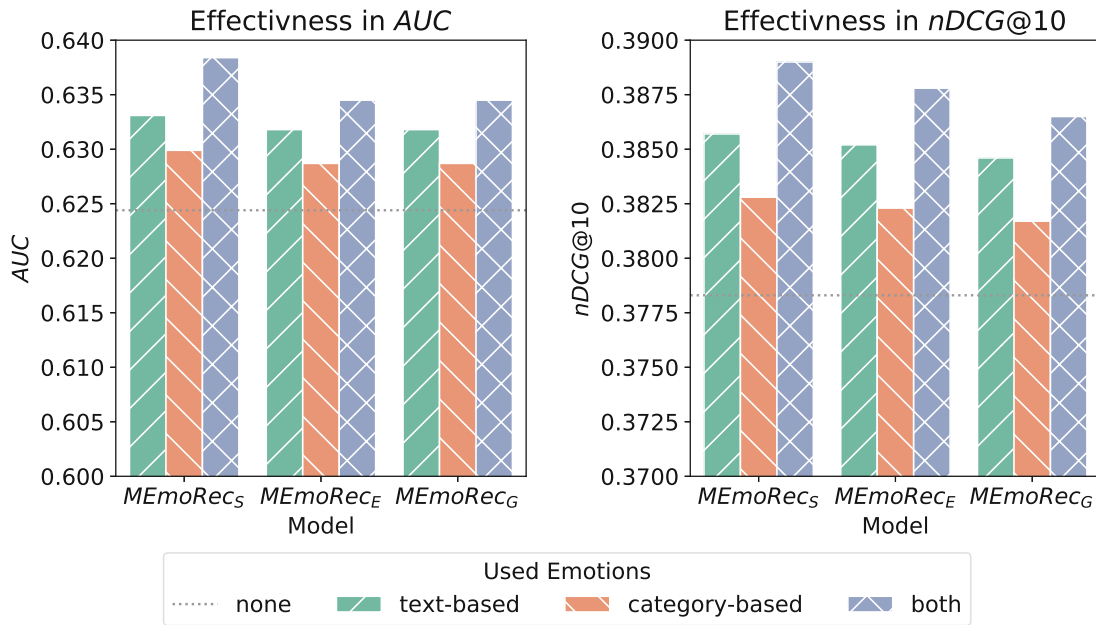


Figure 5.8: Ablation study on different configurations of  $MEmoRec$  - Effectiveness analysis.

Our models generally exhibit less emotional diversity in recommendations compared to purely text-based models (i.e.,  $NRMS$ ,  $LSTUR$ ,  $NAML$ ) with a few exceptions. The  $EmoRec_E$  model only significantly outperforms  $NRMS$  and the random model in terms of emotional diversity when the *Sentiment* taxonomy is used.

In the ablation study, we assess our models' emotional diversity using four configurations: without emotions, utilizing text-based emotions, using category-based emotions, and incorporating both. Across user-centric and intra-list emotional diversity measures (see Figures 5.10 and 5.9), the models exhibit similar behavior. Our results show that the integration of emotions typically diminishes emotional diversity. However, an exception is found in  $EmoRec_E$ , where the inclusion of emotions enhances the diversity of recommendations. Furthermore, we consistently find that recommendations driven by category-based emotions are more diverse than those informed by text-based emotions. Interestingly, the model's full capacity configuration results in the least diverse recommendations.

In addition to emotional diversity, we also explore the topical alignment of the recommended items when emotions are incorporated:

**RQ3.2.3** *To what extent does the incorporation of emotions and the usage of different emotion taxonomies affect topical diversity?*

We evaluate topical diversity using both a user-centric metric ( $TUCD$ ) and an intra-list

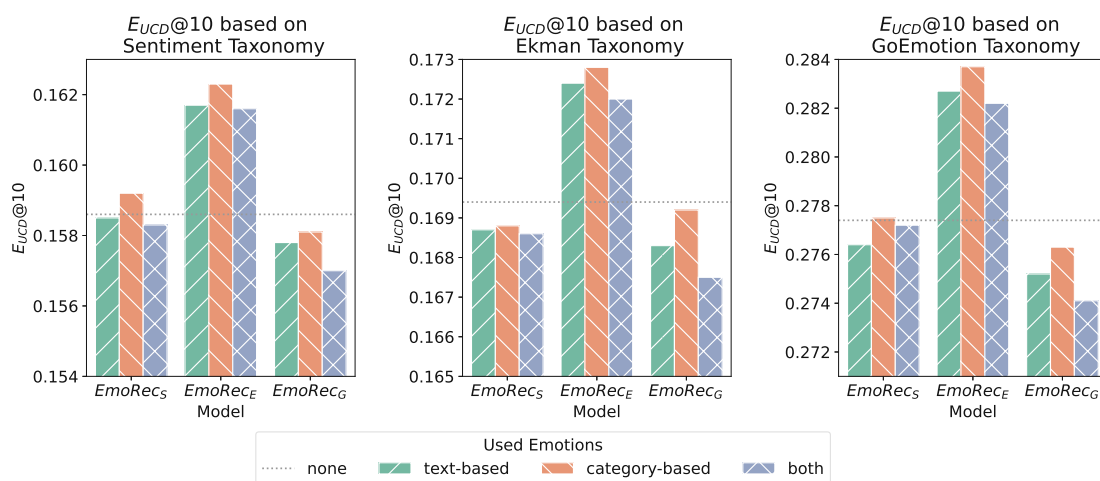


Figure 5.9: Ablation study on different configurations of *EmoRec* - User-centric emotional diversity analysis.

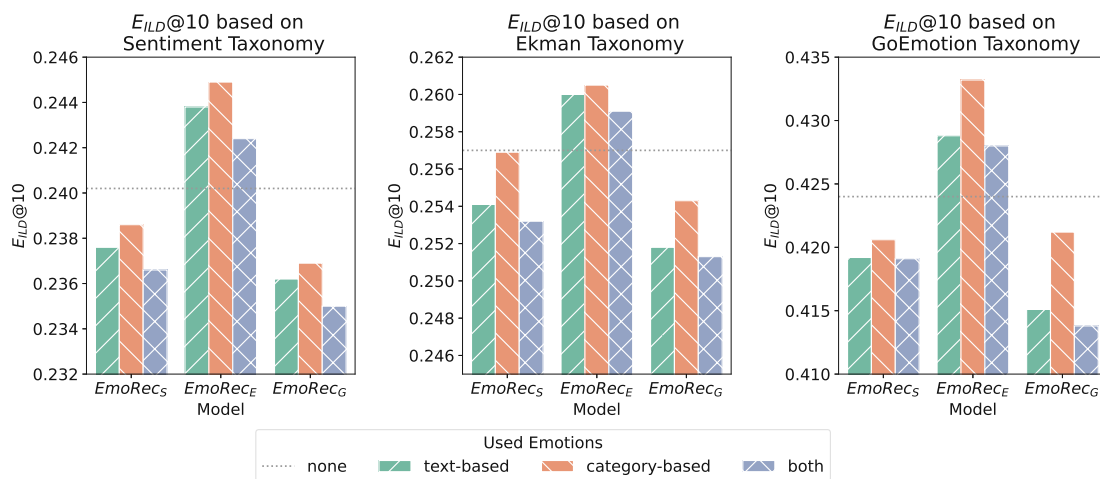


Figure 5.10: Ablation study on different configurations of *EmoRec* - Intra-list emotional diversity analysis.

Table 5.6: Comparing emotional diversity (i.e.,  $E_{UCD}@10$  and  $E_{ILD}@10$ ) of  $EmoRec_S$ ,  $EmoRec_E$ ,  $EmoRec_G$ , and the baselines. Subscripts  $S$ ,  $E$ , and  $G$  indicate the used taxonomy for model training. Column names *Sentiment*, *Ekman*, and *GoEmotion* indicate the taxonomy used for distance calculation. Higher scores indicate more emotionally diverse recommendations. Note, † indicates a statistically significant difference to (our most emotionally diverse model)  $EmoRec_E$  and \* indicates statistically significant difference to the random model, both at alpha 0.05.

Model	Sentiment		Ekman		GoEmotion	
	$E_{UCD}@10$	$E_{ILD}@10$	$E_{UCD}@10$	$E_{ILD}@10$	$E_{UCD}@10$	$E_{ILD}@10$
1 Random	.1604†	.2378†	.1782†	.2665†	.2880†	.4348†
2 <i>SentiRec</i>	.1573*†	.2341*†	.1701*†	.2560*†	.2792*†	.4214*†
3 <i>NRMS</i>	.1607†	.2393†	.1729*†	.2598*†	.2883*†	.4355*†
4 <i>LSTUR</i>	.1666*†	.2516*†	.1762*†	.2659†	.2859*†	.4330*†
5 <i>NAML</i>	<b>.1695*†</b>	<b>.2559*†</b>	<b>.1866*†</b>	<b>.2818*†</b>	<b>.3025*†</b>	<b>.4611*†</b>
6 $EmoRec_E$	.1616*	.2424*	.1720*	.2591*	.2822*	.4280*
7 $EmoRec_S$	.1584*†	.2366*†	.1686*†	.2532*†	.2772*†	.4191*†
8 $EmoRec_G$	.1570*†	.2350*†	.1675*†	.2513*†	.2742*†	.4138*†

Table 5.7: Comparing topical diversity (i.e.,  $T_{UCD}@10$  and  $T_{ILD}@10$ ) of  $EmoRec_S$ ,  $EmoRec_E$ ,  $EmoRec_G$ , and the baselines. Subscripts  $S$ ,  $E$ , and  $G$  indicate the used taxonomy for model training. Higher scores indicate more topically diverse recommendations. Note, † indicates a statistically significant difference to  $EmoRec_E$  and \* indicates statistically significant difference to the random model, both at alpha 0.05.

Model	$T_{UCD}@10$	$T_{ILD}@10$
1 Random	<b>.5572†</b>	<b>.9225†</b>
2 <i>NRMS</i>	.5074*†	.8984*†
3 <i>SentiRec</i>	.5109*†	.8986*†
4 <i>LSTUR</i>	.5133*†	.8974*†
5 <i>NAML</i>	.5563*†	.8367*†
6 $EmoRec_E$	.4997*	.8883*
7 $EmoRec_G$	.4968*	.8895*
8 $EmoRec_S$	.4962*	.8869*

metric ( $T_{ILD}$ ). The results are summarized in Table 5.7. The random model consistently provides the most diverse recommendations according to both metrics. In the context of user-centric emotional diversity, all  $EmoRec$  variants perform significantly worse than all baselines. We find a parallel trend in intra-list topical diversity, although in this instance, the *NAML* baseline performs even more poorly.

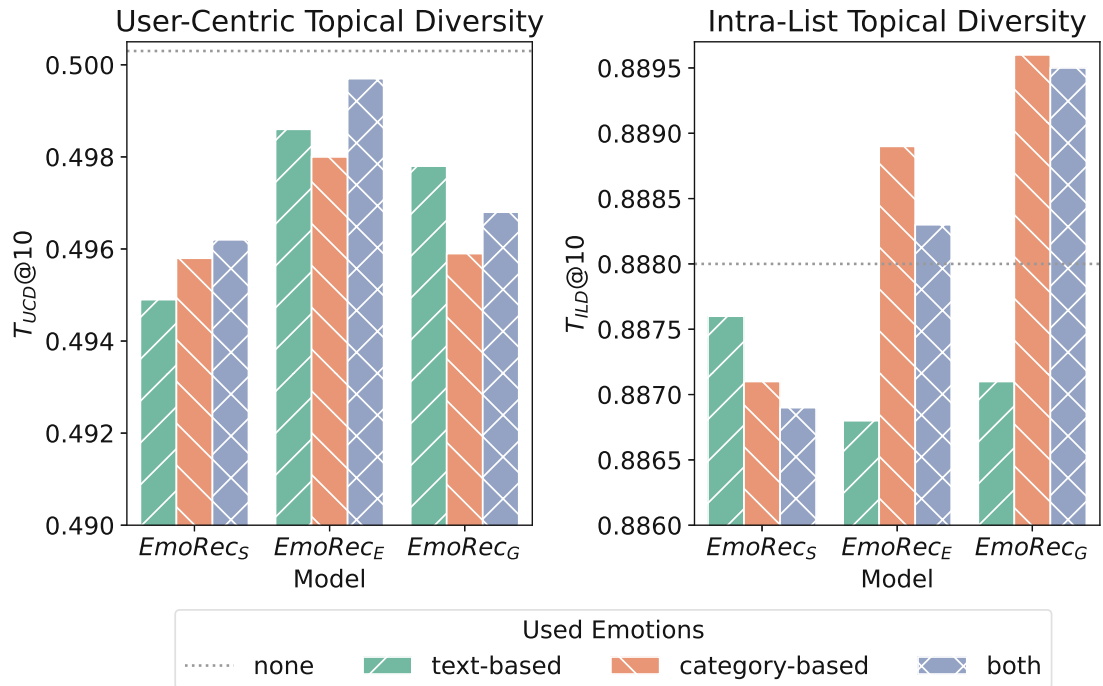


Figure 5.11: Ablation study on different configurations of *EmoRec* - User-centric & intra-list topical diversity analysis.

Figure 5.11 illustrates the topical diversity ablations – we use the same configurations as previously. The evaluation of user-centric topical diversity reveals a decrease in diversity across all configurations that incorporate emotions, whether text-based, category-based, or both. No specific pattern emerges to differentiate the effects of including text-based versus category-based emotions. When considering intra-list topical diversity, the inclusion of text-based emotions consistently leads to less diversity. However, in the models *EmoRec<sub>E</sub>* and *EmoRec<sub>G</sub>*, this decrease is counterbalanced when category-based emotions are included. This results in a more topically diverse recommendation list in the full model, compared to configurations without emotions.

#### 5.2.4 Discussion

We focus on the complex interplay between the content of news articles, the emotions they carry, and their recommendation. Our findings are based on the MIND dataset [124], which primarily contains news articles with a neutral tone, corroborating the findings of Wu et al. [119]. Nevertheless, our model, *EmoRec*, is designed to discern and leverage the subtle emotional nuances within these articles, thereby aligning recommendations with users’ consumption behavior.

We investigated how emotions, extracted at various levels (title, abstract, category, and subcategory) and with varying granularity of emotion taxonomies, influence rec-

ommendation performance and emotional diversity. Notably, we found that a coarser, four-dimensional *Sentiment* taxonomy significantly outperforms finer-grained options. This is in contrast to Deng et al. [28], who found benefits in utilizing finer-grained emotion taxonomies. This discrepancy may stem from the inherently different nature and emotional richness of user-written content, such as social media posts, compared to professionally edited news articles.

Furthermore, we investigate the impact of incorporating emotional signals on diversity, both emotional and topical, within news recommendation models. Though *EmoRec* provides better alignment with users' preferences and yields higher accuracy, it also leads to a significant drop in diversity compared to other baselines. This reduction in diversity raises critical concerns about the potential creation of a self-reinforcing "emotion chamber" over time.

Deep-learning models, increasingly used in recommenders [126], implicitly account for textual nuances and users' tastes. The more proficient these models become, the more they align with users' preferences, potentially further reducing diversity. *EmoRec*, by explicitly modeling emotional dimensions, offers an opportunity to not only communicate and raise awareness – for example, by confronting users with their emotional profiles and corresponding recommendations as illustrated in Figure 5.12 – about this issue but also intervene when necessary.

A critical aspect of our study involves distinguishing between intra-list diversity (within a recommendation list) and user-centric diversity (relative to a user's previous consumption behavior). This leads to a philosophical debate about the approach to recommendations: Should we provide users with diverse options and let them choose, or should we guide them towards more varied content? If the latter, what ethical considerations arise, such as justifying the recommendation of negative news following excessive positive consumption? We also contemplate a more nuanced approach, offering diverse options coupled with insights into a user's overall consumption behavior, enabling more informed decisions.

A recognized limitation in emotion-aware recommenders is the conflation of expressed, perceived, and induced emotions [68]. There are distinct differences between an article's emotional content, how users perceive that emotion, and the emotion actually induced in the reader. Moreover, the automated extraction process we employ adds a layer of complexity. We also urge caution in accepting established emotion taxonomies such as Ekman's [33], as they are highly debated and may be outdated [75]. These issues lead to the overarching question: What exactly are we measuring or considering with the extracted emotions?

In conclusion, our work highlights the complex interplay between accuracy and diversity in emotion-aware news recommendation. While *EmoRec* shows promising results, our findings emphasize the need for a thoughtful and ethically grounded approach to both user choice and emotional representation.

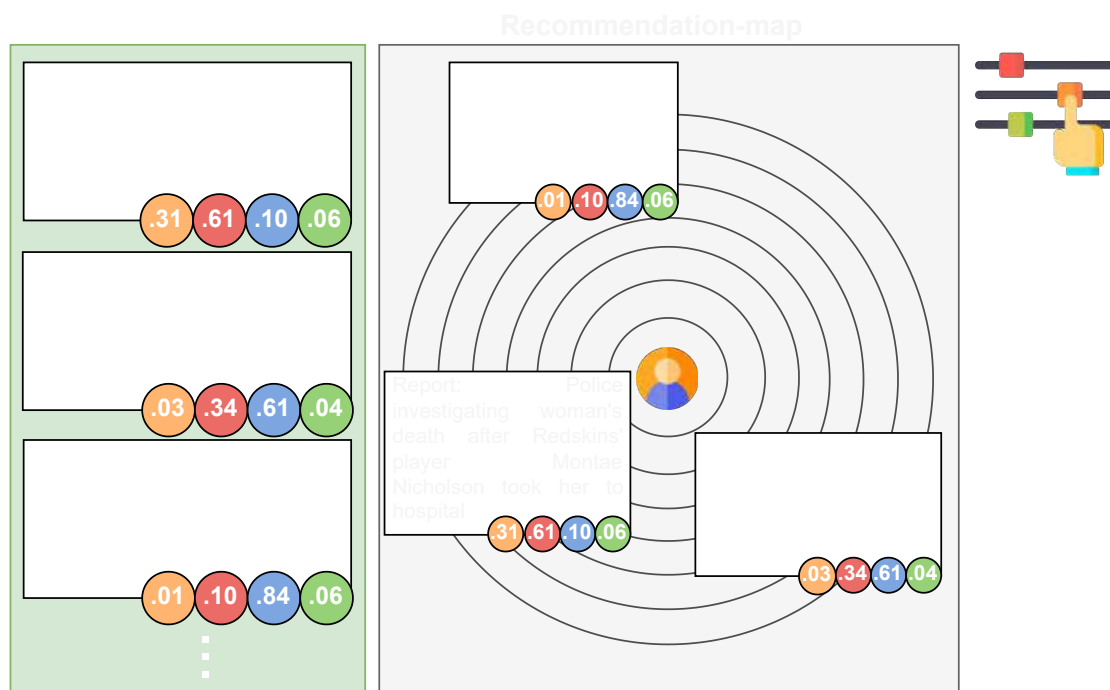


Figure 5.12: Explaining and exploring the recommendation neighborhood – on the left as a list and on the right as a map. News articles are illustrated by their headlines and corresponding sentiment scores are provided in alphabetical order, i.e., ambiguous (orange), negative (red), neutral (blue), positive (green).

### 5.2.5 Conclusions

We investigate the utility of leveraging expressed emotions in news articles to improve recommendation systems. We introduce a multi-level neural news recommendation model, *EmoRec*, designed to discern semantic and emotional content of news articles and users. We leverage emotions expressed in the title and abstract of news articles and aggregate emotions contained in articles of the same category and subcategory.

Through extensive experiments on the real-world MIND dataset [124], our approach outperforms strong non-emotion-aware baselines significantly. Notably, we find that using a coarse emotion taxonomy (i.e., *Sentiment*) significantly enhances performance over models based on more fine-grained emotions. However, the incorporation of emotions into the recommendation process comes at the cost of reduced emotional and topical diversity. Specifically, non-emotion-aware baselines tend to recommend more diverse news articles than our emotion-aware models.

Furthermore, we identify potential ethical challenges posed by a more proficient alignment of recommendations with users' preferences, highlighting the risk of a self-reinforcing "emotion chamber". This raises a critical debate: should recommendation systems provide

users with diverse options and let them choose, or should they actively guide users towards more diverse content? And if the latter, what ethical considerations arise, such as justifying the recommendation of negative news following excessive positive consumption?

Additionally, we recognize a significant limitation in emotion-aware recommenders—the conflation of expressed, perceived, and induced emotions [68]. The automated extraction process we employ adds complexity, emphasizing the need for caution in employing established emotion taxonomies, such as Ekman’s [33], which are debated and may be outdated [75].

In conclusion, our work underscores the complex interplay between accuracy and diversity in emotion-aware news recommendation. While *EmoRec* shows promising results in aligning news recommendations with user preferences, it highlights the need for a thoughtful, ethically grounded approach to both user choice and emotional representation.

In future work, we intend to investigate and compare different intervention strategies and delve into the nuanced differences in expressing, extracting, perceiving, and inducing emotions, as well as critically evaluate the taxonomies employed. This direction will help refine the alignment between user preferences and recommendations, facilitating more diverse and conscious consumption, without sacrificing the quality of the recommendations.

### 5.2.6 Limitations

While our multi-level emotion-aware framework demonstrates improved recommendation performance, several limitations warrant discussion. The reliance on automated emotion extraction from news text introduces potential inaccuracies. The expressed emotion in a news article, as detected by our classifiers, may not perfectly align with the emotion perceived by the reader or the emotion ultimately induced. This discrepancy between expressed, perceived, and induced emotions is a known challenge in emotion-aware systems and is further complicated by the automated nature of our emotion analysis. Furthermore, our approach assumes that the title, abstract, category, and subcategory adequately represent the emotional content of the full article, which may not always be the case, as we did not use the article’s body in our modeling. The use of established emotion taxonomies, while providing a structured framework, is also a potential limitation. These taxonomies (Ekman, GoEmotions, and basic sentiments) are subject to ongoing debate and might not fully capture the complexity of human emotions, particularly in the context of news consumption. Also, emotion taxonomies are inherently subjective. Finally, our experiments are conducted on the MIND dataset, which may have its own biases and characteristics, such as having mainly neutrally written news articles, limiting the generalizability of our findings to other news datasets or domains.

### 5.3 Summary

This chapter comprises two distinct but complementary investigations into emotion-aware and sentiment-aware NRS.

First, we embarked on the reproduction of SentiRec [119], a sentiment diversity-aware neural news recommendation model. We re-implemented SentiRec from scratch using the MIND dataset [124]. Contrary to the original findings, our reproduced model did not exhibit superior sentiment diversity, nor did it consistently outperform baselines, especially the strong NAML baseline [116]. The discrepancies are attributed to dataset differences, spotlighting SentiRec’s generalizability issues. We extended our study to topical diversity and intra-list sentiment diversity, with our reproduction of SentiRec showing notable performance in the intra-list evaluation.

Based on these findings, we proposed RobustSentiRec, a refined model aiming to enhance sentiment diversity robustness. RobustSentiRec outperforms SentiRec in recommendation performance and intra-list sentiment diversity, achieves comparable user-centric sentiment diversity, and offers the benefit of reduced model complexity.

In a parallel research direction, we introduced EmoRec, a multi-level neural news recommendation model that leverages emotions in news articles. Experimenting with the MIND dataset [124], EmoRec significantly outperformed non-emotion-aware baselines. Utilizing a coarse emotion taxonomy (e.g., Sentiment) was found to enhance performance. However, this came at the cost of reduced emotional and topical diversity in recommendations.

Furthermore, both studies collectively raise ethical concerns about the potential creation of self-reinforcing emotional or sentiment chambers due to proficient alignment with users’ preferences. This poses a critical question: Should we provide users with diverse options and let them choose, or should we guide them towards more diverse content? If the latter, what ethical considerations arise, such as justifying the recommendation of negative news following excessive positive consumption?

Additionally, the studies emphasize caution in the conflation of expressed, perceived, and induced emotions [68] and call for scrutiny of the emotion taxonomies employed, given ongoing debates and potential obsolescence [75, 33].

Regarding limitations, both studies relied on automated sentiment or emotion extraction, introducing potential inaccuracies due to discrepancies between expressed, perceived, and induced emotions. The choice of the MIND dataset and its characteristics, along with reliance on potentially limited sentiment/emotion analysis tools and emotion taxonomies, also limits the generalizability and robustness of the findings. Finally, the scope of each study presents further limitations: SentiRec did not explore long-term user satisfaction, while EmoRec did not utilize the full body of the articles.

In conclusion, the juxtaposition of these two research directions highlights the complex interplay between recommendation accuracy and diversity. While both EmoRec and RobustSentiRec promise advancements in aligning news recommendations with user

preferences, they underline the paramount need for an ethically grounded approach balancing user choice, emotional representation, and diversity.

As we look ahead, our unified vision is to integrate diverse auxiliary information, such as personality and normative values, into end-to-end recommendation models. Furthermore, we intend to investigate and compare different intervention strategies. Our aspiration is to craft RSs that enrich society with thoughtful and responsible recommendations without compromising the quality of those recommendations.



CHAPTER **6**

# Conclusions

In this chapter, we provide a comprehensive summary of our research findings, answer the key research questions, and conclude on the effectiveness of incorporating implicit item characteristics in personalized RSs. We discuss the practical implications of our work in both the tourism and news domains, highlighting how our methodologies can enhance user satisfaction and engagement. Additionally, we explore the ethical considerations and potential challenges associated with sentiment and emotion-aware RSs.

We begin by revisiting the main research questions and summarizing the key insights gained from our explorations. This includes the identification and modeling of implicit item characteristics, the validation of our picture-based preference elicitation technique, and the incorporation of sentiment and emotion in NRS. We then outline the practical implications of our findings, offering recommendations for industry applications and potential benefits for academia.

Finally, we discuss future research directions, emphasizing the need for more dynamic and context-aware recommendation systems (CARS), further refinement of user modeling techniques, and addressing the ethical challenges in personalized recommendations. This chapter aims to encapsulate the core contributions of our thesis and provide a roadmap for future advancements in the field.

## 6.1 Summary

In today’s world, characterized by an insatiable thirst for information and a constant influx of data, navigating this overwhelming abundance has become a complex challenge [89]. While many potentially valuable resources are available, identifying what is truly beneficial has become increasingly difficult, often pushing traditional theories of rational decision-making to their limits [58, 22]. At the heart of this dilemma is Herbert Simon’s theory of “Bounded Rationality,” which highlights the limitations of human cognitive capabilities and the need for more efficient decision-making tools [104, 22]. RSs have gained widespread prominence as a solution to this modern-day problem [89, 73]. However, even though they offer increasingly personalized choices, they often fall short in capturing the multi-dimensionality of human preferences, especially the subtle implicit item characteristics that greatly impact user choice and satisfaction. For instance, reading a news article and feeling a strong emotional reaction, or choosing a holiday destination that resonates with one’s inner adventurer—these nuanced factors often escape traditional computational models but significantly shape our choices and biases. This thesis, therefore, explores not just what users overtly express or consume but also the implicit layers of their decision-making processes. Focusing on diverse and challenging domains like tourism and news, this work aims to re-engineer the foundational elements of user modeling by incorporating these implicit item characteristics, thereby striving to deliver more relevant and personally resonant recommendations.

To address the complexities of human decision-making, this thesis employs a dual-strategy approach to enhance RSs with a deeper understanding of implicit item characteristics. The first strategy involves using established domain models like the Seven-Factor Model [76, 77] to blend both overt and covert factors affecting user choices, particularly in the tourism domain. The second strategy goes a step further by explicitly modeling these subtle factors, enriching the semantic representations of items and tailoring the recommendations more closely to individual preferences.

These strategies were supported by a three-fold methodological framework, each corresponding to a different level of user interaction (as highlighted in the introduction – Chapter 1) and addressing a specific research question:

**1. No Interaction (Explorative Analysis) - Addressing RQ1** *How can we systematically identify and expose the implicit item characteristics embedded in structured and unstructured data to enhance recommender systems?*

To address RQ1, our study systematically identified and exposed implicit item characteristics within structured and unstructured data in the tourism and news domains. For tourism, we used exploratory data analysis, cluster analysis, and multiple linear regression. This revealed six distinct destination clusters: *vibrant beach resorts, energetic cities, tranquil seaside resorts, peaceful towns, idyllic island villages, and ordinary towns*. These clusters, encompassing attributes ranging from recreational opportunities to urban density, offered nuanced insights into traveler preferences. Importantly, regression analysis clarified how these attributes correlate with the Seven-Factor Model of travel motivations.

For example, “sun” and “beach” positively influenced the “Sun & Chill-Out” factor, whereas “crowdedness” had a negative impact.

In the news domain, we adopted a tri-layered analytical approach: document, topic, and author-level. Document-level analysis extracted objective knowledge using discriminative terms, enabling recommendations based on lexical similarity. Topic-level analysis uncovered latent themes and temporal trends, facilitating more diverse recommendations. Author-level analysis, achieving 97% accuracy in identifying distinct writing styles, offered the highest potential for diverse and serendipitous recommendations by incorporating stylistic congruence, moving beyond the limitations of document content.

These findings have significant implications for personalized RSs, extending beyond tourism and news. By understanding the relationships between item attributes and user motivations (even without explicit user interaction), more accurate matching can be achieved. In tourism, this could increase visitor attraction, satisfaction, and revenue, while also saving travelers time and enhancing their experiences. In news, personalized recommendations based on writing style, latent topics, and stylistic similarities can enhance user engagement and broaden news consumption. The multi-faceted methodology – combining techniques like TF-IDF, LDA, and author style analysis – is broadly applicable, with potential benefits for e-commerce, entertainment, and social media through the adaptation of user motivation models. Critically, this “no interaction” approach provides a foundation for tackling *cold-start* scenarios and enriching item representations *prior to* any user data collection. It demonstrates that even without user feedback, valuable insights into item characteristics and their potential resonance with users can be derived from underlying domain structures and content features.

## 2. Active User Involvement (Picture-Based Elicitation) - Addressing RQ2

*How can a generic picture-based user and item modeling – not limited to a fixed picture set – improve the efficacy and user satisfaction of tourism recommendation systems?*

A generic picture-based approach enhances TRS by enabling more personalized and engaging user experiences through analyzing user-provided or item-associated images, rather than fixed image sets. Leveraging the Seven-Factor Model, this method, implemented through the *Generic Profiler* (Section 4.1) and validated in the *PicTouRe* user study (Section 4.2), captures nuanced preferences and item characteristics with high accuracy (88%-99% in image classification). The *Aggregator* consolidates these analyses into holistic user or item representations. The user study confirmed the effectiveness, with 65% of participants agreeing with generated profiles and nearly half preferring system recommendations, demonstrating that this approach leads to recommendations better aligned with individual tastes and perceived item qualities, simplifying the preference elicitation process.

The significance of this research extends beyond improved accuracy and the tourism domain; it represents a shift in how preference elicitation and personalization are approached in RSs. The core principles—moving from restrictive, fixed sets to user-driven, visual interaction—are applicable to various domains facing challenges like limited inter-

action data, *cold-start* problems, and high consumption costs associated with traditional preference elicitation. By using a gamified, active method utilizing visual cues, the cognitive burden is reduced, addressing limitations of bounded rationality and fostering more authentic preference profiles. This enables a high degree of personalization and user engagement. The flexibility offered by users leveraging their own image collections supports adaptability across various contexts. This has significant implications beyond tourism, offering potential benefits to recommending items like concerts, events, or jewelry, where visual appeal and personal taste are crucial. This methodology provides a valuable framework for other fields seeking to create more engaging and effective RSs, paving the way for future research, including integrating contextual factors and combining visual with other interaction data.

### **3. Passive Observation (Sentiment and Emotion Analysis) - Addressing RQ3**

*How can sentiment and expressed emotions in news articles be effectively utilized in recommendation systems to improve recommendation performance, and what impact do these factors have on diversity?*

Our research investigates how sentiment and expressed emotions in news articles can be utilized to improve recommendation performance and their impact on diversity. We explored this through both sentiment-aware (RobustSentiRec) and emotion-aware (EmoRec) models. While incorporating these elements showed promise in aligning recommendations with user tonal preferences and potentially increasing satisfaction, it also raised concerns. Traditional sentiment models, and even our improved RobustSentiRec, might inadvertently limit user exposure to a narrow range of sentiments. Similarly, while EmoRec boosted recommendation performance, it risked reducing both emotional and topical diversity, potentially creating “emotion chambers.” This critical examination challenges the generalizability of prior work and underscores the complex interplay between personalization, performance, and diversity when incorporating nuanced textual features like sentiment and emotion.

This research has broad implications for the design and ethical considerations of NRS, specifically concerning how we handle emotional information. For example, our findings on RobustSentiRec suggest that overemphasizing positive sentiment could lead to users primarily seeing “good news,” potentially creating a biased view of the world. Similarly, the EmoRec results highlight a tension: while tailoring recommendations to a user’s current emotional state (e.g., showing more uplifting stories to someone detected as feeling sad) might improve short-term engagement, it could also limit exposure to diverse perspectives and emotions, hindering their ability to develop a nuanced understanding of complex issues. The work contributes to Human-Computer Interaction (HCI) by revealing how users *might react to emotionally targeted recommendations*, and it advances AI-driven personalization by demonstrating the *technical feasibility and limitations of using emotions as signals*. Building upon this research, future work in Computational Social Science might help understand how algorithmically driven news feeds, particularly those personalized based on sentiment and emotion, could shape collective emotions and public opinion; and in Media Studies, this understanding reinforces the need for

responsible platform design and content curation that prioritizes emotional balance and avoids the creation of echo chambers.

Ultimately, a careful integration of sentiment and emotion is needed. This means moving beyond simple positive/negative classifications and considering the full spectrum of human emotion, while actively mitigating the risks of filter bubbles and echo chambers through diversity-promoting mechanisms. This requires a multifaceted approach, considering user engagement, information diversity, long-term well-being, and the ethical implications for both individuals and society. This may, for example, include offering users controls to adjust the emotional tone of their recommendations or incorporating “emotional diversity” metrics into the evaluation of recommendation algorithms.

In summary, this thesis takes a significant step towards integrating implicit item characteristics into RSs. By exploring diverse methodologies—from initial data exploration to active visual elicitation and passive emotional analysis—in both tourism and news, we have demonstrated the potential for richer, more human-centered recommendations. While primarily exploratory, these findings lay the groundwork for future research to refine these approaches, address their limitations, and ultimately develop systems that better understand and respond to the complexities of human preferences. The potential applications extend beyond the studied domains, offering a pathway to more effective and ethically sound RSs across various aspects of our digital lives.

## 6.2 Practical Implications

**Implications for Tourism Destination Recommendations.** Our work lays a foundational framework for the intelligent categorization and recommendation of tourism destinations. By employing the Seven-Factor Model to map out destination attributes, the study brings to light the conceptual significance of characterizing travel locations with respect to individual preferences. This could be instrumental for OTAs, travel platforms, and destination marketing organizations in tailoring their campaigns and offerings.

For academia, this study demonstrates the efficacy of combining cluster analysis and multiple linear regression models to interpret destination attributes. This could guide future research towards enhancing the precision and coverage of RSs in the tourism sector. Additionally, the scalability of our framework lends itself well to the incorporation of more destinations and characteristics as data availability grows, thereby preserving its applicability over time.

**Implications for News Recommendation Systems.** Our examination of news articles from multiple dimensions—document, topic, and author—enriches the existing understanding of news content, opening avenues for a more comprehensive news recommendation strategy. Content providers and platform developers may consider these findings when developing or improving NRS to offer a more tailored, enriched user experience.

The academic community can build upon our multi-layered approach to assess news articles, possibly extending the methodology to other types of content. Additionally, our work offers a preliminary framework that could help address challenges in information diversity, suggesting a line of inquiry for future studies concerned with filter bubbles and information overload.

**Implications for Image-Based Preference Profiling.** The *Generic Profiler* we have introduced serves as a novel tool for capturing nuanced user preferences through image-based data. Its dual applicability to both individual users and destination marketers elevates its utility. While individuals get more personalized recommendations, industry stakeholders can receive a dynamic, data-based characterization of their offerings.

The methodology can also be extended to other sectors like real estate or retail, representing a versatile profiling tool. Academia can benefit from this research by further investigating the performance of image-based profiling against traditional methods and its potential limitations or biases.

**Implications for Emotional Dynamics in News Recommendation.** Our exploration into sentiment- and emotion-aware NRS has a direct bearing on the design and implementation of personalized NRS. While our framework offers a compelling direction for personalization, it also raises ethical considerations by narrowing the diversity of content based on emotional resonance.

Future systems could consider integrating a diversity-aware mechanism to balance personalization with exposure to a wide array of topics and emotions. Academics interested in ethical considerations around RSs can also draw upon our work to explore the ramifications of emotion-based personalization in news consumption.

### 6.3 Future Directions

In the realm of personalized recommendations, the quest for capturing the nuanced human decision-making process remains an enduring challenge. While this thesis successfully integrates content-based methods with an understanding of bounded rationality and decision biases, the landscape of personalized recommendations is far from being fully explored. In particular, the current work lays the groundwork for several future research directions that echo the overarching theme of improving the quality of personalized recommendations by considering the implicit characteristics that are pivotal in user decision-making.

First, in the context of TRS, a critical pathway for future research is the incorporation of CARS. The need to transition from a static snapshot to a dynamic representation is imperative. Real-world scenarios involve ever-changing variables such as weather, local events, and socio-political climates, all of which align well with the theme of bounded rationality and dynamic preferences explored in this thesis. By integrating these contextual elements, future research can better match the complex and emotional

landscape that shapes tourists' choices, thereby offering recommendations that are not just timely but deeply personalized.

Turning our attention to news recommendation, the inclusion of author-level analysis as an additional layer to the recommendation engine forms a promising avenue for further investigation. The intent is to move beyond the confines of content and user preference towards a system that dynamically adapts to include elements of novelty and serendipity. This would counteract uncontrolled decision biases like decoy and position effects by providing a more holistic recommendation landscape that includes both what is shown and how it is presented—a key factor influencing consumer choices, as outlined in this thesis.

For the tourism sector, enhancing the *Aggregator* component within our *Generic Profiler* promises to align closely with our efforts to understand the implicit dynamics of decision-making. The next steps should involve machine learning techniques that can weigh the importance of individual images based on their contextual relevance, bringing us closer to mimicking the complex thought processes that guide human decision-making.

In both domains, we recognize the need for more rigorous user studies to refine the alignment between predicted profiles and user self-perceptions. Future work should aim to integrate additional layers of user-provided and contextually inferred data to mitigate the limitations of sparse interaction matrices. These multi-modal approaches would be tested through extensive user studies to examine their applicability and reliability across a broad spectrum of use cases and diverse user groups.

Privacy is another critical area for future research in RSs. As we gather more detailed and personal data to improve recommendations, ensuring user privacy and data security becomes increasingly important. Future systems must incorporate robust privacy-preserving techniques, such as differential privacy and secure multi-party computation, to protect user data while still enabling effective personalization. Addressing privacy concerns is essential for maintaining user trust and complying with evolving data protection regulations.

Finally, in the field of news recommendations, the challenge of balancing emotional and topical diversity remains unaddressed. The exploration of mechanisms to avoid self-reinforcing “emotion chambers” offers an ethically nuanced extension of the current research. It further deepens our understanding of the role emotions play in the bounded rationality and dynamic preferences of users, hence staying true to the core theme of this thesis.

Overall, the identified future directions not only aim to refine the models developed in this work but also to contribute to the broader dialogue on creating more effective and ethically responsible RSs.



# Bibliography

- [1] Charu C. Aggarwal and ChengXiang Zhai. An Introduction to Text Mining. In Charu C. Aggarwal and ChengXiang Zhai, editors, *Mining Text Data*, pages 1–10. Springer US, Boston, MA, 2012. ISBN 978-1-4614-3223-4. doi: 10.1007/978-1-4614-3223-4\_1. URL [https://doi.org/10.1007/978-1-4614-3223-4\\_1](https://doi.org/10.1007/978-1-4614-3223-4_1).
- [2] Jae-wook Ahn, Peter Brusilovsky, Jonathan Grady, Daqing He, and Sue Yeon Syn. Open user profiles for adaptive news systems: Help or harm? In *Proceedings of the 16th International Conference on World Wide Web, WWW '07*, page 11–20, New York, NY, USA, 2007. Association for Computing Machinery. ISBN 9781595936547. doi: 10.1145/1242572.1242575. URL <https://doi.org/10.1145/1242572.1242575>.
- [3] Xavier Amatriain and Justin Basilico. Recommender systems in industry: A netflix case study. In *Recommender systems handbook*, pages 385–419. Springer, 2015.
- [4] Mingxiao An, Fangzhao Wu, Chuhan Wu, Kun Zhang, Zheng Liu, and Xing Xie. Neural news recommendation with long- and short-term user representations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 336–345, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1033. URL <https://www.aclweb.org/anthology/P19-1033>.
- [5] Ofer Arazy, Nanda Kumar, and Bracha Shapira. Improving social recommender systems. *IT professional*, 11(4):38–44, 2009.
- [6] Nastaran Babanejad, Ameeta Agrawal, Heidar Davoudi, Aijun An, and Manos Papagelis. Leveraging emotion features in news recommendations. In *INRA@RecSys*, pages 70–78, 2019.
- [7] David Beirman et al. Restoring tourism destinations in crisis: A strategic marketing approach. *CAUTHE 2003: Riding the Wave of Tourism and Hospitality Research*, page 1146, 2003.
- [8] Robert M Bell and Yehuda Koren. Lessons from the netflix prize challenge. *Acm Sigkdd Explorations Newsletter*, 9(2):75–79, 2007.

- [9] David Ben-Shimon, Alexander Tsikinovsky, Lior Rokach, Amnon Meisles, Guy Shani, and Lihi Naamani. Recommender system from personal social networks. In *Advances in Intelligent Web Mastering: Proceedings of the 5th Atlantic Web Intelligence Conference–AWIC’2007, Fontainebleau, France, June 25–27, 2007*, pages 47–55. Springer, 2007.
- [10] Bettina Berendt. Text Mining for News and Blogs Analysis. In Claude Sammut and Geoffrey I. Webb, editors, *Encyclopedia of Machine Learning*, pages 968–972. Springer US, Boston, MA, 2010. ISBN 978-0-387-30164-8. doi: 10.1007/978-0-387-30164-8\\_827. URL [https://doi.org/10.1007/978-0-387-30164-8\\_827](https://doi.org/10.1007/978-0-387-30164-8_827).
- [11] Shlomo Berkovsky, Tsvi Kuflik, and Francesco Ricci. Mediation of user models for enhanced personalization in recommender systems. *User Modeling and User-Adapted Interaction*, 18:245–286, 2008.
- [12] Shlomo Berkovsky, Tsvi Kuflik, and Francesco Ricci. Cross-representation mediation of user models. *User Modeling and User-Adapted Interaction*, 19:35–63, 2009.
- [13] Steven Bird, Ewan Klein, and Edward Loper. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O’Reilly Media, Inc.", 2009.
- [14] Jesús Bobadilla, Fernando Ortega, Antonio Hernando, and Abraham Gutiérrez. Recommender systems survey. *Knowledge-based systems*, 46:109–132, 2013.
- [15] Joan Borràs, Antonio Moreno, and Aida Valls. Intelligent tourism recommender systems: A survey. *Expert Systems with Applications*, 41(16):7370–7389, 2014.
- [16] Matthias Braunhofer, Mehdi Elahi, and Francesco Ricci. Usability assessment of a context-aware and personality-based mobile recommender system. In *International conference on electronic commerce and web technologies*, pages 77–88. Springer, 2014.
- [17] Derek Bridge, Mehmet H Göker, Lorraine McGinty, and Barry Smyth. Case-based recommender systems. *The Knowledge Engineering Review*, 20(3):315–320, 2005.
- [18] Robin Burke. Hybrid web recommender systems. *The adaptive web: methods and strategies of web personalization*, pages 377–408, 2007.
- [19] Robin Burke and Maryam Ramezani. Matching recommendation technologies and domains. In *Recommender systems handbook*, pages 367–386. Springer, 2011.
- [20] Pedro G. Campos, Fernando Díez, and Iván Cantador. Time-aware recommender systems: A comprehensive survey and analysis of existing evaluation protocols. *User Modeling and User-Adapted Interaction*, 24(1–2):67–119, feb 2014. ISSN 0924-1868. doi: 10.1007/s11257-012-9136-x. URL <https://doi.org/10.1007/s11257-012-9136-x>.

- [21] Li Chen and Pearl Pu. Critiquing-based recommenders: survey and emerging trends. *User Modeling and User-Adapted Interaction*, 22:125–150, 2012.
- [22] Li Chen, Marco de Gemmis, Alexander Felfernig, Pasquale Lops, Francesco Ricci, and Giovanni Semeraro. Human decision making and recommender systems. *ACM Trans. Interact. Intell. Syst.*, 3(3), oct 2013. ISSN 2160-6455. doi: 10.1145/2533670.2533675. URL <https://doi.org/10.1145/2533670.2533675>.
- [23] Erik Cohen. Toward a sociology of international tourism. *Social research*, pages 164–182, 1972.
- [24] Walter Daelemans. Explanation in computational stylometry. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 451–462. Springer, 2013.
- [25] Abhinandan S. Das, Mayur Datar, Ashutosh Garg, and Shyam Rajaram. Google news personalization: Scalable online collaborative filtering. In *Proceedings of the 16th International Conference on World Wide Web, WWW '07*, page 271–280, New York, NY, USA, 2007. Association for Computing Machinery. ISBN 9781595936547. doi: 10.1145/1242572.1242610. URL <https://doi.org/10.1145/1242572.1242610>.
- [26] Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. GoEmotions: A Dataset of Fine-Grained Emotions. In *58th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2020.
- [27] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.
- [28] Shuiguang Deng, Dongjing Wang, Xitong Li, and Guandong Xu. Exploring user emotion in microblogs for music recommendation. *Expert Systems with Applications*, 42(23):9284–9293, 2015. ISSN 0957-4174. doi: <https://doi.org/10.1016/j.eswa.2015.08.029>. URL <https://www.sciencedirect.com/science/article/pii/S0957417415005746>.
- [29] Maunendra Sankar Desarkar and Neha Shinde. Diversification in news recommendation for privacy concerned users. In *2014 international conference on data science and advanced analytics (DSAA)*, pages 135–141. IEEE, 2014.
- [30] Linus W Dietz. Data-driven destination recommender systems. In *Proceedings of the 26th Conference on User Modeling, Adaptation and Personalization*, pages 257–260. ACM, 2018.
- [31] Linus W Dietz, Mete Sertkan, Saadi Myftija, Sameera Thimbiri Palage, Julia Neidhardt, and Wolfgang Wörndl. A comparative study of data-driven models for travel destination characterization. *Frontiers in big data*, 5:829939, 2022.

- [32] Cynthia Dwork, Ravi Kumar, Moni Naor, and D. Sivakumar. Rank aggregation methods for the web. In *Proceedings of the 10th International Conference on World Wide Web*, WWW '01, page 613–622, New York, NY, USA, 2001. Association for Computing Machinery. ISBN 1581133480. doi: 10.1145/371920.372165. URL <https://doi.org/10.1145/371920.372165>.
- [33] Paul Ekman. Are there basic emotions? *Psychological Review*, 1992.
- [34] Michael D. Ekstrand, F. Maxwell Harper, Martijn C. Willemsen, and Joseph A. Konstan. User perception of differences in recommender algorithms. In *Proceedings of the 8th ACM Conference on Recommender Systems*, RecSys '14, page 161–168, New York, NY, USA, 2014. Association for Computing Machinery. ISBN 9781450326681. doi: 10.1145/2645710.2645737. URL <https://doi.org/10.1145/2645710.2645737>.
- [35] Bruce Ferwerda and Marko Tkalcić. Predicting users' personality from instagram pictures: Using visual and/or content features? In *Proceedings of the 26th Conference on User Modeling, Adaptation and Personalization*, UMAP '18, pages 157–161, New York, NY, USA, 2018. ACM, Association for Computing Machinery. ISBN 9781450355896. doi: 10.1145/3209219.3209248. URL <https://doi.org/10.1145/3209219.3209248>.
- [36] Bruce Ferwerda, Markus Schedl, and Marko Tkalcić. Predicting personality traits with instagram pictures. In *Proceedings of the 3rd Workshop on Emotions and Personality in Personalized Systems 2015*, EMPIRE '15, pages 7–10, New York, NY, USA, 2015. ACM, Association for Computing Machinery. ISBN 9781450336154. doi: 10.1145/2809643.2809644. URL <https://doi.org/10.1145/2809643.2809644>.
- [37] Daniel R Fesenmaier, Karl W Wöber, and Hannes Werthner. *Destination recommendation systems: Behavioral foundations and applications*. Cabi, 2006.
- [38] City Foodsters. Pinenut-herb crusted rack of lamb with portabella mushroom, brussels sprouts, and black garlic. <https://www.flickr.com/photos/cityfoodsters/16468472701/>, 2019. Online, accessed 19-September-2019, licensed under CC BY-SA 2.0 (<https://creativecommons.org/licenses/by-sa/2.0>).
- [39] Linton C Freeman. A set of measures of centrality based on betweenness. *Sociometry*, pages 35–41, 1977.
- [40] Inma Garcia, Laura Sebastia, and Eva Onaindia. On the design of individual and group recommender systems for tourism. *Expert systems with applications*, 38(6): 7683–7692, 2011.

- [41] Suyu Ge, Chuhan Wu, Fangzhao Wu, Tao Qi, and Yongfeng Huang. Graph enhanced representation learning for news recommendation. In *Proceedings of The Web Conference 2020*, pages 2863–2869, 2020.
- [42] Heather Gibson and Andrew Yiannakis. Tourist roles: Needs and the lifecourse. *Annals of tourism research*, 29(2):358–383, 2002.
- [43] Lisa Glatzer, Julia Neidhardt, and Hannes Werthner. Automated assignment of hotel descriptions to travel behavioural patterns. In *Information and Communication Technologies in Tourism 2018*, pages 409–421. Springer, 2018.
- [44] David Goldberg, David Nichols, Brian M Oki, and Douglas Terry. Using collaborative filtering to weave an information tapestry. *Communications of the ACM*, 35(12):61–70, 1992.
- [45] Lewis R Goldberg. An alternative “description of personality”: the big-five factor structure. *Journal of personality and social psychology*, 59(6):1216, 1990. doi: <https://doi.org/10.1037/0022-3514.59.6.1216>.
- [46] John C Gower. A general coefficient of similarity and some of its properties. *Biometrics*, pages 857–871, 1971.
- [47] Ulrike Gretzel, NICOLE Mitsche, Yeong-Hyeon Hwang, Daniel R Fesenmaier, et al. Travel personality testing for destination recommendation systems. *Destination recommendation systems. Behavioural Foundations and Applications. Oxfordshire: CABI*, pages 121–136, 2006.
- [48] Asela Gunawardana and Guy Shani. A survey of accuracy evaluation metrics of recommendation tasks. *Journal of Machine Learning Research*, 10(Dec):2935–2962, 2009.
- [49] Songqiao Han, Hailiang Huang, and Jiangwei Liu. Neural news recommendation with event extraction. *arXiv preprint arXiv:2111.05068*, 2021.
- [50] Sebastian Hofstätter, Mete Sertkan, and Allan Hanbury. Tu wien at trec dl and podcast 2021: Simple compression for dense retrieval. In *Proceedings of Text REtrieval Conference (TREC)*, 2021.
- [51] Myles Hollander, Douglas A Wolfe, and Eric Chicken. *Nonparametric statistical methods*, volume 751. John Wiley & Sons, 2013.
- [52] Linmei Hu, Chen Li, Chuan Shi, Cheng Yang, and Chao Shao. Graph neural news recommendation with long-term and short-term interest modeling. *Information Processing & Management*, 57(2):102142, 2020.
- [53] C. Hutto and Eric Gilbert. Vader: A parsimonious rule-based model for sentiment analysis of social media text. *Proceedings of the International AAAI Conference on Web and Social Media*, 8(1), May 2014. URL <https://ojs.aaai.org/index.php/ICWSM/article/view/14550>.

- [54] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. Linear model selection and regularization. In *An Introduction to Statistical Learning*, pages 203–264. Springer, 2013.
- [55] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. Linear regression. In *An Introduction to Statistical Learning*, pages 59–126. Springer, 2013.
- [56] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. Statistical learning. In *An Introduction to Statistical Learning*, pages 15–57. Springer, 2013.
- [57] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. Tree-based methods. In *An Introduction to Statistical Learning*, pages 303–335. Springer, 2013.
- [58] Dietmar Jannach, Markus Zanker, Alexander Felfernig, and Gerhard Friedrich. *Online consumer decision making*, page 234–252. Cambridge University Press, 2010. doi: 10.1017/CBO9780511763113.012.
- [59] Dietmar Jannach, Markus Zanker, Alexander Felfernig, and Gerhard Friedrich. *Recommender Systems: An Introduction*. Cambridge University Press, USA, 1st edition, 2010. ISBN 0521493366.
- [60] Dietmar Jannach, Markus Zanker, Mouzhi Ge, and Marian Gröning. Recommender Systems in Computer Science and Information Systems – A Landscape of Research. In Christian Huemer and Pasquale Lops, editors, *E-Commerce and Web Technologies*, Lecture Notes in Business Information Processing, pages 76–87. Springer Berlin Heidelberg, 2012. ISBN 978-3-642-32273-0.
- [61] Mozhgan Karimi, Dietmar Jannach, and Michael Jugovac. News recommender systems – survey and roads ahead. *Information Processing & Management*, 54(6):1203–1227, 2018. ISSN 0306-4573. doi: <https://doi.org/10.1016/j.ipm.2018.04.008>. URL <https://www.sciencedirect.com/science/article/pii/S030645731730153X>.
- [62] Mozhgan Karimi, Dietmar Jannach, and Michael Jugovac. News recommender systems – Survey and roads ahead. *Information Processing & Management*, 54(6):1203–1227, November 2018. ISSN 0306-4573. doi: 10.1016/j.ipm.2018.04.008. URL <http://www.sciencedirect.com/science/article/pii/S030645731730153X>.
- [63] Andrej Karpathy. CS231n Convolutional Neural Networks for Visual Recognition. <http://cs231n.github.io/>, 2019. Online, accessed 20-June-2019.
- [64] Leonard Kaufman and Peter J Rousseeuw. Partitioning around medoids (program pam). *Finding groups in data: an introduction to cluster analysis*, pages 68–125, 1990.

- [65] Benjamin Kille, Frank Hopfgartner, Torben Brodt, and Tobias Heintz. The plista dataset. In *Proceedings of the 2013 International News Recommender Systems Workshop and Challenge*, NRS '13, page 16–23, New York, NY, USA, 2013. Association for Computing Machinery. ISBN 9781450323024. doi: 10.1145/2516641.2516643. URL <https://doi.org/10.1145/2516641.2516643>.
- [66] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
- [67] Thomas Ledl. Wien rathaus hochaufloesend. [https://commons.wikimedia.org/wiki/File:Wien\\_Rathaus\\_hochaufl%C3%B6send.jpg](https://commons.wikimedia.org/wiki/File:Wien_Rathaus_hochaufl%C3%B6send.jpg), 2019. Online, accessed 19-September-2019, licensed under CC BY-SA 4.0 (<https://creativecommons.org/licenses/by-sa/4.0>).
- [68] Elisabeth Lex, Dominik Kowald, Paul Seitlinger, Thi Ngoc Trang Tran, Alexander Felfernig, Markus Schedl, et al. *Psychology-informed recommender systems*. Now Publishers, 2021.
- [69] Pasquale Lops, Marco De Gemmis, and Giovanni Semeraro. Content-based recommender systems: State of the art and trends. *Recommender systems handbook*, pages 73–105, 2011.
- [70] Gerald Matthews, Ian J Deary, and Martha C Whiteman. *Personality traits*. Cambridge University Press, 2003.
- [71] Rahul Mazumder, Trevor Hastie, and Robert Tibshirani. Spectral regularization algorithms for learning large incomplete matrices. *Journal of machine learning research*, 11(Aug):2287–2322, 2010.
- [72] Ipsos MediaCT. The 2014 traveler’s road to decision. [https://www.thinkwithgoogle.com/\\_qs/documents/918/2014-travelers-road-to-decision\\_research\\_studies.pdf](https://www.thinkwithgoogle.com/_qs/documents/918/2014-travelers-road-to-decision_research_studies.pdf), June 2014. Online; accessed 14-July-2017.
- [73] Rico Meinel. Recommender Systems: The Most Valuable Application of Machine Learning, towards data science. <https://tinyurl.com/5csbkuy1>, 2020. [Online; accessed 30-August-2023].
- [74] Jan Mizgajski and Mikołaj Morzy. Affective recommender systems in online news industry: how emotions influence reading choices. *User Modeling and User-Adapted Interaction*, 29(2):345–379, 2019.
- [75] Irina Nalis and Julia Neidhardt. Not facial expression, nor fingerprint – acknowledging complexity and context in emotion research for human-centered personalization and adaptation. In *Adjunct Proceedings of the 31st ACM Conference on User Modeling, Adaptation and Personalization*, UMAP '23 Adjunct,

page 325–330, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9781450398916. doi: 10.1145/3563359.3596990. URL <https://doi.org/10.1145/3563359.3596990>.

- [76] Julia Neidhardt, Rainer Schuster, Leonhard Seyfang, and Hannes Werthner. Eliciting the users’ unknown preferences. In *Proceedings of the 8th ACM Conference on Recommender Systems*, RecSys ’14, pages 309–312, New York, NY, USA, 2014. ACM. ISBN 978-1-4503-2668-1. doi: 10.1145/2645710.2645767. URL <http://doi.acm.org/10.1145/2645710.2645767>.
- [77] Julia Neidhardt, Leonhard Seyfang, Rainer Schuster, and Hannes Werthner. A picture-based approach to recommender systems. *Information Technology & Tourism*, 15(1):49–69, 2015. doi: 10.1007/s40558-014-0017-5.
- [78] Office for National Statistics (UK). (n.d.). Share of individuals reading or downloading online news, newspapers or magazines in great britain from 2007 to 2018, Aug 2018. URL <https://www.statista.com/statistics/286210/online-news-newspapers-and-magazine-consumption-in-great-britain/>.
- [79] Özlem Özgöbek, Jon Atle Gulla, and R Cenk Erdur. A survey on challenges and methods in news recommendation. In *International Conference on Web Information Systems and Technologies*, volume 2, pages 278–285. SCITEPRESS, 2014.
- [80] Tulasi K Paradarami, Nathaniel D Bastian, and Jennifer L Wightman. A hybrid recommender system using artificial neural networks. *Expert Systems with Applications*, 83:300–313, 2017.
- [81] PL Pearce et al. The social psychology of tourist behaviour. *The social psychology of tourist behaviour.*, 1982.
- [82] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014. URL <http://www.aclweb.org/anthology/D14-1162>.
- [83] Steven’s Transport Photos. Dscf8575. <https://www.flickr.com/photos/fwc439h/3855936048/>, 2019. Online, accessed 19-September-2019, licensed under CC BY-SA 2.0 (<https://creativecommons.org/licenses/by-sa/2.0>).
- [84] Pearl Pu, Li Chen, and Rong Hu. A user-centric evaluation framework for recommender systems. In *Proceedings of the Fifth ACM Conference on Recommender Systems*, RecSys ’11, page 157–164, New York, NY, USA, 2011. Association for Computing Machinery. ISBN 9781450306836. doi: 10.1145/2043932.2043962. URL <https://doi.org/10.1145/2043932.2043962>.

- [85] Paul Resnick, Neophytos Iacovou, Mitesh Suchak, Peter Bergstrom, and John Riedl. GroupLens: An open architecture for collaborative filtering of netnews. In *Proceedings of the 1994 ACM Conference on Computer Supported Cooperative Work, CSCW '94*, page 175–186, New York, NY, USA, 1994. Association for Computing Machinery. ISBN 0897916891. doi: 10.1145/192844.192905. URL <https://doi.org/10.1145/192844.192905>.
- [86] Michael Reusens, Wilfried Lemahieu, Bart Baesens, and Luc Sels. A note on explicit versus implicit information for job recommendation. *Decision Support Systems*, 98: 26–35, 2017.
- [87] Francesco Ricci. Travel recommender systems. *IEEE Intelligent Systems*, 17(6): 55–57, 2002.
- [88] Francesco Ricci, Lior Rokach, and Bracha Shapira. Recommender systems: introduction and challenges. In *Recommender systems handbook*, pages 1–34. Springer, Boston, MA, 2015. ISBN 978-1-4899-7637-6. doi: 10.1007/978-1-4899-7637-6\_1. URL [https://doi.org/10.1007/978-1-4899-7637-6\\_1](https://doi.org/10.1007/978-1-4899-7637-6_1).
- [89] Francesco Ricci, Lior Rokach, and Bracha Shapira. *Recommender Systems: Techniques, Applications, and Challenges*, pages 1–35. Springer US, New York, NY, 2022. ISBN 978-1-0716-2197-4. doi: 10.1007/978-1-0716-2197-4\_1. URL [https://doi.org/10.1007/978-1-0716-2197-4\\_1](https://doi.org/10.1007/978-1-0716-2197-4_1).
- [90] Peter J Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65, 1987.
- [91] Joy Lal Sarkar and Abhishek Majumder. A new point-of-interest approach based on multi-itinerary recommendation engine. *Expert Systems with Applications*, 181: 115026, 2021.
- [92] J Ben Schafer, Joseph A Konstan, and John Riedl. E-commerce recommendation applications. *Data mining and knowledge discovery*, 5:115–153, 2001.
- [93] J Ben Schafer, Dan Frankowski, Jon Herlocker, and Shilad Sen. Collaborative filtering recommender systems. In *The adaptive web: methods and strategies of web personalization*, pages 291–324. Springer, 2007.
- [94] Mete Sertkan and Julia Neidhardt. Exploring expressed emotions for neural news recommendation. In *Adjunct Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization, UMAP '22 Adjunct*, page 22–28, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450392327. doi: 10.1145/3511047.3536414. URL <https://doi.org/10.1145/3511047.3536414>.

- [95] Mete Sertkan and Julia Neidhardt. On the effect of incorporating expressed emotions in news articles on diversity within recommendation models. In *Proceedings of the 11th International Workshop on News Recommendation and Analytics*, 2023.
- [96] Mete Sertkan, Julia Neidhardt, and Hannes Werthner. Mapping of tourism destinations to travel behavioural patterns. In *Information and Communication Technologies in Tourism 2018*, pages 422–434. Springer, 2018.
- [97] Mete Sertkan, Julia Neidhardt, and Hannes Werthner. From pictures to touristic profiles: A deep-learning based approach. In *Proceedings of the 1st International ‘Alan Turing’ Conference on Decision Support and Recommender Systems, DSRS-Turing’19*, pages 75–78, London, UK, 2019. The Alan Turing Institute. ISBN 978-1-5262-0820-0.
- [98] Mete Sertkan, Julia Neidhardt, and Hannes Werthner. Documents, topics, and authors: Text mining of online news. In *2019 IEEE 21st Conference on Business Informatics (CBI)*, volume 01, pages 405–413, 2019. doi: 10.1109/CBI.2019.00053.
- [99] Mete Sertkan, Julia Neidhardt, and Hannes Werthner. What is the “personality” of a tourism destination? *Information Technology & Tourism*, 21(1):105–133, March 2019. ISSN 1943-4294. doi: 10.1007/s40558-018-0135-6. URL <https://doi.org/10.1007/s40558-018-0135-6>.
- [100] Mete Sertkan, Julia Neidhardt, and Hannes Werthner. Eliciting touristic profiles: A user study on picture collections. In *Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization, UMAP ’20*, page 230–238, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450368612. doi: 10.1145/3340631.3394868. URL <https://doi.org/10.1145/3340631.3394868>.
- [101] Mete Sertkan, Julia Neidhardt, and Hannes Werthner. Pictoure - a picture-based tourism recommender. In *Proceedings of the 14th ACM Conference on Recommender Systems, RecSys ’20*, page 597–599, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450375832. doi: 10.1145/3383313.3411526. URL <https://doi.org/10.1145/3383313.3411526>.
- [102] Mete Sertkan, Julia Neidhardt, and Hannes Werthner. From pictures to travel characteristics: Deep learning-based profiling of tourists and tourism destinations. In Julia Neidhardt and Wolfgang Wörndl, editors, *Information and Communication Technologies in Tourism 2020*, pages 142–153, Cham, 2020. Springer International Publishing. ISBN 978-3-030-36737-4. doi: 10.1007/978-3-030-36737-4\_12. URL [https://doi.org/10.1007/978-3-030-36737-4\\_12](https://doi.org/10.1007/978-3-030-36737-4_12).
- [103] Mete Sertkan, Sofia Althammer, Sebastian Hoftstätter, and Julia Neidhardt. Diversifying sentiments in news recommendation. In *Proceedings of the 2nd Perspectives on the Evaluation of Recommender Systems Workshop*, 2022.

- [104] Herbert A Simon. A behavioral model of rational choice. *The quarterly journal of economics*, pages 99–118, 1955.
- [105] Mark D. Smucker, James Allan, and Ben Carterette. A comparison of statistical significance tests for information retrieval evaluation. In *Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management, CIKM '07*, page 623–632, New York, NY, USA, 2007. Association for Computing Machinery. ISBN 9781595938039. doi: 10.1145/1321440.1321528. URL <https://doi.org/10.1145/1321440.1321528>.
- [106] spacy. spacy: Industrial-strength natural language processing in python, 2024. URL <https://spacy.io>.
- [107] Pang-Ning Tan, Michael Steinbach, Vipin Kumar, et al. Cluster analysis: basic concepts and algorithms. *Introduction to data mining*, 8:487–568, 2006.
- [108] Pang-Ning Tan, Michael Steinbach, and Vipin Kumar. *Introduction to data mining*. Pearson, Harlow, first edition, pearson new international edition edition, 2014. ISBN 978-1-292-02615-2. OCLC: 881290060.
- [109] Erich Christian Teppan and Markus Zanker. Decision biases in recommender systems. *Journal of Internet Commerce*, 14(2):255–275, 2015. doi: 10.1080/15332861.2015.1018703. URL <https://doi.org/10.1080/15332861.2015.1018703>.
- [110] Julián Urbano, Harley Lima, and Alan Hanjalic. Statistical significance testing in information retrieval: An empirical analysis of type i, type ii and type iii errors. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR'19*, page 505–514, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450361729. doi: 10.1145/3331184.3331259. URL <https://doi.org/10.1145/3331184.3331259>.
- [111] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017.
- [112] Sholom M. Weiss, Nitin Indurkha, and Tong Zhang. *Fundamentals of Predictive Text Mining*. Texts in Computer Science. Springer London, London, 2015. ISBN 978-1-4471-6749-5 978-1-4471-6750-1. doi: 10.1007/978-1-4471-6750-1. URL <http://link.springer.com/10.1007/978-1-4471-6750-1>.
- [113] Hannes Werthner and Stefan Klein. *Information technology and tourism: a challenging relationship*. Springer-Verlag Wien, Wien, New York, 1999. ISBN 3-211-83274-2.
- [114] Hannes Werthner and Francesco Ricci. E-commerce and tourism. *Commun. ACM*, 47(12):101–105, December 2004. ISSN 0001-0782. doi: 10.1145/1035134.1035141. URL <http://doi.acm.org/10.1145/1035134.1035141>.

- [115] Amy B Woszczynski, Philip L Roth, and Albert H Segars. Exploring the theoretical foundations of playfulness in computer interactions. *Computers in Human Behavior*, 18(4):369–388, 2002.
- [116] Chuhan Wu, Fangzhao Wu, Mingxiao An, Jianqiang Huang, Yongfeng Huang, and Xing Xie. Neural news recommendation with attentive multi-view learning. *arXiv preprint arXiv:1907.05576*, 2019.
- [117] Chuhan Wu, Fangzhao Wu, Mingxiao An, Yongfeng Huang, and Xing Xie. Neural news recommendation with topic-aware news representation. In *Proceedings of the 57th Annual meeting of the association for computational linguistics*, pages 1154–1159, 2019.
- [118] Chuhan Wu, Fangzhao Wu, Suyu Ge, Tao Qi, Yongfeng Huang, and Xing Xie. Neural news recommendation with multi-head self-attention. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6389–6394, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1671. URL <https://www.aclweb.org/anthology/D19-1671>.
- [119] Chuhan Wu, Fangzhao Wu, Tao Qi, and Yongfeng Huang. SentiRec: Sentiment diversity-aware neural news recommendation. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 44–53, Suzhou, China, December 2020. Association for Computational Linguistics. URL <https://www.aclweb.org/anthology/2020.aacl-main.6>.
- [120] Chuhan Wu, Fangzhao Wu, Tao Qi, and Yongfeng Huang. User modeling with click preference and reading satisfaction for news recommendation. In *IJCAI*, pages 3023–3029, 2020.
- [121] Chuhan Wu, Fangzhao Wu, Tao Qi, and Yongfeng Huang. Empowering news recommendation with pre-trained language models. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1652–1656, 2021.
- [122] Chuhan Wu, Fangzhao Wu, Tao Qi, Chenliang Li, and Yongfeng Huang. Is news recommendation a sequential recommendation task? In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2382–2386, 2022.
- [123] Chuhan Wu, Fangzhao Wu, Yongfeng Huang, and Xing Xie. Personalized news recommendation: Methods and challenges. *ACM Transactions on Information Systems*, 41(1):1–50, 2023.

- [124] Fangzhao Wu, Ying Qiao, Jiun-Hung Chen, Chuhan Wu, Tao Qi, Jianxun Lian, Danyang Liu, Xing Xie, Jianfeng Gao, Winnie Wu, and Ming Zhou. MIND: A large-scale dataset for news recommendation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3597–3606, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.331. URL <https://www.aclweb.org/anthology/2020.acl-main.331>.
- [125] Andrew Yiannakis and Heather Gibson. Roles tourists play. *Annals of Tourism Research*, 19(2):287–303, 1992.
- [126] Shuai Zhang, Lina Yao, Aixin Sun, and Yi Tay. Deep learning based recommender system: A survey and new perspectives. *ACM Comput. Surv.*, 52(1), February 2019. ISSN 0360-0300. doi: 10.1145/3285029. URL <https://doi.org/10.1145/3285029>.
- [127] Qian Zhao, F. Maxwell Harper, Gediminas Adomavicius, and Joseph A. Konstan. Explicit or implicit feedback? engagement or satisfaction? a field experiment on machine-learning-based recommender systems. In *Proceedings of the 33rd Annual ACM Symposium on Applied Computing, SAC '18*, page 1331–1340, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450351911. doi: 10.1145/3167132.3167275. URL <https://doi.org/10.1145/3167132.3167275>.
- [128] Andreas H Zins. Exploring travel information search behavior beyond common frontiers. *Information Technology & Tourism*, 9(3-1):149–164, 2007.