INSTITUTO SUPERIOR DE ENGENHARIA DO PORTO
MESTRADO EM ENGENHARIA DE INTELIGÊNCIA ARTIFICIAL




Hybrid Sentiment-Based Recommender System for E-Commerce

ANA PATRÍCIA MANTA MOREIRA
outubro de 2025

P.PORTO

Hybrid Sentiment-Based Recommender System for E-Commerce

Ana Patrícia Manta Moreira
Student No.: 1222731

A dissertation submitted in partial fulfillment of the requirements for the degree of Master of Science, Specialization Area of Artificial Intelligence Engineering

Supervisor: Joaquim Filipe Peixoto dos Santos, Associate Professor, Institute of Engineering, Polytechnic of Porto

Evaluation Committee:
President:
Diogo Emanuel Pereira Martinho, Assistant Professor, Institute of Engineering, Polytechnic of Porto

Members:
António Constantino Lopes Martins, Associate Professor, Institute of Engineering, Polytechnic of Porto
Joaquim Filipe Peixoto dos Santos, Associate Professor, Institute of Engineering, Polytechnic of Porto

Porto, September 29, 2025

Abstract

This dissertation presents a personalized product recommendation system designed for the use in e-commerce, which has a lot of sentiments inside reviews made by users. Conventionally, most recommendation systems tend to ignore the fine details and emotive sentiments in this user-contributed texts, often relying on numerical ratings or demographical information such as gender or age. This can result in suggestions not entirely in line with a user's genuine interest or emotional reaction to products. A user could have assigned a high numerical rating but still express discontent with a particular aspect such as the durability or the cost in the review, this detail is often lost in the case of most standard methods.

This initiative has been created in partnership with the Cognitively Smart Assistant in Physical-Digital Environment (CAPE) project. CAPE is a joint initiative, subsidized by the European Regional Development Fund (ERDF), dedicated to the triple transformation of the retail market: sustainability, digitalization, and evolution of skills.

In preparation for this system, a systematic literature review took place, which served as essential groundwork for the state of the art of the current e-commerce recommendation systems, including those with emotions, Artificial Intelligence (AI), and Natural Language Processing (NLP). This review was structured around specific research questions about how emotions impact the recommendation, the most efficient methods of segmenting the user for the highest possible sales, and possible weaknesses and the ethics in the recommendation systems. This systematic review followed the PRISMA protocol.

The dataset chosen for the project was the Amazon Electronics dataset, chosen for the wide range of electronic products and the large collection of reviews, which are both ideally suited for the needs of the CAPE project. Because of its large size, there was a critical preprocessing of the data to improve the quality of the data and accelerate the model training.

The suggested system involves the use of the strengths of BERT for learning the context of the words and the LSTM networks for dealing with long-term dependency in sequential information. A range of this architectures (V1 to V6) was created in order to find the optimum point of results and time, which was measured by the Mean Squared Error (MSE) metric.

It was also placed a particular focus on the consideration of ethics, seeking to overcome common problems such as the avoidance of social biases. Data protection legislation such as the GDPR (General Data Protection Regulation) and the AI Act (Artificial Intelligence Act) are complied with by the use of an anonymized public dataset and by the use of the pseudonymous user IDs with the avoidance of sensitive personal attributes in order to reduce bias. The ultimate goal of this research is to improve purchase probability and overall user satisfaction, providing increased customer loyalty by the use of a more accurate and ethically aware recommendation system.

Keywords: Sentiment Analysis, Recommendation System, E-Commerce, BERT, LSTM

Resumo

Esta dissertação apresenta um sistema de recomendação personalizado de produtos, pensado para o comércio eletrónico, que tem em conta os sentimentos presentes nas avaliações feitas pelos utilizadores. Normalmente, a maioria dos sistemas de recomendação ignora estes detalhes e sentimentos nos textos escritos pelos utilizadores, baseando-se apenas em classificações numéricas ou em informações demográficas, como o género ou a idade. Isto pode levar a sugestões que não estão totalmente de acordo com os verdadeiros interesses de um comprador ou com a sua reação emocional ao produto. Um utilizador pode, por exemplo, dar uma nota alta, mas ainda assim mostrar insatisfação com algum aspeto, como a durabilidade ou o preço. Este tipo de detalhe perde-se muitas vezes nos métodos tradicionais.

Este projeto foi desenvolvido em parceria com a iniciativa Cognitively Smart Assistant in Physical-Digital Environment (CAPE). O CAPE é um projeto conjunto, financiado pelo Fundo Europeu de Desenvolvimento Regional (ERDF), que tem como objetivo transformar o mercado retalhista em três áreas: sustentabilidade, digitalização e evolução das competências.

Para preparar este sistema, foi feita uma revisão sistemática da literatura, que serviu como base para conhecer o estado atual dos sistemas de recomendação no comércio eletrónico, incluindo aqueles que usam emoções, Inteligência Artificial (IA) e Processamento de Linguagem Natural (NLP). A revisão centrou-se em perguntas de investigação sobre como as emoções influenciam as recomendações, quais os métodos mais eficazes para segmentar os utilizadores de forma a aumentar as vendas, bem como as limitações e questões éticas destes sistemas. Esta revisão seguiu o protocolo PRISMA.

O conjunto de dados escolhido foi o Amazon Electronics, devido à grande variedade de produtos eletrónicos e ao elevado número de avaliações, que se ajustam bem às necessidades do projeto CAPE. Como o conjunto de dados era muito extenso, foi necessário fazer um pré-processamento para melhorar a qualidade e acelerar o treino do modelo.

O sistema proposto combina as vantagens do BERT, que aprende o contexto das palavras, com as redes LSTM, que conseguem lidar com sequências de informação ao longo do tempo. Foram criadas várias arquiteturas (V1 a V6) para encontrar o melhor equilíbrio entre resultados e tempo, sendo avaliadas com a métrica Mean Squared Error (MSE).

Também foi dada atenção à parte ética, procurando evitar problemas comuns, como os enviesamentos sociais. A legislação de proteção de dados, como o GDPR (Regulamento Geral sobre a Proteção de Dados) e o AI Act (Lei da Inteligência Artificial), é respeitada através da utilização de um conjunto de dados público já anonimizado e de IDs de utilizadores fictícios, sem incluir dados pessoais sensíveis, para reduzir esse enviesamento. O objetivo final desta investigação é aumentar a probabilidade de compra e a satisfação dos utilizadores, promovendo uma maior fidelização dos clientes com um sistema de recomendação mais preciso e consciente em termos éticos.

Palavras-chave: Sentiment Analysis, Recommendation System, E-Commerce, BERT, LSTM

Contents

List of Figures

List of Tables

List of Abbreviations

AI	**A**rtificial **I**ntelligence
BERT	**B**idirectional **E**ncoder **R**epresentations from **T**ransformers
CAPE	**C**ognitively Smart **A**ssistant in **P**hysical-Digital **E**nvironment
EU	**E**uropean **U**nion
LSTM	**L**ong **S**hort-**T**erm **M**emory
NLP	**N**atural **L**anguage **P**rocessing
PRISMA	**P**referred **R**eporting **I**tems for **S**ystematic Reviews and **M**eta-**A**nalyses

List of More Abbreviations

Studies:

ABOM	**A**spect **B**ased **O**pinion **M**ining
CARP	**CA**psule Network-Based Model for **R**ating **P**rediction
CF2	**C**ounterfactual **F**eature-aware **C**ollaborative **F**iltering
EX3	**EX**tract-**EX**pect-**EX**plain
LSWH	**L**inear **S**uperposition of **W**ord2Vec and **H**owNet
MSVA	**M**ultiscale **S**emantic-**V**isual **A**nalysis
PicTouRe	**Pic**ture-based **Tou**rism **Re**commender
REAO	**R**ecommendation system **E**xploiting **A**spect-based **O**pinion mining
SABTMCF	**S**entiment **A**nalysis and **BTM** **C**ollaborative **F**iltering
SAKG	**S**entiment-**A**ware **K**nowledge **G**raph
SDRA	**S**entiment-aware **D**eep **R**ecommender system with neural **A**ttention network
VARS	**V**alue-**A**ware **R**ecommender **S**ystems

Metrics and Other:

CNN	**C**onvolutional **N**eural **N**etwork
CTR	**C**lick-**T**hrough **R**ate
DPIA	**D**ata **P**rotection **I**mpact **A**ssessment
EEG	**E**lectro**e**ncephalo**g**ram
ERDF	**E**uropean **R**egional **D**evelopment **F**und
HR	**H**it **R**atio
MCNN	**M**ultichannel deep **C**onvolutional **N**eural **N**etwork
MSE	**M**ean **S**quared **E**rror
NDCG	**N**ormalized **D**iscounted **C**umulative **G**ain
POS	**P**arts **O**f **S**peech
TF	**T**ensor **F**actorization

Compared Studies:

A2CF	**A**ttribute-**A**ware **C**ollaborative **F**iltering
A3NCF	**A**daptive **A**spect **A**ttention-based **N**eural **C**ollaborative **F**iltering
ADAC	**A**daptive **D**emonstration-**A**ugmented **C**ollaborative Filtering
ALS	**A**lternating **L**east **S**quares
BPR	**B**ayesian **P**ersonalized **R**anking
BTM	**B**iterm **T**opic **M**odel
CF	**C**ollaborative **F**iltering
DRR	**D**ynamic **R**eview-based **R**ecommenders
EFM	**E**xplicit **F**actor **M**odel
HFT	**H**idden **F**actors and **T**opics
KGAT	**K**nowledge **G**raph **AT**tention Network
LDA	**L**atent-**D**irichlet **A**llocation
MF	**M**atrix **F**actorization

MPCN	**M**ulti-**P**ointer **C**o-Attention **N**etworks
MTER	**M**ulti**T**ask **E**xplainable **R**ecommendation
NCF	**N**eural **C**ollaborative **F**iltering
PGPR	**P**ath-based **G**ra**P**h **R**easoning
PROMETHEE II	**P**reference **R**anking **O**rganization **METH**od for **E**nrichment **E**valuation **II**
SACF	**S**entiment **A**nalysis **C**ollaborative **F**iltering
SVD++	**S**ingular **V**alue **D**ecomposition ++
TOPSIS	**T**echnique for **O**rder **P**reference by **S**imilarity to **I**deal **S**olution
VAR	**V**isual **A**dversarial **R**ecommender
VIKOR	**VI**ekriterijumsko **KO**mpromisno **R**angiranje

Chapter 1

Introduction

1.1 Context

Reviews have become an essential tool to buyers when it's time to get products with desired characteristics[1–3]. To address the quality of the products, users usually use reviews, looking at important characteristics that they consider relevant, and sometimes the price, to reduce errors in buying something that they like [4–7]. Sentiments are also used while writing those reviews, revealing user's emotions towards certain aspects of the product that are important to that specific user, for example, a user might rate a product highly overall, but express dissatisfaction with a particular feature, such as durability or price [1, 2]. Traditional recommendation systems, tend to rely heavily on static user-item interactions, like numerical ratings, gender or age, often overlooking the semantics and sentiments found in user reviews [1, 8–10].

This leads to missed opportunities for improving personalization and accuracy. Emerging techniques using reviews to capture sentiments have proven effective when compared with older systems. These techniques might innovate in areas using visual elements to help determine the best recommendations [8, 11], or even using counterfactual reasoning or fuzzy logic [3, 12], to get to the best recommendations in a different way. This use of sentiments in recommendation systems can significantly enhance the user satisfaction, making the likelihood of a purchase also bigger.

1.2 Problem Description

The primary challenge in a recommendation system is the inconsistencies between ratings and sentiments, for example in a review, a high numerical rating might coexist with negative sentiments [13]. In this way, a user can like a product because it's cheap but would still like it better if the battery lasted more. More traditional systems, that often rely solely on the numerical ratings, fail to capture these nuances, resulting in suboptimal recommendations.

To combat this problem, usually new recommender systems use sentiments inside text reviews. The challenge is to figure out how to use this data inside reviews and how to process it to be better. Reviews often contain descriptions and user important opinions about certain aspects of the product, such as "battery life" or "price", which are frequently overlooked in traditional models but very useful to better understand hidden preferences [1]. Also the importance of certain aspects can vary across users and contexts, for example different locations [14] or even different preferences that are often correlated with gender [15].

This research project provides a recommendation system that uses sentiment analysis to improve the precision and applicability of product suggestions with the collaboration of the Cognitively Smart Assistant in Physical-Digital Environment (CAPE) project. By analyzing user reviews, the system aims to create personalized product categories to individual buyers. This approach seeks to increase the chance of purchases and improve user satisfaction, helping e-commerce have more loyal customers.

1.3 CAPE

The CAPE project which is also known as Cognitively Smart Assistant in Physical-Digital Environment was funded by the European Regional Development Fund (ERDF). This project was possible with the collaboration between 15 partners spread across Portugal, Korea, Romania, Singapore, and Turkey, that are operating in technology, retail, and aviation sectors.

Challenges such as resource management, employee tracking, data privacy, customer engagement, and mood recognition can be solved with the use of technologies like Smart Dialogue Systems, Facial Emotion Recognition, IoT-based recommendations, and Federated Learning to comply with the GDPR. CAPE is used usually within the range of smart in-store shopping assistants to AI-powered workplace monitoring for safety and efficiency. Its objective is to develop innovative solutions to improve both customer shopping experiences and employee work environments, by then boosting sales while still supporting sustainable operations.

Chapter 2

State of Art

Natural Language Processing(NLP) has been evolving a lot lately, so naturally a lot of diverse literature research has been conducted in various areas of NLP to improve it more [4, 5, 8]. A comprehensive understanding of the current papers in this domain is crucial for situating this work within the broader academic and practical context [4, 5, 16] as well as focus on the more important aspects. So, in this chapter there will examined the current state of recommendation systems, especially on emotions, using AI and NLP, to improve the experience in e-commerce.

2.1 Systematic Review

A systematic review is a method to identify, evaluate, and compile all of the relevant studies on a certain subject. It will stick to a predetermined methodology and apply certain criteria for study selection and evaluation in an effort to reduce bias. It also aids in identifying areas that need more investigation [17].

2.1.1 Methodology

The Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) approach will be implemented in this systematic review. PRISMA is a precise framework intended to guide researchers as they create systematic reviews and meta-analyses. To keep the process organized and easy to understand for the reader, it outlines the procedures required for finding, vetting, and choosing studies. The flow diagram, which illustrates the research selection procedure at different phases, is an important part of PRISMA. The scientific community also acknowledges this method as an indicator of high reliability and quality [17].

2.1.2 Data Sources

Three popular electronic databases are selected for this study: *IEEE Xplore Digital Library*[18], *ScienceDirect*[19], and *ACM Digital Library*[20].

IEEE Xplore, hosted by the Institute of Electrical and Electronics Engineers, is known for its large repository of articles on electrical engineering, computer science, and related areas; ScienceDirect is an Elsevier database that has a wide range of research on science, technology, and engineering; the ACM Digital Library from the Association for Computing Machinery, focuses on computing and information technology, offering many key publications in computer science.

These databases have been targeted in this study because they are among the most famous and, therefore, offer valid and up-to-date research materials.

2.1.3 Research Questions

Research questions play a crucial role in shaping the focus and direction of a study. These questions aim to explore important aspects that affect user engagement and business outcomes. The emphasis is on understanding the functionality of these systems, pinpointing ways to enhance them, and addressing any potential challenges that might come up during their implementation. Below are the selected questions:

1. How can emotions play a role in recommendation systems for e-commerce to help users and shop owners?
2. What are the most effective ways of segmenting users to maximize sales?
3. What flaws can a recommendation system have in online shopping, and can it be ethical?

The first question reveals how the underlying emotions could be used to build recommendation systems capable of suggesting product categories that, besides being relevant to the user, will enhance the overall satisfaction and therefore will increase the likelihood of a purchase.

The second one looks at how to categorize products or users in the recommendation system, aiming to find the best ways to organize product information which will contribute to the increase in sales and improvement of the customer experience.

The third question explores the importance of ethics in developing recommendation systems. Since these systems influence consumer behavior, it's crucial to include principles of fairness, transparency, and respect for privacy in their design and use, so that they can meet modern ethical requirements referring to technology.

Together, these three questions offer the foundation around which the recommendation system need to be developed, achieving a balance amongst innovation, user involvement, and ethical responsibility.

2.1.4 Search Terms

The search terms used to do the systematic review across three databases are laid out in this subsection. With the objective to target research that focus on recommendation systems, user reviews, and sentiment analysis in the context of e-commerce, the queries were carefully planned to be relevant.

The following terms were chosen considering the explanation below for the creation of the questions of this work:

1. "Recommendation System," "Product Recommendation," or "Recommender System": These terms refer to systems designed to suggest products to users based on their preferences and behavior.
2. "Review": Represents user-generated content about products. Reviews are a key source of information for understanding customer opinions, which are crucial for building effective recommendation systems.

3. “Sentiment,” “Emotion,” or “Feeling”: These terms were selected to obtain studies focused on extracting emotions from user reviews. Sentiment analysis helps to interpret how users perceive products and improves the personalization of recommendations.
4. “E-commerce,” “Online Shopping,” “Online Retail,” or “Marketplace”: These terms ensure that the focus remains on online selling platforms.
5. “Artificial Intelligence,” “AI,” and “NLP”: These terms were included to focus on technologies that revolve around AI and NLP, which are essential for analyzing reviews and generating personalized recommendations.
6. “Product Categories” or “Product Classification”: These terms were added to include studies that address the organization of products into categories or classifications in Recommendation Systems.

The original question was broken down into four smaller searches with different word combinations in ScienceDirect due to the restriction of a maximum of eight logical operators per query. This method ensures that pertinent articles that may have been overlooked because of variations in vocabulary are still found and taken into account. In addition, terms like "Product Categories" and "Product Classification" had to be included in ScienceDirect since, without them, the results would have been too broad and less pertinent to the desired goal.

Here are the queries and their results, displayed in the table below:

Queries	ACM	IEEE Xplore	ScienceDirect
("Recommendation System" OR "Product Recommendation" OR "Recommender System") AND "Review" AND ("Sentiment" OR "Emotion" OR "Feeling") AND ("E-commerce" OR "Online Shopping" OR "Online Retail" OR "Marketplace") AND ("Artificial Intelligence" OR "AI" OR "NLP")	594	11	-
"Recommendation System" AND "Review" AND ("Sentiment" OR "Emotion") AND ("E-commerce" OR "Online Shopping") AND ("AI" OR "NLP") AND ("Product Categories")	-	-	50
"Recommendation System" AND "Review" AND ("Sentiment" OR "Emotion") AND ("E-commerce" OR "Online Shopping") AND ("AI" OR "NLP") AND ("Product Classification")	-	-	6
"Recommendation System" AND "Review" AND ("Sentiment" OR "Emotion") AND ("E-commerce" OR "Online Shopping") AND ("Artificial Intelligence" OR "NLP") AND ("Product Categories")	-	-	57
"Recommendation System" AND "Review" AND ("Sentiment" OR "Emotion") AND ("E-commerce" OR "Online Shopping") AND ("Artificial Intelligence" OR "NLP") AND ("Product Classification ")	-	-	3

Table 2.1: Research Queries and results on the chosen databases, totalling 721 results.

2.1.5 Inclusion and Exclusion Criteria

Inclusion Criteria

In this systematic review, studies were included if they met the following inclusion criteria:

1. The recommendation method takes into consideration emotional aspects;
2. It is specifically designed for online stores;
3. It uses techniques from NLP or AI;
4. It categorizes and sorts products using some characteristics or classes.

Exclusion Criteria

While studies were excluded if they met any of the following exclusion criteria:

1. They are duplicate studies;
2. They are not conference proceedings;

3. It does not make use of user review data to generate recommendations;
4. The approach does not offer personalized recommendations for users.

2.1.6 Quality Assessment

In total, 721 sources were found with the proposed research queries in the different databases. Eliminating the conference proceedings from these brought the number down to 346 papers. After removal of duplicates, there were 293 papers that underwent abstract screening. In this stage, the sources were checked whether they met either the inclusion or exclusion criteria. 260 papers were discarded leaving 33 that were potentially relevant. These potentially relevant sources were subjected to another process, which was reading the introductions and conclusions. This resulted in 11 papers being discarded and 22 classified as relevant. The final 22 papers were further analyzed to answer the research questions. Figure 2.1 displays a flow diagram of the selection process.

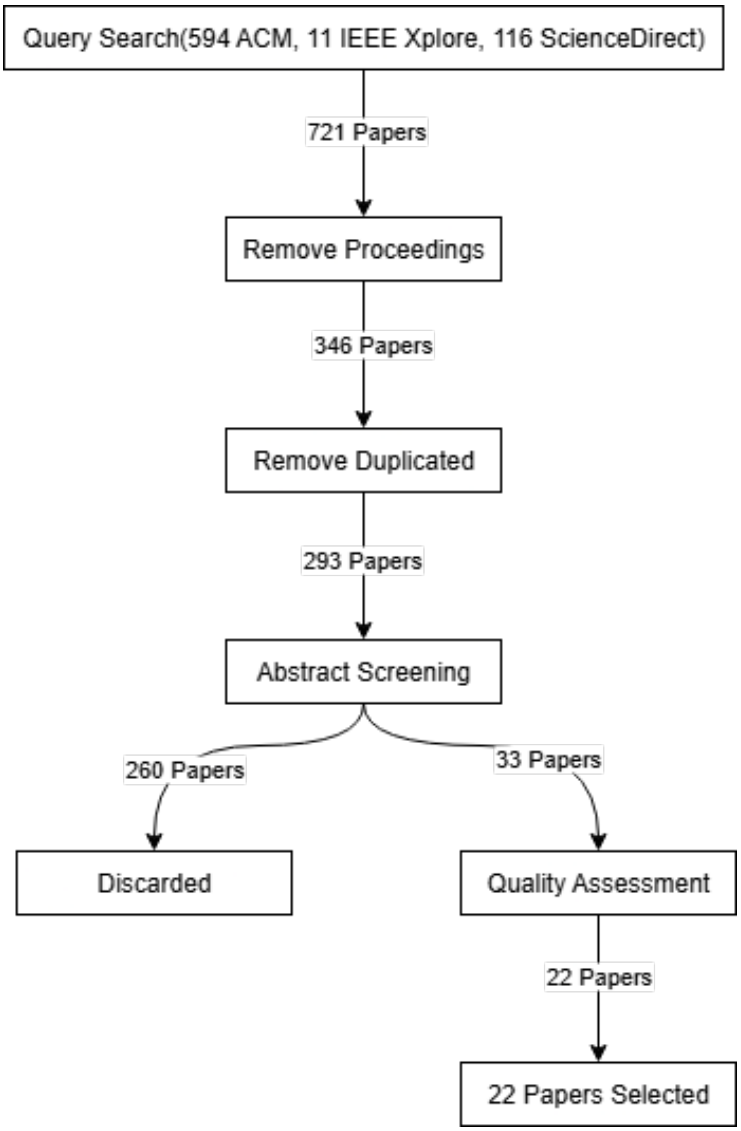

Figure 2.1: Flow Diagram of Systematic Review

2.1.7 Research Insights

In this section all findings related to the research question will be aborded. Later those will be discussed with more detail.

How can emotions play a role in recommendation systems for e-commerce to help users and shop owners?

In the last years there has been some interest in understanding user's sentiments inside of e-commerce environments. When analyzing most e-commerce shops, there will often be recommendation systems acting on the background. Having an enhanced recommendation system that works to favor good emotions can significantly improve the user satisfaction and engagement with the website, leading to more sales.

In the first research, reviews are used to extract emotional sentiments from Aminu Dau and Naomie Salim's Sentiment-aware Deep Recommender system with neural Attention network (SDRA) [1]. To enhance recommendation performance, the SDRA model records both product characteristics and user sensations related to these attributes. Product features and sentiment lexicons are extracted from user reviews using a semi-supervised topic model. These are subsequently integrated into an LSTM encoder using an interactive neural attention mechanism. In order to properly simulate fine-grained user-item interactions and enhance prediction performance, a co-attention technique is also implemented. On several datasets, experiments have demonstrated that SDRA performs better than Matrix Factorization (MF), Hidden Factors and Topics (HFT), DeepCoNN, D-att, and Transnet in terms of Mean Squared Error (MSE).

In another study, Daozhen Min and Lei Huang's Sentiment Analysis and BTM Collaborative Filtering (SABTMCF) [21] model uses reviews to extract emotional sentiments and also product attributes. The SABTMCF system uses short text comments to extract and score emotional words, and, simultaneously, models product feature keywords based on the Biterm Topic Model (BTM), allowing it to capture the user's true emotional preferences. This method generates a scoring matrix that reflects the user's emotional preferences. SABTMCF model outperforms techniques like Collaborative Filtering (CF), Sentiment Analysis Collaborative Filtering (SACF), and LDA-CF(Latent Dirichlet Allocation-Collaborative Filtering).

Xiaoli Wang, Chenxi Zhang, and Zeshui Xu's product recommendation model [6] uses online reviews to evaluate consumer sentiments and product qualities. Their model identifies relevant product attributes and calculates their performance scores, making use of customer sentiment and review usefulness by evaluating review text by applying the Bidirectional Encoder Representations from Transformers (BERT)-Latent Dirichlet allocation (LDA) model. Using this method, the model can provide a rating matrix that takes into consideration the user's preferences and psychological behavior, improving the accuracy and relevance of the recommendations. According to this method, a "ranking matrix" represents a network of items that are compared depending on how well they perform in particular properties. From this network, centrality scores are calculated, directly representing the product rankin scores. This model outperformed methods, such as Technique for Order Preference by Similarity to Ideal Solution(TOPSIS), VIekriterijumsko KOmpromisno Rangiranje (Serbian term for "multi-criteria optimization and compromise solution" - VIKOR), and Preference ranking organization method for enrichment evaluation (PROMETHEE II), showing improvements in recommendation accuracy.

Mehdi Elahi et al.'s hybrid recommendation model [13] significantly enhances the understanding of user preferences by focusing on sentiments extracted from reviews. This model analyzes user reviews using a BERT-based model, specifically BERT-base-uncased, to extract feelings. These sentiments are then combined to create vectors that represent users and items. These combined sentiments are used to compute evaluation metrics and are used as training features into advanced recommendation algorithms such as DeepFM and YoutubeRanker. The researchers carefully assessed their model against innovative sentiment-aware algorithms like DeepCoNN, D-Attn, and ESCOFILT, as well as conventional baselines like item-based collaborative filtering (CF) and alternating let squares (ALS). The findings demonstrated notable gains in recommendation accuracy, with sentiment data integration occasionally more than tripling DeepFM and YoutubeRanker's precision scores. The primary metrics utilized to assess performance were hit rate, precision, and PR-AUC (Precision-Recall Area Under Curve).

The Linear Superposition of Word2Vec and HowNet (LSWH) [2] designed by Wang Xiaoye, Jiang Kaiwen, and Zhou Xiaowen addresses word polarities by determining the polarity of evaluative terms in product reviews. By combining the strengths of Word2Vec and HowNet, the LSWH approach identifies word similarity. Word2Vec converts words into vector forms based on how frequently they appear together in a corpus, while HowNet represents word associations using a lexical tree. Even when words are removed from the sentiment lexicon, LSWH can reliably estimate their polarity by calculating the distance between word vectors and the distance between their meaning in HowNet.In terms of precision, recall, and F-measure, this approach performs better than the earlier approaches, SO-PMI and HowNet. Except for computer cooling, where SO-PMI did marginally better, LSWH showed greater precision in the majority of areas.

In a different representation of the data, the Sentiment-Aware Knowledge Graph (SAKG) [22] by Sung-Jun Park, Dong-Kyu Chae, and Sang-Wook Kim, combines the sentiment analysis with knowledge graphs. These knowledge graphs capture complex dependencies to enhance data comprehension and reasoning by representing connections between different things in an organized manner. To be able to provide a more accurate representation of interactions and item attributes, SAKG distinctively reinforces these connections with sentiment-related labels that are gathered from user reviews and ratings. Regarding a number of metrics, including Normalized Discounted Cumulative Gain (NDCG), Precision, Recall, and Hit Ratio, the experimental results indicated that SAKG, especially when combined with Sentiment-Aware Policy Learning (SAPL), performed better overall than models like Bayesian Personalized Ranking (BPR), BPR with Hidden Factors and Topics (BPR-HFT), RippleNet, and Path-based Graph Reasoning (PGPR). While achieving the highest accuracy on Cellphones and CDs & Vinyls datasets, it showed comparable accuracy to Adaptive Demonstration-Augmented Collaborative Filtering (ADAC) on the Clothings dataset and mostly outperformed Knowledge Graph Attention Network (KGAT).

Continuing with different representation, in this fuzzy logic work, R.V. Karthik and Sannasi Ganapathy's fuzzy recommendation system [12] analyzes customer reviews and computes sentiment scores. The system captures the emotional sentiments expressed in the reviews. To get the reviews fit for sentiment analysis, the data preprocessing technique tokenizes, parses, and tags parts of speech (POS). For each evaluation, sentiment scores are obtained. The total product rating score in different user categories is then calculated using these scores, average customer review scores, and the total number of reviews. Because fuzzy logic enables "varying levels of decisions" as opposed to rigid "recommended" or "not

recommended" outputs, it is used because it is more able to handle the complexity and unpredictability of user interests and emotions than other clear methods.

To address small details in user opinions, Aminu Da'u, Naomie Salim, Idris Rabiu, and Akram Osman developed Recommedation system Exploiting Aspect-based Opinion mining (REAO) [10] using deep learning methods. By concentrating on the fine-grained opinions expressed in user reviews and modeling these with a multichannel deep convolutional neural network (MCNN), which is based on an extended version of the convolutional neural network (CNN) architecture, the system captures the user's specific sentiments regarding various aspects of products. The model can produce rating matrices that are focused on specific characteristics by using this technique. To determine the final product rating, these matrices are then integrated into a Tensor Factorization (TF) model. This approach optimizes the recommendation accuracy by considering both the general rating and the specific user comments on a number of topics. Experimentation demonstrated that the REAO model significantly outperforms methods, such as Matrix Factorization (MF) and Transnets, showing significant improvements in recommendation accuracy.

Asking questions from a different strandpoint is also an effective search mode. Developed by Kun Xiong, Wenwen Ye, Xu Chen, et al., the Counterfactual Feature-aware Collaborative Filtering (CF2) model [3] focuses on counterfactual thinking, which is centered on raising "what if" questions and analyzing other alternatives. "What would be the user's decision if her feature-level preference had been different?" is just one of the specific questions it presents. Or "what would be the user's propensity on a given item pair if her feature-level attentions had been different?". This new approach allows the model to generate new training samples by actively intervening with existing user preferences, thereby aiming to improve the robustness and performance of recommendations. In order to improve the performance and robustness of suggestions, this novel approach enables the model to produce new training samples by actively affecting current user preferences. Counterfactual sample generation in the CF2 model aims to offer recommendations that are more precise and tailored to each user. Several frameworks, including Attribute-Aware Collaborative Filtering (A2CF), Bayesian Personalized Ranking (BPR), Neural Collaborative Filtering (NCF), Multi-Pointer Co-Attention Networks (MPCN), and Explicit Factor approach (EFM), were compared with this approach. It was tested against its own variants, CF2base (without data augmentation) and CF2rand (with random data augmentation), to show the unique efficacy of its learning-based counterfactual approach.The CF2 model consistently achieved the best performance on all evaluated metrics, which included F1 Score, Normalized Discounted Cumulative Gain (NDCG), and Hit Ratio (HR).

When looking at the user's viewpoint we have Chenliang Li et al.'s Capsule Network-Based Model for Rating Prediction (CARP) [23] representing an attempt to understand the reasoning process underlying a rating behavior in e-commerce. It achieves this by jointly considering user viewpoints, item aspects, and sentiments. CARP first extracts viewpoints from user reviews and aspects from item reviews, then pairs them to create "logic units" that serve as a proxy for the causes behind a rating. These logic units are analyzed by a sentiment capsule architecture containing positive and negative capsules. A key innovation is the Routing by Bi-Agreement (RBiA) mechanism, which jointly identifies informative logic units and infers their corresponding sentiments (positive or negative), while also suppressing non-informative units. This approach allows CARP to generate rating predictions that better reflect user preferences and also provides explanations for those ratings by understanding the underlying causes and emotional responses. Mean Square Error (MSE) evaluation indicates that this

method surpasses models like DeepCoNN, TransNet, D-Attn, TARMF, MPCN, and ANR, as well as other review-based methods like RBLT and CMLE, and the traditional PMF model, in terms of prediction accuracy.

As recommender systems investigate more than just semantic qualities, models that include visual elements have been emerging in addition to textual analysis. Mete Sertkan, Julia Neidhardt, and Hannes Werthner developed the Picture-based Tourism Recommender (PicTouRe) [8], taking into consideration the complex nature of recommendation items, including linguistic and visual style characteristics in addition to semantic properties. PicTouRe scans image collections and maps them into the Seven-Factor Model using advanced computer vision models. This model captures dimensions such as Sun & Chill-Out, Knowledge & Travel, Independence & History, Culture & Indulgence, Social & Sports, Action & Fun, and Nature & Recreation. This mapping, when applied to picture collections for users or items, is achieved through a supervised learning approach, requiring labeled datasets to train and test the models. For this particular mapping the focus is on visual attributes from picture collections. It is important to note that the broader research also explores a separate neural end-to-end approach, which is distinct from PicTouRe. This end-to-end approach aims to simultaneously learn user and item representations, incorporating various features. The item-encoder within this end-to-end framework includes components for visual-style features (e.g., color temperature) and text-style features (e.g., sentiment, emotions, subjectivity). Regarding its impact, PicTouRe was introduced and a user study was conducted demonstrating the utility of the approach in the tourism domain.

In a similar way, Zhu Zhan and Bugao Xu's Multiscale Semantic-Visual Analysis (MSVA) [11] model uses reviews and product images to extract user sentiments and preferences. The model is composed of two parallel modules: one for review representation and one for visual representation. In the review representation module, the word-aware attention mechanism identifies important words in reviews that correlate to sentiments and preferences, while the scale-aware attention layer re-weights multiscale features from review texts to select informative n-grams. In the visual representation module, a multiscale visual attention mechanism learns from item images at multiple block sizes. The multiscale scheme and attention mechanisms in both modules help mine fine-grained implicit sentiments and factors. The MSVA model was trained and achieved an average reduction in mean squared error (MSE) compared to Aspect-based Neural Recommender (ANR), Dynamic Review-based Recommenders (DRR), and Visual Adversarial Recommender (VAR).

Finally, Li Chen, Dongning Yan, and Feng Wang's study [4] is focused on enriching recommendation systems by integrating sentiment features extracted from product reviews to generate recommendations and explanations in a category structure. Their approach was to map frequently used noun phrases to pre-established properties as feature candidates, then analyze nearby adjectives to obtain sentiment ratings. An Apriori algorithm was implemented to group products with similar benefits into categories based on their "tradeoff" properties. Methods based solely on static requirements (such as Pref-ORG/Static View) performed worse than this sentiment-enhanced method, notably their "Mixture View" within the Senti-ORG system. The study found that consumers' perceptions regarding the utility of information, purchase intention, preference certainty, product knowledge, and recommendation transparency have all been significantly improved by sentiment-based explanations. Experiments with eye tracking also demonstrated that incorporating sentiment elements encouraged users to engage with recommendations more thoroughly, comparing products across categories and recognizing lower-ranked categories.

What are the most effective ways of segmenting users to maximize sales?

To enhance e-commerce recommendation systems, users must be well understood and classified according to their tastes and behaviors. Effective use of segmentation can maximise revenue creation by enhancing recommendation systems to have more loyal customers and boosting sales..

Sheng Liu and Shixun Yang, in "Machine Learning-Based Market Segmentation and Consumer Behavior Prediction Models" [24] explored the application of machine learning into market segmentation and consumer behavior prediction. To predict future purchasing behavior, they developed a random forest technique following splitting customers using K-means clustering based on characteristics like age, gender, and past purchases. The study concluded that when it came to predicting customer behavior, the random forest model performed better than the K-means clustering approach.

In e-commerce, recommendation systems play a key role in boosting sales and business growth by helping customers explore large product catalogs and choose what suits them best. The method "EX3: Explainable Attribute-aware Item-set Recommendations" [9] by Yikun Xian et al. highlights the value of giving explanations based on product attributes, not just showing relevant items. With this approach, customers can make smarter buying decisions and better understand how the recommendations match their preferences. The system could help users in extending their range of considerations by providing options that are differentiated based on the type of attribute. Through online A/B testing on a sizable e-commerce website, the effectiveness of this explainable recommendation technique in increasing sales has been experimentally confirmed, demonstrating a notable increase in conversion (+0.080%) and revenue (+0.105%).This indicates that providing clear, attribute-based explanations for recommended item sets directly contributes to an improved user experience that translates into tangible business growth and maximized sales.

Another approach focusing on product aspects from reviews was proposed by Aminu Da'u, Naomie Salim, Idris Rabiu, and Akram Osman [10]. Aspect-based Opinion Mining (ABOM) works by breaking down user reviews into specific product features. Focusing on these features is a useful strategy, especially at the start. ABOM can better understand what users like or dislike by identifying these features and creating ratings for each one. Significant improvements in recommendation performance are subsequently shown when these aspect-specific ratings are incorporated into a tensor factorization model to predict overall ratings.

Alvise De Biasio, Andrea Montagna, Fabio Aiolli, and Nicolò Navarin studied value-aware recommender systems (VARS) [17] in a systematic review, focusing on how these systems can increase the economic value of recommendations. They looked at how VARS can take into account the goals of different groups, like customers, service providers, and companies. Their research showed that businesses can perform much better by improving key metrics such as click-through rate (CTR), adoption and conversion rates, sales and revenue, sales distribution, and user engagement. They also found that VARS can boost overall profitability and make recommendation systems more sustainable in areas like e-commerce, advertising, news, and media by balancing the needs of all stakeholders.

Similarly, Xiaoli Wang, Chenxi Zhang, and Zeshui Xu [6] offered a different perspective by introducing a new product recommendation model based on online reviews. This model looks at both how products interact on e-commerce platforms and how customer psychology works, including short attention spans and the habit of accepting "good enough" solutions. It uses a three-step process: first, key product attributes are extracted from reviews with

BERT-LDA; second, attribute performance scores are calculated using review usefulness and consumer sentiment with RNTN; and third, products are ranked and competitors are identified with an improved PageRank algorithm that uses attribute weights and simulates how consumers compare products. When tested on Best Buy's price-segmented mobile phone reviews, the model significantly improved recommendation performance, with the Spearman correlation coefficient rising by up to 18.3% compared to traditional multi-criteria decision-making models.

Interests are always changing as shown by R.V. Karthik and Sannasi Ganapathy's system [12], that dynamically predicts the most relevant products for customers in online shopping, according to their current interests. They recommended that products should be scored differently for different user groups based on their levels of interest, underlining the importance of taking the end-user category into consideration when calculating product rating scores. In order to fulfill consumers' changing expectations, they also highlighted the importance of adding new products to the recommendation list. They also used demographic data, such as age group and delivery location, to further refine their predictions and offer more relevant products.

Continuing on dynamic preferences, Wenhao Guo, Jin Tian, and Minqiang Li [7] proposed a deep learning-based dynamic recommendation model that considers consumers dynamic preferences in both product and price. They designed a review-and-rating-based sequence generator to select products whose prices satisfy consumers, forming a new purchase sequence. Additionally, they developed a multi-level attention mechanism in the transformer layer to explore correlations between consumers price choices and combine those price preferences with product preferences.Retailers can learn what prices customers prefer and make smarter decisions about pricing, discounts, and product bundles.

Lastly, Juan Kong and Chen Lou [14] investigated the role of cultural features in online movie reviews, focusing on cultural differences between China and the United States [P31, 1, 2]. They investigated how cultural orientations, specifically high vs. low-context cultures and uncertainty avoidance, moderate the relationship between review characteristics (such as length, timeliness, title sentiment, and emotional expressions) and the perceived helpfulness of those reviews. According to their findings, review helpfulness in both countries is significantly predicted by these review components. In addition, the correlations between review length, title sentiment, and review helpfulness were found to be strongly moderated by cultural differences. However, the study noticed no significant moderating effect for negative emotional expressions on review helpfulness, and the moderating effect of cultural orientation on review timeliness was not proved in the way that was expected.

What flaws can a recommendation system have in online shopping, and can it be ethical?

Despite the advantages of recommender systems in e-commerce, they still have some flaws. Understanding these potential flaws and knowing how to deal with them is very critical to have a greater recommendation system. One of the first issues encountered was data sparsity.

The cold-start problem in the studied recommendation systems is one of the most talked flaws. One of those researches addresses this challenge by integrating richer user information, in Tran et al. [16] specifically discuss resolving the cold-start problem by incorporating user personality information into recommendation processes. Other studies, such as those

by Da'u and Salim [1] and Da'u et al. [10], use user-generated textual reviews and sentiments to improve overall recommender system performance, often by learning fine-grained, sentiment-aware representations of users and items. In a distinct approach, Mehdi Elahi et al. [13] propose a hybrid system that explicitly accounts for user's textual feedback and sentiment from product reviews to enhance recommendation quality. Furthermore, the MSVA model [11] improves prediction accuracy and recommendation relevance by incorporating both review text and visual analysis of product images. A study by Chen et al. [4] focuses on user evaluations of sentiment-based explanation interfaces for recommender systems, while it discusses its system's ability to serve new users and mentions the "cold-start phenomenon" in the context of high-investment product domains, it does not present specific mitigation strategies for the cold-start problem as a primary contribution.

R.V. Karthik and S. Ganapathy [12] discuss that existing systems often fail to account for the dynamic nature of user interests eventually changing, leading to irrelevant product recommendations in the present. Their fuzzy logic-based recommendation system dynamically predicts relevant products based on current user interests to prevent it. The system also incorporates sentiment analysis and ontology alignment to improve decision accuracy.

Sudhanshu Kumar, Mahendra Yadava and Partha Pratim Roy [25] identify the inability to accurately predict user preferences due to noisy data and unimodal approaches. It proposes the need of multimodal frameworks that combine electroencephalogram (EEG) signals and sentiment analysis to enhance prediction accuracy.

Continuing with over time changes, Wenhao Guo, Jin Tian, and Minqiang Li [7] shows the insufficient consideration of consumers' dynamic price preferences. To overcome this, the authors propose a dynamic recommendation model based on deep learning that considers both consumers' product and price preferences.

Starting with social aspects, Bingkun Wang et al. [5], states that review rating predicitions methods rely on the content of the reviews, ignoring the impact of social relations. It points out that recommendation systems based on social relations have achieved better results than those based on collaborative filtering and user-item matrices.

Christine Pinney, Amifa Raj, Alex Hanna, and Michael D. Ekstrand [15] discuss the problems associated with outdated assumptions about gender and the reliance on a binary understanding of gender. They advocate for a change of current practices and the adoption of methods that acknowledge the diversity of gender identities. It's also addressed ethical concerns surrounding the use of gender, stressing the need to apply gender variables in a responsible and informed way to prevent harm and promote fairness.

EX3 [9] describe that one of the flaws in existing recommendation systems is the lack an explanation of why items are recommended to users, so they propose a system that learns the importance of attributes from users' historical behavior to generate itemset recommendations with attribute-based explanations.

2.1.8 Research Questions' Answers

How can emotions play a role in recommendation systems for e-commerce to help users and shop owners?

Emotions when used in recommendation systems make recommendations more precise. Analyzing reviews may provide important information about the demands of the client, such as whether a user is happy or dissatisfied with the product. By doing this, the system is able to

recommend products that are more in line with their preferences, strengthening their bond with the website and enhancing their enjoyment of their subsequent purchases.

For this project, emotions will be the central piece in the recommendation system. This will allow for custom recommendations with better satisfaction over time. Not only will it predict better suited products but also it will understand some nuances that the user might not even realize when buying a product. It can also boost the website's customer trust and loyalty, increasing the chances of returning for future purchases.

What are the most effective ways of segmenting users to maximize sales?

The studies reviewed show that the most effective way to segment users in order to maximize sales involve analyzing their behaviors, preferences, and needs. The strategies to perform this segmentation include segmenting by purchasing history, price range, and type of product, as well as considering that these segmentations can also dynamic change over time. This allows businesses to create personalized recommendations that are more aligned with what the customers want. This segmenting of users can improve the accuracy of recommendations, enhancing user satisfaction and boosting sales.

For the current work, when understanding how users feel about certain products or product features, it becomes possible to segment them more effectively and make the recommendations based on their preferences. Because customer preferences can shift, both preferences and price ranges need to be taken into account.

What flaws can a recommendation system have in online shopping, and can it be ethical?

Recommendation systems can have many flaws, especially before and during the development phase. One major issue discussed in many papers is data sparsity, especially in cold-start situations where there's not enough information about new users or products to use. Another limitation is the inability of many systems to adapt to evolving user preferences over time, leading to irrelevant recommendations. Both Social biases, such as outdated gender assumptions, and the lack of any kind of transparency are fixes that are often the easiest to be overlooked.

In the current work, there will be an effort to fix gender assumptions as the shift in product preferences. This will allow the recommendation system not only to be more precise but also ethical.

2.1.9 Summary of Models and Metrics

The table below summarizes the presented models and evaluation metrics used in the studies selected for this systematic review.

Paper Ref.	Model Name / Authors	Models Used / Techniques	Metrics
[1]	SDRA	LSTM with interactive attention, semi-supervised topic modeling	MSE
[2]	LSWH	Word2Vec + HowNet for polarity detection	Precision, Recall, F1
[3]	CF^2	Counterfactual reasoning + Collaborative Filtering	F1, NDCG, Hit Ratio
[4]	Senti-ORG	Apriori algorithm + Sentiment scoring + Explanatory system	Qualitative (user preferences)
[5]	Wang et al.	Social relationship weighting + Review content	MAE
[6]	BERT + LDA Model	BERT, LDA	Compared with TOPSIS, VIKOR, PROMETHEE II
[7]	PEDR	Sequence generator + price attention transformer	Precision, Recall, NDCG, MRR
[8]	PicTouRe	CNN + Visual & Textual style mapping with emotion model	Qualitative personalization
[9]	EX3	Attribute-aware embeddings + itemset optimization	NDCG, Precision, Recall
[10]	REAO	Aspect-based opinion mining + MCNN + Tensor Factorization	RMSE, MAE
[21]	SABTMCF	BTM + Sentiment-aware Collaborative Filtering	MAE, Precision, Recall, F-Score
[13]	Elahi et al.	BERT + ALS hybrid recommender	Hit-Rate, Precision, Recall, PR-AUC
[23]	CARP	Capsule networks + Routing-by-agreement	MSE
[11]	MSVA	Multiscale attention + visual & text fusion	MSE
[12]	Fuzzy RS	Fuzzy logic + sentiment scoring + ontology alignment	Recall, Precision, Diversity, Novelty, Serendipity
[17]	VARS	Value optimization over CTR, sales, engagement	CTR, Adoption Rate, Revenue
[22]	SAKG & SAPL	Sentiment-aware Knowledge Graph + entity relationships	NDCG, Precision, Recall, Hit Ratio
[24]	Liu & Yang	K-means + Random Forest for behavior prediction	Precision, Recall, F1 Score
[25]	Kumar et al.	EEG + Sentiment multimodal fusion	R^2, RMSE

Table 2.2: Summary of Models and Evaluation Metrics in the Reviewed Papers

It is evident that a wide variety of models(ex: LSTM) or techniques(ex: Knowledge Graphs) have been explored, although most only once. Evaluations, on the other hand, have some that are more used when compared to others, for example MSE and derivates like RMSE,

as well as Precision and Recall. With this summary it is possible to conclude that the model or techiques used are very broad but the evaluation doesn't change too much.

Chapter 3

Data Collection

3.1 Observed Datasets

This section describes the datasets used on the studies that were previously analysed in Chapter 2. These datasets are mostly from e-commerce platforms, with the summary of their use in those studies below:

Amazon Datasets

Amazon datasets are utilized in a lot of reviewed works due to their coverage across various product categories as well as their size. From all the reviews studies, [1, 3, 4, 7, 9–13, 22, 23] utilize a derivation of a Amazon Dataset. The general categories are:

1. Computers, Laptops, Cellphones or Others: [4, 7, 10, 13];
2. Toys, Video Games, Books, Movies: [1, 3, 7, 10, 12, 13, 23];
3. Clothes: [22];
4. House [3, 9, 10];
5. Several Categories(>20) [11].

Yelp Dataset

The Yelp Datasets provide reviews of local businesses across multiple metropolitan areas in four countries. It was used in four studies, being [1, 3, 10, 23], because of the big size and sparsity.

Movies Dataset

Two of these studies used movies datasets:

1. Douban Movie Reviews: [5, 14];
2. IMDb: [14].

Other Dataset

Some datasets came from various websites but are only used in one study:

1. Jingdong Reviews: Reviews about laptops from the Jingdong e-commerce platform[2];
2. Dangdang: Book reviews collected from the Dangdang platform[21];

3. RateBeer: Contains reviews from the beer domain [23];
4. BestBuy: Data from the e-commerce BestBuy [6].

3.2 Chosen Dataset

In all the datasets observed in the previous section, the Amazon Dataset [26] is the one that stands out. The detailed reasons of the selection are below:

1. This dataset includes a wide range of electronics products, like digital cameras, laptops, cell phones, and other electronics, making it very easy to choose products that better suit CAPE[4, 7, 10, 13];
2. Amazon's Dataset contains a lot of reviews in every category, making it an excellent choice for experimentation in different categories[1, 3, 4, 10–13, 22, 23]. The large dataset size ensures that, even after preprocessing, there is sufficient data to achieve significant results and high model accuracy.
3. This dataset also provides extra information, such as prices, categories, description and others. This extra data can enable detailed analysis of those products;
4. As shown in the chapter before, this Amazon dataset is widely used in academic research [1, 3, 10–13, 22, 23]. It's popularity ensures that the other studies findings are easier to compare.

The chosen category of CAPE was the electronics area, so the chosen Amazon dataset was the Electronics one. This category has 2 files, the first one has reviews and the second has the metadata of the products in the reviews. They have respectively 21GB and 5GB. Since the size is already big, it was chosen only one category to be worked on.

3.3 Data Preprocessing

Given the large size of the Amazon Electronics dataset (approximately 21.1 GB for user reviews and 4.9 GB for product metadata), a preprocessing phase was essential to reduce noise, training time and ensure data quality. The goal was to retain only relevant and useful entries suitable for training and evaluating the recommendation system. After filtering, the reviews dataset was reduced to approximately 1.1 GB, and the metadata to about 90.7 MB. This corresponds to a reduction of nearly 95% in the reviews file size and 98% in the metadata file size.

Both datasets were divided into 2 smaller datasets, 70% for the training and 30% for the evaluating testing and practical testing referred in chapter 4. The first preprocessing was applied to both datasets as described below, and a second preprocessing step was carried out afterward.

3.3.1 User Reviews Dataset Preprocessing

The review dataset was processed in several steps:

1. Identify the Frequency: The first stage involved counting the total number of lines in the dataset and computing the frequency of key identifiers of products: asin and

parent_asin. This was done to later discard products with low counting of the total reviews so the training would not lead to wrong conclusions.

2. Timestamp Validation and Filtering: Reviews with missing or invalid timestamps were excluded. Timestamps were normalized (handling both seconds and milliseconds) and converted into UTC calendar years. This filtering was made to ensure reviews were complete and not corrupted. Also, only reviews from 2019 onwards were retained to ensure up-to-date and relevant content.

3. Filtering for Verified Purchases: To guarantee authenticity, only entries marked as verified_purchase = True were kept. This ensures the reviews come from customers who actually bought the product.

4. ASIN Frequency Threshold: Products with fewer than 50 reviews were excluded to reduce sparsity and facilitate future training. This threshold was selected as a compromise: low enough not to require an excessive number of reviews, yet high enough to provide sufficient data for effective model training.

5. Data Cleaning: Fields such as images, verified_purchase, videos, and bought_together were removed to reduce dimensionality and storage requirements. Only the most relevant fields for this model were kept: rating, title, text, asin, parent_asin, user_id, timestamp, and helpful_vote.

3.3.2 Product Metadata Dataset Preprocessing

The metadata file contained extensive product information, some of which was not aligned with the filtered reviews. To ensure consistency, only metadata entries corresponding to products in the processed reviews file were retained.

1. Valid Parent ASINs Extraction: A set of parent_asin values was extracted from the filtered reviews to serve as a reference for relevance.

2. Metadata Filtering and Field Selection: Each metadata entry was checked for the presence of a parent_asin in the valid set. If found, the entry was kept, otherwise, it was discarded. Non-essential fields such as images, videos, and bought_together were removed.

This resulted in a significantly smaller but semantically richer metadata file, ensuring alignment between products and their associated reviews.

3.3.3 Second Dataset Preprocessing

After training, the results were not satisfactory, as discussed in more detail in Chapter 4. Upon analyzing the potential issues, it appeared that they could be related to token size and the presence of fields with excessive padding, while others were truncated prematurely. The initial values chosen, shown in the table below, were considered reasonable for capturing the desired information early on.

Preprocessing 1

Input Field	Sequence Length
Review Text	128 tokens
Review Title	32 tokens
Product Title	64 tokens
Main Category	32 tokens
Store Name	32 tokens
User ID	32 tokens

Table 3.1: Sequence lengths per input field (Preprocessing 1).

Below are graphs showing the tokens distribution in each field.

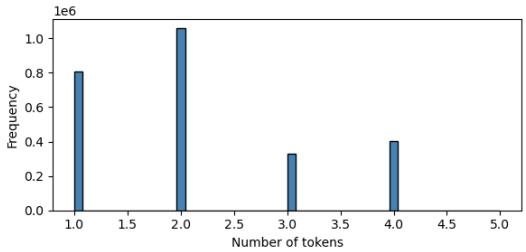

Figure 3.1: Tokens for the field Main Category

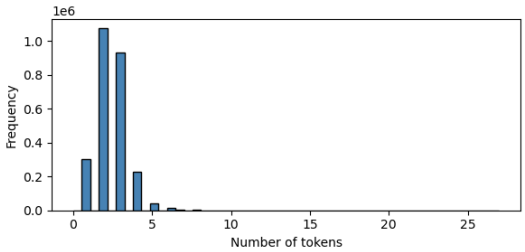

Figure 3.2: Tokens for the field Store Name

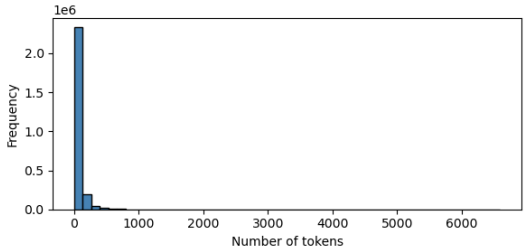

Figure 3.3: Tokens for the field Review Text

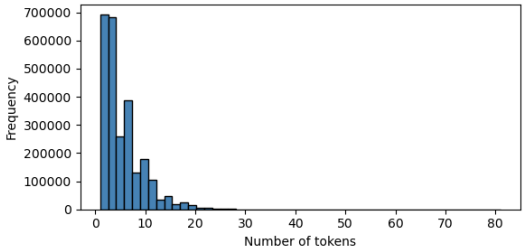

Figure 3.4: Tokens for the field Review Title

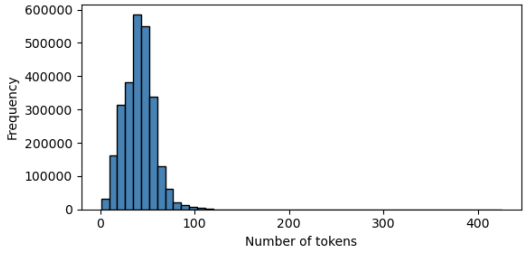

Figure 3.5: Tokens for the field Product Title

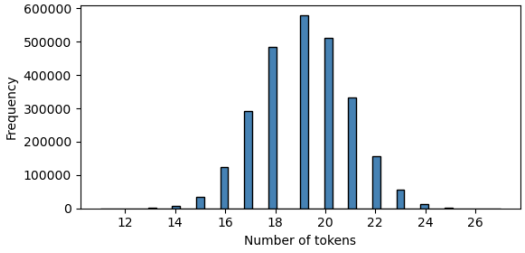

Figure 3.6: Tokens for the field User ID

Below is a summary of the values inside on those graphs, detailing the values for each field, with an average of the total tokens, and where is placed the 95% token, 99% and what is the maximum token number.

Field	Average	95% Percentile	99% Percentile	Max
Review Text	57.99	201.0	443.45	6599
Review Title	5.34	13.0	20.0	81
Product Title	40.79	67.0	86.0	425
Store Name	2.50	4.0	5.0	27
User ID	19.13	22.0	23.0	27
Main Category	2.12	4.0	4.0	5

Table 3.2: Summary of tokens per field

After having this information, the final token values decided were for the Review Title, Product Title and Store Name the 99% value, since the 100%(Max) was very far from this value and was a good compromise. The Main Category was given the max value since it was very small. The User ID has also the full value since having the full text is essential to know what user is being recommended. Finally the Review Text is the biggest one, even on 99% is more than 443. Because of that it's the only one at 95% to optimize the training speed. Below are the summary of these choices.

Preprocessing 2

Input Field	Sequence Length
Review Text	201 tokens
Review Title	20 tokens
Product Title	86 tokens
Main Category	5 tokens
Store Name	5 tokens
User ID	27 tokens

Table 3.3: Sequence lengths per input field (Preprocessing 2).

Below are updated graphs showing the new ignored tokens to the right of the red line.

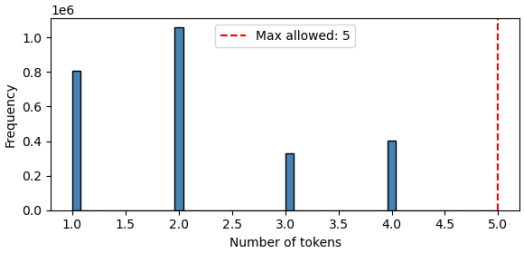

Figure 3.7: Divided Tokens for the field Main Category

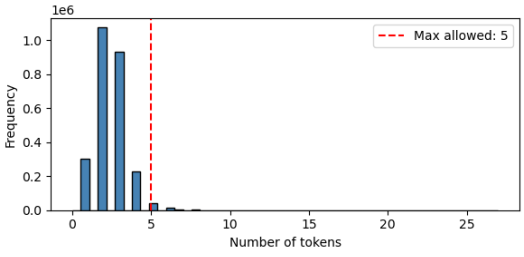

Figure 3.8: Divided Tokens for the field Store Name

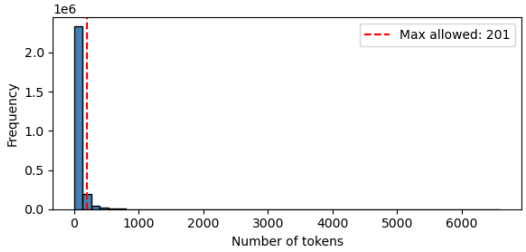

Figure 3.9: Divided Tokens for the field Review Text

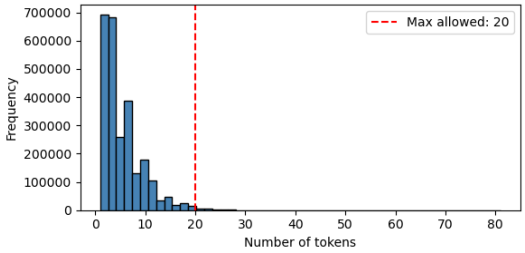

Figure 3.10: Divided Tokens for the field Review Title

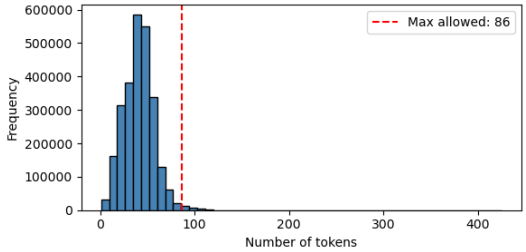

Figure 3.11: Divided Tokens for the field Product Title

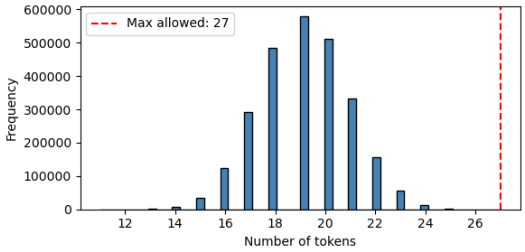

Figure 3.12: Divided Tokens for the field User ID

3.4 Models

Based on the analysis done in Chapter 2, a lot of models and techniques have been utilized in several applications across sentiment analysis, including both more traditional machine learning approaches and more recent deep learning architectures. Hybrid models also show promising results, combining the strengths of each model.

Motivated by this, the present project adopts a hybrid model that combines BERT with LSTM. This choice is based in the observation that both BERT and LSTM have been successfully applied in several of the reviewed studies, as well as the fact that both models are familiar from prior academic experience. Therefore, this project also serves as a chance to explore their implementation and performance in a real-world sentiment analysis task.

One of these two models is BERT [7, 13]. It uses a transformer to determine the meaning of a word by looking at the words that come before and after it in a sentence. BERT was trained on over 2 billion articles from Google News, enabling it to better understand word context. It's usually used for question-answering, sentiment analysis, and text classification tasks.

LSTM [1] neural network is one of the best when it comes to handling long-term relations in data. LSTMs use long-term memory cells, unlike conventional neural networks. These cells have 3 gates, input gate that adds new information, the forget gate that discard the irrelevant information, and finally the output gate that passes relevant information. Particularly tasks where we need to understand the flow of sentiment in a piece of text, as well as calculating anything based on time-series data, can benefit from LSTMs.

By combining BERT and LSTM, the model obtains rich, context-aware embeddings while also capturing long-term dependencies in sentiment flow, leading to more robust and accurate predictions.

3.5 Metrics

When selecting evaluation metrics for this project, two different types were considered: classification-based metrics and regression-based metrics. The classification metrics studied included Accuracy, Recall, and F1 Score, all of which were used in several of the reviewed studies([2], [24], and [7]). These metrics are used in sentiment analysis tasks where the outputs are binary, either positive or negative.

Despite being popular, they were ultimately not chosen for this project. The reason for this is the nature of the desired output: in this case, the model is expected to produce a score indicating the degree of sentiment, or the likelihood of that specific recommendation, rather than a simple class. For example, if the output was a score from 1 to 5 (e.g., a typical review rating), classification metrics would not allow for a differentiation between products that share the same predicted class(for example a score of 5 could have a grade of 0.81 in a scale of 0 to 1, and could be a bad recommendation comparing to a 0.99). This would limit the ability recommend items effectively.

As a result, regression-based metrics were considered more appropriate, particularly those that provide continuous outputs. Among the commonly used ones in the latest analysis, Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE) were considered. MSE penalizes larger errors more heavily due to the squaring of the differences, which helps to push the model toward more accurate predictions. This emphasis on larger deviations is especially useful in recommendation contexts, where recommending a poor product with high confidence could be more damaging than small deviations in predicted ratings.

RMSE, similar to MSE, was not selected primarily because its interpretation is less direct in this context. The square root operation that is done alters the scale of error, which will not add meaningful benefits over MSE. In the same way, MAE, that measures the average absolute difference between predicted and actual values, was considered but ultimately not chosen, as it treats all errors equally. In contrast, MSE's stronger penalization of bigger errors was seen as more aligned with the project's objective of making more accurate recommendations.

These factors led to the selection of MSE as the evaluation metric since it offers a continuous output ranging from 0 to 1, facilitating intricate comparisons. This enables a more accurate recommendation system for the customers by facilitating the recommendation of products based on how close the predicted satisfaction for a particular user score is to 1.

Chapter 4

Implementation, Analysis and Results Discussion

4.1 Introduction

This chapter describes a proposed system, with a first iteration of the workflow, the implementation, testing and evaluation of the said system, and in the end the experimental results. The model's workflow is divided into three main phases: the first is the Data Preprocessing, then the Model Phase, and finally the Deployment.

Below is a detailed description of each step:

Data Preprocessing: This initial phase is crucial for cleaning and preparing raw data, making sure it's relevante for the model training. As discussed in the previous chapter, it is subdivided into two parts: User Reviews Dataset PreProcessing and Product Metadata Dataset PreProcessing. Both PreProcessing cleaned the original datasets in a way that made them smaller and workable on.

Model (using TensorFlow): This is the central phase where the models are built, trained, and evaluated. After the evaluation, a different revision was created to enable the development of a better model, particularly to meet different requirements such as improved performance, reduced computation time, and overall effectiveness. The project adopts a hybrid approach that combines BERT and LSTM.

- BERT Encoding and Tokenizing: Textual fields are processed using a pre-trained BERT tokenizer (bert-base-uncased) to create tokenized sequences. Textual information is then also encoded using BERT.
- LSTM Training: The tokenized sequences are then passed to LSTM layers, which are effective in handling long-term relationships in data, such as the flow of sentiment in a text. Different architectures are explored, including various versions of hybrid models, with more BERT or more LSTM and different specifications between them.
- Test Models: After training, the models are evaluated using the MSE metric.
- Tuning Models: This step involves adjusting model architectures or parameters. The obtained results are compared to the others to see if the paramethers were better to be kept.

Deployment: In this final phase, recommendations were tested in different users, and different filters like category, and price range.

The below diagram illustrates the full workflow, from the dataset preprocessing to final deployment. The following sections details the training configurations, hyperparameters, and results obtained for each model.

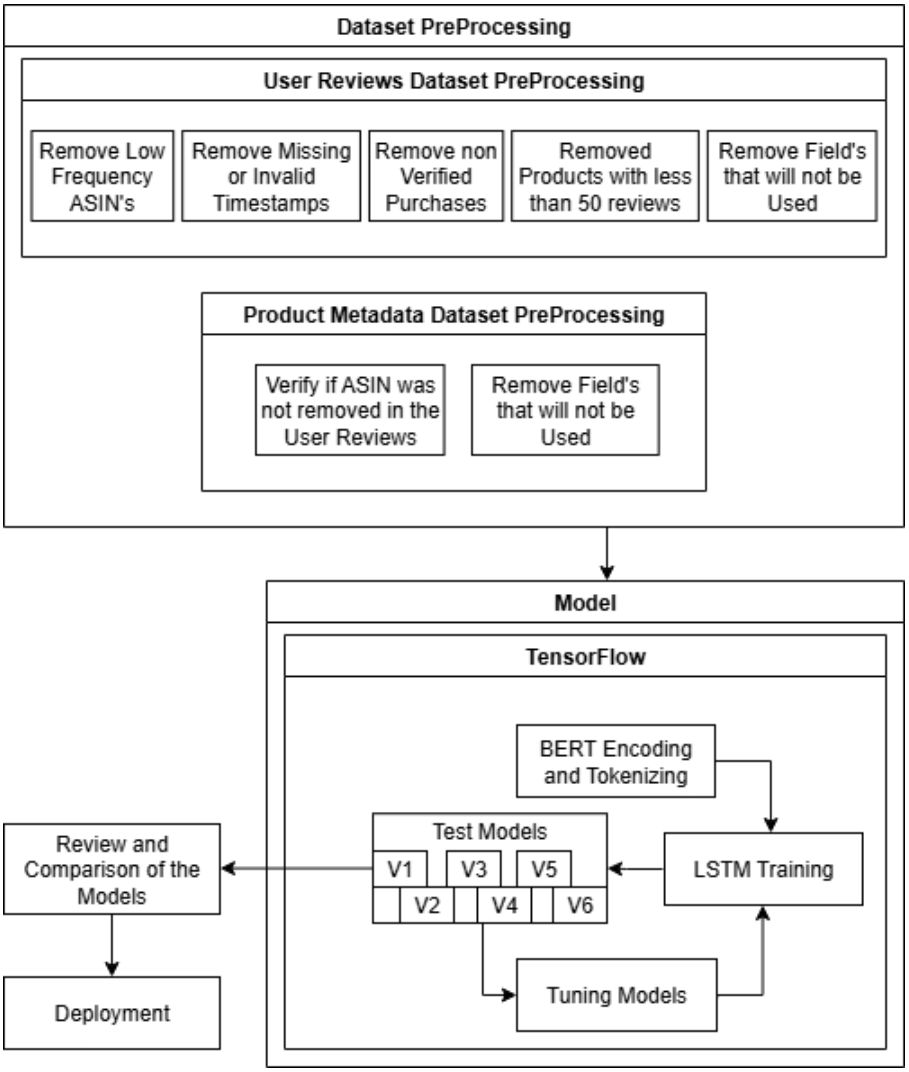

Figure 4.1: System Workflow

4.2 Developed Models

There are different architectures explored in this work(BERT, LSTM, Hybrid), but all models share a unified structure in terms of inputs, vocabulary sizes, and saving.

Input Structure

All models receive the below information on inputs joined from both the user reviews dataset and the metadata dataset by the ASIN(ID of the specific product), which are grouped into three categories:

- **Review-related Inputs:**
  - `review_text`

  - `review_title`

- **Product-related Inputs:**
  - `main_category`
  - `product_title`
  - `store_name`
  - `price`
  - `average_rating`
  - `parent_asin` (ID of the group of products in the same page, usually just in different size or colors)
- **User-related Inputs:**
  - `user_id`

Token Sequence Sizes

Each textual input has a specific length of token sequence. When using transformer-based models like BERT, an additional attention_mask is included for each textual field. These inputs, as discussed in chapter 3, will be using the information contained in the tables 3.1 for the first preprocessing results and the 3.3 for the second preprocessing results. Both will be shown since there is information removed in the second preprocessing that could be important depending on the situation like, for example, a very expensive product. Overall, the second will be better for most products and reviews with high but yet reasonable prices.

Checkpointing and Training State

All models follow a shared mechanism for checkpointing, allowing for the stop of the training when needed without losing much progress. These checkpoints save in an interval that could be changed if the speed in the training was slower or faster. In the end all models were saved in .h5 files, with timestamps for the training time.

4.2.1 Default BERT Model

The created BERT architecture integrates textual and numeric inputs using a modular structure. A pre-trained BERT backbone bert-base-uncased is used to encode all textual fields. Each textual input is processed independently through the shared BERT encoder using both input_ids and attention_masks. The pooled output of the [CLS] token is used as the representation for each field, followed by a dropout layer for regularization.

These representations are joined into a single feature vector and passed through two dense layers with ReLU activation (512 and 256 units) and dropout regularization. The final layer is a single neuron with linear activation, which produces a scalar prediction. This model is trained with the Adam optimizer and a learning rate of 2e-5.

4.2.2 Default LSTM Model

The LSTM model is built to handle both text and numeric inputs using a multi-branch design. Each text input is processed separately with an embedding layer followed by an LSTM layer

with 64 units. Padding tokens in the input sequences are masked, and the outputs of all LSTM layers are joined together to create high-level representations of the different input parts.

These representations, along with the numeric features, are concatenated into a single feature vector, which is passed through two dense layers with ReLU activation (256 and 128 units) and dropout regularization (0.3 and 0.2, respectively). The final output layer consists of a single neuron with linear activation to produce a scalar prediction. The model is trained also using the Adam optimizer with a learning rate of 2e-5.

4.2.3 Hybrid Model(BERT+LSTM)

The hybrid models were created to test different ways of combining transformer-based tokenization with traditional deep learning methods. They all use the same preprocessing pipeline, but each version includes specific changes in architecture or training. The goal of these models is to balance contextual detail, computational efficiency, and interpretability.

Common Structure

All hybrid models use text inputs that are first preprocessed with the BERT tokenizer (`bert-base-uncased`). The tokenized sequences are saved and reused during training to make sure results stay consistent across models. The BERT encoder itself is not always used, in some cases, only tokenization is applied, while in others, the encoder is partly or fully included.

Each text field is usually passed through an embedding layer (size 64), followed by LSTM layers. The features from all branches are then combined and sent through two dense layers (256 and 128 units) with ReLU activation, with dropout layers (0.3 and 0.2) in between. The final layer consists of a single neuron with linear activation to produce a scalar prediction.

Hybrid Model V1: Embedding + LSTM Baseline Hybrid V1 uses the simplest configuration: all BERT-tokenized text sequences are passed through trainable embedding layers and 64-unit LSTMs. The BERT encoder is not used at all. The final output layer uses a linear activation, suitable for regression. The model is trained using the Adam optimizer with a learning rate of 2e-5. This version serves as the baseline for comparison for the other hybrid models.

Hybrid Model V2: Normalization and Sigmoid Output Hybrid V2 keeps the embedding + LSTM setup from V1 but adds LayerNormalization layers for each feature group (review, product, user) before they are combined. This helps make training more stable. The final layer now uses a sigmoid activation, which keeps the output between 0 and 1. The learning rate was raised to 1e-4.

Hybrid Model V3: Frozen BERT Encoder Hybrid V3 uses the BERT encoder again, but it stays frozen during training. For each text field, the output from BERT's second-to-last hidden layer is sent through a 64-unit LSTM. The model keeps the LayerNormalization and sigmoid output from V2 and continues to use a learning rate of 1e-4.

Hybrid Model V4: Dual Branch — BERT + LSTM vs Embedding + LSTM Hybrid V4 introduces a mixed approach. In the review branch, text fields are processed with frozen BERT followed by LSTMs. In the product and user branches, text fields go through standard embedding layers and LSTMs. Each branch is normalized before being combined. The rest of the design stays similar to earlier versions, with dense layers, dropout, and a sigmoid output. The goal is to use BERT only where it is most helpful (the complex review text), making the model smaller and easier to train.

Hybrid Model V5: Global Max Pooling In Hybrid V5 all text fields are processed with embedding layers and 64-unit LSTM layers with return_sequences=True, followed by GlobalMaxPooling1D. This lowers dimensionality while keeping key sequence information. It also uses a linear output activation with a learning rate of 2e-5, the same as V1.

Hybrid Model V6: Attention Pooling over LSTM Outputs Hybrid V6 builds on V5 by replacing GlobalMaxPooling with a learnable attention pooling layer applied to the LSTM outputs. This attention mechanism assigns importance scores to time steps, letting the model focus on the most useful parts of each sequence. All text fields go through embedding layers, LSTMs with return_sequences=True, and attention pooling. The rest of the design stays the same, with dropout layers, a linear output, and the Adam optimizer with a learning rate kept at 2e-5. This version improves interpretability and performance by highlighting informative tokens.

Summary of Main Changes The table below illustrates the main differences between the six versions of code described before.

Version	Main Change	Output	Learning Rate
V1	Baseline: Embedding + LSTM	Linear	2e-5
V2	Added LayerNorm per branch	Sigmoid	1e-4
V3	Frozen BERT encoder + LSTM	Sigmoid	1e-4
V4	Dual branch: Frozen BERT(reviews) + LSTM Embedding(product/user)	Sigmoid	1e-4
V5	GlobalMaxPooling after LSTM	Linear	2e-5
V6	Attention pooling over LSTM outputs	Linear	2e-5

Table 4.1: Summary of differences between versions of all Hybrid Models (BERT+LSTM).

4.3 Variation in Running Times

This section presents the usual time required to train and evaluate each model. All experiments were conducted on a computer equipped with an Intel i5-14600k CPU, an RTX 4070 Super 12GB GPU, and 32GB of RAM, running Windows 10. During this measurements, no other processes were active except for the models themselves, ensuring consistent timing results.

The recorded times include both the training phase, and the evaluation phase,as well as the two preprocessing times described at Chapter 3. These different preprocessing versions were also tested in the same models to determine whether preprocessing had any impact on

runtime, wich it didn't. The average times for each model and stage are summarised in the table below.

Model	Stage	Time
Preprocessing	P1	43m
	P2	45m
BERT	Training	11h42m
	Evaluation	1h44m
LSTM	Training	2h8m
	Evaluation	11m
V1	Training	1h15m
	Evaluation	9m
V2	Training	1h15m
	Evaluation	10m
V3	Training	11h25m
	Evaluation	1h41m
V4	Training	3h02m
	Evaluation	1h
V5	Training	1h14m
	Evaluation	10m
V6	Training	1h16m
	Evaluation	10m

Table 4.2: Running Times for Each Model and Stage

The table shows that preprocessing times for P1 and P2 are almost identical, with the V2 having 2 more minutes because of the minimal extra processing made.

Models V3 and V4, which incorporate a more extensive use of BERT components, have by far the highest training and evaluation times. In particular, V3 requires over 11 hours for training and more than 1.5 hours for evaluation, making it the slowest among all tested models. This extended evaluation time is undesirable for practical applications, as it would also slow down future inference or testing.

In contrast, Models V1, V2, V5, and V6 achieve very similar performance in terms of runtime, both during training and evaluation. This suggests that these models are more efficient and scalable in contexts where speed is also a priority.

4.4 Obtained Results

This section shows the performance results obtained from all the different models tested in this study. Three model categories were tested: BERT-based models, LSTM models, and hybrid architectures combining both approaches. Each model was evaluated under two distinct preprocessing strategies, and the performance was measured using the MSE metric across multiple runs to ensure consistency.

4.4.1 BERT Model Results

The first tested model was a version with a preprocessing and training made entirely with BERT. This one was tested using also the first preprossesing.

1st Preprocessing

BERT	MSE
1 Run	0.0417

Table 4.3: Results with BERT Model in 1st Preprocessing.

From this first test a MSE of 0.0417 was a good starting point.

2nd Preprocessing

With the second preprocessing strategy, the BERT model was tested three times.

BERT	MSE
1 Run	0.0446
2 Run	0.0401
3 Run	0.0482

Table 4.4: Results with BERT Models in 2nd Preprocessing.

With this test the difference between the first and second preprocessing was very small, with the average MSE being 0.0446, and a diference of -0.0029.

4.4.2 LSTM Model Results

The next model tested was also a version with purely LSTM, in the preprocessing as well as the training. As the examples before, firstly it was tested with the fisrt version of the preprocessing.

1st Preprocessing

LSTM	MSE
1 Run	0.0277

Table 4.5: Results with LSTM Models in 1st Preprocessing.

LSTM already shows a great improvement compared to the first BERT model(approximately 33,6% better).

2nd Preprocessing

Now testing with the second preprocessing:

LSTM	MSE
1 Run	0.0301
2 Run	0.0269
3 Run	0.0283

Table 4.6: Results with LSTM Models in 2nd Preprocessing.

As shown before with the BERT model, this second preprocessing made the LSTM average a little bit inferior as compared to the first preprocessing. This average is now at 0.0284, with a difference of -0.0007.

4.4.3 Hybrid Models Results

For the hybrid models, as stated and described in the chapter before(4.2), there are 6 tested versions that are made of a preprocessing using BERT and a training that is more foscused on LSTM. Here are the results for the first preprocessing:

1st Preprocessing

Version	Hybrid	MSE
v1	1 Run	0.0282
v2	1 Run	0.0284
v3	1 Run	0.0417
v4	1 Run	0.0417
v5	1 Run	0.0253
v6	1 Run	0.0222

Table 4.7: Results with Hybrid Models(BERT+LSTM) in 1st Preprocessing.

And for the second preprocessing the results are shown below:

2nd Preprocessing

Version	Hybrid	MSE	Average
v1	1 Run	0.0282	0.0286
	2 Run	0.0287	
	3 Run	0.0288	
v2	1 Run	0.0304	0.0314
	2 Run	0.0316	
	3 Run	0.0321	
v3	1 Run	0.0882	0.0899
	2 Run	0.0928	
	3 Run	0.0886	
v4	1 Run	0.0403	0.0276
	2 Run	0.0214	
	3 Run	0.0212	
v5	1 Run	0.0243	0.0243
	2 Run	0.0241	
	3 Run	0.0245	
v6	1 Run	0.0264	0.0265
	2 Run	0.0269	
	3 Run	0.0261	

Table 4.8: Results with Hybrid Models(BERT+LSTM) in 2nd Preprocessing.

This experimental evaluation provided detailed metrics for each tested version of the model. The following section will discuss these results in depth, as well as compare it with the results from the pure models while also trying to analyse the factors that contributed to the observed performance differences and identifying potential areas for improvement.

4.5 Results Discussion

This section will discuss the experimental results obtained from the various models and both preprocessing strategies. As stated before, the objective is to analyze the performance of the developed models, comparing their effectiveness, and identify factors that contributed to the observed differences, with a focus on their practical implications for a personalized product recommendation system.

4.5.1 Impact of Data Preprocessing

Upon testing the models with Preprocessing 1 in a practical setting with actual recommendations, it was observed that those recommendations did not exhibit sufficient variation, despite demonstrating good performance during evaluation. To investigate this issue, the token distribution graphs presented in Chapter 3.3 were utilized. These visualizations revealed that the data was highly dispersed, indicating the presence of outliers. To address this problem, these outliers were subsequently removed from the dataset by adjusting the sequence lengths based on percentiles, effectively filtering out extreme values. When the

models were re-tested with practical recommendations using Preprocessing 2, more satisfactory results were obtained.

This phenomenon occurred because, with Preprocessing 1, the model could effectively learn an average value across all dispersed data points. For instance, in fields like product price, where extreme values might exist (e.g., millions of dollars), the model's predicted output would remain relatively close to this average, consequently failing to differentiate significantly between varied recommendations. This issue was largely remediated by Preprocessing 2, which by carefully adjusting sequence lengths based on percentiles, effectively reduced the influence of these extreme outliers and allowed the model to focus on more relevant data for typical e-commerce use cases. Although the first preprocessing could be use in very specific scenarios of e-commerce, the second one will be better for most cases and will be the focus.

4.5.2 Model Performance Analysis

Comparing the standalone BERT and LSTM models, the LSTM model generally outperformed the BERT model in terms of MSE. For instance, with Preprocessing 2, the LSTM model achieved an MSE of 0.0269, which was significantly better than the 0.0401 of the BERT model. This suggests that, for this specific task, LSTM's ability to handle long-term dependencies in the textual sentiment flow was highly effective, perhaps because the problem's nature (predicting a rating from diverse inputs) aligns better with LSTM's sequential processing, even with BERT's powerful contextual representations. The standard BERT model, while excellent for contextual understanding, may not have translated that advantage into superior numerical rating prediction when used independently and with the initial preprocessing settings.

The hybrid models, which combined aspects of BERT (for tokenization and contextual embeddings) and LSTM (for sequence processing), showed a variety of performances:

Hybrid V3 consistently exhibited the worst MSE values, averaging 0.0899 with Preprocessing 2, and was also by far the slowest model in terms of training time (11h25m) and evaluation (1h41m). This inefficiency and poor performance, make it impractical for real-world recommendation scenarios where speed and accuracy are critical.

V4 also displayed high MSE values in Preprocessing 1 (0.0417) and a high variance in its MSE results during Preprocessing 2 tests (ranging from 0.0403 to 0.0212, with an average of 0.0276). Despite some runs showing competitive performance, having the best MSE in Preprocessing 2(0.0212), this inconsistency, combined with its relatively long execution times (3h02m training, 1h evaluation), makes it less reliable for consistent performance.

Hybrid V1, Hybrid V2, Hybrid V5, and Hybrid V6 demonstrated similar at the best efficient execution times during training and evaluation. Among these, Hybrid V5 and V6 achieved the best performance in terms of MSE.

Hybrid V6 recorded the lowest MSE of 0.0222 with Preprocessing 1. With Preprocessing 2, its best MSE was 0.0261. As for V5, it closely followed, with the second best MSE of 0.0241 in Preprocessing 2 and an average MSE of 0.0243, showing remarkable consistency (0.0243, 0.0241, 0.0245). Hybrid V5 offered an excellent balance between performance and computational efficiency. Its fast training (1h14m) and evaluation times (10m) are comparable to the best performers, making it a highly viable option for practical deployment.

4.5.3 Performance Comparison

When comparing the different models, it becomes evident how design choices directly impacted performance, stability, and computational efficiency. This section explains the reasons behind the observed variations.

Standalone Models: BERT vs. LSTM

The LSTM worked better than BERT because it is designed to capture sequential patterns and long-term dependencies, which are important when predicting ratings from text and user information. BERT, although very strong in understanding word context, did not convert this strength into good numerical predictions in this setup. Another key difference was efficiency: LSTM trained and evaluated much faster, making it more suitable in practice.

Hybrid Models: V1 to V6

The hybrid models showed how combining BERT and LSTM can be useful, but also risky if the design is not well balanced.

Activation Functions and Learning Rates in Hybrid Models (V1–V6): The choices of activation functions and learning rates were crucial for the behavior of each hybrid model and directly influenced the interpretation of the MSE.

V1, V5, and V6 used a linear activation function in the output layer, with a learning rate of 2e-5 for the Adam optimizer. Linear activation allows the model's predictions to have a continuous and potentially unbounded output, which is suitable for regression of scores without explicit scale restrictions, even though rating_scaled ranges from 0 to 1.

V2, V3, and V4 opted for a sigmoid activation function in the output layer, restricting the output to a range between 0 and 1. In these models, the learning rate was increased to 1e-4. While sigmoid activation is useful for forcing the output into an interpretable scale (such as probability or a normalized rating between 0 and 1), it can limit the model's expressiveness if the underlying data relationships require a wider range of values or a different distribution. The higher learning rate may also have affected training stability and model convergence.

This changes were tested but ended up not changing anything in the final output since what was generated was always between 0 and 1(activation function), as well as the training being too long(learning rate) which made the conversion happen almost always closely.

Lack of Improvement in V2 with Normalization: Although LayerNormalization usually contributes to training stability and accelerates convergence, in this specific case, the input data may already have been on a sufficiently stable scale, making LayerNormalization redundant for MSE improvement, or the data type and architecture in V1 may already have been resilient to fluctuations, not requiring extra normalization for optimization.

V3's Complete Failure: In V3, the BERT encoder was applied to all text fields, including ones such as Main Category (e.g., "Electronics"), Store Name (e.g., "Amazon"), and User ID. Some of these fields are short, categorical in nature, or pseudo-random identifiers. BERT is a pretrained model for understanding context and relationships in complex natural language. Applying it to short strings with low semantic content, or to identifiers without linguistic

structure, does not utilize BERT strengths. Instead, it resulted in noise(observed in its high MSE and long runtime).

Comparison

- V1 and V2: Both used embeddings and LSTMs without heavily relying on the BERT encoder. They were efficient and stable yet not the best, showing that simpler designs can already deliver good results. Normalization in V2 did not bring improvements.
- V3: This version failed because it used BERT on all text inputs, including fields with little semantic meaning (such as Main Category, Store Name, User ID), making the model overly complex, and very slow.
- V4: By freezing BERT and applying it only to richer review fields (such as review_text and review_title), V4 sometimes achieved the best results. However, its outcomes were inconsistent across runs and still very slow, which made it very bad in its reliability.
- V5: This design simplified the process by using LSTMs with global max pooling. It was consistent, accurate, and highly efficient, making it a strong candidate for real-world implementation.
- V6: This time the model used attention pooling. This made the model slightly more accurate and more interpretable, since attention highlights which parts of the sequence matter most. Like V5, it was also efficient but this time it was for niche cases.

4.5.4 Results Conclusion

In summary, while V4 showed the best performance in a single run, it lacked in consistency and time efficency, showing that complexity is not always beneficial. Models that used BERT without clear restrictions (like V3) performed poorly. Models that balanced simplicity and complexity delivered the best compromise between accuracy, consistency, and runtime. V5 is the most reliable and practical option, while V6 offers an additional boost in accuracy and interpretability for cases where these are prioritized.

4.5.5 Practical Recommendation System

After the evaluation of the models, a practical implementation was developed to test their usefulness in a real recommendation setting. For this purpose, the 30% of the dataset that was previously set aside for testing (as explained in Chapter 3) was also used here.

A script was created that loads the trained model and generates personalized recommendations for a given user. It works by first identifying products that the user has not reviewed yet and filtering them, according to some optional conditions, such as main category, price range or number of recommendations. Then, the trained model is applied to predict the rating score for each candidate product. The products with the highest predicted scores are selected and presented as recommendations.

Each recommendation given includes not only the predicted score, but also key product details such as the title, category, store, price, and average rating, which are decoded. An example of the recommendation output of a existing user is shown below:

```
Recommendations for: AEVPPTMG43C6GWSR7I2UGRQN7WFQ
Parent_ASIN: B01BFJ0MDE
Title: hqrp 2 pack ceiling fan capacitor cbb61 8uf 2 wire ul listed
Main Category: amazon home
Store: hqrp
Price: $6.91
Average Rating: 5.00
Score: 0.9956
```

Figure 4.2: Example of a practical recommendation

Overall, this step demonstrates how the trained models can be integrated into real and interpretable recommendations to users.

Chapter 5

Data Protection, Security, Ethics

5.1 GDPR

General Data Protection Regulation (GDPR) [27] has been around since 2016, setting rules inside the European Union (EU) to protect the privacy of personal data. It's applied to every data that is derived from EU residents, independently from when the company that is using the data is located. This data can only be processed to a certain extent and for specific purposes, like with explicit consent or in cases of public interest.

To avoid discrimination, the data processing should be transparent and fair, avoiding processes that lead to bias in race, gender, or religion. Users should also be well-informed when their data is collected, processed, and stored. Organizations need to provide how AI systems work clearly, as well as their capabilities and limitations. These AI systems need to have full technical documentation, including data sources, algorithms, and validation processes, to ensure compliance.

5.2 AI Act

Another EU regulation is the AI Act [28], that also gives some rules for the use of artificial intelligence in a secure and ethical way. This regulation is more recent and directed to AI. These systems are required to ensure that data processing, for example, user reviews, is limited to what is absolutely required for the purpose. Data anonymization is mandated, and the processing of biometric data is forbidden unless in exceptional cases.

Automated decisions like sentiment classification can't be done by AI systems without any human involvement, especially if it impacts in individual rights. In the AI Act, systems also need to be designed in a way to not make decisions that reinforce any discrimination in any attributes like gender, racial origin, or religion. The results should also be easy to interpret for any of the consumers.

Just like GDPR, these systems should provide full information about the capabilities and limitations of the system. This documentation should include the origin and nature of the dataset used and a detailed description of the algorithms.

5.3 Data Protection Impact Assessment (DPIA)

From the begining of this project, data protection, security, and ethical considerations were integrated into the design and implementation process. The use of the Amazon Electronics public dataset, which is openly available and anonymised by its provider, ensured compliance

with regulations as the ones stated above. After this anonymization, any user possible identifiers are also pseudonymised, preventing any direct identification of individuals.

5.3.1 Description of the data processing

The dataset contains product reviews, ratings, product titles, and metadata. Only a small and carefully selected set of fields was used: review text, title, rating, product ID, and pseudonymized user ID. Extra fields like photos, videos, or other personal details were removed. The data is processed only to train and evaluate the diverse recommender system models (BERT, LSTM and Hybrid). The system's output is a numerical score representing predicted sentiment or recommendation strength, and no raw reviews or identifiers are stored or shared after processing.

5.3.2 Necessity and proportionality

The project follows the principle of data minimization: only the data strictly needed for training was kept. Irrelevant or sensitive fields were excluded, reducing privacy risks and keeping processing proportional to the research goals.

5.3.3 Assessment of risks to data subjects

Re-identification risk: Even though the dataset is anonymized, there is a small chance of re-identification if those recommendations are combined with external information.

Bias and fairness risk: The models may reflect or amplify biases if presented in a considerable amount in the original reviews.

Ethical risk: Automated recommendations could influence user purchasing decisions without being given full transparency.

5.3.4 Measures to mitigate risks

Data protection: The use of anonymized and pseudonymized data, and exclusion of sensitive attributes.

Security measures: Data was processed only in a secure local environment. Temporary files were deleted immediately, and model checkpoints contained only learned parameters, not raw data.

Ethical principles: No demographic attributes (e.g., gender, age, ethnicity) were included, reducing risks of discrimination. The full workflow was documented to ensure transparency and reproducibility.

Accountability: All datasets, methods, and limitations are reported in this dissertation so that the results can be remade and critically assessed.

5.3.5 Conclusion

By combining anonymized public data, secure workflows, and ethical design choices, this project follows practices for responsible AI development. The residual risk to data subjects is considered low, and the research complies with GDPR principles of fairness, transparency, and accountability, as well as it follows the AI Act regulations.

Chapter 6

Conclusions

6.1 Conclusions

Using sentiment analysis from customer reviews, this thesis designed a customized product recommendation system for e-commerce that increases relevance and accuracy. This approach considers emotions and the details hidden in user-written reviews, in contrast to conventional recommendation systems that primarily depend on ratings or demographic information. This solves a key problem: customers may give a product a high overall rating but still be unhappy with certain aspects (like price or durability). By recognizing these details, the system can suggest products that better match what the user truly wants and feels.

The foundation of this work was a structured literature review carried out using the PRISMA protocol. This review examined researches on e-commerce recommendation systems, focusing on those that integrate emotions with Artificial Intelligence (AI), and Natural Language Processing (NLP). With this review there was addressed key questions such as how can emotions influence recommendations, the best ways to group users, and the ethical challenges in recommendation systems.

The Amazon Electronics dataset was chosen for this project because it has a wide range of products and a very large number of reviews. It was also well-suited for the needs of the CAPE project, which this work is collaborating with. Since the dataset was so large, a careful data-cleaning process was carried out. This reduced the dataset by about 95% for review's one and 98% for product metadata, while still keeping the most useful information. After the cleaning, there were 2 preprocessing stages performed.

The "2nd Preprocessing" method gave the best and most precise recommendations. It adjusted the text sequence lengths based on percentiles, which removed extreme outliers from the data. Without this step (as in "1st Preprocessing"), the model tended to average results and make less personalized suggestions, even with most of them still being relevant. While "Preprocessing 1" may work for niche cases, "Preprocessing 2" was clearly better for general e-commerce use.

The system's architecture combines two powerful models: BERT, which comprehends word context, and LSTM, which captures long-term patterns in text. This hybrid take was chosen because both models had proven successful in past studies. Several hybrid versions (V1 to V6) were tested, and their performance was measured using the MSE metric.

The LSTM model on its own performed better than BERT, achieving a lower MSE (0.0269 with Preprocessing 2 vs. BERT's 0.0401). This shows LSTM was especially good at following the flow of sentiment in text when predicting ratings.

Of the hybrid versions, Hybrid V5 was the best overall. Although Hybrid V4 reached the lowest MSE (0.0212), Hybrid V5 offered more consistent results with Preprocessing 2 (average MSE of 0.0243 instead of 0.0276) and trained much faster (1h14m vs. 3h02m). This balance of accuracy, stability, and speed makes Hybrid V5 the most practical for real-world use. For the niche cases, the V6 was the best model, since it could achieve 0.0222, while still being the fastest.

Ethical responsibility was a key part of the project. The system was designed to reduce common issues such as bias and to fully comply with GDPR (General Data Protection Regulation) and the AI Act. This was done by using only anonymized public data, and then even assigning pseudonymous IDs to users, and avoiding sensitive and personal details. The project's approach followed principles of fairness, transparency, and responsible AI use.

This study backs up the three retail objectives of the larger CAPE project: talent development, digital transformation, and sustainability. The system boosts the probability of a purchase, enhances consumer satisfaction, and develops long-term loyalty by offering fair and accurate recommendations.

6.1.1 Future Work

A key area for future improvements is making the system more adaptable to user preferences changing over time. One of the initial objectives in this work was to provide suggestions that gave less weight to older recommendations. The time needed to implement and test prevented this feature from being completely deployed. However, adding a way to assign less weight to older reviews is very important. This would address one of the biggest challenges in recommendation systems: keeping up with constantly changing interests and new product trends, which would improve satisfaction and long-term engagement.

Additionally, to better test the model, it would be useful to run new versions of the models using alternative metrics such as Precision, Recall, and F1-score, in combination with discrete sentiment categories. Instead of relying only on a continuous output, predictions could be mapped to fixed categories, for example, a scale from 0 to 1 or sentiment labels such as "positive," "negative," "angry," or "happy." This approach would allow the testing in cases where multiple products fall into the same category (and therefore are not strictly "better" or "worse" than one another), the system can still deliver acceptable results and potentially achieve even better performance on terms of speed, by focusing on classification accuracy rather than regression error.

Another crucial point for future work is performing additional fine-tuning on the existing models. Unfortunately again because of time constraints, it was not possible to dedicate more resources to this extra training. However, it would be highly advisable to carry out a substantial number of fine-tuning iterations across all model versions (BERT, LSTM, and Hybrids). With this, a model could always be better that what it is now, and it would allow to explore a broader parameter space and identify configurations that could significantly improve the results, and with them create more versions, optimizing the performance and efficiency of each architecture.

Finally a consideration to take in the near future is the computational demand of these recommendation systems. Due to the immense computational demand of the models, it is very difficult to achieve such immediate results at this dimension, even with the most advanced hardware currently available on the market. This limitation highlights the importance of not only exploring optimization techniques if the products in the store are as numerous as the

ones present in the tested dataset, but also waiting for better and cheaper hardware to be available, so it could be considered better than the existing recommendation systems right now.

Bibliography

[1] Aminu Da'u and Naomie Salim. "Sentiment-Aware Deep Recommender System With Neural Attention Networks". In: *IEEE Access* 7 (2019), pp. 45472–45484. doi: 10.1109/ACCESS.2019.2907729.

[2] Wang Xiaoye, Jiang Kaiwen, and Zhou Xiaowen. "Polarity Discrimination of Evaluation Words in The Product's Review". In: ICCAI '20 (2020), pp. 481–486. doi: 10.1145/3404555.3404567. url: https://doi.org/10.1145/3404555.3404567.

[3] Kun Xiong et al. "Counterfactual Review-based Recommendation". In: CIKM '21 (2021), pp. 2231–2240. doi: 10.1145/3459637.3482244. url: https://doi.org/10.1145/3459637.3482244.

[4] Li Chen, Dongning Yan, and Feng Wang. "User Evaluations on Sentiment-based Recommendation Explanations". In: *ACM Trans. Interact. Intell. Syst.* 9.4 (Aug. 2019). issn: 2160-6455. doi: 10.1145/3282878. url: https://doi.org/10.1145/3282878.

[5] Bingkun Wang et al. "Review rating prediction based on the content and weighting strong social relation of reviewers". In: UnstructureNLP '13 (2013), pp. 23–30. doi: 10.1145/2513549.2513554. url: https://doi.org/10.1145/2513549.2513554.

[6] Xiaoli Wang, Chenxi Zhang, and Zeshui Xu. "A product recommendation model based on online reviews: Improving PageRank algorithm considering attribute weights". In: *Journal of Retailing and Consumer Services* 81 (2024), p. 104052. issn: 0969-6989. doi: https://doi.org/10.1016/j.jretconser.2024.104052. url: https://www.sciencedirect.com/science/article/pii/S0969698924003485.

[7] Wenhao Guo, Jin Tian, and Minqiang Li. "Price-aware enhanced dynamic recommendation based on deep learning". In: *Journal of Retailing and Consumer Services* 75 (2023), p. 103500. issn: 0969-6989. doi: https://doi.org/10.1016/j.jretconser.2023.103500. url: https://www.sciencedirect.com/science/article/pii/S0969698923002473.

[8] Mete Sertkan. "Modeling Users and Items for Recommenders:There Is More than Semantics". In: RecSys '21 (2021), pp. 873–877. doi: 10.1145/3460231.3473898. url: https://doi.org/10.1145/3460231.3473898.

[9] Yikun Xian et al. "EX3: Explainable Attribute-aware Item-set Recommendations". In: RecSys '21 (2021), pp. 484–494. doi: 10.1145/3460231.3474240. url: https://doi.org/10.1145/3460231.3474240.

[10] Aminu Da'u et al. "Recommendation system exploiting aspect-based opinion mining with deep learning method". In: *Information Sciences* 512 (2020), pp. 1279–1292. issn: 0020-0255. doi: https://doi.org/10.1016/j.ins.2019.10.038. url: https://www.sciencedirect.com/science/article/pii/S0020025519310060.

[11] Zhu Zhan and Bugao Xu. "Analyzing review sentiments and product images by parallel deep nets for personalized recommendation". In: *Information Processing & Management* 60.1 (2023), p. 103166. issn: 0306-4573. doi: https://doi.org/10.1016/j.ipm.2022.103166. url: https://www.sciencedirect.com/science/article/pii/S0306457322002679.

[12] R.V. Karthik and Sannasi Ganapathy. "A fuzzy recommendation system for predicting the customers interests using sentiment analysis and ontology in e-commerce". In: *Applied Soft Computing* 108 (2021), p. 107396. issn: 1568-4946. doi: https://doi.org/10.1016/j.asoc.2021.107396. url: https://www.sciencedirect.com/science/article/pii/S1568494621003197.

[13] Mehdi Elahi et al. "Hybrid recommendation by incorporating the sentiment of product reviews". In: *Information Sciences* 625 (2023), pp. 738–756. issn: 0020-0255. doi: https://doi.org/10.1016/j.ins.2023.01.051. url: https://www.sciencedirect.com/science/article/pii/S0020025523000518.

[14] Juan Kong and Chen Lou. "Do cultural orientations moderate the effect of online review features on review helpfulness? A case study of online movie reviews". In: *Journal of Retailing and Consumer Services* 73 (2023), p. 103374. issn: 0969-6989. doi: https://doi.org/10.1016/j.jretconser.2023.103374. url: https://www.sciencedirect.com/science/article/pii/S0969698923001212.

[15] Christine Pinney et al. "Much Ado About Gender: Current Practices and Future Recommendations for Appropriate Gender-Aware Information Access". In: CHIIR '23 (2023), pp. 269–279. doi: 10.1145/3576840.3578316. url: https://doi.org/10.1145/3576840.3578316.

[16] Thi Ngoc Trang Tran, Alexander Felfernig, and Nava Tintarev. "Humanized Recommender Systems: State-of-the-art and Research Issues". In: *ACM Trans. Interact. Intell. Syst.* 11.2 (July 2021). issn: 2160-6455. doi: 10.1145/3446906. url: https://doi.org/10.1145/3446906.

[17] Alvise De Biasio et al. "A systematic review of value-aware recommender systems". In: *Expert Systems with Applications* 226 (2023), p. 120131. issn: 0957-4174. doi: https://doi.org/10.1016/j.eswa.2023.120131. url: https://www.sciencedirect.com/science/article/pii/S0957417423006334.

[21] Daozhen Min and Lei Huang. "Research on Recommendation Methods Based on Sentiment Analysis and BTM Topic Modeling". In: CSAI '18 (2018), pp. 425–430. doi: 10.1145/3297156.3297229. url: https://doi.org/10.1145/3297156.3297229.

[22] Sung-Jun Park et al. "Reinforcement Learning over Sentiment-Augmented Knowledge Graphs towards Accurate and Explainable Recommendation". In: WSDM '22 (2022), pp. 784–793. doi: 10.1145/3488560.3498515. url: https://doi.org/10.1145/3488560.3498515.

[23] Chenliang Li et al. "A Capsule Network for Recommendation and Explaining What You Like and Dislike". In: SIGIR'19 (2019), pp. 275–284. doi: 10.1145/3331184.3331216. url: https://doi.org/10.1145/3331184.3331216.

[24] Sheng Liu and Shixun Yang. "Machine Learning-Based Market Segmentation and Consumer Behavior Prediction Models". In: ICDSM '24 (2024), pp. 122–126. doi: 10.1145/3686081.3686100. url: https://doi.org/10.1145/3686081.3686100.

[25] Sudhanshu Kumar, Mahendra Yadava, and Partha Pratim Roy. "Fusion of EEG response and sentiment analysis of products review to predict customer satisfaction". In: *Information Fusion* 52 (2019), pp. 41–52. issn: 1566-2535. doi: https://doi.org/10.1016/j.inffus.2018.11.001. url: https://www.sciencedirect.com/science/article/pii/S1566253517304918.

Webgraphy

[18] IEEE Xplore. *IEEE Xplore Digital Library*. 2025. url: https://ieeexplore.ieee.org/Xplore/home.jsp.

[19] ScienceDirect. *ScienceDirect - Scientific Articles and Journals*. 2025. url: https://www.sciencedirect.com/.

[20] ACM Digital Library. *Association for Computing Machinery Digital Library*. 2025. url: https://dl.acm.org/.

[26] Yupeng Hou et al. "Bridging Language and Items for Retrieval and Recommendation". In: *arXiv preprint arXiv:2403.03952* (2024).

[27] European Union. *Regulation (EU) 2016/679 of the European Parliament and of the Council. General Data Protection Regulation (GDPR)*. OJ L 119, 4.5.2016, p. 1–88; em vigor desde 25.5.2018. 2016. url: https://eur-lex.europa.eu/eli/reg/2016/679/oj/eng (visited on 03/10/2025).

[28] European Union. *Regulation (EU) 2024/1689 of the European Parliament and of the Council. Artificial Intelligence Act (AI Act)*. OJ L 131, 24.7.2024, p. 1–49. 2024. url: https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng (visited on 01/26/2025).