



Universidade Federal de Campina Grande  
Centro de Engenharia Elétrica e Informática  
Coordenação de Pós-Graduação em Ciência da Computação

Dissertação de Mestrado

DigAI: A Chatbot Interface for Fashion  
Recommender Systems

André Ricardo Dantas Bezerra Landim

Campina Grande, Paraíba, Brasil

©André Ricardo Dantas Bezerra Landim, 08/02/2024

Universidade Federal de Campina Grande  
Centro de Engenharia Elétrica e Informática  
Coordenação de Pós-Graduação em Ciência da Computação

# DigAI: A Chatbot Interface for Fashion Recommender Systems

André Ricardo Dantas Bezerra Landim

Dissertação submetida à Coordenação do Curso de Pós-Graduação em  
Ciência da Computação da Universidade Federal de Campina Grande -  
Campus I como parte dos requisitos necessários para obtenção do grau  
de Mestre em Ciência da Computação.

Área de Concentração: Ciência da Computação  
Linha de Pesquisa: Inteligência Artificial

José Antão Beltrão Moura  
(Orientador)

Evandro de Barros Costa  
(Co-orientador)

Campina Grande, Paraíba, Brasil

©André Ricardo Dantas Bezerra Landim, 08/02/2024

L257d Landim, André Ricardo Dantas Bezerra.  
DigAI: a chatbot interface for fashion recommender systems / André Ricardo Dantas Bezerra Landim. – Campina Grande, 2024.  
61 f. : il. color.

Dissertação (Mestrado em Ciência da Computação) – Universidade Federal de Campina Grande, Centro de Engenharia Elétrica e Informática, 2024.  
"Orientação: Prof. Dr. José Antônio Beltrão Moura, Prof. Dr. Evandro de Barros Costa".  
Referências.

1. Artificial Intelligence. 2. Dialog System. 3. User Experience.  
4. Fashion E-commerce. I. Moura, José Antônio Beltrão. II. Costa, Evandro de Barros. III. Título.

CDU 004.8(043)



MINISTÉRIO DA EDUCAÇÃO  
UNIVERSIDADE FEDERAL DE CAMPINA GRANDE  
POS-GRADUACAO EM CIENCIA DA COMPUTACAO

Rua Aprígio Veloso, 882, Edifício Telmo Silva de Araújo, Bloco CG1, - Bairro Universitário, Campina Grande/PB,  
CEP 58429-900

Telefone: 2101-1122 - (83) 2101-1123 - (83) 2101-1124

Site: <http://computacao.ufcg.edu.br> - E-mail: [secretaria-copin@computacao.ufcg.edu.br](mailto:secretaria-copin@computacao.ufcg.edu.br) / [copin@copin.ufcg.edu.br](mailto:copin@copin.ufcg.edu.br)

## FOLHA DE ASSINATURA PARA TESES E DISSERTAÇÕES

**ANDRÉ RICARDO DANTAS BEZERRA LANDIM**

### **DigAI: A CHATBOT INTERFACE FOR FASHION RECOMMENDER SYSTEMS**

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação como pré-requisito para obtenção do título de Mestre em Ciência da Computação.

Aprovada em: 08/02/2024

Prof. Dr. JOSÉ ANTÃO BELTRÃO MOURA, UFCG, Orientador

Prof. Dr. EVANDRO DE BARROS COSTA, UFAL, Orientador

Prof. Dr. CLÁUDIO ELÍZIO CALAZANS CAMPELO, UFCG, Examinador Interno

Prof. Dr. THALES MIRANDA DE ALMEIDA VIEIRA, UFAL, Examinador Externo



Documento assinado eletronicamente por **Evandro de Barros Costa, Usuário Externo**, em 09/02/2024, às 08:29, conforme horário oficial de Brasília, com fundamento no art. 8º, caput, da [Portaria SEI nº 002, de 25 de outubro de 2018](#).



Documento assinado eletronicamente por **JOSE ANTAO BELTRAO MOURA, PROFESSOR 3 GRAU**, em 09/02/2024, às 10:44, conforme horário oficial de Brasília, com fundamento no art. 8º, caput, da [Portaria SEI nº 002, de 25 de outubro de 2018](#).



Documento assinado eletronicamente por **CLAUDIO ELIZIO CALAZANS CAMPELO, PROFESSOR(A) DO MAGISTERIO SUPERIOR**, em 15/02/2024, às 09:41, conforme horário oficial de Brasília, com fundamento no art. 8º, caput, da [Portaria SEI nº 002, de 25 de outubro de 2018](#).



A autenticidade deste documento pode ser conferida no site <https://sei.ufcg.edu.br/autenticidade>, informando o código verificador **4180629** e o código CRC **875B2219**.

## Resumo

A indústria da moda passou por uma transformação significativa nos últimos anos com o advento da tecnologia da informação e a proliferação das plataformas digitais. Essa mudança criou uma necessidade urgente de se comunicar efetivamente com os usuários e atender às suas necessidades de maneira personalizada e significativa. No entanto, o enorme tamanho dos catálogos de itens de moda e o número explosivo de combinações de produtos e preferências dos clientes levaram a um fenômeno conhecido como problema de sobrecarga de informações, que tende a degradar a experiência online dos clientes. Para mitigar os efeitos desse problema e melhorar a experiência online dos clientes, muitas empresas de moda implementam Sistemas de Diálogo (DS) como uma solução. Esses sistemas permitem que os usuários interajam com a plataforma e resolvam dúvidas sobre produtos, servindo como uma interface. No entanto, a complexidade da linguagem humana representa um desafio significativo para a eficácia e aceitação desses sistemas, particularmente em cenários orientados a tarefas e limitados ao contexto. O sucesso de um DS em entender a intenção de um usuário impacta diretamente sua experiência com o sistema. Por exemplo, um DS sofisticado, mas com baixo desempenho, pode ser pior do que uma solução muito mais simples (i.e. uma interface gráfica de usuário (GUI)). Como tal, projetar um sistema eficiente e confiável é fundamental para proporcionar experiências de usuário satisfatórias. Para enfrentar esse desafio, este trabalho tem como objetivo projetar, desenvolver e avaliar um chatbot chamado DigAI que sirva de interface para um sistema de recomendação que auxilie os usuários na busca de roupas. Para avaliar o desempenho, a usabilidade, os valores hedônicos e pragmáticos do chatbot, os usuários potenciais no Brasil avaliaram sua satisfação geral com a eficácia do chatbot em fornecer recomendações personalizadas em comparação com uma GUI mais simples. Essa avaliação nos permitirá identificar áreas de melhoria e refinar os recursos do chatbot. Este trabalho contribui para um esforço mais amplo de melhoria e personalização da experiência de compra online, aumentando assim a satisfação do cliente e impulsionando o crescimento dos negócios.

## **Abstract**

The fashion industry has undergone a significant transformation in recent years with the advent of information technology and the proliferation of digital platforms. This change has created an urgent need to effectively communicate with users and address their needs in a personalized and meaningful way. However, the massive size of fashion item catalogs and the explosive number of product combinations and customer preferences have led to a phenomenon known as the information overload problem, which tends to degrade customers' online experience. To mitigate the effects of this problem and improve customers' online experience, many fashion businesses have implemented Dialog Systems (DS) as a solution. These systems allow users to interact with a platform and resolve product queries by serving as an interface. However, the complexity of human language poses a significant challenge to the effectiveness and acceptance of these systems, particularly in task-oriented and context-limited scenarios. The success of a DS in understanding a user's intent directly impacts their experience with the system. For instance, a sophisticated but poorly performing DS may be worse than a much simpler solution (e.g., a Graphical User Interface (GUI)). As such, designing an efficient and reliable system is critical to delivering satisfactory user experiences. To address this challenge, this work aims to design, develop and evaluate a chatbot called DigAI that serves as an interface for a recommendation system that assists users in finding clothing. To evaluate the chatbot's performance, usability, hedonic and pragmatic values, potential users in Brazil assessed their overall satisfaction with the chatbot's effectiveness in providing personalized recommendations as compared to a simpler GUI. This evaluation contributes to the broader effort to improve and personalize the online fashion shopping experience, thereby enhancing customer satisfaction and driving business growth.

## **Agradecimentos**

Expresso meu agradecimento a Deus, pela sua graça. Agradeço a minha esposa Giselly pelo apoio incondicional e paciência durante todo o processo. Agradeço também aos meus pais e familiares pelo suporte, que me permitiu seguir em frente com meus estudos e alcançar este objetivo tão importante na minha vida. Agradeço ao professor Antão Moura, meu orientador, pelo incentivo e dedicação. Agradeço também aos professores Evandro Costa e Thales Vieira, meus co-orientadores, por sua ajuda e contribuição significativa para este trabalho. Agradeço aos colegas do grupo Projeto Moda: Artur Maia, Gabriel Barbosa, Luiz Sales e Robério Santos, pela ajuda em meu crescimento neste programa. Agradeço aos membros da banca examinadora, por suas críticas construtivas e sugestões valiosas. Por fim, gostaria de agradecer à Universidade Federal de Campina Grande, especialmente Paloma Porto, os professores e funcionários do Departamento de Sistemas e Computação, pela oportunidade de realizar este trabalho e pela formação acadêmica e profissional que recebi durante todos esses anos.

# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
<b>2</b>	<b>Background</b>	<b>5</b>
2.1	Recommendation Systems . . . . .	5
2.2	Dialog Systems (Chatbots) . . . . .	7
2.2.1	User Intent Classification . . . . .	7
2.2.2	Dialog Control . . . . .	8
2.2.3	Input Processing and Natural Language Understanding . . . . .	9
2.2.4	Evaluation of Chatbots . . . . .	9
<b>3</b>	<b>Methodology</b>	<b>11</b>
3.1	Systematic Literature Review . . . . .	12
3.2	Chatbot Design . . . . .	12
3.2.1	User Intent Classification . . . . .	13
3.2.2	Dialog Policy . . . . .	15
3.3	Chatbot Experiment . . . . .	16
3.3.1	Platform . . . . .	16
3.3.2	Selection Criteria . . . . .	17
3.3.3	Catalog . . . . .	17
3.3.4	Experiment . . . . .	18
3.3.5	Analysis . . . . .	22
3.4	OpenScience Practices . . . . .	23
<b>4</b>	<b>Literature Review</b>	<b>24</b>
4.1	Categorization . . . . .	24



---

4.2	Findings . . . . .	27
4.3	Discussion . . . . .	29
4.3.1	Computational perspective . . . . .	30
4.3.2	Non-computational perspective . . . . .	31
4.4	Literature Review Findings & This Work . . . . .	33
<b>5</b>	<b>Results for DigAI</b>	<b>34</b>
5.1	Chatbot Design . . . . .	34
5.1.1	Entity Extraction . . . . .	35
5.1.2	User Intent Classification . . . . .	36
5.1.3	Dialog Control . . . . .	40
5.1.4	Response Generator . . . . .	40
5.2	Experiment . . . . .	41
<b>6</b>	<b>Conclusion and Future Work</b>	<b>45</b>
6.1	Conclusion . . . . .	45
6.2	Threats to Validity . . . . .	47
6.3	Future Work . . . . .	48
<b>A</b>	<b>Questionnaire</b>	<b>59</b>
A.1	Concerning the quality or accuracy of the recommendations: (Likert) . . . . .	59
A.2	Concerning the diversity or variety of the recommendations: (Likert) . . . . .	59
A.3	Concerning the control you had on the flow of the recommendations: (Likert)	60
A.4	Concerning the effectiveness of the recommendations: (Likert) . . . . .	60
A.5	Concerning the trust you had in the system: (Likert) . . . . .	60
A.6	Concerning your overall satisfaction with the system: (Likert) . . . . .	61
A.7	Concerning your experience with the system: (adjectives) . . . . .	61

# List of Figures

3.1	Three Phase Methodology . . . . .	11
3.2	Turns by intent. . . . .	14
3.3	Dialog Policy Flow Chart . . . . .	16
3.4	Experiment first screen. . . . .	18
3.5	First Scenario: Recommendation System only. . . . .	19
3.6	Second Scenario: Chatbot and Recommendation System. . . . .	20
3.7	Likert question. . . . .	21
3.8	Adjectives question. . . . .	22
4.1	Categorization of e-commerce chatbot research. . . . .	25
5.1	F1 (95% CI) by Model. . . . .	37
5.2	F1 (95% CI) of LSTM and BiLSTM in a dropout x units grid. . . . .	38
5.3	Training time by Model. . . . .	39
5.4	Prediction throughput by Model. . . . .	40
5.5	Experiment time (95% CI). . . . .	41
5.6	Likert difference Part B - Part A (95% CI). . . . .	42
5.7	Difference Part B - Part A (95% CI). . . . .	44

# List of Tables

- 3.1 Turns' Intents. . . . . 14
- 3.2 LSTM and BiLSTM hyperparameters. . . . . 15
- 4.1 Computational main categories . . . . . 26
- 5.1 DigAI Categorization. . . . . 35
- 5.2 Entity categories. . . . . 36

# Acronyms

**AI** Artificial Intelligence. 2, 3

**AIML** Artificial Intelligence Mark-up Language. 9

**AR** Augmented Reality. 31

**AUC** apparel use context. 6

**BERT** Bidirectional Encoder Representations from Transformer. 3, 7, 8, 14, 15, 36–39, 46

**BiLSTM** Bidirectional LSTM. 3, 7, 8, 14, 15, 36–38

**BR** Brazil. 20

**CBH** customer’s browsing history. 6

**CM** Customer Model. 5–7

**CNN** Convolutional Neural Networks. 9, 28

**CPC** customer’s physical characteristics. 6

**CPT** customer’s personality traits. 6, 7

**CV** Computer Vision. 7

**DL** Deep Learning. 3, 8, 9, 26–28, 30

**DM** Dialogue Management. 1

**DNNs** Deep Neural Networks. 7

- 
- DSS** Decision Support Systems. 5
- DSTC11** The Eleventh Dialog System Technology Challenge. 3
- FKC** Fashion-Knowledgeable Component. 29
- GAN** Generative Adversarial Networks. 9, 28
- GUI** Graphical User Interface. 2, 22, 41, 42, 45–47
- IR** Information Retrieval. 28
- IT** Information Technology. 20
- KB** Knowledge-based. 6
- LSTM** Long Short-term Memory. 3, 7, 8, 14, 15, 36–38
- ML** Machine Learning. 1, 3, 9, 26, 28
- NL** Natural Language. 1, 25
- NLG** Natural Language Generation. 1, 25, 35
- NLP** Natural Language Processing. 7–9, 24, 26, 30
- NLU** Natural Language Understanding. 1, 9, 25, 35
- NUI** Natural User Interface. 30
- PF** products' features. 6, 7
- RNN** Recurrent Neural Network. 7, 9, 28
- RQ** Research Question. 2, 33, 43, 46
- RS** Recommender System. 1, 3, 5, 6, 18, 29, 42, 43, 46, 47
- SVC** Support Vector Classification. 3, 7, 8, 14, 15, 36–39, 46

**SVM** Support Vector Machine. 8, 33, 35

**TAM** Technology Acceptance Model. 29, 31

**TF-IDF** Term Frequency–Inverse Document Frequency. 9, 14

**U&G** Use & Gratification Theory. 29

**UK** United Kingdom. 20

**VFR** Virtual Fitting Room. 31

# Chapter 1

## Introduction

The prevalence of interactive Recommender System (RS) in the form of chatbots and virtual assistants is on the rise in today's society. In 2021, it was estimated that chatbots were used by approximately 1.4 billion people worldwide, with the United States, India, Germany, the United Kingdom, and Brazil being the top users [6]. Moreover, it was projected that over 67% of global consumers utilized chatbots for customer service in 2019 [65].

In the realm of e-commerce, the incorporation of chatbots and virtual assistants on websites and social media platforms has seen a significant increase. This has resulted in benefits for both customers and business owners, in areas that include 24/7 customer support, automated purchase recommendations, and enhanced customer engagement. Furthermore, chatbots have the potential to help companies save up to 30% on customer support costs and improve response times by addressing up to 80% of routine inquiries [65].

While chatbots based on pattern matching and simple "Q&A" style are still common, the transition to more human-like conversations remains a challenge. This evolution is heavily reliant on Machine Learning (ML) algorithms. Modern chatbots necessitate three components: an Natural Language Understanding (NLU) component to discern the user's intent; a Dialogue Management (DM) to formulate a response based on the conversation's context; and an Natural Language Generation (NLG) to generate a response in Natural Language (NL) [66].

This work concentrates on the identification and classification of user intent, a crucial task for the operation of a chatbot. As Schuurmans and Frasincar [64] suggest, the objective of intent classification is to comprehend the motives behind the customer's interaction with

the company and the goals they aim to achieve through this interaction. To fulfill its objectives, the chatbot must have access to a database and data model that aligns with the specific domain of its application. However, despite significant advancements in the field of natural language processing and the creation of robust databases, many domains still have limited data available for training chatbot classifiers and limited research conducted on them.

To facilitate the continued development and widespread adoption of conversational agents in society, it is crucial to carry out comprehensive research and comparative evaluations of the various Artificial Intelligence (AI) models commonly used in training these agents. This will offer a deeper understanding of their capabilities and limitations, highlight areas for improvement, and ultimately lead to the creation of more advanced and efficient chatbots.

In this work, we address the following Research Questions (RQ):

- RQ 1: What are the existing methods and technologies used in Dialog Systems for fashion businesses, and how have they addressed the information overload problem to improve customers' online experience?
- RQ 2: Which Machine Learning algorithm can most effectively understand and classify user intent in the context of a Dialog System for a fashion recommendation platform?
- RQ 3: a) How effective is the developed chatbot, as compared to a simpler GUI, serving as an interface for a recommendation system, in terms of user satisfaction and its ability to provide personalized recommendations? b) What areas of improvement can be identified from usability testing and applying a hedonic and pragmatic questionnaire with potential users?

In order to provide a more comprehensive response to RQ 1 within the forthcoming literature review (Chapter 4), a strategic approach was employed, involving the subdivision of RQ 1 into the following specific questions:

- RQ 1.1: Which are the current proposed chatbot design approaches for e-commerce in general, and particularly for fashion applications?



- 
- RQ 1.2: How could research on e-commerce chatbot be categorized in an integrated manner?
  - RQ 1.3: What are the research opportunities of chatbots design to deal with the specificities of fashion e-commerce applications?

The objectives of this study consist in answering the above RQs and revolve around addressing the challenges posed by the information overload problem in the online fashion industry and enhancing customers' online experience through the development and evaluation of a chatbot named DigAI to serve as an interface for a recommendation system in fashion e-commerce. The study aims to evaluate DigAI's performance, usability, hedonic, and pragmatic values in comparison to a simpler Graphical User Interface (GUI). The assessment involves potential users in Brazil, measuring overall satisfaction and effectiveness in providing personalized recommendations.

This study contributes to the field of online fashion retail by the prevalent issue of information overload in the fashion industry, exacerbated by extensive catalogs and varying customer preferences. A key contribution lies in the meticulous annotation of a dataset used to train DigAI, enhancing DigAI's proficiency in understanding the catalog for more effective personalized recommendations. The experiment findings, comparing DigAI to a simpler GUI, provide insights into fashion chatbots' performance, usability, and user satisfaction in general.

The remainder of this document is organized as follows. Chapter 2 introduces concepts of AI, RS and Chatbot to facilitate reading subsequent chapters. Then, chapter 3 presents the methodology adopted to answer the research questions. Chapter 4 details the systematic literature review conducted to answer RQ 1. Chapter 5 answers RQ 2 and 3. It brings the results of the experiment with both traditional classifiers and cutting-edge Deep Learning (DL) approaches, as discussed in Section 3.2. The ML classifiers considered in this investigation included the Support Vector Classification (SVC) [13], Bidirectional Encoder Representations from Transformer (BERT) [32], Long Short-term Memory (LSTM) [27], and Bidirectional LSTM (BiLSTM) [23]. We utilized a dataset provided by Facebook in The Eleventh Dialog System Technology Challenge (DSTC11), which consists of task-oriented dialogues in the field of fashion shopping [36]. However, the full database is not accessible,

making it impractical to perform fair comparisons with the published results of the challenge [35]. Then, we conducted an experiment with the proposed chatbot detailed in Section 3.3. Finally, chapter 6 concludes the study, summarizing the key findings and their implications, and providing future work for mitigating the threats of the study.

# Chapter 2

## Background

The contents of this chapter serve the purpose of establishing a standardized knowledge foundation for the subsequent chapters, particularly focusing on aspects related to AI and Chatbot usage. The intention is to provide a comprehensive overview that facilitates a smoother reading experience in the subsequent sections. This chapter is designed to be optional for readers who are already acquainted with the topic.

### 2.1 Recommendation Systems

In a broad sense, RS fall under the umbrella of Decision Support Systems (DSS). A defining characteristic of RS is the incorporation of a user model, which encapsulates vital information about an individual or group. This user model is crucial for a DSS to exhibit adaptive behavior, tailoring its responses for different users. In essence, RSs encompass a class of well-established software tools and techniques designed to offer suggestions based on user preferences. Resnick et al. [62] note that recommendations play a pivotal role in consumer decision-making across various domains, including products, services, and general content. Current applications of RSs span music services, news, restaurants, and the realm of fashion e-commerce, where the "user" signifies the customer seeking fashion products or services. The combination of Customer Model (CM) and RS (CM-RS) proves instrumental in assisting customers by recommending suitable purchasing options.

The prominence of RSs surged with the rise of e-commerce and the availability of extensive catalogs, addressing the challenge of information overload. RSs employ various tech-

niques, such as content-based filtering, collaborative filtering, knowledge-based methods, or hybrid approaches combining these strategies [2].

Content-based filtering techniques utilize predictive algorithms to align the characteristics of a product or service with a customer's profile. The process involves cataloging items in a straightforward manner, while CMs are gathered either explicitly or implicitly. Various methods can be integrated to collect, construct, and update a CM. Typically, these systems gather data from customer interactions, which may include activities such as rating items, ranking items, and selecting items from an item gallery. Implicit collection might involve analyzing customer views (and the duration of views) of store items, examining purchase history, and conducting social network analysis, among other methods. However, these systems may encounter a 'cold start' problem when a new and unknown customer begins interacting with a RS [4].

Collaborative filtering operates on the assumption that a group of customers with similar ratings or consumer behavior preferences will likely share common preferences for other items [22]. In both content-based and collaborative filtering approaches, robust CMs are imperative for delivering high-quality recommendations.

Knowledge-based (KB) recommenders derive suggestions based on domain knowledge about how items align with user preferences, encompassing knowledge about users, items, and the match between an item and the user's needs [2].

CMs for RSs may be static, considering only long-term preferences, or dynamic, incorporating both long-term and short-term preferences. Identifying a customer's short-term and long-term preferences is crucial, exemplified by scenarios where a customer from a tropical country predominantly explores and purchases summer outfits (long-term preference) but may occasionally seek a winter coat for a vacation trip (short-term preference). For an in-depth exploration of RSs, readers are referred to [4].

Pereira et al. [56] identified five feature categories for short-term and long-term preferences: i) products' features (PF); ii) apparel use context (AUC); iii) customer's browsing history (CBH); iv) customer's physical characteristics (CPC); and v) customer's personality traits (CPT). Category ii) represents transient needs valid for specific time-space situations (short-term features), while category v) signifies much slower-varying features reflecting long-term preferences. Category iv) mainly includes body measurements and color (of

skin, eyes, hair). Categories i) and iii) may be interconnected, representing customer needs through product characteristics.

The prevalence of PF in CMs is not surprising, given that customers, retailers, and algorithms all need to consider the products themselves when making purchase recommendations. In contrast, CPT aspects were found to be present in only one study [52]. This could be due to the greater difficulty in acquiring and verifying customer personality traits compared to their physical characteristics. However, CPT aspects could play a crucial role in personalized online recommendations. Each of the remaining categories was addressed in a third (15) of the papers investigated, indicating that these categories are not mutually exclusive.

Preprocessing techniques are often necessary to extract high-level structured data for constructing a customer model. Techniques vary based on data modality, with image data typically processed using state-of-the-art Deep Neural Networks (DNNs) and, less frequently, classical Computer Vision (CV) algorithms. For textual data, classical Natural Language Processing (NLP) algorithms are prevalent, though text-based DNNs are increasingly utilized. Recent work also explores multimodal approaches, combining visual and textual data to obtain comprehensive customer information [28; 24; 63].

## 2.2 Dialog Systems (Chatbots)

### 2.2.1 User Intent Classification

In this study, we perform tests on several classifiers frequently utilized for intent classification, namely: LSTM, BiLSTM, BERT, and SVC. The selection of these classifiers was based on their successful implementation in numerous related studies, as outlined in chapter 4.

Recurrent Neural Networks (RNNs) stand apart from other neural network types due to their ability to retain information while handling sequential data. However, they encounter challenges with the transmission of “long-term” information. To put it differently, the farther the necessary information is within the network, the more strenuous it becomes for RNNs to access it.

LSTM represents a unique architecture within the realm of recurrent neural net-

works, specifically engineered to circumvent issues related to long-term dependencies. Its widespread application in NLP tasks can be attributed to its proficiency in handling sequential data.

BiLSTM, also known as bidirectional LSTM, is a model that integrates two interconnected LSTMs, thereby augmenting the volume of information accessible to the network. Unlike the traditional LSTM network, which is limited to utilizing information from preceding contexts or layers, BiLSTM is capable of simultaneous training in both temporal directions. In essence, it can "read" and interpret sequential text data bidirectionally (both forward and backward).

BERT is a potent DL technique that employs the Transformer mechanism to proficiently comprehend the contextual associations between words or subwords in a text, thereby enhancing performance across a range of NLP tasks. BERT Transformer encoder processes the sequence of words in a text collectively, and is thus deemed bidirectional. This characteristic enables the model to understand the context of a word based on its surrounding elements (either to the right or left of the word).

SVC is a type of supervised machine learning algorithm, falling under the broader category of Support Vector Machine (SVM). It is adept at resolving classification issues via maximum margin separating hyperplanes. Specifically, Linear SVC, a variant of SVM with a linear kernel, exhibits commendable performance in numerous NLP tasks.

### 2.2.2 Dialog Control

Each individual contribution that forms a dialog is referred to as a turn [29]. Reactive agents operate by responding only after a user's turn, thereby solely leveraging information explicitly provided by the user's initiative. In contrast, proactive chatbots utilize an engagement strategy to interact with and influence users, drawing upon predictions about users' needs.

To assist users in accomplishing a specific task, chatbots must implement a dialog policy. This policy is tasked with determining the system's next action. While numerous distinct approaches have been proposed, we have identified three primary subcategories in the literature:

- Predict the user's intent, and subsequently select a specific response from a finite set

of predefined responses;

- Assess the similarity between the user's queries and questions in a dataset, and choose a corresponding response; and
- Sequence-to-sequence DL approaches first decode the user's intent, and then generate a response.

### 2.2.3 Input Processing and Natural Language Understanding

When user interactions are facilitated through buttons or multiple-choice interfaces, the chatbot can readily comprehend the expressed intents and/or other pertinent information. However, when interactions involve human natural language in the form of unstructured data (potentially extracted from speech), specific NLU algorithms must be implemented. Rule-based algorithms utilize handcrafted rules, for example, the Artificial Intelligence Mark-up Language (AIML). Traditional ML algorithms typically combine the extraction of handcrafted features from unstructured textual data, such as n-gram counts or the Term Frequency–Inverse Document Frequency (TF-IDF) statistical measure. DL algorithms, which employ deep neural network architectures for sequence processing, offer state-of-the-art performance. These primarily include variations of Convolutional Neural Networks (CNN) and RNN to identify complex patterns within data.

The responses generated by chatbots also adhere to approaches that mirror NLU algorithms. Thus, they can be classified as rule-based, where sentences are generated from pre-written templates, or neural-based, where sentences are generated from textual training data. In the latter case, variants of CNN, RNN, and Generative Adversarial Networks (GAN) are among the most prevalent approaches. For a thorough introduction to NLP algorithms, we direct the reader to Jurafsky and Martin [29].

### 2.2.4 Evaluation of Chatbots

Chatbots are evaluated based on two key aspects: hedonic and pragmatic features [19]. Hedonic characteristics are tied to the emotional reactions, such as happiness or excitement, that a chatbot can provoke, influencing its perceived appeal. Pragmatic characteristics, con-

versely, are linked to the chatbot's practical elements, like its usefulness, precision, and overall quality.

AttrakDiff [25], a tool developed by UID in partnership with academic institutions, is often used to measure these features. It allows for the anonymous evaluation of a product by its users or customers. The data gathered from these evaluations helps determine the product's perceived attractiveness in terms of usability and aesthetics, and whether any improvements are needed.

In relation to chatbots, examples of pragmatic attributes include efficient assistance and interpretation issues, while hedonic attributes might include entertainment value and unusual or impolite responses. Tools like AttrakDiff provide developers with important insights into user experience, enabling them to fine-tune their chatbots to better cater to both hedonic and pragmatic needs.



# Chapter 3

## Methodology

In this chapter, we present a three-phase methodology (see Figure 3.1) adopted to investigate the feasibility of using chatbot as a interface for a recommendation system. The first phase involves conducting a systematic literature review to gather relevant research papers and existing work in conversational agents (see chapter 4). This comprehensive review provides a foundation for subsequent phases. The second phase focuses on implementing the chatbot named DigAI (from this point on, chatbot and DigAI will be mentioned interchangeably). Finally, the third phase involves conducting an experiment with real users to evaluate the chatbot’s performance, measure user satisfaction, and gather feedback for further improvements.

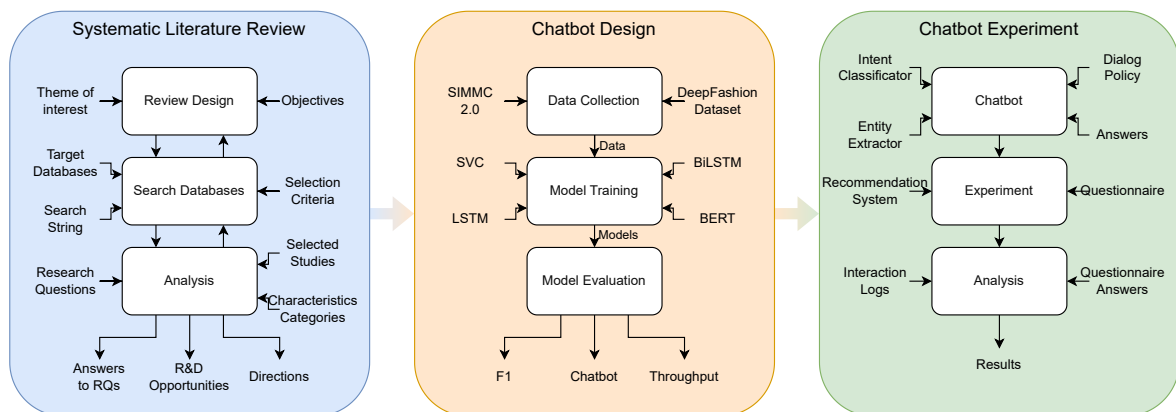


Figure 3.1: Three Phase Methodology

## 3.1 Systematic Literature Review

We conducted a theme-based literature review using a flexible three-step methodology adapted from Okoli [51]. The steps, which are not strictly sequential, allowed us to refine the review's design, execution, and analysis based on the results. This comprehensive review aims to answer specific research questions:

To address the three mentioned research questions, we used the following search string:

(Chatbot OR Dialog System OR Conversational Agent OR Virtual Assistant OR Digital Assistant) AND ((E-comm\* OR Ecomm\* OR Retail) OR (Fashion OR Garment OR Apparel OR Clothing OR Cosmetics OR Makeup OR Make-up))

This string was used to gather papers published between 2011 and 2021 from various databases, including ACM Digital Library, Emerald, Google Scholar, and others. This spanned a decade of research in this field, with no studies on the theme found prior to 2014.

We also used a second search string that wasn't limited to 'e-commerce' and 'fashion'. We included variations for both 'fashion' and 'chatbots'. Studies were excluded if they were not in English, unavailable online, or not focused on digital retail chatbots. The included works were journal papers, conference papers, and book chapters.

The results of this phase of the methodology are presented in chapter 4.

## 3.2 Chatbot Design

The architecture of the DigAI chatbot is structured around 4 essential components, each contributing to its overall functionality:

- The intent classifier plays a pivotal role in understanding user queries and discerning their underlying purpose;
- Meanwhile, the entity extractor focuses on identifying specific entities or pieces of information within the user input, enhancing the chatbot's ability to comprehend and respond accurately;

- In addition, the chatbot incorporates a dialog control system. This component is designed to manage the flow of conversation, ensuring a coherent and contextually relevant exchange between the user and the chatbot;
- Furthermore, a response generator is integrated into the system, enabling the chatbot to generate appropriate and context-aware responses based on the information processed by the intent classifier and entity extractor.

### 3.2.1 User Intent Classification

In this section, we delve into the process of training the intent classification model. This training procedure involves four essential steps:

1. **Dataset Selection:** To begin, we carefully choose the dataset that serves as the foundation for our model's training.
2. **Model Training:** Next, we train the selected models.
3. **Performance Evaluation:** Then, we evaluate the performance of the trained models to gauge their effectiveness and accuracy (F1).
4. **Dialog Policy:** Finally, we propose a dialog policy that decides the chatbot's next action.

#### Data Collection

Regarding data for model training and testing, the dataset SIMMC 2.1 [34] was used. It contains 9,557 dialogs between humans and virtual assistants totaling 50,230 turns in the fashion-related shopping domain. The dataset was made available in three parts, called Train, Dev and Test-Dev. The Test-Std part is not publicly available and is used for internal evaluation of the model in the challenges. Out of the limited datasets available, this one stood out as the optimal choice due to its substantial size and relevance to the fashion domain. Figure 3.2 shows the distribution of turns by user intent

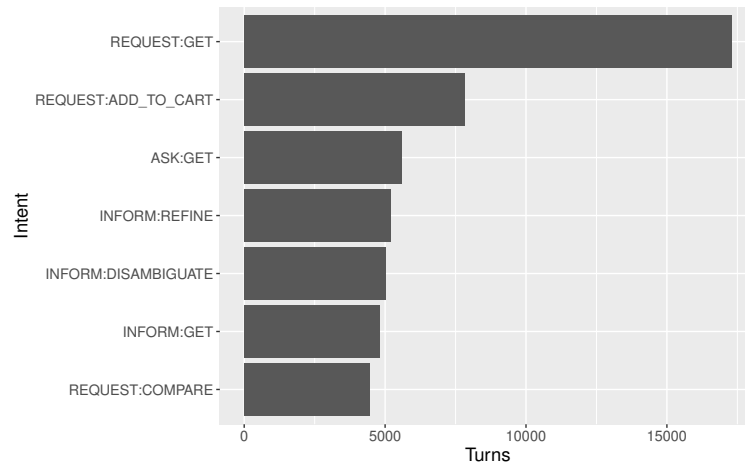


Figure 3.2: Turns by intent.

As a first step, the dataset was cleaned. Symbols were removed and sentences were converted to lowercase. Then each user’s utterance was either tokenized, for the models LSTM and BiLSTM, or encoded as TF-IDF, for the SVC model. Note that, since the BERT model has a text pre-processor, the text is passed as-is to the model. Finally, we split the turns of the dataset into three chunks: training (70%), validation (15%) and testing (15%).

The user intent defines the user’s objectives or requests. This information plays a crucial role in guiding the chatbot’s subsequent actions and responses. Table 3.1 presents a compilation of illustrative user phrases along with their corresponding intents, providing tangible examples that demonstrate the association between user input and the intended outcomes or purposes.

Table 3.1: Turns’ Intents.

Turn	Intent
What do you think of the grey pair on the left?	INFORM:GET
Do you have any plain jeans?	REQUEST:GET
Which one do you mean?	REQUEST:DISAMBIGUATE
That brown one should work for me.	REQUEST:ADD_TO_CART

## Model Training

As already informed in Section 2.2.1, four classifiers were chosen for evaluation: LSTM, BiLSTM, BERT and SVC. Each of them was trained with the training data and the validation data were used to keep track of the current loss and F1. The validation loss is monitored with the early stopping method, which was configured with a patience of 10 epochs, default min delta (0) and returning the best weights achieved.

The hyperparameters experimented for the LSTM and BiLSTM models are shown in Table 3.2.

Table 3.2: LSTM and BiLSTM hyperparameters.

Hyperparameter	Values
Units	128, 256 and 512
Dropout	10%, 20% and 30%
Batch size	32, 64 and 128

Regarding BERT, we used the bert-en-uncased pre-trained model (L=8, H=256, A=4, v2) as initialization and the same hyperparameters, with the exception of Units.

For the SVC, the regularization strength parameters (C) used were: 0.1, 0.5, 1, 2, 5, 10, 20 and 100.

## Model Evaluation

The training was repeated 10 times for each chosen hyperparameter configuration and the F1 averaged with a 95% confidence interval. Finally, the F1 score of the trained models on the test set was observed to rank the classifiers.

### 3.2.2 Dialog Policy

The proposed dialog policy is given as follows. The language of the user's request is identified. If the language is determined to be English, the system proceeds to discern the intent

of the request. If the intent is to request a product or inform the user’s preference then it fills any pertinent slots accordingly. If the user wants to ask about the product, disambiguate a missing information or compare products, then the system tries to answer. In cases where a slot remains unfilled, the system prompts the user to provide the necessary information.

For non-English language requests, a translation process precedes the identification of intent, and subsequent slot filling is carried out in a language-agnostic manner. The overall workflow is systematically represented in a flowchart (Figure 3.3), which outlines conditions for direct question responses and inquiries about the provision of assistance based on the timing of the last message.

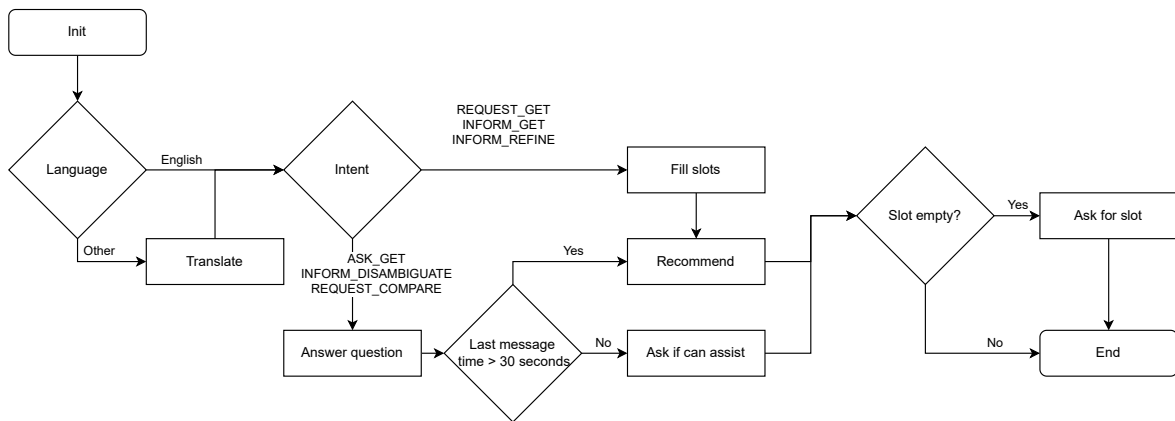


Figure 3.3: Dialog Policy Flow Chart

## 3.3 Chatbot Experiment

In this section we present the details of the experiment, its platform, selection criteria, fashion (clothing) catalog, scenarios, questionnaire and data analysis method.

### 3.3.1 Platform

The platform is designed as a web system utilizing a microservice architecture, comprising the DigAI chatbot and a recommendation system. We adopted the recommendation system proposed by Pereira et al. [57], tailoring it for seamless integration with the updated web interface. It is essential to highlight that the recommendation system, as implemented, lacks features for catalog filtering. Catalog filtering is beyond the scope of this study but could

be explored in future research efforts. Should the recommendation system offer filtering facilities, the chatbot experiment to evaluate its performance may differ from those reported in section 5.2.

The DigAI chatbot is unveiled in the second scenario (refer to section 3.3.4). Users have the flexibility to seek information, discuss the catalog, and solicit clothing recommendations from the chatbot at any given moment. The chatbot seamlessly interfaces with the recommendation system, allowing it to search the catalog based on user-provided information and initiate requests for new clothing items through the recommender interface.

### 3.3.2 Selection Criteria

The experiment had the active participation of a total of 25 individuals, initially gathered by convenience -i.e., the author's contacts, and who had some familiarity with online fashion shopping. Additionally, a snowball sampling method was employed to broaden the participant pool. Through this approach, initial participants were asked to refer others who might be interested, creating a cascading effect that facilitated the inclusion of a diverse range of individuals.

### 3.3.3 Catalog

To conduct the experiment, the Deep Fashion [42] database was selected, housing approximately 300,000 images of various clothing items. Given the extensive dataset, we narrowed our focus to skirts, yielding a subset of 13,000 images. However, initial examinations revealed a notable prevalence of mislabeled data, prompting a comprehensive re-annotation of the entire database.

In the initial phase, images featuring items other than skirts were systematically removed, resulting in a refined dataset of 11,000 skirt images. Subsequently, 99 individuals were recruited from our network of contacts to undertake the annotation process. These annotators were tasked with categorizing images across five dimensions: color, fabric, pattern, size, and shape. A consensus criterion was established, requiring agreement from at least two annotators for each category. Over a span of three months, a total of 1,175 images underwent meticulous annotation, forming the curated catalog utilized for the subsequent stages of the

experiment.

### 3.3.4 Experiment

Initially, the user accesses the platform<sup>1</sup>, which outlines the experiment's objective of identifying a suitable medium-sized striped skirt for purchase. On the initial screen (Figure 3.4) of the experiment, the user is provided with the opportunity to review instructions on utilizing the tool and commencing the experiment.

**UFCEG**

**UFAL**

## USER EXPERIENCE IN CLOTHING FASHION E-COMMERCE - EVALUATION EXPERIMENT

**What is the research about?**

I am André Landim, a Computer Science Master's student at the Federal University of Campina Grande (UFCEG) in collaboration with the Federal University of Alagoas (UFAL) in Brazil. We would like to invite you to participate in a study regarding chatbot and recommender system for fashion e-commerce. This study was approved by the Ethics Committee at UFAL (CAAE 43950621.4.0000.5013).

**What will happen to me if I take part?**

You are kindly requested to help us evaluate scenarios of Recommender Systems' applications to Fashion E-commerce. For that, we designed an evaluation website in which you try to select a skirt you like and would possibly buy from a fashion e-commerce site in three different scenarios. The website interactive experiment should take you less than 15 minutes to complete (there are 2 parts). Results will be kept anonymous. After the experiment, you will be asked to complete an anonymous questionnaire which should take you another 15 minutes approximately. Thus, altogether (experiment + questionnaire) you might spend around 39min in this study.

**What is the "evaluation experiment" via website interaction and what am I supposed to do?**

In the experiment you will find a list of skirts to choose. No personal information will be collected, only your interactions with the website via clicks. Your goal is to find a striped mid-size skirt that you like and would "buy". To simulate your buying action and end the experiment you click on the (shopping cart) button. The website interaction will take place in three parts. In Part A recommendations are made to you without the assistance of a chatbot and in Part B, you may use a chatbot after an initial, random set of skirt suggestions

**What information will be collected?**

The questions in this survey seek to identify various aspects of the user's experience, including the quality, accuracy, variety, control, and efficiency of the recommendations provided by the platform, as well as the overall trust and experience of the user. We also collect logs of tool usage, which record each user iteration and the timestamp of each action taken. We do not collect any personal information, only user analytics. We do not collect any information, such as email, that may identify you in any way. Some of the survey questions contain textboxes where you will be asked to type in your own answers. Please note that in order for this survey to be anonymous, you should not include in your answers any information from which you, or other people, could be identified.

Please click on the "Instructions" button below before proceeding. Then, return to this page, check the "Consent Box" and press the "Start Experiment" button. Thank you for your time.

By continuing with this survey you confirm that you are at least 18 years of age and that you consent to participate. If you do not consent to participate, please exit this survey or close your browser.

**INSTRUCTIONS** **START EXPERIMENT**

Figure 3.4: Experiment first screen.

The experiment is segmented into two distinct scenarios. The first scenario involves utilizing the RS exclusively to identify the desired striped skirt. The second scenario incorporates the Chatbot alongside the RS to facilitate a more interactive and personalized shopping experience.

<sup>1</sup><https://digai.andrebezerra.com>



### Scenario A

The first scenario (Figure 3.5) serves as the baseline for the experiment, as it exclusively utilizes the recommender system. In this scenario, the user is presented with a random selection of skirts and is provided with the option to indicate their preference by either liking or disliking the garments. As the user progresses through the experiment and provides feedback on the clothes presented, the system will adapt and refine its recommendations based on the customer's taste. By utilizing the "more skirts recommendations" button, the system will present similar clothes based on the user's preferences.

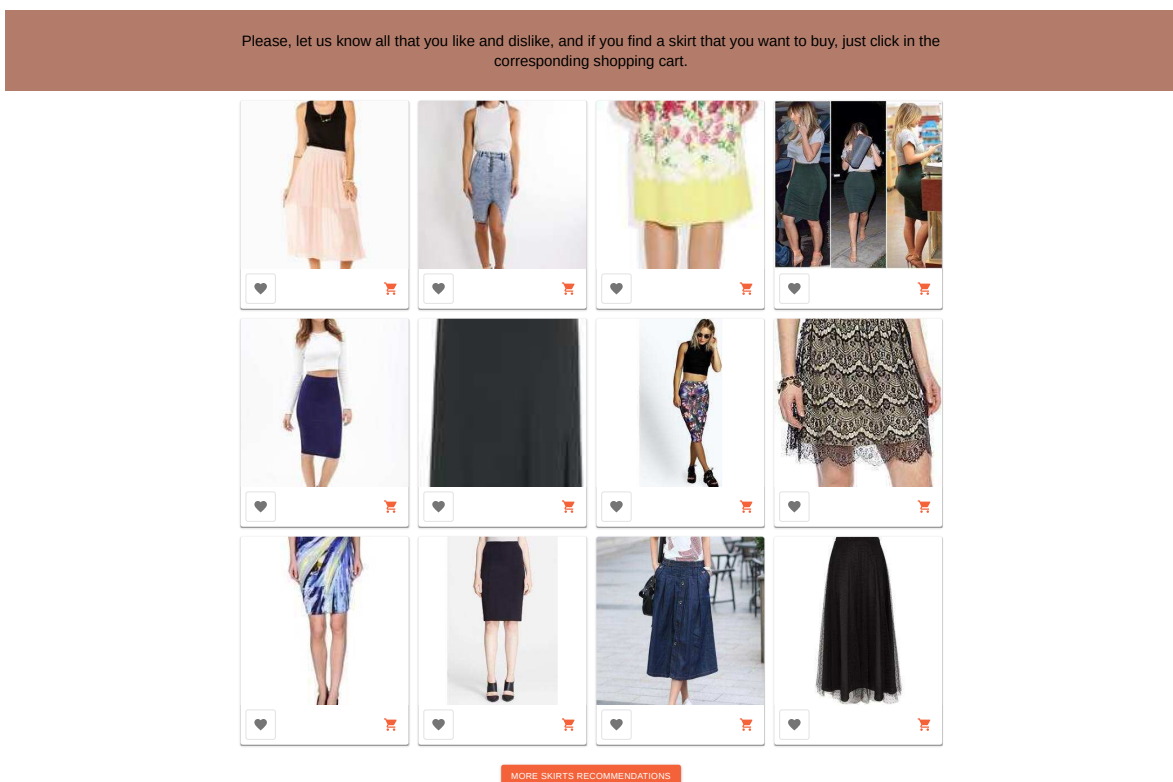


Figure 3.5: First Scenario: Recommendation System only.

### Scenario B

The second scenario (Figure 3.6) seeks to identify the impact of incorporating a chatbot as an interface for the same recommendation system. In this scenario, the user is presented with a selection of random skirts and can similarly indicate their preference by liking or disliking the garments. The main difference in this scenario is that the chatbot provides an additional

layer of interactivity, allowing the user to filter the results by conveying their specific search criteria through natural language processing. This feature allows for a more personalized and intuitive shopping experience, as the chatbot can accurately interpret the user's search criteria and tailor recommendations accordingly.

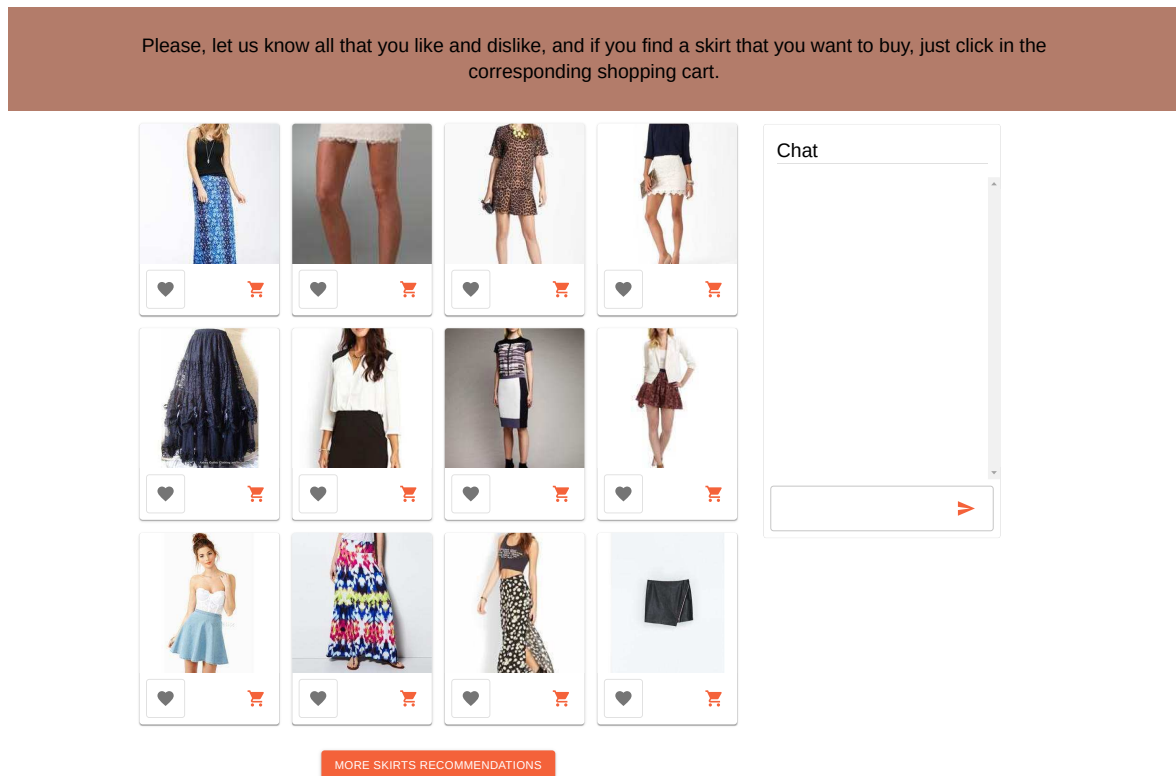


Figure 3.6: Second Scenario: Chatbot and Recommendation System.

### Questionnaire

The method used was a survey with the application of a questionnaire (Refer to Appendix A) to collect user feedback for scenarios A and B. Before being formally administered, the questionnaire underwent a pre-testing process involving collaboration with fashion experts in the United Kingdom (UK) and professionals in Information Technology (IT), as well as fashion experts in Brazil (BR).

This form consists of a series of questions answered on a 5 level Likert scale (Figure 3.7), which was produced based on the questions proposed by Tsai & Brusilovsky [68] adapted for the fashion domain. It seeks to identify various aspects of the user's experience, including

the quality, accuracy, variety, control, and efficiency of the recommendations provided by the platform.

Questionnaire

MECG UFAL

1 2 3 4 5 6 7

Concerning the quality or accuracy of the recommendations:

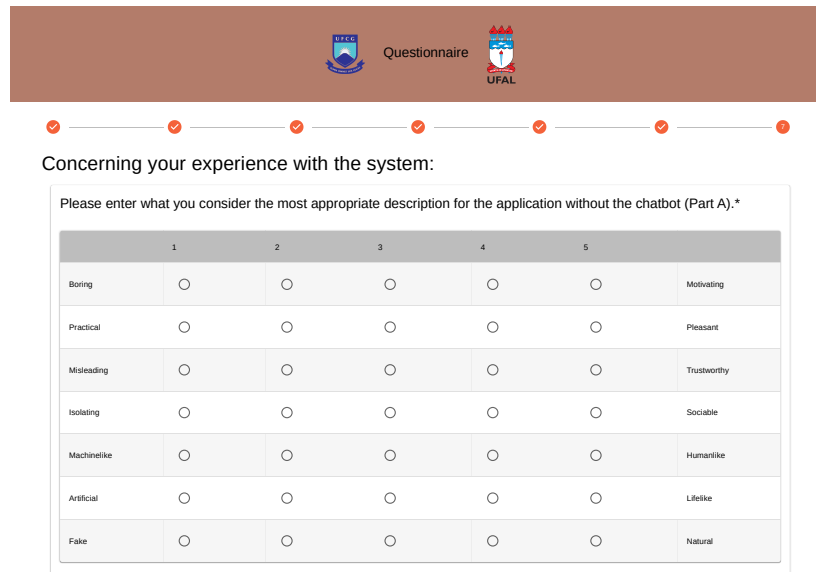
Q1: The recommender provided good recommendations.\*

Part	Strongly agree	Agree	Neutral	Disagree	Strongly disagree
Part A	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Part B	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Q1: You may skip this. But any comments explaining your marks for Part A or B above would help us much.

Figure 3.7: Likert question.

Questions that are answered on a scale between antonymous adjectives were also used based on the methodologies of Hassenzahl et al. [25] and Ho & MacDorman [26] to evaluate the overall trust and experience of the user (Figure 3.8). In addition, for all experiments, we collect logs of tool usage, which record each user iteration and the timestamp of each action taken.



Concerning your experience with the system:

Please enter what you consider the most appropriate description for the application without the chatbot (Part A).\*

	1	2	3	4	5	
Boring	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Motivating
Practical	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Pleasant
Misleading	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Trustworthy
Isolating	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Sociable
Machinelike	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Humanlike
Artificial	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Lifelike
Fake	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Natural

Figure 3.8: Adjectives question.

### 3.3.5 Analysis

Our goal is to assess user satisfaction and analyze metrics such as the total time spent on the experiment and the effectiveness of the chatbot's use, as indicated by the collected logs. For this we measure the average time spent on the platform with the chatbot and just the simple GUI, calculating the 95% confidence interval using bootstrap [17] for each scenario.

We also conducted an analysis comparing questionnaire responses difference for Likert questions and antonymous adjective scales ( $Part_B - Part_A$ ), incorporating a 95% confidence interval calculated using bootstrap. Specifically, a positive value signifies superior chatbot performance, whereas negative values indicate better performance by the simple GUI. A zero value denotes no discernible difference between the two scenarios. By analyzing these metrics, we can gain valuable insights into the effectiveness of the experiment and identify areas for improvement to enhance the user experience and increase overall satisfaction.

## 3.4 **OpenScience Practices**

To ensure the reproducibility of this work, all data and tools necessary to carry out the experiment are available at our public repository <sup>2</sup>.

---

<sup>2</sup><https://github.com/arbezerra/digai>

# Chapter 4

## Literature Review

The material in this chapter has been published as an article in the International Journal of Fashion Design, Technology and Education by Taylor & Francis [37]. In the subsequent sections, we will present the findings from this literature review.

### 4.1 Categorization

Chatbot studies can be primarily divided into two categories: computational and non-computational aspects. Computational aspects pertain to areas like Computer Science or Information Technologies, including the use of NLP. Non-computational aspects encompass all other areas, such as the study of consumer acceptance. This categorization, initially presented in works by Jurafsky and Martin [29], and Diederich, Brendel, and Kolbe [16], is further elaborated upon in this study, as depicted in Figure 4.1.

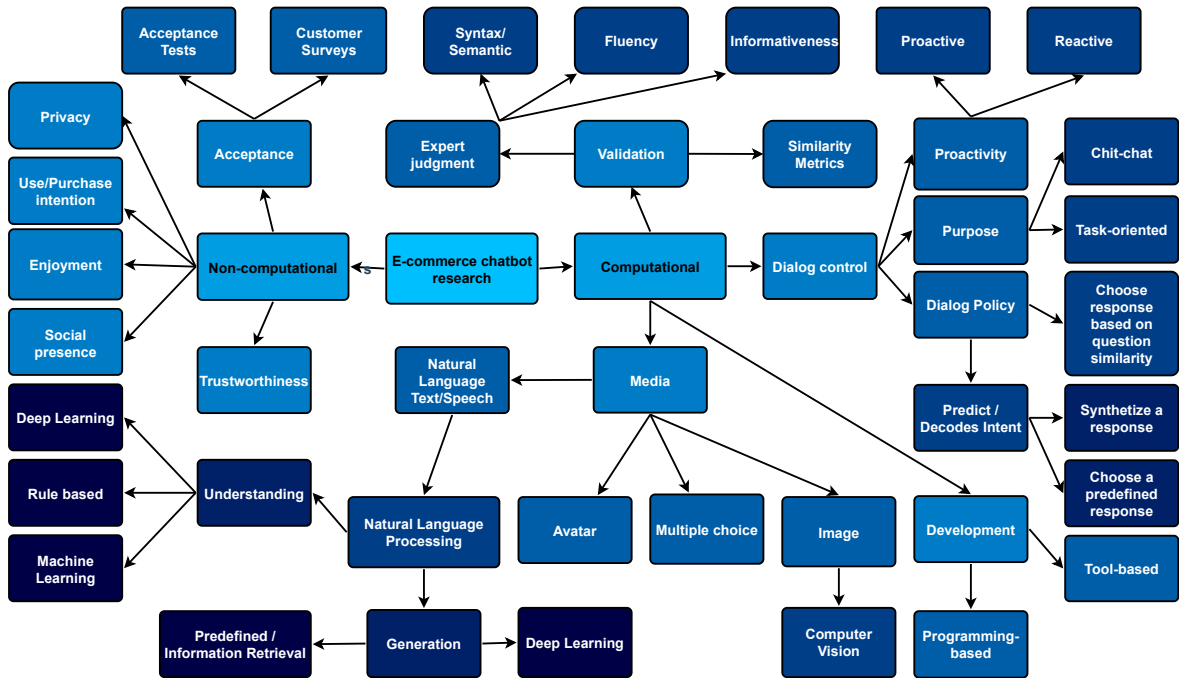


Figure 4.1: Categorization of e-commerce chatbot research.

Computational papers have different categorizations. Here we focus on the high-level categories (Table 4.1).

Chatbots, when used as conversational recommender systems, offer a more advanced approach with a broader range of interactions that enhance preference elicitation and user engagement through NL. They can capture contextual information, a feature intended by context-aware recommender systems. In the realm of fashion e-commerce, chatbots may require additional features related to online sales, such as engaging and persuading customers to purchase a product, potentially involving negotiation dialogues akin to a salesperson interacting with a customer [30]. These systems tackle the primary challenges of natural language interaction, including NLU, information extraction, and NLG, which can be tailored specifically for the fashion industry.

Table 4.1: Computational main categories

Category	Description
Domain	Chatbots may contribute to the consumer journey in different e-commerce domains. The literature focuses on the retail industry – e.g. electronics [55], makeup [46], or clothing [48].
Development	Chatbot tools are available as open-source or paid products - some may not provide flexibility to implement novel functionalities; others support implementation.
Chatbot language	Language considered in the study – e.g. US English language; NLP algorithms perform differently according to each language peculiarities, such as their morphological structure [43].
Media of input/output	Most literature considers raw text, but speech is also present. Typically, speech is automatically converted to text and viceversa. Avatars are also proposed as a visual humanised interface alternative. More natural interactions between customers and chatbots is detailed in Section 4.3.
Dialog control	Encompasses aspects of proactivity regarding task-orientation (objective control) to help users complete a specific task; and non-task-oriented (subjective), also known as ‘chit-chat’, when the chatbot presents skills to seamlessly talk with humans in a natural and informal manner [71].
Natural Language Processing	NLP methodologies for understanding and generation of text or speech. Algorithms are herein categorized as rule-based, classical ML-based, and DL
Validation aspects related to experimentation	Experimentation to assess quantitative and qualitative indicators to evaluate automatically generated sentences (see Section 3.2.1).



After applying the proposed methodology, 76 out of 5959 papers were selected: 46 describing research on chatbot computational aspects; and 30 being surveys about the state of the art or the user experience using chatbots. Such increasing interest may be explained by the advent of DL and the subsequent interest to comprehend non-computational aspects of chatbots. Papers were classified according to the categorisations proposed and the results, together with their references, appear in Tables X1 and X2 in an external public repository <sup>1</sup>.

## 4.2 Findings

Most research on chatbot computational aspects had English as their primary language (76.3%), followed by papers on Indonesian chatbots (6.8%) and other languages like Chinese and Bangla. However, the resulting papers were mostly not fashion-specific (87.7%). Contrastingly, a few papers like Liao et al. [40] and Vaccaro, Agarwalla, Shivakumar, and Kumar [69] were fashion-specific.

It is also worth mentioning that, while non-computational research mainly employed a diversity of ready-to-use chatbot tools like Amazon Alexa, computational papers usually focus on chatbot development using a specific programming language.

The majority of recent studies (79.4%) use natural language text as the primary input method. However, exceptions include the works of Aarathi [1], Pricilla, Lestari, and Dharma [59], and Wintersberger, Klotz, and Riener [70], which provide users with a predefined list of options in certain scenarios. A minority of studies (15.1%) are voice-based, while others utilize avatars to boost user confidence and system perception [18].

Dialog turns in most systems are reactive, with the exception of those in Aarathi [1] and Liu et al. [41], which can control the dialog flow. The dialog control options in Table 4.1 have been thoroughly explored, with neural-based controls emerging as a popular research direction. Most of these systems are designed to perform specific tasks, such as answering product-related questions or facilitating purchases. While the majority of papers focus on retail tasks, a few address 'chit-chat' chatbots, an area that could be further explored in fashion-specific chatbot research.

The dialog policies investigated are primarily based on variations of DL architectures

---

<sup>1</sup><http://doi.org/10.6084/m9.figshare.14519700>

like RNN and CNN. Interest in DL began to surge in 2017, but traditional ML approaches and rule-based methods continue to be used in cutting-edge research.

Algorithms for generating or selecting predefined chatbot responses mostly employ rule-based approaches, where responses are chosen from a predefined set. However, neural-based methods using neural architectures such as RNN, CNN, and GAN have been proposed. Specifically, Information Retrieval (IR) and RNN methods are prevalent, with over 93.1% of the selected papers addressing them.

To measure the effectiveness of chatbots' sentence generation, the literature primarily uses two approaches: similarity metrics to compare a chatbot's response to a human's, and/or expert evaluation. To assess the quality of a chatbot-generated sentence, Nie, Wang, Hong, Wang, and Tian [49] and Gao et al. [20] used the BLEU Score [54], a metric for evaluating similarity in automatic language translations based on word position. Chen et al. [8] proposed using the Distinct-1/2 score [38], which is based on the ratios of distinct unigrams and bigrams.

Nevertheless, similarity metrics do not account for syntax and semantics. Thus, another approach to chatbot validation involves linguistic experts evaluating the meaning of the chatbot's sentences. In the works of Nie et al. [49], Chen et al. [8], and Qiu et al. [60], experts assessed whether a response fits the question context in terms of fluency, relevance, informativeness, semantic consistency, and syntax precision. Nazir et al. [48] followed Turing's Loebner principle [44], where experts interact with the chatbot for ten minutes to evaluate it based on context orientation, dynamism, and grammar structure.

Multimodal chatbots, which integrate visual media with textual chatbot interfaces, have been proposed. Liao et al. [40] proposed a multimodal domain knowledge enriched fashion chatbot that understands product image semantics, modifies attributes during back-end retrieval, offers matching suggestions, and generates multimodal responses. De Carolis, De Gemmis, and Lops [14; 15] use multimodality to consider extra-rational consumer factors such as attitudes, emotions, likes, and dislikes, to provide refined recommendations accordingly.

Virtual assistants have the potential to emulate the role of a salesperson [47]. This concept has been explored by Sapna et al. [63], who conducted research on chatbots and ML techniques to recreate the experience of offline shopping.

Another area for future research is the incorporation of fashion-specific knowledge into chatbots, which is currently a significant gap in the field. The ontology-driven chatbot model developed by Nazir et al. [48], which is based on the semantic web, provides a glimpse into the potential outcomes of such research.

Non-computational studies spanned 20 categories, encompassing aspects such as consumer trust, acceptance, satisfaction, and experience. It was observed that chatbot acceptance has been evaluated using models like the Technology Acceptance Model (TAM) and/or the Use & Gratification Theory (U&G), with a primary focus on perceived usefulness, ease of use, and enjoyment [31; 61]. In addition to enjoyment and utilitarian factors [61], chatbot acceptance also hinges on the social presence typically found in human-human interactions [7]. Factors such as privacy concerns and demographic variations also influence chatbot acceptance [61]. The review highlighted anthropomorphism [21] and privacy [9] as two factors impacting chatbot trustworthiness.

To address the lack of training datasets specific to the fashion domain, Sapna et al. [63] integrated a RS with a Fashion-Knowledgeable Component (FKC) to create a chatbot named 'Athena'. Athena's RS utilizes the product inventory of the e-commerce site, while its FKC gathers fashion information from social media, model photographs, and stylist-curated fashion items. Athena's recommendations are based on inventory-style associations, making them impersonal. However, the fashion database needs to be regularly updated to align with the inventory.

A trend was identified towards using chatbots as value co-creators during purchases, assuming roles specific to fashion, such as a salesperson or stylist. For instance, in-person fashion stylists were used to derive chatbot requirements that aid in building trust and managing uncertainty in e-commerce [69]. Rese et al. [61] also discovered gender differences in fashion e-commerce consumers, indicating a higher usage of chatbots among females.

### 4.3 Discussion

The main research gaps and needs the literature review unearthed and that serve to base suggestions of chatbot research directions in general and for the fashion domain in particular are presented here along the computational and non-computational perspectives.

### 4.3.1 Computational perspective

Research for innovatively applying chatbots to e-commerce settings in general will move forward in a three-pronged way:

1. Evolution of DL approaches for sales assistance: DL approaches have emerged and become the trending research direction for chatbot tasks related to NLP;
2. Availability of databases to enhance DL training: DL algorithms strongly rely on huge datasets for training. Our literature review points to scarcity of public datasets for chatbots training. Alternatively, Li, Li, and Ji [39] used a dataset of conversations between customers and customer service; Chen et al. [8], a dataset with user reviews from e-commerce; and Yu et al. [72], a generic question pair dataset. The publication of such datasets, for several domains not just fashion, is a promising research direction, including how to deal with data privacy; and,
3. Investigation of audio-enabled chatbots in Natural User Interfaces (NUIs): using audio cues to evaluate user experience of NUIs is still limited. Prajwal et al. [58] and Palma, Seeger, and Heinzl [50] consider voice-only chatbots; and De Carolis et al. [14], Eisman et al. [18] and Tan and Liew [67] besides voice, have an avatar. Most of these studies convert speech to text for the chatbot, except for De Carolis et al. [14; 15] that uses audio input directly to enable recognition of voice intonation and prosody to better determine users' intent by reducing loss of information on cues due to voice-text conversion.

#### **Fashion-domain specific computational research opportunities**

Four major opportunities may be highlighted:

1. More encompassing, real-life, professional-grade datasets on fashion items are needed if fashion chatbots are to be more extensively trained and trusted: given that no large public dataset is readily available, Nazir et al. [48] alternatively collected data manually from clothing brands websites and from Facebook posts and comments. Others like Liao et al. [40] proposed to transfer knowledge from richer domains.

2. Applying chatbots to fashion e-commerce specific needs such as multimedia conversations (e.g. text, voice and images): Pantano, Passavanti, Priporas, and Verteramo [53] also revealed a lack of innovative technologies in the fashion luxury industry. De Carolis et al. [14] recommended new clothes based on users' visual cues, Sapna et al. [63] make the chatbot ask users their preferences and Liao et al. [40] proposed a multimodal chatbot to gather user's visual and textual clothing needs. Future work could explore chatbot's retrieval of users' short- and long-term preferences to better recommend fashion products.
3. Integration of chatbots to other fashion applications in different points within the consumer journey: since chatbots can provide personalised information for consumers across their journey [69], research that explores such integration, e.g. virtually trying a recommended item on, shows potential.
4. Another opportunity is the integration with Augmented Reality (AR) – e.g. in Virtual Fitting Room (VFR) applications. Moriuchi, Landers, Colton, and Hair [46] compare e-commerce apps that use chatbots and those which use AR, but they do not propose integration of both to serve the fashion domain.

### 4.3.2 Non-computational perspective

Future research on non-computational aspects of chatbots for e-commerce that can be applied in the fashion domain appears more promising in the following 4 areas:

1. Chatbot acceptance and design for different demographics: Rese et al. [61] found that females had a negative attitude towards chatbots due to the technology immaturity and privacy issues. Models that measure consumer acceptance should be addressed (e.g. TAM and U&G, mentioned earlier) in future research for chatbot contexts, particularly when considering differences in Womenswear and Menswear categories.
2. Consumer autonomy and identity in chatbot consumer experience: consumer autonomy [3] is related to the perceived sense of control that consumers have over the interaction with chatbots and it can be attached to motivational factors (e.g. Self-Determination Theory). Future studies might address the role that consumer autonomy and identity

play in consumer trust and acceptance, for example, by measuring chatbots' design approaches that can trigger these states.

3. Chatbot design, consumer trust and privacy: since consumers might see chatbots with negative eyes if there are privacy concerns, consumer trust is another factor that can be explored further. Aspects such as transparent advice [69] and problem-solving [10], could be investigated in future research, addressing the role that design plays in this context and if other factors (e.g. social elements, cultural values, self-identity) influence consumers' trust on chatbots.
4. Perceived enjoyment, usability, and usefulness: perceived enjoyment and the utilitarian nature of the interaction (e.g. whether useful or not) also influence consumer acceptance. User experience and usability were also factors highlighted by this review. Further studies that analyse the influence of these elements in chatbot design for e-commerce are needed. Other aspects that might be added into this area is the influence of playfulness and gamification in perceived enjoyment of chatbots.

#### **Fashion-domain specific non-computational research opportunities**

Promising fashion chatbot research that is essentially non-computational and that can be inferred from the review's results:

1. Culture and gender-aware chatbots: fashion retail differs across cultures and individual fit (womenswear vs. menswear), chatbots need to be tailored accordingly. Inclusive studies such as cultural and gender-related research in this field are in order.
2. Multiuser chatting: Merrilees and Miller [45] observed that traditional shopping with a companion influences the consumer experience. Alone consumers tend to be more price sensitive. Future studies may explore the way fashion consumers seek for advice from chatbots that could be experimented with by adjusting social factors (e.g. including a friend in the conversation), evaluating the impact of these factors on user acceptance levels, and bring said factors into the dialog when indicators (rejected number of recommendations, say) cross thresholds.

3. Designbots: Fashion chatbots may be made to behave as a fashion designer, providing a platform to support co-creation of value [11]. Conversational platforms can provide insights for brands to recognise consumer value [12], which means that future research in this area can also enhance the consumer experience.
4. Persuasion: e-commerce chatbots should be able to function as a persuasive salesperson [30]. Building persuasion capabilities into fashion chatbots, given the scarcity of real-life training datasets, will require research to elicit expertise and tactics from fashion offline stylists and sales experts.
5. Replicating other offline fashion experiences: Research that mines insights from consumer-stylists' conversations, similarly to the work by Vaccaro et al. [69], could be also useful to other parts of the fashion supply chain - e.g. design and marketing of new fashion collections.

## 4.4 Literature Review Findings & This Work

In addressing RQ 1, it becomes evident that advanced deep learning techniques are significantly enhancing the capabilities of chatbots in understanding and adeptly responding to user queries. The integration of recommendation systems facilitates personalized product recommendations, thereby elevating the overall online shopping experience. Additionally, the integration with various fashion applications broadens the scope of chatbots, offering users a more comprehensive service. These technological advancements play a pivotal role in mitigating the challenges of information overload by delivering tailored and pertinent information, thereby further refining the online shopping journey.

In the literature review, two chatbots, proposed by Catapang et al [5]. and by Khan [33], were found. These chatbots, similar to DigAI, employ SVM for discerning the user intent. The ensuing response selection is then governed by predefined rules, influencing the nature of their interactions based on identified user intents.

However, there is still potential for improvement, particularly in analyzing perceived enjoyment, usability, and usefulness. This study aims to address the multifaceted challenge of consumer acceptance in the context of chatbot design for e-commerce.

# Chapter 5

## Results for DigAI

In this chapter, we present a detailed analysis of the impact observed when integrating the DigAI chatbot into the existing recommendation system. The focus is on objectively assessing the effects on the recommendation system's performance and functionality. As anticipated in subsection 3.3.1, should the recommender system used in the experiment have filtering features, the results might differ from those reported in this chapter. The use of a recommender system with filters for items in the catalog could be the focus of future research.

### 5.1 Chatbot Design

The implementation of the DigAI chatbot took place in the Python programming language, primarily driven by the selection of Tensorflow as the preferred training library for artificial intelligence models. Based on the classification framework introduced in Section 4.1, DigAI can be categorized as:



Table 5.1: DigAI Categorization.

Category	Value
Domain	Fashion
Implemented / Tool	Implemented
Language	English
Media of input/output	Text
Free Text / Buttons	Free Text
Turns	Hybrid
Dialog Policy	Identify/predict the user's intent, then choose a response in a finite predefined set of responses
Purpose	Task / Mediator
Input understanding/NLU	SVM
Answers selection/NLG	Information Retrieve

### 5.1.1 Entity Extraction

In light of the relatively limited set of categories present in our catalog, a strategic decision was made to implement the entity extractor employing a regular expression approach. This choice stems from a pragmatic consideration of the specific context in which the system operates, where a more intricate or resource-intensive method may not be warranted.

The regular expression rules were crafted to account for a broad spectrum of linguistic expressions, covering both the specific categories and values delineated in Table 5.2, as well as their analogous forms and synonymous representations. This approach aimed to facilitate comprehensive data validation and processing, ensuring the effective recognition of diverse

expressions pertaining to the information in the table. The incorporation of similar forms and synonyms within the regex rules sought to enhance the system's adaptability and precision in handling varied inputs and linguistic nuances.

Table 5.2: Entity categories.

<b>Color</b>	<b>Fabric</b>	<b>Pattern</b>	<b>Size</b>	<b>Shape</b>
Black	Denim	Animal Print	Mini	Pleated
Blue	Knitted	Geometric	Midi	Straight
Brown	Laced	Camouflage	Maxi	Asymmetric
Beige	Glossy (Leather)	Checked		
Gray	General	Floral		
Green	Velvet	Paisley		
Orange		Plain		
Pink		Polka Dot		
Purple		Striped		
Red		Tie Dye		
White				
Yellow				

### 5.1.2 User Intent Classification

The training was carried out using the selected dataset in order to evaluate the proposed intent classifier. The results, as depicted in Fig. 5.1, demonstrate that BERT outperformed all other models, including LSTM, BiLSTM, and SVC. BERT demonstrated the highest

F1 score in intent classification, which confirms its superiority over other state-of-the-art models. Surprisingly, the F1 difference between the top-performing BERT and the worst-performing SVC was less than 0.03.

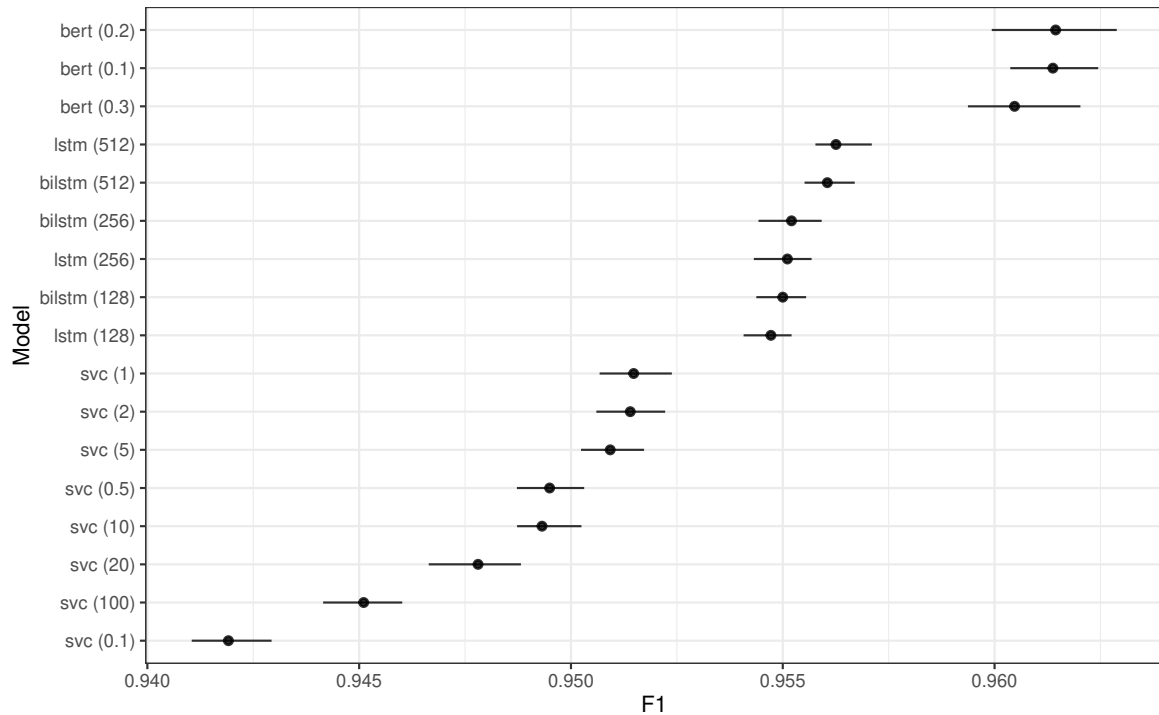


Figure 5.1: F1 (95% CI) by Model.

As shown in Fig. 5.2 the performance achieved by LSTM and BiLSTM, regardless of hyperparameter variation, had similar F1 scores for intent classification, which means that both models have similar levels of precision and recall for the task at hand. This result suggests that both models were able to effectively capture the patterns and relationships in the data and make accurate predictions.

These results reveal a stable behavior among all evaluated classifiers. Three conclusions may be drawn. First, the user intent prediction in dialogs with chatbots in a fashion e-commerce context is feasible. Second, it is relatively straightforward for all evaluated techniques. Third, techniques are not sensible to hyperparameter tuning.

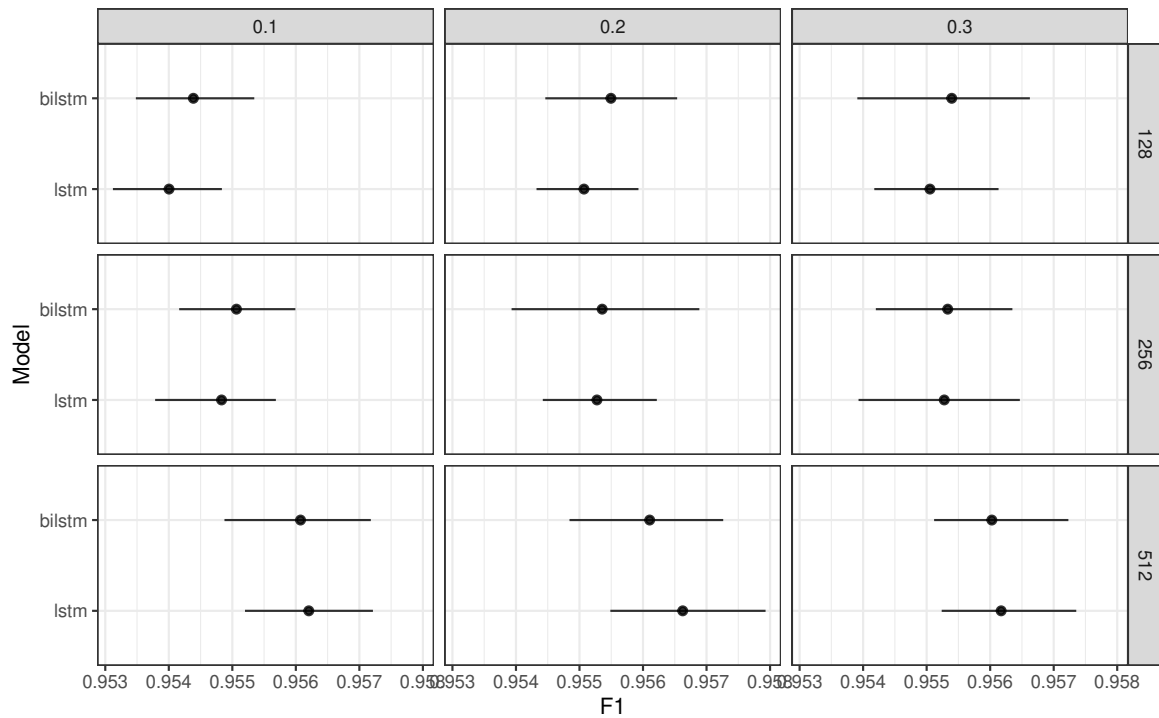


Figure 5.2: F1 (95% CI) of LSTM and BiLSTM in a dropout x units grid.

The findings depicted in Fig. 5.3 underscore the anticipated longer training time of BERT compared to the experimented classifiers, namely BiLSTM and LSTM neural networks. This observation aligns seamlessly with the inherent complexity reflected in the number of trainable weights associated with these neural networks. Following BERT, both BiLSTM and LSTM exhibited relatively shorter training times. Notably, SVC classifiers demonstrated remarkable efficiency by training nearly in real-time, despite yielding the lowest overall F1 score.

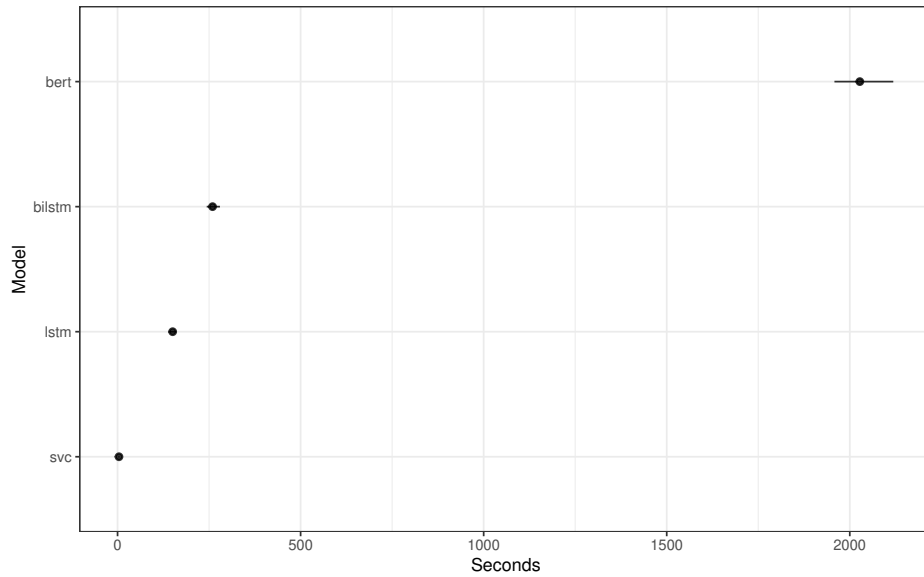


Figure 5.3: Training time by Model.

This trade-off between training speed and predictive performance is further substantiated by Fig. 5.4, which portrays a consistent pattern in prediction throughput. Surprisingly, although the F1 score of the SVC model is only marginally inferior to the best performing BERT model, we contend that the SVC approach is the most pragmatic choice for user intent classification in fashion e-commerce chatbots, considering its admirable balance between training efficiency and predictive accuracy. Consequently, we have opted for SVC as the intent classification model for DigAI.

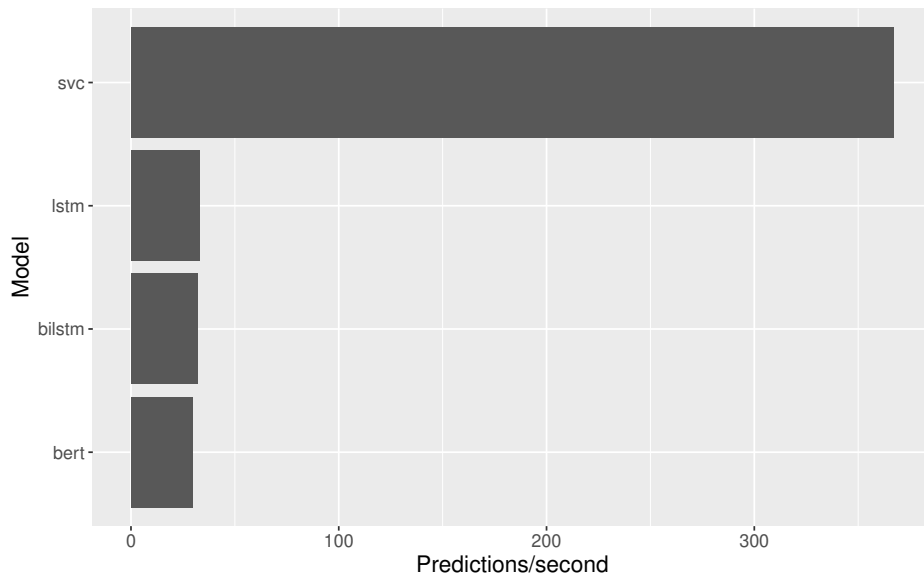


Figure 5.4: Prediction throughput by Model.

### 5.1.3 Dialog Control

The dialog control module receives user intent and extracted entities as input, determining subsequent chatbot actions based on a predefined flowchart outlined in Section 3.2.2. It performs four primary types of actions: recommending clothing items, accessing the recommendation system API, and updating user preferences; asking for clarification, if it has a low confidence level in understanding the user's intention; request the user's preference, using the entropy level of each category in the catalog; and responding to user queries in which it triggers the answer generator.

### 5.1.4 Response Generator

The implementation of the response generator involves utilizing a set of predefined templates, which are selected based on factors such as the user's intention, filled slots, and the level of confidence in interpreting the user's intention.

## 5.2 Experiment

Overall, the two distinct scenarios offer valuable insights into the role of both recommendation systems and chatbots in facilitating a more personalized and engaging shopping experience for customers.

The average experiment duration in Part B, featuring a recommendation system with a chatbot interface, is less than 150 seconds, while in Part A, focused solely on a recommendation system, it slightly exceeds 300 seconds (Figure 5.5). The selection (or purchase decision) of a clothing item (skirt) using DigAI is accomplished within a timeframe comparable to the minimum duration observed with the simple GUI (Scenario A). Lower volatility during this time is preferable, as purchasing decisions are made more quickly with the chatbot.

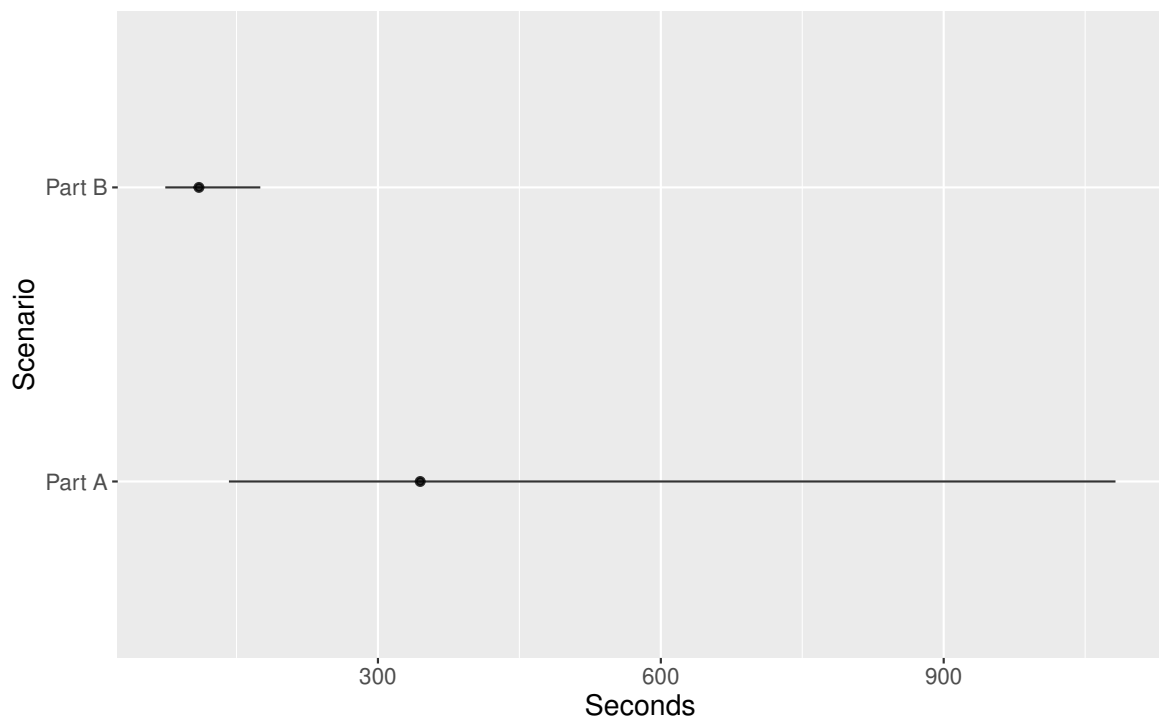


Figure 5.5: Experiment time (95% CI).

The reduced time observed in Part B may stem from the chatbot interface's potential to streamline the interaction process, or it could suggest that the chatbot interface enables quicker and more efficient utilization of the recommendation system. Further empirical studies are necessary to substantiate these initial observations and to gain a more comprehensive

understanding of the intricate dynamics between user interaction patterns and the integration of recommendation systems with chatbot interfaces. Additional research efforts would contribute to a more objective evaluation of the factors influencing experiment duration and user engagement in these two setups.

Figure 5.6 offers a comparative analysis between a recommendation system with DigAI's interface and a standalone recommendation system - i.e., a RS with a GUI as illustrated in Figure 3.5. Noteworthy distinctions were observed in key dimensions: Trust (T), Satisfaction (S), Quality (Q), Effectiveness (E) and Control (C). Users consistently reported higher levels of trust, satisfaction, perceived quality, and a sense of control when utilizing the chatbot interface, indicating its outperformance in these areas.

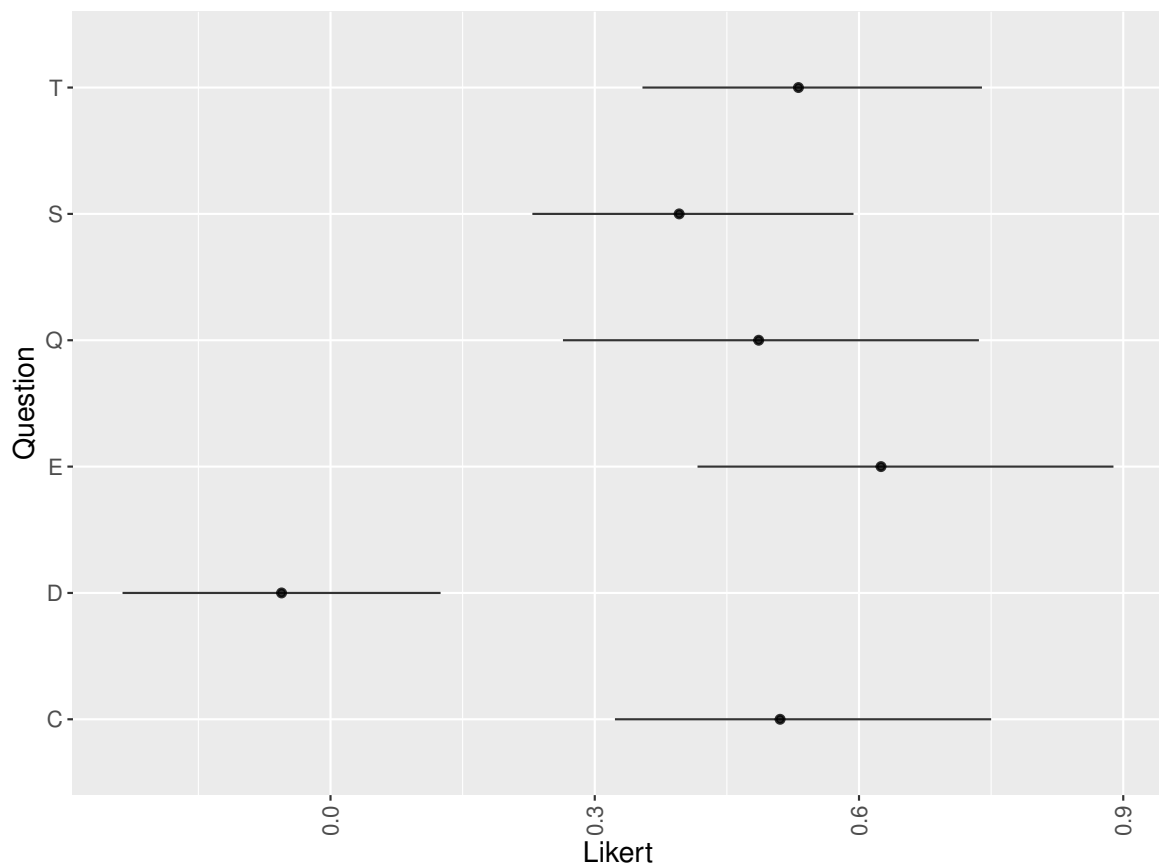


Figure 5.6: Likert difference Part B - Part A (95% CI).

Conversely, both systems exhibited comparable performance in terms of Diversity (D), suggesting that the addition of a chatbot interface did not significantly influence the diversity of recommendations generated. This finding implies that users experienced a similar range



and variety of suggestions, regardless of the presence of the chatbot interface. This was expected, as the chatbot effectively refined the catalog by filtering it to the preferences of the user, thereby resulting in a more focused selection of options. Nevertheless, a cohort of users voiced their discontent, asserting that certain clothing items they liked were no longer taken into account in the RS.

To delve into a solution, the exploration of RQ 3b sheds light on a potential avenue for improvement. One plausible enhancement lies in the development of a segregated user model that not only discerns short- and long-term preferences. By incorporating such user model, the RS could offer a more comprehensive and personalized experience, addressing the concerns raised by users who found their favored clothing items overlooked.

Survey participants emphasize the superior performance of DigAI in the realm of "Trust". In scenario A, participants opt to make decisions independently, suggesting a potential lack of confidence in their choices or a desire for psychological reassurance through the additional "opinion" provided by DigAI. Exploring this aspect in future research could provide valuable insights into how participants navigate decision-making processes, offering a more objective understanding of the role trust and external validation play in shaping their choices.

The chatbot interface demonstrated strengths in attributes like naturalness, pleasantness, and motivation, as shown in Figure 5.7. However, it's important to note that while averages for lifelikeness, human-likeness, sociability, and trustworthiness were positive, the lower limit of the confidence interval fell below 0. Consequently, it cannot be definitively concluded that the chatbot outperformed the recommendation system. The lower confidence interval indicates a level of uncertainty in user perceptions, emphasizing the need for cautious interpretation of the results. Further analysis and user feedback may offer additional insights into the comparative effectiveness of the chatbot interface and the recommendation system.

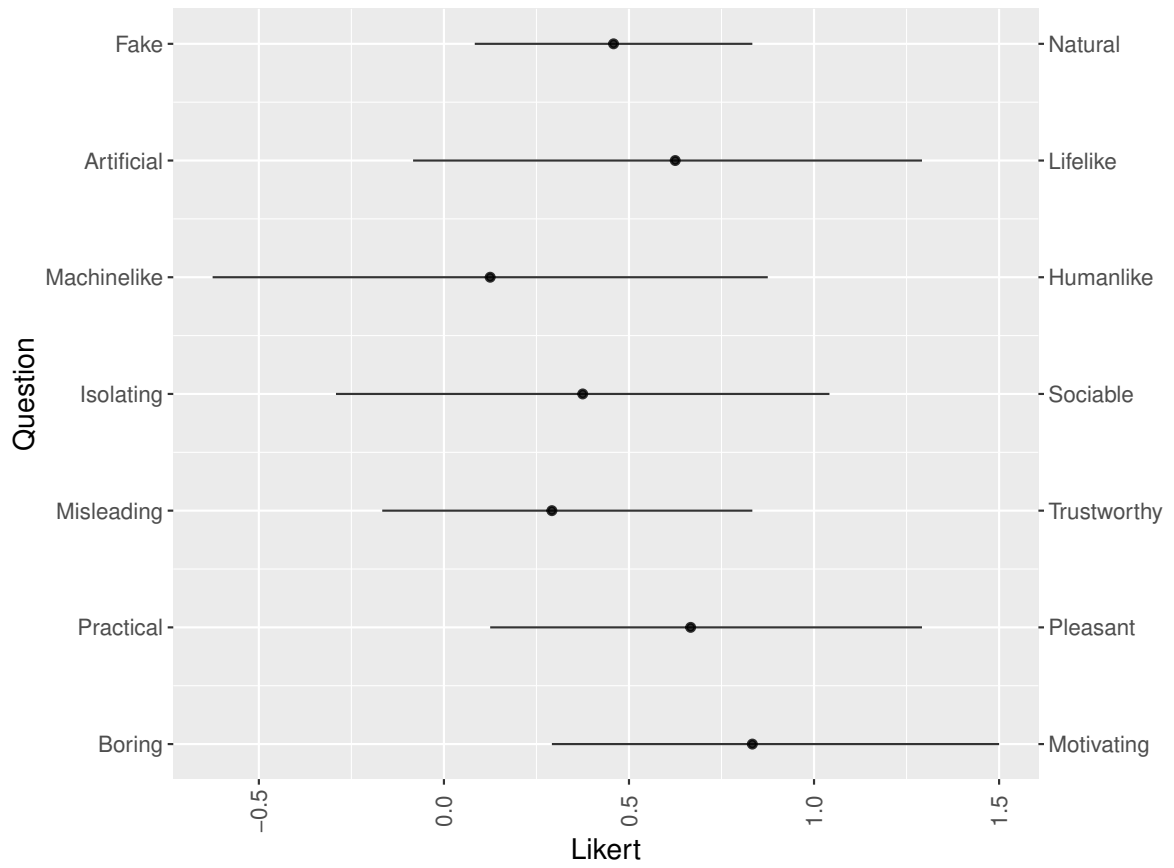


Figure 5.7: Difference Part B - Part A (95% CI).

# Chapter 6

## Conclusion and Future Work

### 6.1 Conclusion

In summary, this research addressed the challenges posed by information overload in the fashion industry, brought about by the integration of information technology and digital platforms. The proliferation of fashion catalogs and the diverse array of product combinations and customer preferences has led to an increased emphasis on effective communication and personalized user experiences.

To tackle these challenges, many fashion businesses have turned to Dialog Systems as a solution, enabling users to interact with platforms and obtain product-related information. However, the complexity of human language, particularly in task-oriented and context-limited scenarios, presents a notable hurdle to the optimal performance and acceptance of these systems.

This study focuses on designing and developing a chatbot as an interface for a recommendation system aimed at assisting users in finding clothing. The evaluation of the chatbot's performance, usability, hedonic and pragmatic values, by potential users in Brazil assessed their overall satisfaction with the chatbot's effectiveness in providing personalized recommendations as compared to a simpler GUI. This objective assessment will guide future refinements to enhance the chatbot's capabilities.

The primary contribution of this research lies in the application of a comprehensive three-phase methodology, which aimed to address various aspects of recommendation chatbots within the fashion industry. The initial phase involved an extensive systematic review of the

literature, meticulously examining the current state of the art in implementing recommendation chatbots. This foundational step allowed us to gain insights into existing methodologies, challenges, and advancements in the field.

Moving on to the second phase, we focused on the development of the DigAI chatbot. Our emphasis was on utilizing an intention classification model that not only demonstrated a high level of performance but also maintained a robust performance-accuracy relationship. The objective was to create a chatbot that not only recommended fashion items effectively but also responded to user queries in a manner that significantly improved their acceptance and trust in the system.

In the final phase of our methodology, we conducted a comprehensive survey to gauge user experiences with the DigAI chatbot. This evaluation was carried out in two contexts: firstly, the users' interaction with the RS in isolation through a simple GUI, and secondly, their experience with the chatbot. Consequently, we provide answers to the RQs as outlined below:

- RQ 1: Advanced deep learning techniques are notably enhancing chatbot capabilities in understanding and responding effectively to user queries. The integration of recommendation systems facilitates personalized product suggestions, improving the overall online shopping experience. Additionally, integrating with various fashion applications broadens chatbots' scope, offering users a more comprehensive service. These technological advancements contribute significantly to mitigating challenges related to information overload by delivering tailored and relevant information, thereby refining the online shopping journey.
- RQ 2: The research findings reveal that BERT, a state-of-the-art language model, achieved the highest F1 score among the considered Machine Learning algorithms. However, it is noteworthy that SVC, while exhibiting a slightly lower F1 score, presents a compelling alternative due to its significantly reduced training time and faster prediction speed. The marginal difference in F1 scores is outweighed by the practical advantages of SVC in terms of computational efficiency.
- RQ 3a: The developed chatbot exhibits higher user-reported levels of trust, satisfaction, perceived quality, effectiveness and a sense of control compared to a simpler

GUI. Users' preference for the chatbot in decision-making scenarios aligns with its perceived psychological reassurance and potential superiority.

- RQ 3b: Upon analyzing the questionnaire, we observed a discrepancy between the preferences identified by the RS and those directly communicated to the chatbot. The conflict arises from the chatbot filtering items that were previously liked, creating a need for enhancement. A potential solution involves establishing an independent user model that considers preferences communicated through both channels, effectively addressing this conflict.

In contributing to the ongoing efforts to improve the online shopping experience, this research aligns with the broader industry goal of improving user satisfaction and promoting business growth. The insights derived from the study aim to contribute objectively to the discourse on leveraging technology in the fashion sector to meet evolving consumer needs.

## 6.2 Threats to Validity

Firstly, the effectiveness of the chatbot in understanding user intent may be influenced by the diversity of user language and expressions, potentially leading to biases in the system's performance. For example, consider a scenario where a fashion e-commerce chatbot is trained predominantly on data from a specific demographic group, such as young adults in a particular region. If the chatbot is then tested with users from a different age group or cultural background, the effectiveness of the chatbot in understanding user intent may be compromised. The diverse language and expressions used by users from varied demographics could introduce biases in the system's performance, impacting its ability to accurately interpret and respond to a broader range of user inputs. This could be mitigated by continuous monitoring and iterative improvement mechanisms are implemented, allowing for the identification of biases and subsequent refinement of the model based on real user feedback.

Secondly, the usability testing conducted with a specific group of potential users may not fully represent the broader range of customers, raising concerns about the generalizability of the findings. To address this concern, we plan to enhance the external validity of our study by conducting a replication of the experiment. This time, we will involve a diverse

group of participants, specifically undergraduate and graduate fashion students in the UK through our R&D collaboration with the University of Southampton, Winchester School of Arts. This approach aims to ensure a more comprehensive representation of potential users, thus addressing the potential limitations associated with a narrow focus on a specific group.

Additionally, the dynamic nature of fashion trends and user preferences poses a challenge, as the chatbot's recommendations may become outdated over time. For instance, the chatbot might suggest certain clothing items or styles that were popular during its training period but have since fallen out of fashion. Users relying on the chatbot may end up with outdated fashion advice, leading to dissatisfaction and a lack of trust in the chatbot's recommendations. To alleviate this issue, regularly feeding the model with the latest data on current fashion trends, user preferences, and market dynamics helps keep the recommendations up-to-date and reflective of the ever-changing landscape. Additionally, incorporating real-time feedback loops from users allows the chatbot to adapt swiftly to emerging trends and refine its suggestions based on immediate user reactions.

Furthermore, the study's focus on user satisfaction and effectiveness may not capture other important aspects of user experience, such as sustainability and privacy concerns or ethical considerations related to data handling.

### **6.3 Future Work**

In future work, it is crucial to address the identified threats to validity to enhance the robustness and generalizability of the study's findings. Firstly, to mitigate the potential bias introduced by diverse user language and expressions, incorporating natural language processing techniques that continuously adapt to evolving linguistic patterns could be explored. This adaptive approach may enhance the chatbot's ability to comprehend a broader range of user intents, thereby improving overall performance.

Secondly, expanding the scope of usability testing to include a more diverse and representative sample of potential users can help ensure that the chatbot's effectiveness and user satisfaction are evaluated across different demographic groups. Employing user segmentation based on factors such as age, cultural background, and online shopping habits can provide valuable insights into the system's performance variations among distinct user

profiles.

To address the dynamic nature of fashion trends and user preferences, implementing a mechanism for real-time updates and continuous learning within the recommendation system can help keep the chatbot's suggestions current and aligned with the rapidly changing fashion landscape.

Furthermore, future research should extend beyond user satisfaction and effectiveness metrics to encompass a more comprehensive evaluation of user experience. This includes investigating potential privacy concerns and ethical considerations related to data handling, ensuring that the implementation of Dialog Systems aligns with user expectations and industry regulations.

Moreover, we can explore enhancing the chatbot's responsiveness by experimenting with its reactivity. This involves elevating its proactive capabilities, enabling it to initiate conversations with users and provide tailored recommendations derived from insights gathered from user data.

As one last consideration, it would be interesting to evaluate how DigAI would fare against a recommender system that filters items in the fashion catalog according to users' specifications.

By systematically addressing these considerations in future research endeavors, we can not only enhance the validity and reliability of the study's outcomes but also contribute to the ongoing refinement and improvement of Dialog Systems in the dynamic context of the fashion industry.

# Bibliography

- [1] N Ganitha Aarthi, G Keerthana, A Pavithra, and K Pavithra. Chatbot for retail shop evaluation. *International Journal of Computer Science and Mobile Computing*, 9(3):69–77, 2020.
- [2] G. Adomavicius and A. Tuzhilin. Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions. *TKDE*, 17(6):734–749, June 2005.
- [3] Nisreen Ameen, Sameer Hosany, and Ali Tarhini. Consumer interaction with cutting-edge technologies: Implications for future research. *Computers in Human Behavior*, 120:106761, 2021.
- [4] Jesús Bobadilla, Fernando Ortega, Antonio Hernando, and Jesús Bernal. A collaborative filtering approach to mitigate the new user cold start problem. *KBS*, 26:225–238, February 2012.
- [5] Jasper Kyle Catapang, Geoffrey A Solano, and Nathaniel Oco. A bilingual chatbot using support vector classifier on an automatic corpus engine dataset. In *2020 International Conference on Artificial Intelligence in Information and Communication (ICAIIIC)*, pages 187–192. IEEE, 2020.
- [6] Tapsi Chadha. 50+ vital chatbot statistics for 2021 to know post pandemic, Jan 2023.
- [7] Ana Paula Chaves and Marco Aurelio Gerosa. How should my chatbot interact? a survey on social characteristics in human–chatbot interaction design. *International Journal of Human–Computer Interaction*, 37(8):729–758, 2021.



- 
- [8] Shiqian Chen, Chenliang Li, Feng Ji, Wei Zhou, and Haiqing Chen. Driven answer generation for product-related questions in e-commerce. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, pages 411–419, 2019.
- [9] Yang Cheng and Hua Jiang. How do ai-driven chatbots impact user experience? examining gratifications, perceived privacy risk, satisfaction, loyalty, and continued use. *Journal of Broadcasting & Electronic Media*, 64(4):592–614, 2020.
- [10] Minjee Chung, Eunju Ko, Heerim Joung, and Sang Jin Kim. Chatbot e-service and customer satisfaction regarding luxury brands. *Journal of Business Research*, 117:587–595, 2020.
- [11] Chiara Colombi, Pielah Kim, and Nioka Wyatt. Fashion retailing “tech-gagement”: engagement fueled by new technology. *Research Journal of Textile and Apparel*, 22(4):390–406, 2018.
- [12] Jonathan Copulsky. Do conversational platforms represent the next big digital marketing opportunity? *Applied Marketing Analytics*, 4(4):311–316, 2019.
- [13] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, sep 1995.
- [14] Berardina De Carolis, Marco de Gemmis, and Pasquale Lops. A multimodal framework for recognizing emotional feedback in conversational recommender systems. In *Proceedings of the 3rd Workshop on Emotions and Personality in Personalized Systems 2015*, pages 11–18, 2015.
- [15] Berardina De Carolis, Marco de Gemmis, Pasquale Lops, and Giuseppe Palestra. Recognizing users feedback from non-verbal communicative acts in conversational recommender systems. *Pattern Recognition Letters*, 99:87–95, 2017.
- [16] Stephan Diederich, Alfred Benedikt Brendel, and Lutz M Kolbe. On conversational agents in information systems research: analyzing the past to guide future work. In *Proceedings of Internationale Tagung Wirtschaftsinformatik*, Siegen, Germany, 2019.
- [17] Bradley Efron. Bootstrap methods: another look at the jackknife. In *Breakthroughs in statistics: Methodology and distribution*, pages 569–593. Springer, 1992.

- 
- [18] Eduardo M Eisman, María Navarro, and Juan Luis Castro. A multi-agent conversational system with heterogeneous data sources access. *Expert Systems with Applications*, 53:172–191, 2016.
- [19] Asbjørn Følstad and Petter Bae Brandtzaeg. Users’ experiences with chatbots: findings from a questionnaire study. *Quality and User Experience*, 5(1):3, 2020.
- [20] Shen Gao, Zhaochun Ren, Yihong Zhao, Dongyan Zhao, Dawei Yin, and Rui Yan. Product-aware answer generation in e-commerce question-answering. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, pages 429–437, 2019.
- [21] Eun Go and S Shyam Sundar. Humanizing chatbots: The effects of visual, identity and conversational cues on humanness perceptions. *Computers in Human Behavior*, 97:304–316, 2019.
- [22] Ken Goldberg, Theresa Roeder, Dhruv Gupta, and Chris Perkins. Eigentaste: A constant time collaborative filtering algorithm. *JIR*, 4(2):133–151, Jul 2001.
- [23] A. Graves and J. Schmidhuber. Framewise phoneme classification with bidirectional lstm networks. In *Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005.*, volume 4, pages 2047–2052 vol. 4, 2005.
- [24] Xintong Han, Zuxuan Wu, Yu-Gang Jiang, and Larry S Davis. Learning fashion compatibility with bidirectional lstms. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1078–1086, 2017.
- [25] Marc Hassenzahl, Michael Burmester, and Franz Koller. Attrakdiff: Ein fragebogen zur messung wahrgenommener hedonischer und pragmatischer qualität. *Mensch & Computer 2003: Interaktion in Bewegung*, pages 187–196, 2003.
- [26] Chin-Chang Ho and Karl F MacDorman. Revisiting the uncanny valley theory: Developing and validating an alternative to the godspeed indices. *Computers in Human Behavior*, 26(6):1508–1518, 2010.
- [27] Sepp Hochreiter and Jürgen Schmidhuber. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780, 11 1997.

- [28] Sang-Young Jo, Sun-Hye Jang, Hee-Eun Cho, and Jin-Woo Jeong. Scenery-based fashion recommendation with cross-domain generative adversarial networks. In *2019 IEEE International Conference on Big Data and Smart Computing (BigComp)*, pages 1–4. IEEE, 2019.
- [29] Dan Jurafsky and James H Martin. *Speech and language processing* (3rd (draft) ed.), 2019.
- [30] Shaidah Jusoh. Intelligent conversational agent for online sales. In *2018 10th International Conference on Electronics, Computers and Artificial Intelligence (ECAI)*, pages 1–4. IEEE, 2018.
- [31] Dharun Lingam Kasilingam. Understanding the attitude and intention to use smartphone chatbots for shopping. *Technology in Society*, 62:101280, 2020.
- [32] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186, 2019.
- [33] Mohammad Monirujjaman Khan. Development of an e-commerce sales chatbot. In *2020 IEEE 17th International Conference on Smart Communities: Improving Quality of Life Using ICT, IoT and AI (HONET)*, pages 173–176. IEEE, 2020.
- [34] Satwik Kottur and Seungwhan Moon. Overview of situated and interactive multimodal conversations (simmc) 2.1 track at dstc 11. In *Proceedings of The Eleventh Dialog System Technology Challenge*, pages 235–241, 2023.
- [35] Satwik Kottur, Seungwhan Moon, Alborz Geramifard, and Babak Damavandi. Overview of situated and interactive multimodal conversations (simmc) 2.0 track at dstc 10. *DSTC10 challenge workshop at AACL*, 2021.
- [36] Satwik Kottur, Seungwhan Moon, Alborz Geramifard, and Babak Damavandi. SIMMC 2.0: A task-oriented dialog dataset for immersive multimodal conversations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4903–4912, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

- [37] ARDB Landim, AM Pereira, Thales Vieira, E de B. Costa, JAB Moura, V Wanick, and Eirini Bazaki. Chatbot design approaches for fashion e-commerce: an interdisciplinary review. *International Journal of Fashion Design, Technology and Education*, 15(2):200–210, 2022.
- [38] Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. A diversity-promoting objective function for neural conversation models. *arXiv preprint arXiv:1510.03055*, 2015.
- [39] Lu Li, Chenliang Li, and Donghong Ji. Deep context modeling for multi-turn response selection in dialogue systems. *Information Processing & Management*, 58(1):102415, 2021.
- [40] Lizi Liao, You Zhou, Yunshan Ma, Richang Hong, and Tat-seng Chua. Knowledge-aware multimodal fashion chatbot. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 1265–1266, 2018.
- [41] Che Liu, Junfeng Jiang, Chao Xiong, Yi Yang, and Jieping Ye. Towards building an intelligent chatbot for customer service: Learning to respond at the appropriate time. In *Proceedings of the 26th ACM SIGKDD international conference on Knowledge Discovery & Data Mining*, pages 3377–3385, 2020.
- [42] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1096–1104, 2016.
- [43] Fiammetta Marulli, Marco Pota, and Massimo Esposito. A comparison of character and word embeddings in bidirectional lstms for pos tagging in italian. In *Intelligent Interactive Multimedia Systems and Services: Proceedings of 2018 Conference 11*, pages 14–23. Springer, 2019.
- [44] Michael L Mauldin. Chatterbots, tinymuds, and the turing test: Entering the loebner prize competition. In *AAAI*, volume 94, pages 16–21, 1994.

- [45] Bill Merrilees and Dale Miller. Companion shopping: The influence on mall brand experiences. *Marketing Intelligence & Planning*, 37(4):465–478, 2019.
- [46] Emi Moriuchi, V Myles Landers, Deborah Colton, and Neil Hair. Engagement with chatbots versus augmented reality interactive technology in e-commerce. *Journal of Strategic Marketing*, 29(5):375–389, 2021.
- [47] Elena Morotti, Lorenzo Donatiello, and Gustavo Marfia. Fostering fashion retail experiences through virtual reality and voice assistants. In *2020 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*, pages 338–342. IEEE, 2020.
- [48] Aisha Nazir, Muhammad Yaseen Khan, Tafseer Ahmed, Syed Imran Jami, and Shaukat Wasi. A novel approach for ontology-driven information retrieving chatbot for fashion brands. *Int. J. Adv. Comput. Sci. Appl. IJACSA*, 10(9), 2019.
- [49] Liqiang Nie, Wenjie Wang, Richang Hong, Meng Wang, and Qi Tian. Multimodal dialog system: Generating responses via adaptive decoders. In *Proceedings of the 27th ACM international conference on multimedia*, pages 1098–1106, 2019.
- [50] Maria del Carmen Ocón Palma, Anna-Maria Seeger, and Armin Heinzl. Mitigating information overload in e-commerce interactions with conversational agents. In *Information Systems and Neuroscience: NeuroIS Retreat 2019*, pages 221–228. Springer, 2020.
- [51] Chitu Okoli. A guide to conducting a standalone systematic literature review. *Communications of the Association for Information Systems*, 37, 2015.
- [52] Shweta Pandey and Deepak Chawla. Evolving segments of online clothing buyers: an emerging market study. *Journal of Advances in Management Research*, 15(4):536–557, 2018.
- [53] Eleonora Pantano and Gabriele Pizzi. Forecasting artificial intelligence on online customer assistance: Evidence from chatbot patents analysis. *Journal of Retailing and Consumer Services*, 55:102096, 2020.

- [54] Kishore Papineni, Salim Roukos, Todd Ward, and WBLEU Zhu. A method for automatic evaluation of machine translation”. *the Proceedings of ACL-2002, ACL, Philadelphia, PA, July 2002*, 2001.
- [55] Anirudha Paul, Asiful Haque Latif, Foysal Amin Adnan, and Rashedur M Rahman. Focused domain contextual ai chatbot framework for resource poor languages. *Journal of Information and Telecommunication*, 3(2):248–269, 2019.
- [56] Artur M Pereira, J Antao B Moura, Evandro De B Costa, Thales Vieira, Andre RDB Landim, Eirini Bazaki, and Vanissa Wanick. Customer models for artificial intelligence-based decision support in fashion online retail supply chains. *Decision Support Systems*, 158:113795, 2022.
- [57] Artur Maia Pereira, Thales Vieira, and Evandro de Barros Costa. Balancing exploration and exploitation: An image-based approach to item retrieval with enhanced diversity. *Computers & Electrical Engineering*, 84:106605, 2020.
- [58] SV Prajwal, G Mamatha, P Ravi, D Manoj, and Shri Krishna Joisa. Universal semantic web assistant based on sequence to sequence model and natural language understanding. In *2019 9th International Conference on Advances in Computing and Communication (ICACC)*, pages 110–115. IEEE, 2019.
- [59] Catherine Pricilla, Dessi Puji Lestari, and Dody Dharma. Designing interaction for chatbot-based conversational commerce with user-centered design. In *2018 5th International Conference on Advanced Informatics: Concept Theory and Applications (ICAICTA)*, pages 244–249. IEEE, 2018.
- [60] Minghui Qiu, Feng-Lin Li, Siyu Wang, Xing Gao, Yan Chen, Weipeng Zhao, Haiqing Chen, Jun Huang, and Wei Chu. Alime chat: A sequence to sequence and rerank based chatbot engine. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 498–503, 2017.
- [61] Alexandra Rese, Lena Ganster, and Daniel Baier. Chatbots in retailers’ customer communication: How to measure their acceptance? *Journal of Retailing and Consumer Services*, 56:102176, 2020.

- [62] Paul Resnick and Hal R. Varian. Recommender systems. *Communications of the ACM*, 40(3):56–58, March 1997.
- [63] Sapna, Ria Chakraborty, Kartikeya Vats, Khyati Baradia, Tanveer Khan, Sandipan Sarkar, and Sujoy Roychowdhury. Recommendation and fashion sense: Online fashion advisor for offline experience. In *Proceedings of the ACM India Joint International Conference on Data Science and Management of Data*, pages 256–259, 2019.
- [64] Jetze Schuurmans and Flavius Frasincar. Intent classification for dialogue utterances. *IEEE Intelligent Systems*, 35(1):82–88, 2020.
- [65] Ayat Shukairy. Chatbots in customer service – statistics and trends [infographic], May 2022.
- [66] Prissadang Suta, Xi Lan, Biting Wu, Pornchai Mongkolnam, and JH Chan. An overview of machine learning in chatbots. *International Journal of Mechanical Engineering and Robotics Research*, 9(4):502–510, 2020.
- [67] Su-Mae Tan and Tze Wei Liew. Designing embodied virtual agents as product specialists in a multi-product category e-commerce: The roles of source credibility and social presence. *International Journal of Human–Computer Interaction*, 36(12):1136–1149, 2020.
- [68] Chun-Hua Tsai and Peter Brusilovsky. The effects of controllability and explainability in a social recommender system. *User Modeling and User-Adapted Interaction*, 31:591–627, 2021.
- [69] Kristen Vaccaro, Tanvi Agarwalla, Sunaya Shivakumar, and Ranjitha Kumar. Designing the future of personal fashion. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–11, 2018.
- [70] Philipp Wintersberger, Tobias Klotz, and Andreas Riener. Tell me more: Transparency and time-fillers to optimize chatbots’ waiting time experience. In *Proceedings of the 11th Nordic Conference on Human-Computer Interaction: Shaping Experiences, Shaping Society*, pages 1–6, 2020.

- 
- [71] Rui Yan. "chitty-chitty-chat bot": Deep learning for conversational ai. In *IJCAI*, volume 18, pages 5520–5526, 2018.
- [72] Jianfei Yu, Minghui Qiu, Jing Jiang, Jun Huang, Shuangyong Song, Wei Chu, and Haiqing Chen. Modelling domain relationships for transfer learning on retrieval-based question answering systems in e-commerce. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, pages 682–690, 2018.



# Appendix A

## Questionnaire

After each question there is an open question for additional comments.

### **A.1 Concerning the quality or accuracy of the recommendations: (Likert)**

- Q1 :The recommender provided good recommendations.
- Q2 :I liked the recommendations provided by the system.
- Q3 :The recommended skirts fitted my preference.

### **A.2 Concerning the diversity or variety of the recommendations: (Likert)**

- D1: The recommender helped me discover new skirts.
- D2: The skirts that were recommended to me are diverse.
- D3: The list of recommendations included skirts of many different types.

### **A.3 Concerning the control you had on the flow of the recommendations: (Likert)**

- C1: I became familiar with the system very quickly.
- C2: The layout of the recommendations on the screen was adequate
- C3: The recommender allowed me to inform my preference easily
- C4: The recommender helped me decide on subsequent options faster than I would looking at a catalog of skirts

### **A.4 Concerning the effectiveness of the recommendations: (Likert)**

- E1: Using the system is a pleasant experience.
- E2: I made better choices with the recommender.
- E3: I found better items using the recommender.

### **A.5 Concerning the trust you had in the system: (Likert)**

- T1: The recommendations the system made were convincing.
- T2: The recommender made me more confident about my final selection/decision
- T3: I am confident I will like other fashion items the system recommends me in the future
- T4: The recommender can be trusted.

## **A.6 Concerning your overall satisfaction with the system: (Likert)**

- S1: I will use this recommender again.
- S2: I am likely to recommend my friends use fashion e-commerce sites with more efficient recommendation tools.
- S3: Overall, I am satisfied with the recommender.
- S4: The recommender helped me find a skirt I really liked.

## **A.7 Concerning your experience with the system: (adjectives)**

- A1: Please enter what you consider the most appropriate description for the application without the chatbot (Part A).
  - Boring - Motivating
  - Practical - Pleasant
  - Misleading - Trustworthy
  - Isolating - Sociable
  - Machinelike - Humanlike
  - Artificial - Lifelike
  - Fake - Natural
- A2: Please enter what you consider the most appropriate description for the application with the chatbot (Part B).