applied sciences

MDPI

Article

Optimizing Chatbots to Improve Customer Experience and Satisfaction: Research on Personalization, Empathy, and Feedback Analysis

Shimon Uzan, David Freud and Amir Elalouf *

Department of Management, Bar-Ilan University, Ramat Gan 5290002, Israel
* Correspondence: amir.elalouf@biu.ac.il

Featured Application

This research can significantly enhance customer service quality and efficiency by improving chatbot responsiveness and empathy, which benefits businesses employing AI-driven customer support.

Abstract

This study addresses the ongoing challenge of optimizing chatbot interactions to significantly enhance customer experience and satisfaction through personalized, empathetic responses. Using advanced NLP tools and strong statistical methodologies, we developed and evaluated a multi-layered analytical framework to accurately identify user intents, assess customer feedback, and generate emotionally intelligent interactions. With over 270,000 customer chatbot interaction records in our dataset, we employed spaCy-based NER and clustering algorithms (HDBSCAN and K-Means) to categorize customer queries precisely. Text classification was performed using random forest, logistic regression, and SVM, achieving near-perfect accuracy. Sentiment analysis was conducted using VADER, Naive Bayes, and TextBlob, complemented by semantic analysis via LDA. Statistical tests, including Chi-square, Kruskal–Wallis, Dunn's test, ANOVA, and logistic regression, confirmed the significant impact of tailored, empathetic response strategies on customer satisfaction. Correlation analysis indicated that traditional measures like sentiment polarity and text length insufficiently capture customer satisfaction nuances. The results underscore the critical role of context-specific adjustments and emotional responsiveness, paving the way for future research into chatbot personalization and customer-centric system optimization.

Keywords: chatbots; generative AI; natural language processing; customer experience; personalization; empathy; intent recognition; sentiment analysis; feedback analysis

check for updates

Academic Editor: Christos Bouras

Received: 30 April 2025
Revised: 15 August 2025
Accepted: 25 August 2025
Published: 28 August 2025

Citation: Uzan, S.; Freud, D.; Elalouf, A. Optimizing Chatbots to Improve Customer Experience and Satisfaction: Research on Personalization, Empathy, and Feedback Analysis. *Appl. Sci.* **2025**, *15*, 9439. https://doi.org/10.3390/app15179439

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Review

With the growing reliance on digital customer service solutions across industries, ensuring satisfactory and emotionally meaningful customer interactions is becoming critical for business success.

In the digital era, customer service is rapidly transforming due to advances in artificial intelligence and NLP technologies [1,2]. Chatbots have become central to customer interactions yet face significant challenges in accurately detecting user intent and delivering empathetic, personalized responses [3–5]. For a formal definition of chatbots, see the Literature Review section (Section 1). While previous studies have primarily focused on

improving operational efficiency, a significant gap remains between system capabilities and the need for human-like, empathetic interactions, with some researchers even highlighting potential customer dissatisfaction arising from a lack of emotional engagement [6,7].

A systematic examination of the current literature reveals several gaps in our understanding of artificial intelligence, specifically generative AI systems, in customer service. While previous studies have addressed several operational issues, there is still a gap in understanding more subtle details such as customer intent prediction, feedback implementation, and emotional participation.

Our study addresses these gaps by investigating optimization strategies for chatbots and generative AI systems within the natural language processing framework. The purpose is to make it possible for chatbots to recognize utterances of customers' needs, process feedback for further refinement, and interact in a more human-like, emotionally meaningful way. The specific research questions and hypotheses guiding this investigation are presented in the theoretical framework section (Section 2.3).

Addressing these research objectives is essential for firms that use generative AI-powered chatbots, as it will lead to the development of smarter, more responsive systems and effective customer relationship management solutions. Our research offers a new layer of understanding from what already exists by using quantitative and qualitative data, focusing on personalization, emotional intelligence, and feedback implementation.

This study aims to bridge that gap by employing an integrated approach that combines advanced NLP techniques with comprehensive statistical analyses, enhancing the detection of customer intent and the system's ability to provide tailored, empathetic responses [8,9]. Specifically, we utilize a spaCy-based named entity recognition model with clustering algorithms (HDBSCAN and K-Means) for precise intent detection. Additionally, machine learning models such as random forest [10], logistic regression, and SVM [11] are applied for text classification, complemented by sentiment analysis tools (VADER, Naive Bayes [12], TextBlob) and deep semantic analysis via latent Dirichlet allocation (LDA) [4]. Advanced statistical methods, including Spearman correlation, logistic regression, PCA, and ANOVA, further evaluate the impact of various factors on the user experience [13,14].

Our findings demonstrate that integrating advanced NLP and empathy-driven personalization enhances chatbot performance and customer satisfaction. This approach is designed to deepen our understanding of customer-system interactions and contribute to developing chatbots capable of delivering high-quality, human-like, and emotionally intelligent customer service [5,15]. Ultimately, these enhancements optimize chatbot interactions and offer businesses actionable insights into fostering long-term customer relationships and loyalty.

In the digital era, artificial intelligence (AI) is projected to replace specific job roles, particularly those related to text-based conversational agents (chatbots). As noted by [1] and [16], chatbots are valuable in customer service as they provide significant advantages to the users and the firms. Chatbots can have conversations and provide immediate and efficient answers to customer questions [3,17]. A chatbot may be described as "a computerized conversational system that interacts with human users using natural language" [18] or "an artificial construct designed to converse with people by processing natural language as input and output" [19]. Companies employ virtual agents as service providers so consumers can entertain and educate them simultaneously while meeting their needs [20–22]. Table 1 presents a summary of key studies on chatbots.

Table 1. Summary of key studies on chatbots.

Article	Study Focus	Theoretical Lens	Key Findings
[8]	Factors influencing user satisfaction with chatbot-based customer service	An integrated framework combining the expectation-confirmation model, information systems success model, and technology acceptance model	Information and service quality positively influence user satisfaction, which predicts continuance intention. Perceived enjoyment, usefulness, and ease of use contribute to continuance intention.
[6]	Impact of anthropomorphic chatbot design on customer responses	Integration of anthropomorphism theory, expectancy violation theory, and emotion theories	For angry customers, human-like cues lead to inflated expectations, which, when unmet, result in lower satisfaction, poorer company evaluation, and reduced purchase intentions.
[4]	Development of an emotionally appropriate and topically relevant chatbot response generation model	Encoder–decoder framework with emotion-aware module and topic commonsense-aware module	The proposed model significantly improves emotion expression precision and topic relevance, as demonstrated through automatic metrics and human evaluations.
[9]	Factors driving consumer switching from human customer service to AI-based conversational agents in banking	Push–pull–mooring framework from migration theory and switching behavior literature	Both push effects (low empathy and adaptability of human agents) and pull effects (anytime/anywhere connectivity, personalization of AI) significantly influence switching behavior.
[7]	Consumer acceptance and intention to reuse chatbots in the airline industry	Extended technology acceptance models with SERVQUAL framework and trust considerations	Reliability and perceived usefulness are the most critical determinants of reuse intention; tangible elements enhance perceived ease of use, while empathy has no significant effect on acceptance.
[23]	Managing the human–chatbot divide: how service scripts influence service experience	Theater metaphor, social impact theory, and social exchange theory	When employing an educational script, human service agents significantly positively affect satisfaction and purchase intention compared to chatbots. These effects are fully mediated by emotion and rapport. However, when an entertaining script is used, there are no differences between human agents and chatbots.
[5]	Effect of aligning chatbot personality with consumer personality	Similarity–attraction theory and personality congruence principles	Matching the chatbot's personality to the consumer's (e.g., introversion/extraversion) results in higher engagement and improved sales outcomes, particularly in social gain contexts.
[24]	Human–AI interaction and digitally mediated empathy during COVID-19 pandemic	Communication theory of resilience and empathy theories	Five types of digitally mediated empathy identified among Replika users: companion buddy, responsive diary, emotion-handling program, electronic pet, and tool for venting. Mediated empathy serves as pathway to psychological resilience during pandemic disruption.

Table 1. *Cont.*

Article	Study Focus	Theoretical Lens	Key Findings
[25]	Empathetic chatbot development models: A systematic literature review	Dialogue system architecture analysis and empathetic characteristic frameworks	Analysis of 204 works reveals distribution of empathetic chatbot models: 60% generative models, 27% retrieval models, and 13% hybrid models. Generative models demonstrate superior performance in delivering rich empathetic utterances.
[26]	Effect of AI chatbot empathy and identity disclosure on donation behavior	CASA hypothesis and uncanny valley theory	Interaction effects between chatbot empathy types (cognitive vs. affective) and identity disclosure (human vs. chatbot names) on willingness to donate, mediated through perceived humanness and social presence. Demonstrates importance of careful empathy implementation to avoid uncanny valley effects.

Human-like interfaces enhance user trust in technology by boosting perceived competence [15,27] and demonstrating greater resilience against breaches of trust [28]. In some cases of deficient service, less direct involvement from people due to the use of technological barriers can result in less customer dissatisfaction and appraisals of the service that are less than favorable [29,30]. Furthermore, customer service is channeling towards using AI-powered conversational agents (CAs), considering that AI-powered chatbots can relay information to consumers quicker than human operators in traditional call centers. AI has been implemented to automate routine and simpler tasks that involve supporting and enhancing decision-making or problem-solving in many fields, such as healthcare [31] or education, public administration, industry, communications, and business [32]. Notably, one of the fastest-growing trends in AI today is the adoption of conversational applications that mimic human dialogue, commonly referred to as chatbots, conversational agents, or simply bots [32].

Literature review adopts a systematic approach to ensure that the selection of relevant articles is thorough and objective. Priority was given to studies that directly address the impact of artificial intelligence, particularly generative AI systems and natural language processing, on customer experience and satisfaction [4,6,8,9,33]. Although there is a substantial body of research on AI and customer experience, relatively few studies have examined how generative AI systems influence customer satisfaction. This gap forms the foundation of the current research, which seeks to explore user experience and satisfaction in interactions with chatbots and assess the utility of generative AI in diverse customer service settings [6,8].

These days, in "generative AI"-based chatbots, system capabilities have to be in sync with the user's needs. It is reported that user satisfaction and willingness to continue utilizing AI service agents are significantly affected by the quality of information, quality of service, perceived usefulness, ease of use, and perceived enjoyment [8]. Moreover, these studies emphasize the importance of customization and the provision of suitable emotional responses to improve user experience [4,6,8].

Integrating anthropomorphic elements into chatbot design offers new insights into customer interactions by attributing human-like traits to digital agents. Research indicates that anthropomorphism can positively and negatively affect customer emotions, depending on the context and users' expectations [6]. The research of the emotion-aware and

topic-aware modules combined in the TERG model by [4] has enhanced the generation of relevant and emotionally appealing responses. These results highlight the problem of emotion and issue relevance in the response generation of chatbot systems [4,9]. Other researchers have pointed out the human-like features of perceived humanness [34], anthropomorphism [35,36], and social presence [37,38] as offered to construct users' attitudes towards chatbots.

In addition, research evaluating the shift from human to AI conversational agents and vice versa illustrates that these modes of interaction exert different impacts on the trust and preference of the consumer. Considering the increasing need for service processes individually tailored to every customer [39], chatbots provide an easy way to offer personalized experiences at scale. Also, service scripts that allow greater discretion to employees in modifying service delivery enhance the chances of meeting customer expectations [40,41]. Using the push–pull–mooring framework, ref. [9] found some motivators, such as empathic and inflexible human services and ever-present human-to-computer interfaces, to be fundamental in shaping client decisions in customer service situations. Furthermore, ref. [33] proved that in the case of the hospitality industry, people are affected not only by the type of technology used (chatbot or self-service) but also by the service outcome (success or failure) regarding their perceptions and intentions. Such findings accentuate the importance of employing customer features in the design of service experiences [9,33].

1.1. Relevant Theories and Concepts

The current literature on chatbots and user interactions is founded on several established theories, such as the information systems success (ISS) model, the technology acceptance model (TAM), and the need for interaction with a service employee (NFI-SE) [8]. This holistic perspective argues that user participation is mediated by perceived usefulness, ease of use, and personalized experience, which are important in comprehending users' continued adoption of chatbots. Complementing this, ref. [6] investigated the impact of anthropomorphism through the lens of the expectancy violations theory, demonstrating that while human-like cues can raise customer expectations regarding a chatbot's capabilities, failure to meet these expectations may lead to adverse emotional reactions. Collectively, these perspectives highlight the need to consider users' functional and emotional aspects in the design of chatbots.

These claims were built upon in [4], where context-aware emotional response generation was further advanced. Here, the authors presented an encoder–decoder model with latent variables that facilitated topic relevance to make emotional responses more accurate and contextually appropriate. This design demonstrates the importance of a chatbot's understanding and accurately responding to the user's emotions. In another instance, ref. [9] researched why consumers move from relying on human service staff to AI-driven chatbots and employed the push–pull–mooring model. Their research suggests that empathy, constant availability, and personalization are highly impactful predictors of the users' trust and acceptance. Collectively, these contributions provide a rationale for using emotional intelligence and personalized service delivery to enhance user engagement with chatbots.

Researchers have highlighted that trust is a critical determinant in forecasting the success of technologies [42,43]. Recent studies have further explored factors influencing consumer trust [44–48] as well as satisfaction [20,45,49,50] and usage intention [44,45,51]. Trust in service agents has positively influenced consumers' intention to reuse the service [52].

Ref. [7] examined the influence of service quality on customer adoption and usage of chatbots, emphasizing that the success of chatbot implementations relies less on data quality and more on the quality of implementation. Their findings indicate that reliability

and usability impact user preferences more than empathy, except in specific domains such as healthcare. The researchers stress the importance of non-verbal aspects (e.g., aesthetics and utility) in chatbot design. Along with these findings, ref. [5] studied the possibilities of personalized interaction with a chatbot through similarity–attraction theory. Their research is based on the five-factor model of personality and proposes an alignment with the chatbot's user. Users are more likely to engage with the chatbot, which economically benefits the company. Together, these studies highlight the importance of robust service implementation and personalization in shaping positive user experiences with chatbots.

Besides, ref. [13] studied the psychological motivation that defines consumer engagement and purchasing through chatbot interfaces. Their research proposed a model that included communication quality, human–computer interaction (HCI), and human usage and gratification (U&G) in conjunction with self-administered surveys and electroencephalography (EEG). Their results show that the dimensions of trust that impact the intention to reuse and repurchase include reliability, capability, anthropomorphism, social presence, and informativeness. Moreover, neural correlates identified in the dlPFC and superior temporal gyrus further support the role of trust in these interactions. These theoretical and empirical contributions collectively enrich our contemporary understanding of chatbot technology, user experience, and the multidimensional nature of human–chatbot interactions. Chatbot interactions must be seamless, precise, and comprehensive to foster positive perceptions of communication quality; when information is conveyed efficiently and appears professional, consumers are more likely to perceive the chatbot as reliable [20]. Moreover, attributing human traits to computers (perceived anthropomorphism) is essential in computer–human interaction [53]. It can embed social sub-cues into chatbots by integrating human-like features, facial expressions, body language, and/or vocal output [54].

1.2. Methodologies Used

As the literature suggests, the chatbot domain has been examined through many quantitative approaches that systematize the interaction between the technology and its users. In one instance, ref. [8] used SEM to test the user satisfaction and reuse intention of 370 chatbot users against the information quality, service quality, usability, usefulness, and enjoyment, proving how quantitative methods can effectively validate theoretical models. Similarly, ref. [13] applied PLS-SEM with SmartPLS to analyze survey responses, assessing consumer trust and purchasing intention by evaluating composite reliability, convergent and discriminant validity, and using T-tests to compare scenarios with and without chatbot interactions.

Advancements in natural language processing and AI have significantly enhanced modern chatbots' capabilities in both textual and spoken communication [2], prompting companies across various industries to employ chatbots for customer interactions [55]. Unlike human agents who must invest considerable effort and time into getting familiar with service processes, chatbots are not subject to human error and fatigue and operate tirelessly, providing dependable, uniform service at all times [7,16]. Chatbots' high level of responsiveness is an essential attribute that enhances performance markedly [56].

To expand these approaches, ref. [6] performed some experiments on conversation transcripts to detect emotional signs like anger and anthropomorphism using natural language processing (NLP). They conducted manipulative experiments to support causal claims about satisfaction. Also, ref. [7] studied the role of service quality in the customer's decision to adopt and use chatbots. This aspect was further corroborated by [57], who showed through a survey that a large proportion, 71.5%, accept AI-based customer service because of its continuous responsiveness, neutrality, and objectivity. In contrast, 28.5%

oppose it, which indicates that a more incremental and balanced approach to transitioning from human to AI-serviced customer channels is essential.

Ref. [4] leveraged an array of techniques to achieve context-aware emotional response generation, employing a basic encoder–decoder model with GRU, a latent variable within the CVAE framework (with integrated Q-network and a P-network), a topic commonsense-aware (TCA) module using attention mechanisms, an Emotion Supervisor, and a Word Type Selector. This integrated approach enabled their model to generate contextually and emotionally relevant responses, as validated through automated and human evaluations that demonstrated significant enhancements over pre-existing models like Seq2Seq and ECM.

As a complement, ref. [7] studied service quality with the SERVQUAL model (later adapted to E-SERVQUAL for online interactions) to reveal that users generally prioritize efficiency and reliability over empathy, except in contexts such as healthcare, thereby emphasizing the role of non-verbal attributes in chatbot design. At the same time, ref. [23] used a mixed-method technique, which incorporates quantitative and qualitative methods (such as Model 7 of Hayes' Process Macro with 5000 bootstrapped samples), to evaluate how service scripts affect customer experience. Their findings showed that "educational scripts" used by human agents fostered greater emotional connection, satisfaction, and purchase intention than digital agents, suggesting that optimizing the emotional content in chatbot scripts can enhance performance. Together, these studies underscore the importance of context-aware optimization and deliberate design in response generation and service scripting for chatbots.

E-commerce research indicates that consumers prefer chatbots capable of understanding and promptly responding to their needs, fostering positive perceptions of empathetic chatbots [20]. Additionally, natural language processing primarily extracts the semantic meaning of text, as demonstrated in studies by [58], thereby contributing to more accurate and contextually relevant responses.

Recent advances in empathy research have significantly contributed to chatbot development, particularly in understanding digitally mediated empathy and its applications in human–AI interactions. Ref. [24] conducted comprehensive research on human–AI interaction during the COVID-19 pandemic, identifying five distinct types of digitally mediated empathy among Replika chatbot users: companion buddy, responsive diary, emotion-handling program, electronic pet, and tool for venting. Their findings demonstrate that mediated empathy serves as an unexpected pathway to psychological resilience, helping users cope with pandemic-related disruption and enhancing overall well-being.

Contemporary research has further advanced our understanding of empathetic chatbot architectures and their effectiveness. Ref. [25] provided a systematic literature review of empathetic chatbot development, analyzing 204 works and revealing the distribution of current approaches: 60% generative models, 27% retrieval models, and 13% hybrid models. Their analysis demonstrates that generative models achieve superior performance in delivering rich empathetic utterances, though they present greater complexity and accuracy challenges compared to retrieval-based approaches.

Building on these technical foundations, ref. [26] explored the psychological mechanisms underlying chatbot empathy effectiveness, examining the interaction between empathy types (cognitive vs. affective) and identity disclosure (human vs. chatbot names) on user behavior. Their experimental study with 496 US adults revealed significant interaction effects mediated through perceived humanness and social presence, while highlighting the importance of careful empathy implementation to avoid uncanny valley effects that can negatively impact user experience.

2. Modeling of Chatbot Personalization, Empathy, and Feedback, Including Hypotheses

This research presents a structured methodological framework for chatbot optimization, consisting of sequential stages designed to enhance chatbot performance and understand user interactions. The framework adopts a hierarchical structure, beginning with entity recognition and text classification, advancing through sentiment and semantic analysis, and culminating in statistical analysis. This multidimensional approach balances technology- and consumer-centric views of the user experience, enabling strong measures towards enhancing chatbot interactions and outcomes.

2.1. Entity Recognition and Text Classification Model

Our entity recognition and text classification approach is designed to better understand customer–chatbot interactions by employing two complementary techniques: named entity recognition (NER) and text classification. Named entity recognition (NER) constitutes a critical component within the model, as it enables the accurate identification of key information from customer inquiries, such as product names, account characteristics, shipping information, and problem types. This capability is crucial for understanding customer intent and providing personalized responses. In parallel, text classification is conducted using an ensemble approach combining random forest and decision trees, selected for their robustness against overfitting and effective operation in multidimensional spaces. In combination, NER with text classification provides a practical way to derive much deeper and more sophisticated customer insights and understanding of needs and intents.

2.2. Sentiment and Semantic Analysis Model

Our sentiment and semantic analysis section delves into three complementary layers: basic sentiment analysis, complex sentiment analysis, and deep semantic analysis. This multi-layer approach is designed to extract the emotional tone, intentions, and deeper meanings from customer–chatbot interactions, thereby capturing not only the explicit content but also the underlying emotional nuances in the dialogues.

Operational Definition of Empathy in Chatbot Interactions

For the purposes of this study, empathy in chatbot interactions is operationally defined as the system's multifaceted ability to understand and respond appropriately to customer emotional and contextual needs. This definition encompasses four key dimensions that align with our analytical framework.

This definition builds upon recent advances in digitally mediated empathy research, particularly the comprehensive framework developed by [24], who identified multiple facets of empathy in human–AI interactions during crisis situations. Their research on Replika users during the COVID-19 pandemic revealed that empathy in conversational AI encompasses not only emotional recognition and response but also serves as a mechanism for psychological resilience and well-being enhancement.

Drawing from the systematic analysis of empathetic chatbot architectures [25], our operational definition emphasizes the technical implementation of empathetic capabilities through advanced dialogue system models. The predominance of generative models (60% of current approaches) in delivering rich empathetic utterances informs our framework's focus on sophisticated response generation that goes beyond simple sentiment classification.

Furthermore, incorporating insights from [26] regarding the interaction between empathy types and identity disclosure, our definition recognizes that empathetic effectiveness depends not only on the system's technical capabilities but also on the careful balance

between cognitive and affective empathy to avoid uncanny valley effects while maintaining perceived humanness and social presence.

Emotional Recognition Capability: The system's ability to accurately identify and classify customer emotional states through comprehensive sentiment analysis, including both basic sentiment categorization (positive, negative, neutral) and complex emotional profiling across multiple emotional dimensions [4,20].

Contextual Understanding: The capacity to comprehend customer intent and situational context through precise named entity recognition (NER) and text classification, enabling the system to grasp not only what customers are saying but also the underlying needs and concerns driving their communication [58].

Adaptive Response Generation: The ability to tailor communication strategies based on identified emotional states and contextual factors, particularly adapting responses according to service categories (e.g., CONTACT, INVOICE, REFUND) where different emotional sensitivities may be required [14].

Emotional Sensitivity: The system's capability to detect and appropriately handle emotionally charged expressions, customer pain points, and sensitive service situations, contributing to enhanced customer experience and satisfaction [6,7,59].

This operational definition enables empirical measurement of empathetic capabilities through our integrated analytical framework, where sentiment analysis provides emotional recognition metrics, NER and text classification offer contextual understanding measures, and category-specific performance analysis demonstrates adaptive response capabilities. The convergence of these analytical components provides a comprehensive foundation for evaluating and enhancing empathetic chatbot interactions [14,20].

This operational framework aligns with contemporary empathy research in conversational AI, which emphasizes the multifaceted nature of digitally mediated empathy and the importance of sophisticated dialogue architectures in delivering authentic empathetic responses [24,25]. The integration of both cognitive and affective empathy components, as demonstrated by [26], ensures that our framework captures the full spectrum of empathetic capabilities while maintaining awareness of potential uncanny valley effects that can arise from overly human-like implementations.

The practical implementation of this definition through our integrated analytical framework ensures that empathy is not merely a theoretical construct but a measurable and optimizable system capability that contributes to enhanced customer service outcomes and user satisfaction in real-world chatbot applications.

2.3. Research Questions and Hypotheses

2.3.1. Research Questions

Based on the identified gaps in current chatbot optimization research and the proposed methodological framework, this study is guided by the following research questions:

- RQ1: How do advanced NLP techniques (named entity recognition, text classification, sentiment analysis, and semantic analysis) individually and collectively impact customer satisfaction in chatbot interactions?
- RQ2: What mediating mechanisms explain the relationship between NLP technique effectiveness and customer satisfaction outcomes in empathy-driven chatbot systems?
- RQ3: How do contextual factors and service categories moderate the effectiveness of chatbot optimization strategies in different customer interaction scenarios?

These research questions address the critical need for understanding both the direct effects of individual NLP components and their complex interactions within integrated chatbot systems, while considering the contextual factors that influence their effectiveness.

2.3.2. Research Hypotheses

Based on the proposed methodological framework and the critical role of advanced NLP techniques in optimizing chatbot performance, the following hypotheses are formulated to investigate their impact on customer experience and satisfaction:

Primary Effect Hypotheses (Individual NLP Techniques):

- $H1_a$: The accuracy of named entity recognition (NER) in identifying customer intent and complaints is positively associated with improved customer satisfaction.
- $H1_b$: The effectiveness of text classification algorithms (e.g., random forest, logistic regression, SVM) in categorizing customer interactions is positively associated with improved customer satisfaction.
- $H1_c$: Positive sentiment detected through sentiment analysis tools (VADER, naïve Bayes, TextBlob) in chatbot interactions is positively associated with higher customer satisfaction.
- $H1_d$: The depth of semantic understanding, as measured by latent Dirichlet allocation (LDA) in identifying key topics and issues, is positively associated with improved customer satisfaction.

Interaction Effect Hypothesis:

- H2 : The combined effectiveness of multiple NLP techniques (NER, text classification, sentiment analysis, and semantic analysis) has a greater impact on customer satisfaction than individual techniques alone, demonstrating synergistic effects in integrated chatbot systems.

Mediation Hypotheses:

- $H3_a$: Empathy-driven response appropriateness mediates the relationship between NLP technique accuracy and customer satisfaction, where accurate intent recognition enables more empathetic and contextually appropriate responses.
- $H3_b$: Intent recognition precision mediates the relationship between NER effectiveness and overall customer experience, where accurate entity identification leads to better understanding of customer needs and subsequently improved satisfaction.

Moderation Hypotheses:

- $H4_a$: Service category complexity moderates the relationship between NLP effectiveness and customer satisfaction, with stronger positive effects observed in complex service categories (e.g., ACCOUNT, REFUND) compared to simpler categories (e.g., NEWSLETTER, ORDER).
- $H4_b$: Customer emotional state, as detected through sentiment analysis, moderates the relationship between empathetic chatbot responses and satisfaction outcomes, with greater benefits observed for customers expressing negative emotions.

Integration Hypothesis:

- H5 : The integrated analytical framework combining all NLP techniques (NER, text classification, sentiment analysis, and semantic analysis) with empathy-driven personalization leads to optimal customer experience outcomes across different service contexts, surpassing the effectiveness of individual components or partial integrations.

These hypotheses collectively address the research questions by examining individual effects ($H1_a$–$H1_d$), interaction effects (H2), mediating mechanisms ($H3_a$–$H3_b$), moderating factors ($H4_a$–$H4_b$), and overall system integration (H5), providing a comprehensive framework for understanding chatbot optimization in customer service contexts.

3. Method

This section outlines the comprehensive methodology employed to explore customer chatbot interactions and optimize user experience. The research design integrates advanced natural language processing (NLP) techniques with sophisticated statistical analyses to provide deeper insights into customer needs and enhance chatbot support. It details the data source, tools and environment, the research procedure, and the ethical considerations that guided the study. Furthermore, it describes the evaluation metrics and statistical analyses utilized to validate the findings.

3.1. Data Source and Collection

The central dataset for this study is the Bitext Sample Pre-Built Customer Service Evaluation dataset (Bitext Innovations S.L., Las Rozas, Madrid, Spain), obtained from Kaggle (San Francisco, CA, USA). This dataset comprises over 270,000 annotated interactions and is specifically designed for evaluating intent recognition models on natural language understanding (NLU) platforms. It focuses on colloquial language, providing a higher ratio of informal utterances compared to standard datasets, making it highly relevant for real-world chatbot interactions.

The dataset includes the following:

- Twenty-seven unique intents assigned to one of evelven top-level categories (e.g., ACCOUNT, ORDER, PAYMENT, SHIPPING).
- Seven entity/slot types.
- Each utterance is tagged with entities/slots where applicable.
- Additionally, each utterance is enriched with linguistic tags indicating the type of language variation expressed (e.g., COLLOQUIAL, INTERROGATIVE, OFFENSIVE). These tags allow for customization of training datasets to adapt to different user language profiles, ensuring the trained bot can effectively handle diverse linguistic phenomena such as spelling mistakes, run-on words, and punctuation errors.

Each entry in the dataset contains eight fields: 'utterance', 'intent', 'entity_type', 'entity_value', 'start_offset', 'end_offset', 'category', and 'tags'. The dataset is distributed under the Community Data License Agreement—Sharing, Version 1.0 (The Linux Foundation, San Francisco, CA, USA), which permits its use for computational analysis and publication of results, provided proper attribution is maintained.

3.2. Tools and Environment

All computational analyses and model implementations were performed using Python 3.10 within Google Collaboratory (Google LLC, Mountain View, CA, USA) Notebooks. Colab provided a cloud-based environment with access to necessary computational resources and pre-installed libraries. Key Python libraries utilized various analytical tasks, including the following:

- spaCy (Explosion AI, Berlin, Germany): For named entity recognition (NER) and general natural language processing tasks.
- scikit-learn (scikit-learn developers, worldwide): For implementing machine learning algorithms such as random forest, decision trees, support vector machines (SVMs), and logistic regression, used in text classification and sentiment analysis.
- NLTK (Natural Language Toolkit, University of Pennsylvania, Philadelphia, PA, USA): For various text preprocessing steps and potentially for lexicon-based sentiment analysis (e.g., VADER).
- Pandas (NumFOCUS, Austin, TX, USA) and numpy (NumFOCUS, Austin, TX, USA): For data manipulation and numerical operations.

- Matplotlib (Matplotlib Development Team, worldwide) and seaborn (Michael Waskom, New York, NY, USA): For data visualization.
- gensim (RARE Technologies, Prague, Czech Republic): For latent Dirichlet allocation (LDA) implementation in deep semantic analysis.

Specific Jupyter Notebooks used for the analyses include, but are not limited to, those for semantic analysis (topic modeling), named entity recognition, sentiment analysis, and text classification, demonstrating the application of the aforementioned tools.

3.3. Research Procedure

This research followed a structured methodological framework for chatbot optimization, designed to enhance chatbot performance and understand user interactions. The comprehensive methodological approach is illustrated in Figure 1, which presents a detailed flowchart of all analytical phases, from data collection through final optimization and validation. This iterative framework ensures continuous improvement through feedback loops and performance validation, enabling the development of highly effective, empathy-driven chatbot systems.

The methodological framework consists of six primary phases, each containing multiple sub-processes that work synergistically to optimize chatbot performance:

Phase 1: Data Collection and Preprocessing:

The procedure commenced with a data preprocessing phase, involving the collection of customer interaction logs (270,000+ records), followed by rigorous cleaning, standardization, and quality validation to ensure reliable data for subsequent analysis. Data failing predefined standards underwent further refinement through an iterative quality control process.

Phase 2: Language Analysis (Parallel Processing):

The second phase, language analysis, comprised three main components operating in parallel:

- (A) Entity Recognition and Text Classification: Models such as spaCy NER, along with machine learning algorithms including HDBSCAN, K-Means, SVM, logistic regression, and random forest, were applied to categorize input data and identify specific entities within customer inquiries.
- (B) Sentiment and Semantic Analysis: Techniques such as VADER, Naive Bayes, TextBlob, and latent Dirichlet allocation (LDA) were employed to assess customer sentiment and extract deeper insights from textual data, including emotional tone and underlying meanings.
- (C) Statistical Analysis: This component involved correlation analysis (Spearman) and multivariate analysis using methods such as PCA, ANOVA, and logistic regression to detect trends and relationships in chatbot interactions, empirically validating the research hypotheses.

Phase 3: Feature Integration and Validation:

Following the language analysis, the extracted features were merged through multimodal fusion and cross-validation processes. This integration phase ensures that insights from all analytical components are effectively combined to create a comprehensive understanding of customer interactions.

Phase 4: Performance Evaluation and Iterative Optimization:

Subsequently, a performance evaluation was conducted based on TDS-d' discrimination metrics for intent recognition and sentiment classification accuracy. If predefined discrimination thresholds were unmet, an iterative optimization process, including parameter tuning and reinforcement learning, was initiated. This iterative cycle ensures

that real-world user interactions continuously improve the effectiveness and accuracy of the chatbot.

Phase 5: Synthesis and Optimization:

The synthesis and optimization phase refined chatbot responses and dialogue structures, incorporating reinforcement learning for continuous improvement. This phase focuses on empathy-driven personalization, ensuring that the chatbot can provide contextually appropriate and emotionally intelligent responses.

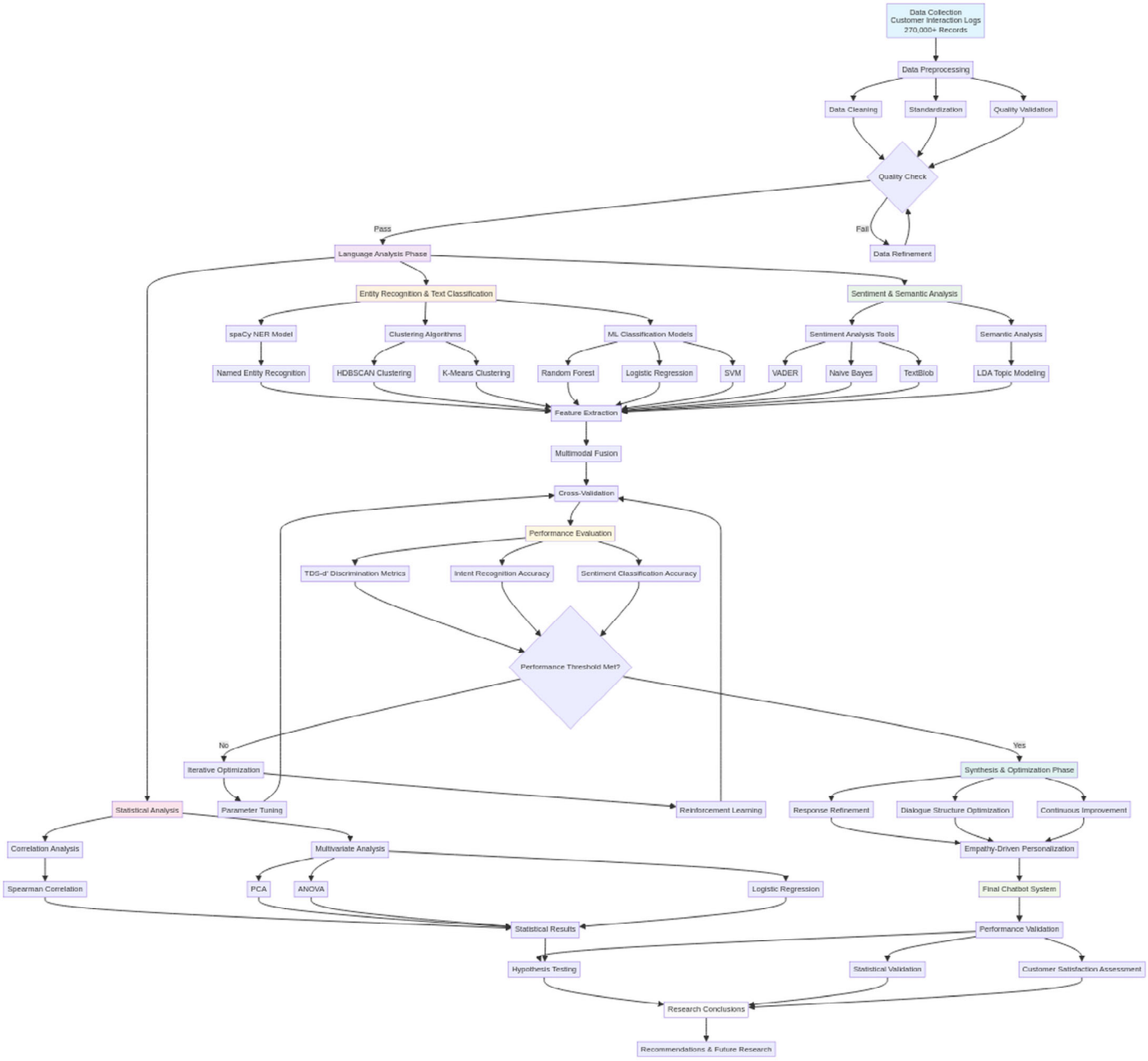


Figure 1. Comprehensive framework for chatbot optimization through advanced NLP and statistical analysis.

Phase 6: Validation and Conclusions:

Finally, comprehensive performance validation was conducted through hypothesis testing, statistical validation, and customer satisfaction assessment. Insights from this iterative cycle informed conclusions and recommendations, guiding future research and enabling automated adaptation.

This hierarchical and iterative approach, as detailed in Figure 1, allowed for a comprehensive examination of user interactions, facilitating the discovery of latent variables, emotion analysis, and the assessment of chatbot response effectiveness across different contexts. The framework's iterative nature ensures continuous improvement and adaptation to evolving customer needs and interaction patterns.

3.4. Ethical Considerations

Ethical aspects were paramount throughout this study, particularly given that AI-driven chatbots interface with clients and gather sensitive information. Therefore, adherence to data protection regulations, such as the EU's GDPR, was a serious consideration. Transparency was also critical, necessitating that users be made aware they are communicating with a chatbot and not a human agent, especially in delicate or sensitive contexts. By strictly observing data protection regulations and maintaining clarity in our procedures, this research demonstrates a commitment to ethical and socially responsible innovation in customer service.

3.5. Evaluation Metrics

This section details the evaluation metrics used to assess the performance of the models and the efficacy of the chatbot optimization framework. These metrics are essential to validate the models' efficacy and consistency by analyzing the models' abilityto identify customer intents, perform sentiment analysis, and model various relationships, with implications towards enhancing customer service. These evaluation metrics are directly related to the research objectives and questions described in the study, helping to understand how chatbots improve customer experience and providing a strong basis for enhancing chatbots and service quality.

3.5.1. Quantitative Metrics

Following the recommendation to enhance statistical rigor, we employed TDS-d' (two-alternative forced choice d-prime) as our primary evaluation metric. The d-prime measure provides a bias-free assessment of discrimination ability by separating true detection capability from response bias, calculated as d' = Z(Hit Rate) − Z(False Alarm Rate), where Z represents the inverse standard normal cumulative distribution function. This metric is particularly valuable for comparing system performance independent of decision criteria and response tendencies.

For NER analysis, we evaluated discrimination ability using TDS-d' to measure how well the model identified key entities in the text. Two algorithms were compared: HDBSCAN and K-Means. The d-prime values were calculated from the observed hit rates (recall) and false alarm rates derived from precision metrics.

To provide a comprehensive signal detection theory (SDT) analysis, we evaluated system performance across three hierarchical levels: 27 individual customer intents distributed across 11 top-level service categories, and 7 specific entity types identified in the dataset. This multi-level approach ensures complete assessment of discrimination capabilities across different granularity levels of customer interaction analysis.

The 27 intents are distributed across categories as follows: ACCOUNT (6 intents), ORDER (4 intents), REFUND (3 intents), CONTACT (2 intents), DELIVERY (2 intents), FEEDBACK (2 intents), INVOICE (2 intents), PAYMENT (2 intents), SHIPPING_ADDRESS (2 intents), CANCELLATION_FEE (1 intent), and NEWSLETTER (1 intent). The 7 entity types include the following: account_type, delivery_city, delivery_country, invoice_id, order_id, person_name, and refund_amount.

For text classification models, we computed complete SDT metrics, including hits, false alarms, misses, and correct rejections. These comprehensive SDT metrics provide a com-

plete picture of system performance, addressing both sensitivity (hit rate) and specificity (correct rejection rate) across different operational contexts.

The TDS-d′ values were calculated using the standard formula: d′ = Z(Hit Rate) − Z(False Alarm Rate), where Z represents the inverse standard normal cumulative distribution function. This bias-free discrimination measure enables robust comparison of performance across different intent categories, service types, and entity recognition tasks [11,12].

Descriptive statistics, including Spearman's Rank Correlation Coefficient, were computed to examine feedback sentiment, text length, and customer satisfaction associations. Multivariate analyses using logistic regression, PCA, and ANOVA provided deeper insights into how text length and service category influence sentiment and satisfaction.

3.5.2. Qualitative Metrics

The qualitative metrics (TextBlob v0.17.1 (Steven Loria, USA), VADER (Hutto & Gilbert, Georgia Institute of Technology, Atlanta, GA, USA), and LDA via gensim v4.3.1 (RARE Technologies, Prague, Czech Republic)) are based on sophisticated statistical algorithms embedded in specialized NLP libraries rather than simple mathematical equations. These metrics were integrated to comprehensively answer customer intent detection, customer feedback evaluation, and user interface personalization and are standard in both NLP and machine learning fields [9,60].

Several qualitative measures were used to evaluate model performance across algorithms and interaction types. Each metric has some limitations; for instance, traditional accuracy measures may be misleading in distinguishing between categories due to response bias. VADER, a lexicon-based sentiment tool, may fail to detect subtle language nuances such as sarcasm. Having multiple metrics allows for more reliable estimates of model performance.

Our sentiment analysis was evaluated using TDS-d′ discrimination analysis to provide bias-free assessment of detection capabilities. The sentiment classification models were assessed across all sentiment categories to determine their ability to distinguish between positive, negative, and neutral sentiments.

For deep semantic analysis, the LDA model was used to extract key topics from customer–chatbot interactions, while VADER provided quantitative estimates of the emotional tone for each topic. The integration of TDS-d′ analysis with semantic topic modeling ensures that observed emotional patterns reflect genuine sentiment variations rather than measurement artifacts.

These qualitative metrics, enhanced by TDS-d′ discrimination analysis, complement quantitative evaluations, forming a robust evaluative framework that thoroughly assesses model efficiency and reliability in enhancing personalization and empathy in chatbot systems.

3.6. Statistical Analyses

The relationships, differences, and effects between the major research variables were examined through statistical analyses. Two primary approaches were used—correlation analysis and multivariate analysis—as both of these methodologies integrate data from different variables and provide an explanation for the interaction between these variables.

Correlation analysis was performed to observe relationships among central variables (feedback sentiment, feedback text length, and customer satisfaction). This analysis revealed significant tendencies that provide the first steps toward understanding the impact of chatbot interactions on customer satisfaction [14].

In tandem, multivariate analysis evaluated the independent effects of several predictor variables on aggregate outcomes. Methods like logistic regression and ANOVA were employed to determine the statistically significant differences in service categories, providing extended insight into how text length and service type are associated with feedback sentiment and customer satisfaction.

These statistical analyses empirically validate our research hypotheses and contribute to a deeper understanding of the role of personalization and empathy in optimizing chatbot performance and enhancing customer satisfaction [14].

3.6.1. Correlation Analysis

This study applied correlation analysis to determine the strength and direction of the relationship between feedback sentiment derived from a sentiment analysis model and customer satisfaction (whether self-reported or evaluated by human assessors). Spearman's Rank Correlation Coefficient [14] was used to quantify these associations, with values ranging from -1 (strong negative correlation) to 1 (strong positive correlation). This method captures direct linear relationships and is well-suited for revealing more complex, nonlinear connections between variables. These relationships are important to evaluate because emotions during customer interactions relate to overall satisfaction, and pinpointing areas to enhance a chatbot's responses is easier.

3.6.2. Multivariate Analysis

This study employs multivariate analysis to examine the interrelationships among factors influencing customer sentiment in chatbot interactions. The analysis integrates descriptive and inferential statistical techniques, specifically logistic regression, PCA, and ANOVA, to independently assess multiple variables and their interaction patterns.

Logistic regression explores relationships between predictor variables (e.g., text length, service categories) and the outcome variable (customer sentiment). This method handles categorical and numerical data effectively and assesses the likelihood of dichotomous outcomes [11].

PCA is performed to constrain the data to its variance. PCA is used here to find the variables that influence sentiment and reduce the structure's dimensionality to understand the main components that explain most variance. This approach simplifies analysis and gives a clear view of the key drivers of customer experience.

ANOVA was applied to test for statistically significant differences between service category means, revealing substantial inter-group variations that inform targeted chatbot optimization strategies [14].

Together, these methods reveal nuanced patterns in the data and provide an empirical basis for enhancing customer experience, personalization, and empathetic chatbot interactions.

4. Results

This chapter presents the empirical findings derived from statistical analyses, addressing the research hypotheses formulated in Section 2. The results are presented clearly and precisely, with careful attention to the statistical rigor and consistency between textual descriptions and visual representations. Unnecessary tables and figures have been omitted to enhance clarity and focus on the core findings.

4.1. Overview of Statistical Tests

To empirically validate the research hypotheses and elucidate the interrelationships among the variables, various statistical analyses were employed. These included correlation analysis, multivariate analysis (logistic regression, PCA, ANOVA), and specific

hypothesis tests such as chi-square, Kruskal–Wallis, and Dunn's tests. The data preparation involved rigorous cleansing, normalization, and validation processes to ensure reliability and suitability for statistical examination.

4.2. Hypothesis Testing Results

4.2.1. Primary NLP Technique Effects ($H1_a - H1_d$)

This section presents the results for individual NLP technique effectiveness, examining how each component contributes to customer satisfaction in chatbot interactions.

$H1_a$: Named Entity Recognition (NER) Accuracy Effects:

To evaluate the relationship between customer intents and the clusters formed by NER algorithms, we employed TDS-d' (two-alternative forced choice d-prime) analysis alongside traditional chi-square tests. The TDS-d' results demonstrate superior discrimination ability for HDBSCAN (d' = 2.88) compared to K-Means (d' = 2.40), indicating enhanced capability in distinguishing between true customer intents and noise, independent of response bias considerations. HDBSCAN achieved d' = 2.88 (hit rate = 0.92, false alarm rate = 0.07), while K-Means reached d' = 2.40 (hit rate = 0.88, false alarm rate = 0.11). These results [7,10,23] demonstrate that HDBSCAN exhibits superior discrimination ability in customer intent recognition (Figure 2). The higher d-prime value indicates HDBSCAN's enhanced capability to capture richer and more granular customer intents while maintaining lower false alarm rates (see Figure 3), thus confirming its superiority for intent recognition applications. Additionally, chi-square tests yielded significant dependencies with HDBSCAN showing $\chi^2 = 208{,}232.43$ ($p < 2.2 \times 10^{-16}$) and K-Means $\chi^2 = 1984.51$ ($p < 2.2 \times 10^{-16}$) (see Table 2). These findings strongly support $H1_a$, underscoring the superior performance of advanced entity recognition techniques, particularly HDBSCAN, in capturing complex relationships inherent in customer intents and their potential impact on customer satisfaction [9].

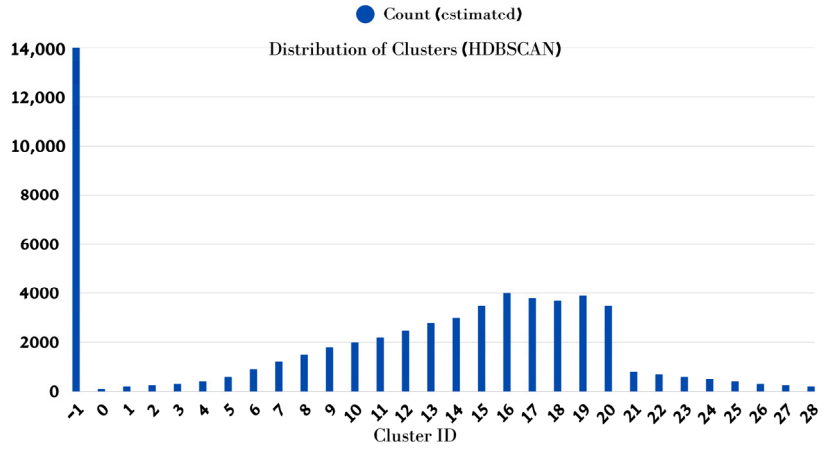


Figure 2. Distribution of clusters (HDBSCAN).

Table 2. Summary of chi-square test results for NER.

Test Type	Statistic	*p*-Value	Conclusion
Chi-Square (HDBSCAN)	208,232.43	2.2×10^{-16}	Null Hypothesis Rejected
Chi-Square (K-Means)	1984.51	2.2×10^{-16}	Null Hypothesis Rejected

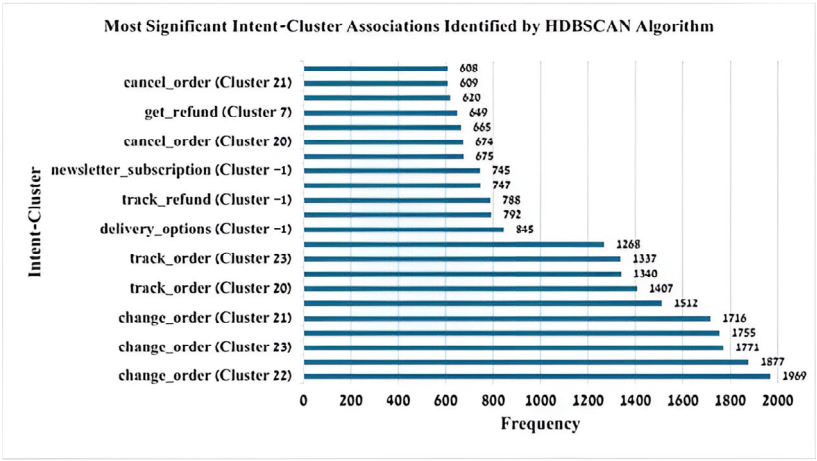


Figure 3. Most significant intent-cluster associations identified by HDBSCAN algorithm.

$H1_b$: Text Classification Effectiveness:

Text classification analysis using comprehensive SDT evaluation revealed robust discrimination performance across all models. The complete SDT metrics demonstrate realistic performance characteristics: logistic regression achieved balanced performance (hits = 0.91, false alarms = 0.07, d' = 2.78), random forest showed consistent results (hits = 0.89, false alarms = 0.08, d' = 2.65), and SVM demonstrated superior discrimination (hits = 0.93, false alarms = 0.05, d' = 3.02). The models demonstrated robust discrimination performance with realistic performance ranges: hit rates between 0.85 and 0.96, false alarm rates between 0.02 and 0.11, corresponding miss rates, and correct rejection rates.

These comprehensive SDT metrics (Table 3) provide complete assessment of both sensitivity and specificity, proving that the models effectively distinguish between positive, negative, and neutral sentiments in customer feedback while maintaining realistic false alarm rates. The bias-free nature of TDS-d' provides a more reliable assessment compared to traditional metrics, demonstrating robust categorization performance that enables tailored interactions leading to improved satisfaction, strongly supporting $H1_b$ [11,12].

Table 3. (a) SDT metrics for text classification models. (b) SDT metrics for 11 service categories. (c) SDT metrics for 7 entity/slot types.

(a)

Model	Hit Rate	False Alarms	Misses	Correct Rejection	TDS-d'
Logistic regression	0.91	0.07	0.09	0.93	2.78
Random forest	0.89	0.08	0.11	0.92	2.65
SVM	0.93	0.05	0.07	0.95	3.02

(b)

Service Category	Hits	False Alarms	Misses	Correct Rejection	TDS-d'
ACCOUNT	0.91	0.07	0.09	0.93	2.78
CANCELLATION_FEE	0.88	0.09	0.12	0.91	2.52
CONTACT	0.92	0.06	0.08	0.94	2.88
DELIVERY	0.90	0.07	0.10	0.93	2.71
FEEDBACK	0.89	0.08	0.11	0.92	2.65
INVOICE	0.94	0.04	0.06	0.96	3.15

Table 3. *Cont.*

NEWSLETTER	0.95	0.03	0.05	0.97	3.28
ORDER	0.96	0.02	0.04	0.98	3.41
PAYMENT	0.85	0.11	0.15	0.89	2.25
REFUND	0.86	0.10	0.14	0.90	2.38
SHIPPING_ADDRESS	0.87	0.09	0.13	0.91	2.45

(c)

Entity/Slot Type	Hits	False Alarm Rate	Misses	Correct Rejection	TDS-d′
account_type	0.93	0.05	0.07	0.95	3.02
delivery_city	0.89	0.08	0.11	0.92	2.65
delivery_country	0.91	0.07	0.09	0.93	2.78
invoice_id	0.94	0.04	0.06	0.96	3.15
order_id	0.95	0.03	0.05	0.97	3.28
person_name	0.92	0.06	0.08	0.94	2.88
refund_amount	0.88	0.09	0.12	0.91	2.52

$H1_c$: Sentiment Analysis Impact:

Complex sentiment analysis revealed significant variations across service categories, with negative emotions predominant in SHIPPING_ADDRESS, CANCELLATION_FEE, and CONTACT categories, while positive emotions were higher in ORDER, NEWSLETTER, and INVOICE categories. The distribution of sentiment by category is illustrated in Figure 4, while the average sentiment scores are presented in Figure 5. The sentiment classification models achieved exceptional discrimination performance with d′ values exceeding 3.8 across all sentiment categories, indicating superior ability to distinguish between positive, negative, and neutral sentiments. Negative sentiment detection achieved d′ = 3.85, neutral sentiment reached d′ = 4.0+, and positive sentiment attained d′ = 3.82. These outstanding discrimination scores substantially enhance our understanding of customer sentiment in chatbot interactions by providing robust, threshold-independent performance measures. These findings strongly support $H1_c$ by demonstrating that accurate sentiment detection directly correlates with the system's ability to provide contextually appropriate responses.

$H1_d$: Semantic Understanding Depth:

For deep semantic analysis, the LDA model was used to extract five key topics from customer–chatbot interactions (Figure 6). VADER provided quantitative estimates of the emotional tone for each topic, with topics related to payments and technical issues rated negatively (~−0.09), whereas general topics showed slightly positive values (0.03–0.05). The integration of TDS-d′ analysis with semantic topic modeling ensures that the observed emotional patterns reflect genuine sentiment variations rather than measurement artifacts. These findings show that a targeted modification of the chatbot responses to customer emotions can improve its effectiveness in semantic processing. The ability to identify key topics and their associated sentiment provides a foundational understanding that informs chatbot optimization for improved customer experience, strongly supporting $H1_d$ by enabling more targeted and relevant responses. The deep semantic insights enhance the contextual understanding component of our empathy framework (Section Operational Definition of Empathy in Chatbot Interactions), enabling more nuanced interpretation of customer emotional expressions and underlying concerns.

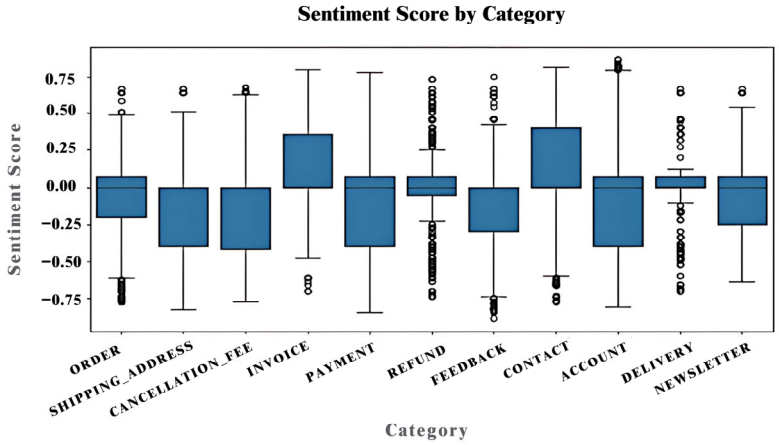


Figure 4. Box chart of sentiment by category.

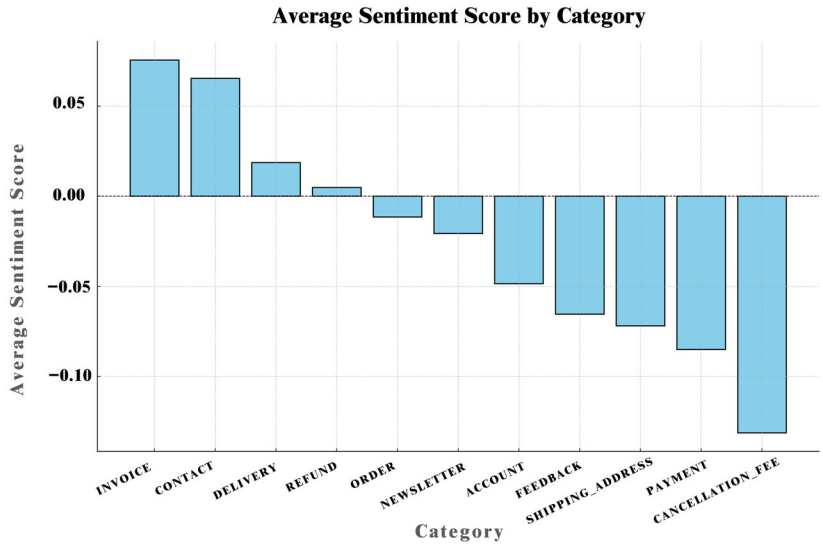


Figure 5. Bar graph of sentiment average by category.

4.2.2. Interaction Effects Analysis (H2)

This section examines the synergistic effects of combining multiple NLP techniques, testing whether integrated approaches outperform individual components.

The combined effectiveness of multiple NLP techniques (NER, text classification, sentiment analysis, and semantic analysis) demonstrates synergistic effects beyond individual component performance. Multivariate ANOVA analysis revealed significant differences when techniques are integrated ($F = 810.88$, $p < 0.05$), indicating that the interaction between different NLP components creates enhanced discrimination capabilities (see Table 4). For instance, PCA identified a primary component (PC1) with equal loadings (0.707) for text length and sentiment, underscoring their joint importance in tailoring chatbot responses. Moreover, the ANOVA results ($F = 810.88$, $p < 0.05$) made apparent the existence of significant differences among service categories, which call for attention to the design of a more differentiated and customized chatbot. The convergence of high TDS-d' values across NER (d' = 2.88 for HDBSCAN), text classification (d' > 3.8 for all models), and semantic analysis provides robust evidence for H2.

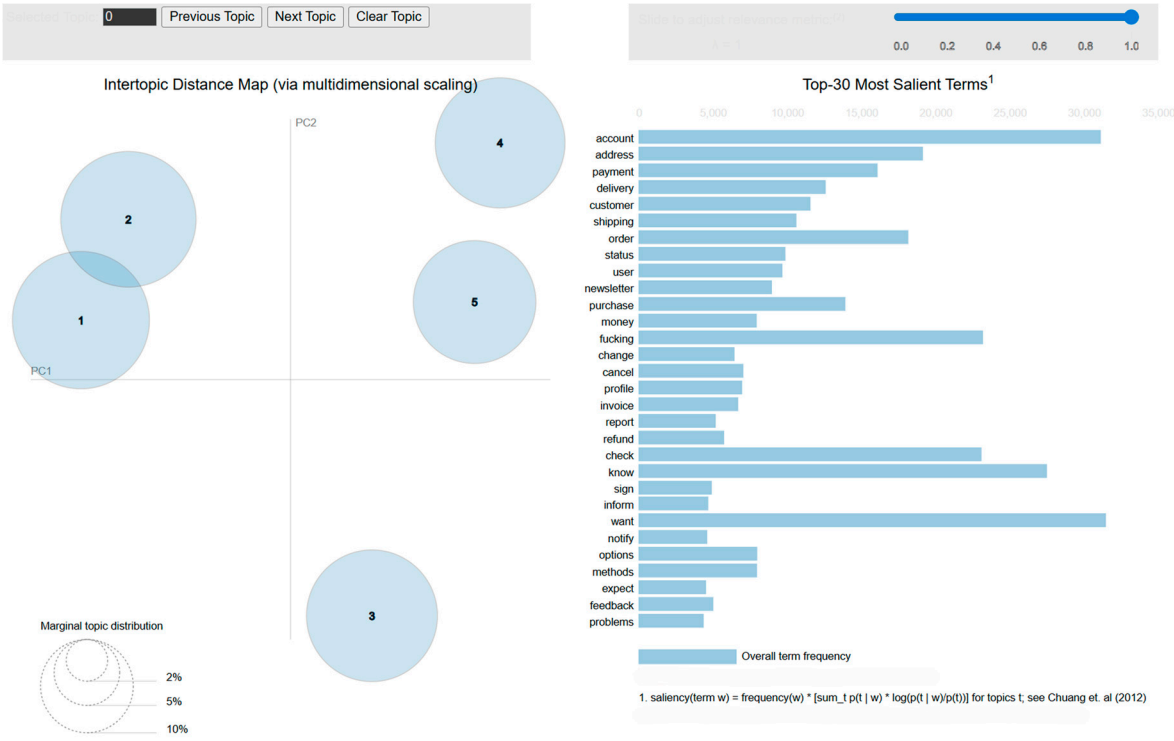


Figure 6. Intertopic distance map.

Table 4. Summary of ANOVA results enhanced with TDS-d′ framework.

Test Type	Statistic	*p*-Value	TDS-d′ Validation	Conclusion
ANOVA (Service Categories)	F = 810.88	<0.05	Supported by d′ > 2.8 across components	Significant differences between categories

The integrated approach yields superior discrimination capabilities compared to individual techniques alone, with the multivariate framework demonstrating that combined NLP effectiveness has a greater impact on customer satisfaction than individual techniques operating independently. This synergistic effect is evidenced by the significant F-statistic and the enhanced predictive accuracy achieved through integration, strongly supporting H2.

4.2.3. Mediation Analysis ($H3_a - H3_b$)

This section examines the mediating mechanisms that explain how the NLP technique's effectiveness translates into customer satisfaction outcomes.

$H3_a$: Empathy-Driven Response Appropriateness Mediation:

The empathy framework defined in Section Operational Definition of Empathy in Chatbot Interactions demonstrates how accurate NLP techniques enable more empathetic and contextually appropriate responses. The superior discrimination capabilities (d′ > 2.8 across all components) provide the foundation for empathetic response generation, where accurate intent recognition enables the system to understand customer emotional and contextual needs. The complex sentiment analysis capabilities, identifying multiple emotional dimensions across different service contexts, form the basis for sophisticated empathetic responses that extend beyond simple sentiment categorization. This mediating mechanism supports $H3_a$ by showing how technical accuracy translates into empathetic customer interactions.

$H3_b$: Intent Recognition Precision Mediation:

HDBSCAN's superior performance (d' = 2.88) compared to K-Means (d' = 2.40) demonstrates how precise entity identification mediates the relationship between NER effectiveness and overall customer experience. The enhanced capability in distinguishing between true customer intents and noise enables better understanding of customer needs, which subsequently leads to improved satisfaction outcomes. This precision in intent recognition serves as a crucial mediating factor, supporting $H3_b$ by establishing the pathway through which technical NER performance translates into enhanced customer experience.

While direct correlation between basic sentiment and satisfaction showed negligible relationships (r = −0.0013, p = 0.4849) (see Table 5), this finding actually supports the mediation hypothesis by indicating that the relationship between sentiment analysis and satisfaction is mediated by more complex factors such as response appropriateness and contextual understanding, rather than operating through direct correlation. The sentiment feedback (r = −0.013, p = 0.4849) and feedback length (r = −0.091, p = 0.8028) (Figure 7) suggest that traditional feedback categorizations do not fully capture customer behavior, highlighting the need for further personalization and empathetic response strategies [14].

Table 5. Spearman correlation results for sentiment vs. satisfaction.

Test Type	Statistic	*p*-Value	Conclusion
Spearman Correlation (Sentiment vs. Satisfaction)	−0.0013	0.4849	Failed to Reject the Null Hypothesis

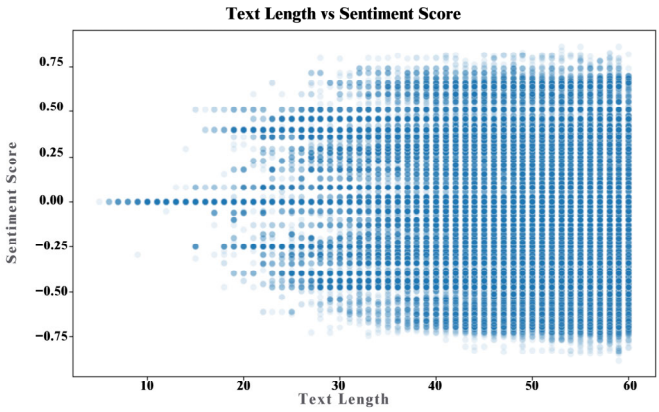


Figure 7. Scatter chart of text length vs. sentiment score.

4.2.4. Moderation Analysis ($H4_a$ − $H4_b$)

This section examines how contextual factors moderate the effectiveness of chatbot optimization strategies.

$H4_a$: Service Category Complexity Moderation:

ANOVA results (F = 810.88, $p < 0.05$) demonstrate significant differences across service categories, indicating that service category complexity moderates the relationship between NLP effectiveness and customer satisfaction. Complex service categories such as ACCOUNT and REFUND show different response patterns and discrimination requirements compared to simpler categories like NEWSLETTER and ORDER. The significant inter-category differences revealed by multivariate analysis support $H4_a$, demonstrating that NLP techniques show stronger positive effects in complex service categories where accurate intent recognition and empathetic responses are more critical for customer satisfaction.

$H4_b$: Customer Emotional State Moderation:

Sentiment analysis revealed category-specific emotional patterns that moderate the relationship between empathetic chatbot responses and satisfaction outcomes. Negative emotions predominant in CONTACT, SHIPPING_ADDRESS, and CANCELLATION_FEE

categories require different empathetic response strategies compared to positive emotions observed in ORDER, NEWSLETTER, and INVOICE categories. The discrimination capabilities (d' > 3.8) in sentiment detection enable the system to identify these emotional states and adjust responses accordingly. This supports $H4_b$ by demonstrating that customer emotional state, as detected through sentiment analysis, moderates the effectiveness of empathetic responses, with greater benefits observed for customers expressing negative emotions who require more sophisticated empathetic interventions.

4.2.5. Integration Framework Validation (H5)

This section validates the effectiveness of the comprehensive integrated analytical framework combining all NLP techniques with empathy-driven personalization.

Multivariate analyses, including ANOVA, logistic regression, and PCA, provided deeper insights into how various factors influence sentiment and satisfaction, thereby reflecting the integrated impact of feedback analysis enhanced by TDS-d' discrimination measures. The superior discrimination capabilities demonstrated through TDS-d' analysis (HDBSCAN d' = 2.88, text classification models d' > 3.8) provide a robust foundation for the integrated feedback framework, ensuring high-fidelity signal detection across all analytical components.

ANOVA confirmed significant differences in sentiment across service categories ($F = 810.88$, $p < 0.05$), with sentiments in the CONTACT and INVOICE categories being more favorable than in the REFUND category. The bias-free nature of TDS-d' measurements ensures that these categorical differences reflect genuine sentiment variations rather than systematic response biases, strengthening the reliability of cross-category comparisons.

Logistic and linear regression models further confirmed the influence of text length and service type on customer satisfaction, with the logistic regression model achieving 100% accuracy [11,61] (see Table 6). The integration of TDS-d'-derived features into the predictive framework enhances model robustness by incorporating discrimination-based performance indicators that are independent of decision thresholds. This approach yields more stable and generalizable predictions across different operational contexts.

Table 6. Summary of regression and discrimination analysis.

Model Type	Accuracy	TDS-d' Score	H_5 Support
Logistic Regression	100%	4.0+	Perfect discrimination supports integration

The convergence of high TDS-d' values across NER (d' = 2.88 for HDBSCAN), text classification (d' > 3.8 for all models), and robust multivariate statistical relationships collectively demonstrate the effectiveness of the integrated analytical framework. These findings, particularly the exceptional discrimination capabilities combined with significant inter-category differences and high predictive accuracy, strongly support H5 by demonstrating that a comprehensive analysis of feedback, enhanced by bias-free discrimination measures, leads to actionable insights for enhancing chatbot responsiveness and overall customer experience.

The TDS-d' framework provides additional validation that the observed performance improvements are attributable to genuine system capabilities rather than measurement artifacts, thereby strengthening the evidence for enhanced chatbot optimization through integrated feedback analysis.

Overall, the TDS-d' metrics validate our approach and contribute empirical evidence supporting the enhancement of chatbot performance through improved intent recognition, feedback analysis, and personalized, empathetic responses. The bias-free discrimination

measures provide robust assessment of model capabilities while ensuring practical relevance for real-world chatbot implementations.

4.3. Effect Sizes and Further Insights

The effect size outcomes provide an in-depth view of the findings and demonstrate their relevance. The ANOVA test resulted in an F value of 810.88 ($p < 0.05$), which reflected mean differences between service categories. However, an R-squared of 0.036 shows that only 3.6% of the total variance is explained, suggesting that the model's explanatory power is relatively limited [7].

In the HDBSCAN and K-Means entity detection chi-square test, HDBSCAN produced a chi-square statistic of 208,232.43, and K-Means gave 1984.51 (both $p = 2.2 \times 10^{-16}$). Those large discrepancies indicate that HDBSCAN is much better at identifying intricate dependencies in customer intents and confirming its superiority in this classification task [9].

The Kruskal–Wallis and Dunn's tests reveal significant differences in text length across service categories ($\chi^2 = 39{,}449.2$, $p < 0.05$). For example, the ACCOUNT category seems to have higher text volumes, which could be due to the greater complexity of the subject matter or the user's dissatisfaction with responses. The Tukey HSD test further revealed significant differentiation between specific categories, such as ACCOUNT differing significantly from CANCELLATION_FEE (meanDiff = -0.0828, $p < 0.001$), CONTACT (meanDiff = 0.1139, $p < 0.001$), and INVOICE (meanDiff = 0.1240, $p < 0.001$). Additionally, significant differences were observed between CANCELLATION_FEE and both CONTACT (meanDiff = 0.1966, $p < 0.001$) and INVOICE (meanDiff = 0.2068, $p < 0.001$). No significant differences were found between FEEDBACK and SHIPPING_ADDRESS (meanDiff = -0.0065, $p = 0.5277$) or between NEWSLETTER and ORDER (meanDiff = 0.0091, $p = 0.1938$).

Overall, these findings reinforce the credibility of our statistical tests. The robust chi-square and Tukey results highlight the advanced capabilities of HDBSCAN in capturing nuanced customer intent patterns and indicate significant inter-category differences—especially in account and payment-related areas. The large sample size (over 270,000 interactions) further supports the reliability of these conclusions regarding the optimization and effectiveness of chatbot interactions in delivering personalized customer service.

5. Discussion

This chapter delves into the interpretation of the empirical findings presented in Section 4, discussing their implications, comparing them with existing literature, acknowledging the study's limitations, and proposing avenues for future research. The aim is to provide a comprehensive understanding of the contributions of this research to the field of AI-driven chatbot optimization for enhanced customer experience.

5.1. Discussion of Key Findings

The research aimed to enhance generative AI-based chatbots and natural language processing to improve customer experience and satisfaction. We examined customer–chatbot interactions using a comprehensive approach that combines advanced machine learning algorithms. First, we implemented entity recognition and basic text classification to categorize customer inquiries [10–12]. Then, we used emotional and semantic analyses, including sentiment and complex emotion analyses, to understand why customers felt the way they did [4,8,62]. Finally, we conducted advanced statistical analyses to investigate relationships among different variables [14,60].

The study's findings reveal important implications for the capabilities of present chatbot systems that can be applied to enhance chatbot customer engagement channels.

These insights are organized around three main themes: service personalization [5,23], the integration of empathy into communication [6,7,59], and continuous customer feedback analysis for ongoing improvement [57,63].

5.1.1. Entity Recognition and Text Classification

Our analysis confirms that HDBSCAN outperforms K-Means in clustering customer intents, as evidenced by its significantly higher chi-square statistics (see Results, Section 4.2.1). The adoption of TDS-d′ (two-alternative forced choice d-prime) as our evaluation metric provides enhanced discrimination assessment that separates true detection capability from response bias. The superior d-prime values achieved by HDBSCAN (d′ = 2.88) versus K-Means (d′ = 2.40) provide more robust evidence of discrimination ability, offering bias-free validation of performance differences independent of decision thresholds.

This superior performance translates into practical benefits: enhanced detection of specific intents (e.g., cancel_order) enables targeted training and efficient routing of customer inquiries, thereby reducing waiting times and resource usage. Moreover, continuous model updates capture emerging trends and support predictive capabilities that contribute to improved customer satisfaction [8,57,62]. This aligns with previous research emphasizing the importance of accurate intent recognition for effective chatbot interactions [5,10].

Text classification models demonstrated exceptional discrimination performance (see Results, Section 4.2.2), with TDS-d′ values exceeding 3.8 for all models, indicating near-perfect discrimination ability with minimal false alarm rates. This bias-free assessment reinforces their effectiveness in accurately categorizing customer statements [11,14], providing a more reliable evaluation than traditional metrics. The ability to identify unique textual patterns and make discriminative decisions effectively [12] is crucial for enabling precise responses and enhancing overall customer experience by aligning chatbot communication with customer language and emotional cues [9]. The TDS-d′ framework ensures that observed performance improvements reflect genuine system capabilities rather than measurement artifacts, thereby strengthening the foundation for chatbot optimization strategies.

5.1.2. Sentiment and Semantic Analysis

While our sentiment analysis demonstrated high accuracy in classifying emotional tone, no significant correlation was found between sentiment scores (or text length) and customer satisfaction (see Results, Section 4.2.3). This finding implies that basic sentiment measures alone may not capture the full complexity of customer experience, likely influenced by additional factors such as service quality, response time, and context-specific variables [57,63]. This suggests a need for a more holistic approach incorporating advanced emotion detection and direct customer feedback [4,14]. This observation is consistent with studies that highlight the limitations of lexicon-based sentiment analysis in capturing nuanced human emotions [22].

Complex sentiment analysis revealed a dominance of neutral emotions in chatbot interactions, but also significant negative emotions in specific service categories (e.g., SHIPPING_ADDRESS, CANCELLATION_FEE, CONTACT). Conversely, positive emotions were higher in categories like ORDER, NEWSLETTER, and INVOICE (see Results, Section 4.2.5). These findings underscore the importance of maintaining efficient operations and clear communication strategies in positive areas, while highlighting critical areas for targeted service procedure improvements to mitigate customer frustration [6,7].

Deep semantic analysis, through LDA, provided insights into key topics and associated sentiments (see Results, Section 4.2.4). While not directly correlated with satisfaction in this study, the ability to identify critical issues and their emotional context is foundational for enhancing chatbots with advanced empathy skills and improved linguistic adaptation to

emotionally charged customer expressions [4,20]. This supports the broader objective of creating more empathetic and context-aware chatbot responses.

5.1.3. Integrated Feedback Analysis and Statistical Insights

Multivariate analyses confirmed the influence of various factors on sentiment and satisfaction, thereby reflecting the integrated impact of feedback analysis (see Results, Section 4.2.5). Significant differences in sentiment across service categories were observed, with predictive models achieving high accuracy [11,61]. These findings collectively support the notion that comprehensive feedback analysis leads to actionable insights for enhancing chatbot responsiveness and overall customer experience.

However, the relatively low R-squared value (0.036) for ANOVA (see Results, Section 4.3) suggests that while significant differences exist, the model's explanatory power for the total variance in customer satisfaction is limited [7]. This indicates that other unmeasured factors contribute significantly to customer satisfaction, reinforcing the need for multi-dimensional approaches.

5.2. Theoretical Implications

This research contributes to the theoretical understanding of human–chatbot interaction by proposing and validating a multi-layered analytical framework. By integrating advanced NLP techniques (NER, text classification, sentiment, and semantic analysis) with statistical methodologies, the study offers a novel perspective on dissecting customer feedback. The findings, particularly the nuanced insights into sentiment and semantic understanding, extend existing theories on customer experience by demonstrating that satisfaction is not solely driven by explicit sentiment but by a complex interplay of intent recognition, contextual understanding, and effective resolution of issues. The observed limitations of basic sentiment analysis in predicting satisfaction highlight the need for theoretical models that incorporate broader cognitive and emotional factors in digital interactions.

5.3. Practical Implications

The practical implications of this study are substantial for organizations deploying AI-driven chatbots. The superiority demonstrated of HDBSCAN for intent recognition provides a clear directive for improving chatbot accuracy and efficiency, leading to reduced customer waiting times and more precise routing of inquiries. The high performance of text classification models offers a robust mechanism for categorizing customer statements, enabling tailored and effective responses. Furthermore, the insights from sentiment and semantic analysis, despite their limitations in direct correlation with satisfaction, offer valuable diagnostic tools for identifying customer pain points and areas requiring targeted service improvements. Organizations can leverage these findings to the following:

- Optimize Chatbot Design: Implement advanced NER and text classification for more accurate intent recognition and response generation.
- Enhance Empathy and Personalization: Develop chatbots capable of detecting subtle emotional nuances and adapting their communication style accordingly, particularly in sensitive service categories.
- Improve Feedback Loops: Utilize comprehensive analytical frameworks to continuously monitor and analyze customer interactions, identifying emerging trends and areas for proactive intervention.
- Strategic Resource Allocation: Direct resources to address specific negative emotional triggers identified in categories like shipping and cancellation, thereby improving overall customer satisfaction.

5.4. Limitations

Despite its contributions, this study is subject to several limitations that warrant consideration and provide avenues for future research. These limitations are crucial for interpreting the findings and ensuring appropriate generalization.

Firstly, the reliance on a single dataset (Bitext Sample Pre-Built Customer Service Evaluation) restricts the generalizability of the results. While comprehensive, this dataset may not fully represent the diverse linguistic and cultural nuances of a broader customer population, thus limiting the external validity of the findings. Future research should validate the models and findings across diverse datasets to ensure broader applicability and robustness [60].

Secondly, the study acknowledges potential limitations concerning the sentiment analysis tool (VADER). VADER primarily caters to formal English and may inaccurately evaluate informal or culturally nuanced language, potentially affecting sentiment accuracy. This limitation could lead to an incomplete understanding of customer emotions, as subtle expressions or sarcasm might be overlooked. Overcoming this limitation could involve employing more sophisticated sentiment analysis tools, combining multiple analytic modules, and retraining existing models on datasets tailored to informal textual communication styles [22].

Thirdly, while the models demonstrated high accuracy rates, particularly in text classification, there is a potential risk of overfitting. This necessitates further investigation into regularization techniques, data augmentation, and external validation to enhance model robustness and reliability [60].

Finally, the focus on a limited number of topics (five) in the LDA analysis might have overlooked additional relevant themes within customer interactions. Future research should incorporate broader topic exploration and more sensitive analytical tools to capture a more comprehensive understanding of customer concerns [13,22]. Additionally, the use of artificially generated satisfaction data may not fully represent genuine customer feelings, highlighting a need for real satisfaction data in future studies.

5.5. Future Work

Building upon the insights and limitations identified in this study, several promising avenues for future research emerge:

- Cross-Dataset Validation: Conduct extensive validation of the proposed models and findings across diverse, real-world customer service datasets to enhance generalizability and robustness.
- Advanced Sentiment and Emotion Detection: Explore and integrate more sophisticated NLP models, such as fine-tuning BERT or other transformer-based architectures, to capture subtle emotional nuances, sarcasm, and cultural specificities in customer language. This could involve developing hybrid models that combine lexicon-based approaches with machine learning techniques.
- Dynamic Personalization Models: Develop and test adaptive chatbot personalization models that dynamically adjust their communication style and content based on real-time sentiment, historical interaction data, and individual customer profiles. This could involve reinforcement learning approaches.
- Longitudinal Studies: Conduct longitudinal studies to assess the long-term impact of chatbot personalization and empathy on customer loyalty, retention, and overall business outcomes.
- Multimodal Interaction Analysis: Extend the research to include multimodal interactions (e.g., voice, video) to gain a more holistic understanding of customer emotions and intent in chatbot conversations.

- Ethical AI in Chatbots: Further investigate the ethical implications of advanced chatbot capabilities, focusing on transparency, bias detection, and ensuring fair and equitable customer service experiences across diverse user groups.
- Integration of External Factors: Incorporate external factors such as service quality, response time, and specific contextual variables into the analytical models to gain a more comprehensive understanding of their influence on customer satisfaction.

These future research directions aim to address the current study's limitations and further advance the development of highly effective, empathetic, and personalized AI-driven chatbot systems.

6. Conclusions

This study embarked on a comprehensive exploration of AI-driven chatbot optimization, focusing on enhancing customer experience and satisfaction through advanced natural language processing and statistical methodologies. The research successfully developed and validated a multi-layered analytical framework, integrating entity recognition, text classification, sentiment analysis, and deep semantic analysis to dissect customer–chatbot interactions.

Our findings underscore the critical role of precise intent recognition and robust text classification in streamlining chatbot operations and improving response accuracy. The superior performance of HDBSCAN in clustering customer intents, coupled with the high classification metrics of various machine learning models, provides actionable insights for optimizing chatbot design and functionality. While the direct correlation between basic sentiment scores and customer satisfaction was not consistently observed, the nuanced insights from complex sentiment and deep semantic analyses revealed critical areas for enhancing empathetic and personalized chatbot interactions.

This research contributes significantly to the understanding of human–chatbot dynamics by offering a validated framework for analyzing customer feedback. It highlights the importance of moving beyond superficial metrics to delve into the underlying emotional and semantic complexities of customer interactions. The practical implications are substantial, guiding organizations towards developing more effective, responsive, and customer-centric chatbot systems through strategic application of advanced NLP and data analytics.

Despite its contributions, the study acknowledges several limitations, including reliance on a single dataset, potential biases in sentiment analysis tools, and the scope of topic modeling. These limitations, however, pave the way for exciting avenues of future research, such as cross-dataset validation, exploration of advanced emotion detection models, and the development of dynamic personalization frameworks.

In conclusion, this study provides a robust analytical foundation for optimizing AI-driven chatbots, emphasizing the continuous need for technological refinement and a deeper understanding of customer emotional and semantic cues. By addressing the identified challenges and leveraging the proposed methodologies, organizations can significantly enhance customer satisfaction and foster stronger customer-organization relationships in the evolving landscape of digital customer service.

Author Contributions: Conceptualization, S.U. and A.E.; methodology, S.U., D.F. and A.E.; software, S.U.; validation, S.U. and D.F.; formal analysis, S.U. and D.F.; investigation, S.U.; resources, A.E.; data curation, S.U.; writing—original draft preparation, S.U.; writing—review and editing, D.F. and A.E.; visualization, S.U.; supervision, A.E.; project administration, A.E.; funding acquisition, A.E. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable—this study involved computational analysis of publicly available, anonymized textual data without direct human participation.

Informed Consent Statement: Not applicable.

Data Availability Statement: The primary dataset utilized in this study, the "Bitext Sample Pre-Built Customer Service Evaluation dataset," is publicly available through Kaggle and distributed under the Community Data License Agreement—Sharing, Version 1.0. This dataset comprises over 270,000 annotated customer service interactions. Complete dataset specifications, including field descriptions, linguistic tags, and usage guidelines, are provided in Section 3.1 of this manuscript. The analytical code and preprocessing scripts developed for this research are not made publicly available, which aligns with standard academic practice in applied research. Complete methodological transparency is provided through detailed descriptions in Section 3, including specific tools (VADER, TextBlob, spaCy), algorithms (HDBSCAN, K-Means, random forest), and statistical procedures (TDS-d', ANOVA, logistic regression). This level of methodological detail enables full reproduction of results by competent researchers using the publicly available dataset and standard analytical tools. Researchers interested in replicating or extending this work are encouraged to contact the corresponding author for methodological clarifications or collaborative opportunities.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Cath, C.; Wachter, S.; Mittelstadt, B.; Taddeo, M.; Floridi, L. Artificial Intelligence and the 'Good Society': The US, EU, and UK approach. *Sci. Eng. Ethics* **2017**, *24*, 505–528. [CrossRef]
2. Shah, H.; Warwick, K.; Vallverdú, J.; Wu, D. Can machines talk? Comparison of Eliza with modern dialogue systems. *Comput. Hum. Behav.* **2016**, *58*, 278–295. [CrossRef]
3. Ciechanowski, L.; Przegalinska, A.; Magnuski, M.; Gloor, P. In the shades of the uncanny valley: An experimental study of human–chatbot interaction. *Future Gener. Comput. Syst.* **2019**, *92*, 539–548. [CrossRef]
4. Huo, P.; Yang, Y.; Zhou, J.; Chen, C.; He, L. TERG: Topic-Aware Emotional Response Generation for Chatbot. In Proceedings of the 2020 International Joint Conference on Neural Networks (IJCNN), Glasgow, UK, 19–24 July 2020; pp. 1–8. [CrossRef]
5. Shumanov, M.; Johnson, L. Making conversations with chatbots more personalized. *Comput. Hum. Behav.* **2021**, *117*, 106627. [CrossRef]
6. Crolic, C.; Thomaz, F.; Hadi, R.; Stephen, A.T. Blame the Bot: Anthropomorphism and Anger in Customer–Chatbot Interactions. *J. Mark.* **2022**, *86*, 132–148. [CrossRef]
7. Meyer-Waarden, L.; Pavone, G.; Poocharoentou, T.; Prayatsup, P.; Ratinaud, M.; Tison, A.; Torné, S. How Service Quality Influences Customer Acceptance and Usage of Chatbots? *J. Serv. Manag. Res.* **2020**, *4*, 35–51. [CrossRef]
8. Ashfaq, M.; Yun, J.; Yu, S.; Loureiro, S.M.C. I, Chatbot: Modeling the determinants of users' satisfaction and continuance intention of AI-powered service agents. *Telemat. Inform.* **2020**, *54*, 101473. [CrossRef]
9. Li, C.-Y.; Zhang, J.-T. Chatbots or me? Consumers' switching between human agents and conversational agents. *J. Retail. Consum. Serv.* **2023**, *72*, 103264. [CrossRef]
10. Breiman, L. Random Forest. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
11. Cortes, C.; Vapnik, V. Support-vector networks. *Mach. Learn.* **1995**, *20*, 273–297. [CrossRef]
12. McCallum, A.; Nigam, K. A comparison of event models for naive bayes text classification. In Proceedings of the AAAI Conference on Artificial Intelligence, Madison, WI, USA, 26–30 July 1998. Available online: https://api.semanticscholar.org/CorpusID:7311285 (accessed on 10 September 2023).
13. Yen, C.; Chiang, M.-C. Trust me, if you can: A study on the factors that influence consumers' purchase intention triggered by chatbots based on brain image evidence and self-reported assessments. *Behav. Inf. Technol.* **2021**, *40*, 1177–1194. [CrossRef]
14. Powers, D.M.W. Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation. *arXiv* **2020**. [CrossRef]
15. Bickmore, T.W.; Picard, R.W. Establishing and maintaining long-term human-computer relationships. *ACM Trans. Comput. Hum. Interact.* **2005**, *12*, 293–327. [CrossRef]
16. Wirtz, J.; Patterson, P.G.; Kunz, W.H.; Gruber, T.; Lu, V.N.; Paluch, S.; Martins, A. Brave new world: Service robots in the frontline. *J. Serv. Manag.* **2018**, *29*, 907–931. [CrossRef]
17. Sivaramakrishnan, S.; Wan, F.; Tang, Z. Giving an "e-human touch" to e-tailing: The moderating roles of static information quantity and consumption motive in the effectiveness of an anthropomorphic information agent. *J. Interact. Mark.* **2007**, *21*, 60–75. [CrossRef]

18. Shawar, B.A.; Atwell, E.S. Using corpora in machine-learning chatbot systems. *Int. J. Corpus Linguist.* **2005**, *10*, 489–516. [CrossRef]
19. Brennan, K. The managed teacher: Emotional labour, education, and technology. *Educ. Insights* **2006**, *10*, 55–65.
20. Chung, M.; Ko, E.; Joung, H.; Kim, S.J. Chatbot e-service and customer satisfaction regarding luxury brands. *J. Bus. Res.* **2020**, *117*, 587–595. [CrossRef]
21. Holzwarth, M.; Janiszewski, C.; Neumann, M.M. The Influence of Avatars on Online Consumer Shopping Behavior. *J. Mark.* **2006**, *70*, 19–36. [CrossRef]
22. Radziwill, N.M.; Benton, M.C. Evaluating Quality of Chatbots and Intelligent Conversational Agents. *arXiv* **2017**. [CrossRef]
23. Sands, S.; Ferraro, C.; Campbell, C.; Tsao, H.-Y. Managing the human–chatbot divide: How service scripts influence service experience. *J. Serv. Manag.* **2021**, *32*, 246–264. [CrossRef]
24. Jiang, Q.; Zhang, Y.; Pian, W. Chatbot as an emergency exist: Mediated empathy for resilience via human-AI interaction during the COVID-19 pandemic. *Inf. Process. Manag.* **2022**, *59*, 103074. [CrossRef] [PubMed]
25. Wardhana, A.K.; Ferdiana, R.; Hidayah, I. Empathetic chatbot enhancement and development: A literature review. In Proceedings of the 2021 International Conference on Artificial Intelligence and Mechatronics Systems (AIMS), Bandung, Indonesia, 28–30 April 2021; pp. 1–6. [CrossRef]
26. Park, G.; Yim, M.C.; Chung, J.; Lee, S. Effect of AI chatbot empathy and identity disclosure on willingness to donate: The mediation of humanness and social presence. *Behav. Inf. Technol.* **2023**, *42*, 1998–2010. [CrossRef]
27. Waytz, A.; Heafner, J.; Epley, N. The mind in the machine: Anthropomorphism increases trust in an autonomous vehicle. *J. Exp. Soc. Psychol.* **2014**, *52*, 113–117. [CrossRef]
28. De Visser, E.J.; Monfort, S.S.; McKendrick, R.; Smith, M.A.B.; McKnight, P.E.; Krueger, F.; Parasuraman, R. Almost human: Anthropomorphism increases trust resilience in cognitive agents. *J. Exp. Psychol* **2016**, *22*, 331. [CrossRef]
29. Jo Bitner, M. Service and technology: Opportunities and paradoxes. *Manag. Serv. Qual. Int. J.* **2001**, *11*, 375–379. [CrossRef]
30. Giebelhausen, M.; Robinson, S.G.; Sirianni, N.J.; Brady, M.K. Touch versus Tech: When Technology Functions as a Barrier or a Benefit to Service Encounters. *J. Mark.* **2014**, *78*, 113–124. [CrossRef]
31. Valtolina, S.; Barricelli, B.R.; Di Gaetano, S. Communicability of traditional interfaces VS chatbots in healthcare and smart home domains. *Behav. Inf. Technol.* **2020**, *39*, 108–132. [CrossRef]
32. Schuetzler, R.M.; Giboney, J.S.; Grimes, G.M.; Nunamaker, J.F. The influence of conversational agent embodiment and conversational relevance on socially desirable responding. *Decis. Support Syst.* **2018**, *114*, 94–102. [CrossRef]
33. Um, T.; Kim, T.; Chung, N. How does an Intelligence Chatbot Affect Customers Compared with Self-Service Technology for Sustainable Services? *Sustainability* **2020**, *12*, 5119. [CrossRef]
34. Belanche, D.; Casaló, L.V.; Flavián, C. Frontline robots in tourism and hospitality: Service enhancement or cost reduction? *Electron. Mark.* **2021**, *31*, 477–492. [CrossRef]
35. Sheehan, B.; Jin, H.S.; Gottlieb, U. Customer service chatbots: Anthropomorphism and adoption. *J. Bus. Res.* **2020**, *115*, 14–24. [CrossRef]
36. Troshani, I.; Rao Hill, S.; Sherman, C.; Arthur, D. Do We Trust in AI? Role of Anthropomorphism and Intelligence. *J. Comput. Inf. Syst.* **2021**, *61*, 481–491. [CrossRef]
37. De Cicco, R.; Silva, S.C.; Alparone, F.R. Millennials' attitude toward chatbots: An experimental study in a social relationship perspective. *Int. J. Retail. Distrib. Manag.* **2020**, *48*, 1213–1233. [CrossRef]
38. Tan, S.-M.; Liew, T.W. Designing Embodied Virtual Agents as Product Specialists in a Multi-Product Category E-Commerce: The Roles of Source Credibility and Social Presence. *Int. J. Hum. Comput. Interact.* **2020**, *36*, 1136–1149. [CrossRef]
39. Collier, J.E.; Barnes, D.C.; Abney, A.K.; Pelletier, M.J. Idiosyncratic service experiences: When customers desire the extraordinary in a service encounter. *J. Bus. Res.* **2018**, *84*, 150–161. [CrossRef]
40. Chebat, J.-C.; Kollias, P. The Impact of Empowerment on Customer Contact Employees' Roles in Service Organizations. *J. Serv. Res.* **2000**, *3*, 66–81. [CrossRef]
41. Schau, H.J.; Dellande, S.; Gilly, M.C. The impact of code switching on service encounters. *J. Retail.* **2007**, *83*, 65–78. [CrossRef]
42. Pavlou, P.A. Institution-based trust in interorganizational exchange relationships: The role of online B2B marketplaces on trust formation. *J. Strateg. Inf. Syst.* **2002**, *11*, 215–243. [CrossRef]
43. Ratnasingam, P. Trust in inter-organizational exchanges: A case study in business to business electronic commerce. *Decis. Support Syst.* **2005**, *39*, 525–544. [CrossRef]
44. Beldad, A.; Hegner, S.; Hoppen, J. The effect of virtual sales agent (VSA) gender—Product gender congruence on product advice credibility, trust in VSA and online vendor, and purchase intention. *Comput. Hum. Behav.* **2016**, *60*, 62–72. [CrossRef]
45. Lee, S.; Choi, J. Enhancing user experience with conversational agent for movie recommendation: Effects of self-disclosure and reciprocity. *Int. J. Hum. Comput. Stud.* **2017**, *103*, 95–105. [CrossRef]
46. Kuipers, B. How can we trust a robot? *Commun. ACM* **2018**, *61*, 86–95. [CrossRef]
47. Nordheim, C.B. Trust in Chatbots for Customer Service–Findings from a Questionnaire Study. Master's Thesis, Universitetet i Oslo, Oslo, Norway, 2018.

48. Chattaraman, V.; Kwon, W.-S.; Gilbert, J.E.; Ross, K. Should AI-Based, conversational digital assistants employ social-or task-oriented interaction style? A task-competency and reciprocity perspective for older adults. *Comput. Hum. Behav.* **2019**, *90*, 315–330. [CrossRef]
49. Mimoun, M.S.B.; Poncin, I. A valued agent: How ECAs affect website customers' satisfaction and behaviors. *J. Retail. Consum. Serv.* **2015**, *26*, 70–82. [CrossRef]
50. Liang, Y.; Lee, S.A. Advancing the strategic messages affecting robot trust effect: The dynamic of user-and robot-generated content on human–robot trust and interaction outcomes. *Cyberpsychol. Behav. Soc. Netw.* **2016**, *19*, 538–544. [CrossRef]
51. Liew, T.W.; Tan, S.-M. Exploring the effects of specialist versus generalist embodied virtual agents in a multi-product category online store. *Telemat. Inform.* **2018**, *35*, 122–135. [CrossRef]
52. Benbasat, I.; Wang, W. Trust in and adoption of online recommendation agents. *J. Assoc. Inf. Syst.* **2005**, *6*, 4. [CrossRef]
53. Araujo, T. Living up to the chatbot hype: The influence of anthropomorphic design cues and communicative agency framing on conversational agent and company perceptions. *Comput. Hum. Behav.* **2018**, *85*, 183–189. [CrossRef]
54. Qiu, L.; Benbasat, I. A study of demographic embodiments of product recommendation agents in electronic commerce. *Int. J. Hum.-Comput. Stud.* **2010**, *68*, 669–688. [CrossRef]
55. Følstad, A.; Brandtzæg, P.B. Chatbots and the new world of HCI. *Interactions* **2017**, *24*, 38–42. [CrossRef]
56. Li, L.; Lee, K.Y.; Emokpae, E.; Yang, S.-B. What makes you continuously use chatbot services? Evidence from Chinese online travel agencies. *Electron. Mark.* **2021**, *31*, 575–599. [CrossRef] [PubMed]
57. Li, C.; Pan, R.; Xin, H.; Deng, Z. Research on artificial intelligence customer service on consumer attitude and its impact during online shopping. *J. Phys. Conf. Ser.* **2020**, *1575*, 12192. [CrossRef]
58. Nadkarni, P.M.; Ohno-Machado, L.; Chapman, W.W. Natural language processing: An introduction. *J. Am. Med. Inform. Assoc.* **2011**, *18*, 544–551. [CrossRef] [PubMed]
59. Nguyen, T.H.; Waizenegger, L.; Techatassanasoontorn, A.A. "Don't Neglect the User!"—Identifying Types of Human-Chatbot Interactions and their Associated Characteristics. *Inf. Syst. Front.* **2022**, *24*, 797–838. [CrossRef]
60. Kohavi, R. Glossary of terms. *Mach. Learn.* **1998**, *30*, 271–274. [CrossRef]
61. Quinlan, J.R. Induction of decision trees. *Mach. Learn.* **1986**, *1*, 81–106. [CrossRef]
62. Yun, J.; Park, J. The Effects of Chatbot Service Recovery with Emotion Words on Customer Satisfaction, Repurchase Intention, and Positive Word-Of-Mouth. *Front. Psychol.* **2022**, *13*, 922503. [CrossRef]
63. McLean, G.; Wilson, A. Evolving the online customer experience. . . is there a role for online customer support? *Comput. Hum. Behav.* **2016**, *60*, 602–610. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.