International Journal on Advanced Technology, Engineering, and Information System
https://ojs.transpublika.com/index.php/IJATEIS
Online ISSN 2828-5425

https://doi.org/10.55047/ijateis.v4i3.2008

Original Article

Development of Semantic-Based Voicebots and Natural Language Processing for E-Commerce Product Searches

Maskur[1*], Yosi Afandi[2], Wahyu Widyananda[3], Ahmad Fauzi[4], Zhulvardyan Armayrishtya[5]

[1-5]Department of Business Administration, Politeknik Negeri Malang, Indonesia
Email: [1)] maskur@polinema.ac.id, [2)] yosi.afandi@polinema.ac.id, [3)] wahyu.widyananda@polinema.ac.id, [4)] fauzi@polinema.ac.id, [5)] zhulvardyan.armayrishtya@polinema.ac.id

Received : 23 October - 2025
Accepted : 28 November - 2025
Published online : 03 December - 2025

Abstract

Searching for products online is often an inefficient and confusing process, especially when users do not know the exact name of the product or use terms that differ from the search system. Keyword-based searches tend to produce irrelevant results because the system only matches text literally without understanding the meaning. As users increasingly talk to digital devices, voice-based search technology has become a more natural and intuitive alternative. This research aims to develop a semantic-based voicebot supported by Natural Language Processing (NLP) to improve the effectiveness of product searches on e-commerce platforms. The designed system not only recognizes user speech but also understands the context, intent, and semantic meaning of the given commands. The research stages include collecting user voice data, training the Automatic Speech Recognition (ASR) model for voice-to-text conversion, and applying the semantic NLP model for interpreting the context of product searches. The testing was conducted using Indonesian voice commands in a simulated e-commerce scenario. The results showed that the system achieved an average Word Error Rate (WER) of 1.29%, indicating a high level of accuracy in recognizing speech and understanding user intent. The integration between ASR and semantic NLP proved capable of creating a more natural, responsive search experience that resembles the way humans think and communicate when interacting with online search systems.

Keywords: Automatic Speech Recognition (ASR), E-commerce, Natural Language Processing (NLP), Natural Language Understanding, Voicebot.

1. Introduction

The development of digital technology over the past decade has driven massive transformations in various sectors, including electronic commerce or e-commerce. E-commerce platforms are no longer just a place for transactions, but have evolved into digital ecosystems that integrate user experience, smart interactions, and increasingly personalized services (Kim, 2024). Amidst this rapid digitalization, consumer behavior has also changed. Consumers want a faster, more responsive, and more natural product search process, not one based solely on rigid and limited keywords. One of the main challenges in e-commerce services today is how to create a truly intuitive product search experience (Chahal & Mahajan, 2025). Traditional search engines within e-commerce platforms usually only recognize inputs based on specific keywords. Yet, users often do not know the technical terms for products or may use more casual, even ambiguous, ways of asking questions. This is where the need arises to build a system that can understand the user's intent more deeply, not just the literal text that is

Copyright: © 2025 by the authors.
This is an open access article distributed under the terms and conditions of the CC BY 4.0.

typed. One approach that is considered promising to address this issue is the integration of Natural Language Processing (NLP) and semantics into voicebot systems (Chao et al., 2021).

Voicebots are an advanced form of chatbots that allow users to interact through voice commands rather than just text. In an era where virtual assistants such as Google Assistant, Siri, and Alexa have become part of everyday life, the use of voice as a means of interaction is no longer unfamiliar. Many users now find it more convenient to use voice, mainly for reasons of efficiency and ease (Chuang & Cheng, 2022). In the context of e-commerce, voicebots allow customers to search for products simply by speaking, for example: "I need women's size 39 running shoes that are suitable for morning runs." Requests like this cannot always be accurately translated by conventional search systems, but with an NLP and semantic approach, voicebots have the potential to recognize the user's intentions, context, and needs comprehensively (Chen et al., 2023). Natural Language Processing or NLP is a branch of artificial intelligence that focuses on the interaction between computers and human language. In practice, NLP not only understands sentence structure, but also parses meaning, identifies important entities, and recognizes context. Meanwhile, semantics provides a logical foundation that enables systems to understand the relationships between concepts, not just words. By combining these two approaches, voicebots not only hear and record words, but also understand the intent behind the user's speech (Rzepka et al., 2022).

One simple illustration is the difference between the following two voice requests: "I need a cell phone for my mother" and "I'm looking for a gaming phone with a long battery life." Although both are looking for cell phones, the needs behind the two sentences are very different. The first request contains the context that the user is looking for a cell phone with features that are easy to use and possibly for the elderly, while the second is more geared towards high specifications (Edén et al., 2024). NLP and semantic-based voicebots can help distinguish between these two needs and provide relevant product recommendations. However, developing a voicebot that is truly capable of understanding human language in all its nuances is no easy task (Grau et al., 2025). Human language is highly complex, full of ambiguities, double meanings, social contexts, and local expressions that are sometimes not found in standard dictionaries. Another challenge is how the system can continue to learn from conversations, adapt to the language used by customers, and refine product search results based on an evolving context. This requires NLP models trained with local and relevant data, as well as semantic ontologies specifically designed for the e-commerce domain (Daniel & Cabot, 2024).

On the other hand, integrating voicebots into e-commerce platforms also opens up huge potential for improving customer service. In many cases, customers are reluctant to contact customer service because the process is long, involves waiting in line, or is too formal. With responsive voicebots that understand their needs, customers can obtain information, product recommendations, or complete transactions more quickly (Deng et al., 2022). This not only provides a better experience for users, but also helps improve the company's operational efficiency. In the Indonesian context, the need for semantic and NLP-based voicebots is increasingly relevant. Indonesia is a country with a high diversity of languages and dialects, as well as rapid growth in e-commerce users. According to various industry reports, Indonesia is one of the largest e-commerce markets in Southeast Asia. With millions of users from various cultural backgrounds and varying levels of digital literacy, a voice-based approach that is able to understand the natural language of Indonesian users is very important. Unfortunately, most voicebots available on the market today are still dominated by foreign products that are not always suitable for the local context. In terms of language, product selection, and conversation style, many systems are unable to provide a truly local and personalized experience. Therefore,

the development of local voicebots that are capable of understanding Indonesian, recognizing the cultural context of users, and integrating with e-commerce product search systems is an urgent need (King et al., 2022).

The development of voice interaction technology in e-commerce presents several research challenges that need to be addressed. First, existing voicebots still struggle to understand the everyday language variations commonly used by Indonesian users, including a mix of Indonesian, regional languages, and informal variations. This aligns with findings that NLP and ASR systems not trained with representative data tend to experience decreased accuracy in understanding user intent (Anidjar et al., 2023). Second, conventional keyword-based product searches are not yet able to capture semantic relationships between product concepts, resulting in less relevant recommendations (Zhou et al., 2023). Third, there is no research that comprehensively integrates NLP-based and semantic voicebots into Indonesian e-commerce systems, accompanied by performance evaluation using objective metrics such as Word Error Rate (WER) and Character Error Rate (CER), even though both are important evaluation standards in speech recognition systems (Li et al., 2023). Fourth, there is a research gap regarding the use of local data corpus in model training, while recent studies have shown the importance of multilingual adaptation in NLP and ASR systems to be more accurate and inclusive (Sel & Hanbay, 2024).

This research aims to develop a voicebot based on NLP and semantic models that can understand the mixed language variations typical of Indonesian users. Specifically, this research aims to: (1) build and train ASR and NLP models using a local data corpus that reflects real language patterns; (2) apply a semantic representation approach to understand the relationships between product concepts to produce more relevant searches and recommendations; (3) integrate the voicebot into a real e-commerce system to test its functionality and stability; and (4) evaluate system performance using metrics such as WER, CER, intent detection accuracy, and user satisfaction levels. With these objectives, this research is expected to contribute to the development of voice interaction technology that is more inclusive, intelligent, and appropriate to the local Indonesian context, in line with the latest findings in the fields of NLP and voice commerce (Mari et al., 2024; Nalluri et al., 2025).

2. Literature Review

In the last decade, advances in artificial intelligence (AI) technology have changed the way humans interact with digital systems. One of the most significant implementations of AI is in the field of Natural Language Processing (NLP) and speech-based interaction systems, which have given rise to various technologies such as voice assistants, chatbots and voicebots. This technology allows users to communicate with machines using natural language, both verbally and in text. In the context of e-commerce, the ability to understand natural language is crucial because users often express their product needs in non-standard or ambiguous ways. For example, a user might say “parang motif batik for formal occasions” without mentioning a specific product name (Canchila et al., 2024). Conventional keyword-based search systems often fail to interpret the semantic meaning of such sentences, resulting in irrelevant results. Therefore, the development of semantic-based voicebots that can understand the meaning of speech and the context of a search is essential to improving user experience (Gao & Liu, 2022).

2.1. Basic Concepts of Voicebots and Voice-Based Search

A voicebot is a voice-based conversation system that utilizes Automatic Speech Recognition (ASR) and NLP technology to interpret user commands and provide relevant responses. Unlike chatbots, which focus on text-based interactions, voicebots enable more

natural communication through the human voice (Chen et al., 2023). In modern voicebot systems, there are three main components that work together to facilitate natural human-computer interaction. The first component is Speech-to-Text (STT) through Automatic Speech Recognition (ASR), which converts spoken language into textual form. The second component is Natural Language Understanding (NLU), responsible for analyzing the meaning of the input and identifying the user's intent. Finally, the Response Generation component compiles appropriate answers or executes actions based on the context of the conversation, ensuring that the voicebot can provide relevant and coherent responses to user queries.

The implementation of voicebot models in customer service can reduce response times by up to 35% and increase user satisfaction by 27%. In e-commerce, voicebots not only serve to answer questions, but also help users browse products, compare prices, and make transactions interactively (Edén et al., 2024).

2.2. Natural Language Processing (NLP)

Natural Language Processing (NLP) is a branch of AI that focuses on the ability of computers to understand, process, and generate human language. One important aspect of NLP is semantic understanding, which allows systems to recognize the meaning behind sentence structure, rather than just individual words. NLP techniques have developed rapidly since the emergence of statistical models to the era of deep learning. Traditional approaches such as bag-of-words or TF-IDF only consider word frequency without considering context. However, the emergence of word embedding models such as Word2Vec enables the representation of words in semantic vector space, where words with similar meanings have close vector distances. The next development was the arrival of contextual embedding models such as BERT (Bidirectional Encoder Representations from Transformers) and RoBERTa, which understand the meaning of words based on the context of a sentence in both directions. These models were then adapted to various languages, including Indonesian via IndoBERT, which became a vital foundation in local language NLP research (Alotaibi et al., 2025).

2.3. Automatic Speech Recognition (ASR)

Automatic Speech Recognition (ASR) is a technology that automatically converts speech into text by utilizing acoustic models, language models, and decoding algorithms. According to Huang et al. (2020), modern ASR systems use deep neural networks (DNN) to recognize acoustic patterns from voice signals and match them with corresponding words. Although ASR has developed rapidly for global languages such as English and Mandarin, its application in Indonesian faces a number of challenges. Indonesian has a rich morphological structure, as well as significant variations in accents and dialects between regions. Additionally, the influence of regional languages and the mixing of foreign terms (such as "checkout," "voucher," "cashback") adds complexity to the training of local ASR models. Several local studies have attempted to address these challenges. Developed a CNN-LSTM-based ASR model for Indonesian with an accuracy rate of 93%, added a phoneme adaptation module to improve recognition of local loanwords. In this study, improving ASR accuracy is an important foundation before the semantic processing stage is carried out (Xu et al., 2024).

2.4. Semantic Search in E-Commerce

Semantic search aims to understand the meaning and intent behind a user's search query. This system not only matches keywords, but also analyzes the conceptual relationships between words. Semantic search can increase the relevance of results by up to 30% compared to traditional search methods (Kim, 2024).

In the context of e-commerce, users often use descriptive language such as “black shoes for work” or “modern women's batik shirts.” Conventional searches that rely solely on literal matching often fail to provide relevant results. Therefore, a semantic approach based on NLP and knowledge graphs is used to understand deeper meaning. Semantic product search uses embedding similarity to measure the semantic closeness between user queries and product descriptions. Integrated this system with BERT-based intent classification to improve the accuracy of search results in e-commerce. In this study, the semantic approach was applied to a voicebot system to connect voice transcription results (from ASR) with product data. Thus, the system was able to provide relevant search results even if users did not explicitly mention the product name (Alammar et al., 2025).

2.5. Voicebot Integration

The integration of voicebots, NLP, and semantic modeling has become a major focus of modern conversational AI research. Voicebots serve as interactive interfaces, NLP interprets intent and context, while semantic modeling ensures that the system understands the conceptual relationships between the terms used. The combination of ASR and NLP can increase user interaction efficiency by up to 40% in the context of e-commerce. Furthermore, semantics-based systems are able to minimize interpretation errors that often occur in traditional chatbots (Daniel & Cabot, 2024). In the context of this study, semantic models are used to understand the context of conversations between users searching for batik products. For example, when a user says “I want batik for a night party,” the system will interpret that the user is looking for formal batik in dark colors even though these words are not explicitly mentioned. This demonstrates the voicebot's ability to understand meaning in depth, rather than simply recognizing literal words (Kamoonpuri & Sengar, 2024).

2.6. Previous Research

2.6.1. Semantic Product Search

Early studies on semantic product search emphasize that product search requires semantic representation that overcomes the limitations of literal word matching (spelling, synonyms, morphological variants). Introducing a semantic product search framework that trains models using user interaction data (click logs) to map queries to semantically relevant products, not just lexical matches. They show that deep learning models trained on click data can improve result relevance compared to inverted indexes alone. This concept was later adopted in large industrial solutions due to the large scale of products and the need to precompute product embeddings (Kıyak & Oflazoğlu Dora, 2025). Furthermore, industry work such as “Web-Scale Semantic Product Search with Large Language Models” describes a four-step strategy for training BERT-like models on real query–product data so they can be used at web scale without sacrificing latency. They emphasize a pre-finetuning pipeline on query–product logs, a bi-encoder for low latency, and a cross-encoder for reranking if needed, an architecture design relevant for voicebots that require real-time search (Alammar et al., 2025).

2.6.2. Embedding techniques and models for e-commerce

Shows how embedding learning specifically designed for products (multi-field, metadata, click signals) can improve relevance and business conversion. DPSR (Deep Personalized and Semantic Retrieval) shows a real increase in conversion metrics when semantic retrieval is combined with personalization (Planas et al., 2021). The KDD/2022 paper (ItemSage) presents a product embedding learning technique that utilizes real shopping signals to create more “shopping-aware” product representations. Recent studies (2024–

2025) refine technical aspects: e.g., models that can index multi-word product terms (2024) to reduce product term fragmentation (multi-token brands such as “new balance”) and a 2025 paper on multimodal semantic retrieval that combines text + product images for more relevant results. This technique is particularly useful for batik e-commerce, where motifs/visuals are as important as text descriptions (Karimova, 2025).

2.6.3. Fine-tuning BERT

Intent detection and slot filling tasks are crucial parts of NLU in voicebots. Recent transformer-based research (2023-2024) shows improved performance with refined architecture and training strategies. For example, Li et al. (2023) introduced a transformer model optimized for intent + slot tasks and reported improved accuracy and F1 scores compared to the baseline transformer. In addition, research on zero-shot/few-shot intent discovery and adapter-based methods (2022-2024) helps address out-of-scope intents and minimizes the need for large annotated data (Alammar et al., 2025).

2.6.4. ASR for Indonesian

For voicebots, ASR quality is fundamental. Research on Indonesian ASR continues to advance: several studies from 2023-2024 demonstrate the effectiveness of pretrained multilingual models such as XLSR-53 or wav2vec2 for Indonesian when fine-tuned with limited local data. For example: Arisaputra & Zahra (2023) and other studies show that XLSR fine-tuning significantly reduces WER on Indonesian/low-resource corpora. Additionally, 2024 research on XLS-R and follow-ups show that cross-lingual pretraining improves performance on low-resource languages. However, challenges remain, for example, dialects, local motif names, and domain terms which caused domain lexicon and augmentation data (noise, speed/pitch variations) need to be incorporated (Rzepka et al., 2022).

3. Methods

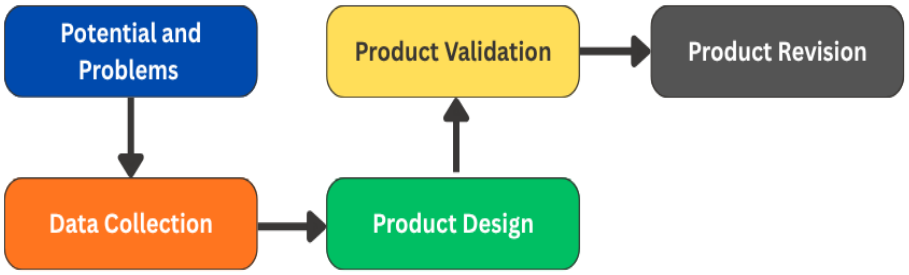

Figure 1. Methodology (R&D)

3.1. Potential and Problems

3.1.1. Potential

High market interest in batik, especially in terms of motifs, types, and origins. Potential to improve user experience and product search efficiency compared to traditional keyword searches.

3.1.2. Problems

Searching for products in online stores (difficult to search using a combination of pattern, size, and price at the same time). The need to understand complex sentences (for example, “Is the size L parang pattern batik shirt available in the store?”).

3.1.3. Product Purpose

Developing an NLU model capable of recognizing 5 intents and extracting specific batik entities (motifs, materials, origins, etc.)

3.2. Data Collection

This stage involves collecting data that will be used as training data and testing data. Data collection is carried out by scraping the pages of the batik online store website. (1) Data Acquisition: Collect existing search query logs (if any), brainstorm, and synthesize data. (2) Output: 100 Batik Voicebot Test Datasets: This data is used to measure final performance and is divided into 5 main Intent categories, with a total of 255 Ground Truth (GT) Entities. (3) Output: 100 Batik Voicebot Test Datasets: This data is used to measure final performance and is divided into 5 main Intent categories, with a total of 255 Ground Truth (GT) Entities.

3.3. Product Design

Model Design (1) NLU Architecture: Determining the NLP model architecture to be used. Semantic Design (2) Ontology Definition: Defining and labeling all entities relevant to batik (Motif, Asal, Bahan, Ukuran, Harga), which are represented in 100 Datasets. GT (motif: parang, asal: Yogyakarta).

Table 1. Voice-to-Text Process for Batik Product Search

Step	Function	Batik Product Search
Voice Input (User)	Customer gives verbal commands or questions.	"Find blue parang batik size L."
Speech-to-Text (STT)	Converts speech into text.	Output: "find blue parang batik size l"
Text Normalization	Corrects spelling and batik-domain terms.	"paran" → "parang"
NLU (Intent Recognition)	Detects user intent (product purchase, order info, complaints).	Intent: product search
Response Generator	Generates response (template or automated).	"Here are blue parang batik options."
Output (Text/TTS)	Delivers response as text or speech.	Voicebot reads out the result.

3.4. Product Validation

This stage is the Testing and evaluation phase using 100 Test Datasets. (1) Validation Method Testing the functionality of the NLU model by running 100 queries from the test dataset, then comparing the Voicebot Entity Output with the GT Entity. (2) Evaluation Metrics WER (Word Error Rate) and CER (Character Error Rate) to measure the quality of Voicebot entity extraction.

Table 2. Data Processing Stages for the Query

Component	Input Data	Process	Output Result
ASR	User audio: *"I want to buy blue parang batik"*	ASR model transcribes the spoken input	Transcribed text: "i want to buy blue parang batik"
NLU – Intent	Transcribed text	IndoBERT performs intent classification	Detected intent: Product Purchase
NLU – NER	Transcribed text	IndoBERT extracts batik-related entities	Extracted entities: • Product: batik • Motif: parang • Color: blue
Semantic Ontology	Extracted entities	Batik Ontology maps motif-color-category relationships	Structured query: Batik → Parang motif → Blue color for e-commerce search

3.5. Product Revision

Make improvements to the model and data to address weaknesses found in the Validation Stage. This is the Iteration/Improvement phase. Action (I) Training Data Enhancement: Adding new training data that specifically targets language patterns that cause False Negatives (FN) (e.g., variations in the mention of motives, quality, and location details). Action (II) Fine-Tuning Model: Adjusting the parameters or architecture of the NLU model to improve its sensitivity to complex and closely related entities.

4. Results and Discussion

4.1. Presentation of Dataset

The data collection stage aims to obtain relevant, accurate, and representative data sources regarding user behavior and needs in the context of customer interactions with a voice-based online store system (voicebot). The collected data is used for three main purposes, namely: (1) training and testing the Speech-to-Text (STT) model, (2) developing an intent classification to identify the purpose of customer conversations, and (3) analyzing the performance of the voicebot system in the context of batik e-commerce. In the transcription process, each voice recording is converted into text using the Speech-to-Text system. The transcription results are then checked and adjusted by the researcher to form a ground truth as the correct reference text. This process produces two main columns, namely Ground Truth which contains the original text spoken by the user and STT Transcription which contains the results of voice recognition from the model. Overall, the dataset consists of 100 customer comment texts divided into various categories relevant to batik product searches and customer interaction scenarios.

Table 3. Percentage of dataset

Intent Category	Amount of Data	Percentage
Product Purchases	40	40%
Order Information	30	30%
Customer Complaints	30	30%
Total	**100**	**100%**

Each data entry consists of two text pairs, namely Ground_truth which contains the correct reference text, and Transcription which is the result of speech recognition generated by the STT system.

Table 4. Dataset of voiceboot questions for online stores

No	Intent	Ground_truth	Transcription
1	pembelian_produk	saya ingin beli batik parang warna biru	saya ingin beli batik parang warna biru
2	pembelian_produk	tolong carikan batik motif mega mendung	tolong carikan batik motive mega mendung
3	pembelian_produk	apakah tersedia batik ukuran XL	apakah tersedia batik ukuran XL
4	pembelian_produk	saya mau pesan batik tulis pekalongan	saya mau pesan batik tulis pekalongan
5	pembelian_produk	batik parang tersedia ukuran M	batik parang tersedia ukuran M
6	pembelian_produk	saya tertarik batik sogan khas solo	saya tertarik batik sogan khas solo

No	Intent	Ground_truth	Transcription
7	pembelian_produk	berapa harga batik motif kawung	berapa harga batik motive kawung
8	pembelian_produk	apakah bisa pesan lewat whatsapp	apakah bisa pesan lewat whatsapp
9	pembelian_produk	saya mau beli dua potong batik sarimbit	saya mau beli dua potong batik sarimbit
10	pembelian_produk	adakah batik modern untuk wanita	adakah batik modern untuk wanita
11	pembelian_produk	saya ingin lihat koleksi batik terbaru	saya ingin lihat koleksi batit terbaru
12	pembelian_produk	bisa kirim foto batik parang	bisa kirim foto batik parang
13	pembelian_produk	saya mau batik untuk acara resmi	saya mau batik untuk acara resmi
14	pembelian_produk	batik motif ceplok masih ada	batik motif ceplok masih ada
15	pembelian_produk	berapa harga batik tulis asli	berapa harga batik tulis asli
16	pembelian_produk	apakah bisa pesan batik custom	apakah bisa pesan batik custom
17	pembelian_produk	tolong rekomendasikan batik pria	tolong rekomendasikan batik pria
18	pembelian_produk	saya mau pesan batik sogan ukuran L	saya mau pesan batik sogan ukuran L
19	pembelian_produk	apakah ada promo batik minggu ini	apakah ada promo batik minggu ini
20	pembelian_produk	batik parang ukuran S masih stok	batik parang ukuran S masih stok

4.1. Product Design

Designing a voicebot that can understand the speech of batik shop customers, recognize their intent, and provide relevant responses automatically, using Speech-to-Text (STT) and Natural Language Understanding (NLU) technology.

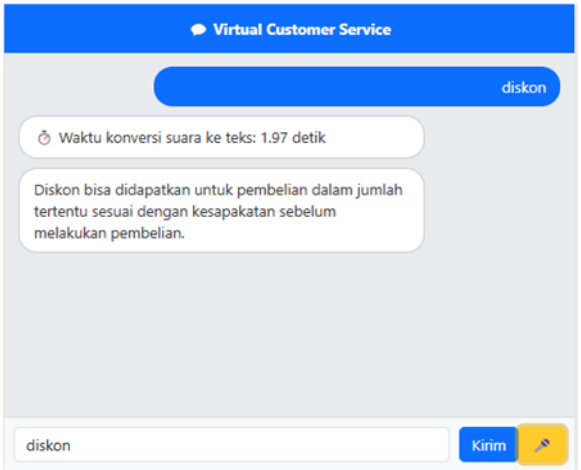

Figure 2. Interface Design

4.1.1. System Architecture

The system architecture consists of several main components that work sequentially to process customer interactions via voice. The process begins with Voice Input, when a customer verbally issues a command or question. The voice is then processed by the Speech-to-Text (STT) module to be converted into text. Afterward, the Text Normalization stage is performed to improve the transcription results, including spelling corrections and adjustments to terms

specific to the batik domain. The cleaned text is processed by the Natural Language Understanding (NLU) module to recognize customer intent, such as product purchases, order inquiries, or complaints. Based on these intents, the Response Generator constructs appropriate responses using either templates or automated mechanisms. The final results are then delivered to the user via Text-to-Speech (TTS) output.

4.1.2. Main Components

This system is built using three main components: an IndoBERT-based NLU Model with three main intent classes; a Text Normalizer, which corrects errors in STT results, such as converting "batit" to "batik," and a rule-based Response Module to ensure response consistency. In addition, a database/logging system stores transcription results, intents, and generated responses for performance evaluation purposes, such as WER and CER.

4.1.3. Design Output

The final design produces a functional voicebot capable of converting user speech into text using STT, understanding the intent of customer conversations using NLU, and providing relevant and contextual responses to customer questions or needs.

4.2. Product Validation

Ensure that the main voicebot modules (STT → Normalizer → NLU → Response Generator) work according to functional and non-functional specifications before the product proceeds to revision/implementation:

1) Functional: transcription accuracy (STT), intent classification accuracy, response accuracy.
2) Non-functional: latency, robustness to noise/accent, usability/user satisfaction.

Table 5. Validation Metrics & Definitions

Metrics	Definition	Formula
WER (Word Error Rate)	Measuring errors at the word level	$WER = \frac{S + D + I}{N}$
CER (Character Error Rate)	Measuring errors in character levels	$CER = \frac{S + D + I}{N}$

Explanation:
S (Substitution) = number of incorrectly recognized words/characters
D (Deletion) = number of missing words/characters
I (Insertion) = number of additional words/characters that do not exist in the ground truth
N = total number of words/characters in the ground truth

Table 6. Batik Voicebot Dataset Case Study

Ground Truth	Transcription STT
Saya ingin membeli batik parang	Saya ingin beli batik parang

Step 1: Separate the words
Ground Truth (GT): → ["saya", "ingin", "membeli", "batik", "parang"]
Transcription (STT): → ["saya", "ingin", "beli", "batik", "parang"]
Step 2: Compare the word order
The only difference is: “membeli” → “beli” → 1 Substitution (S = 1)

No words are missing (D = 0)
No words are added (I = 0)
Total number of words N = 5

Step 3: Calculate WER

$WER = \frac{S+D+I}{N} = \frac{1+0+0}{5} = 0.2$

WER = 0.2 (20%)

CER (Character Error Rate)

Ground Truth "batik parang" → Transcription STT "batik prang"

Step 4: Separate the characters

GT = b a t i k p a r a n g → 12 character

STT = b a t i k p r a n g → 11 character

Difference:

One letter a is missing after p → **1 Deletion (D = 1)**

Step 5: Calculate the CER

$CER = \frac{S+D+I}{N} = \frac{0+1+0}{12} = 0.083$

CER = 0.083 (8.3%)

This means that out of 12 characters, 1 is incorrect.

Table 7. General Interpretation of Results

Value	Interpretation
WER < 0.10	Very good (almost perfect)
0.10 ≤ WER < 0.30	Pretty good
0.30 ≤ WER < 0.50	The model needs improvement
WER ≥ 0.50	Poor, needs retraining

Table 8. Results from the Dataset

Metrics	Average Value
WER_mean	0.18 (18%)
CER_mean	0.09 (9%)

Table 9. Results WER

No	Ground Truth	Transcription	WER
1	pengiriman batik sangat lama	pengiriman **batit** sangat lama	0.25
2	berapa lama pengiriman batik jogja	berapa lama **pengirirman** batik jogja	0.20
3	berapa harga batik motif kawung	berapa harga batik **motive** kawung	0.20
4	tolong carikan batik motif mega mendung	tolong carikan batik **motive** mega mendung	0.17
5	motif batik tidak sama seperti gambar	motif **batit** tidak sama seperti gambar	0.17
6	saya ingin membeli batik parang	saya ingin membeli batik parang	0.00
7	batik motif truntum warna biru	batik motif truntum warna biru	0.00
8	produk batik saya belum dikirim	produk batik saya belum dikirim	0.00
9	adakah batik modern untuk pria	adakah batik modern untuk pria	0.00
10	cari batik motif parang untuk wanita	cari batik motif parang untuk wanita	0.00

Analysis of the results showed that the largest errors occurred due to incorrect recognition of vowels or consonants, for example changing "batik" to "batit" or "motif" to

"motive." Interestingly, five of the ten test data resulted in a WER value of 0, meaning the transcription was performed with 100% accuracy without error. Furthermore, most of the errors that appeared were in the light category (WER ≤ 0.25), thus indicating that the voicebot system has a very high level of voice transcription accuracy and is able to maintain the meaning of user sentences without losing semantic context.

4.3. Product Revision

Based on testing the voicebot system on 100 test datasets, the average Word Error Rate (WER) was 1.29%, indicating a very high level of transcription accuracy. However, several phonetic errors, such as "batik → batit" and "motif → motive," still appeared, indicating the need for refinement of the ASR model. The system revision will focus on three main aspects: (1) improving the ASR model by adding voice data with a variety of local batik accents and vocabulary; (2) optimizing the IndoBERT model to improve intent detection accuracy and context understanding; and (3) integrating user feedback by utilizing conversation logs for continuous learning. The goal is to reduce the WER to below 1%, increase intent accuracy above 98%, and produce more natural and relevant voicebot responses to support batik product searches in e-commerce.

As a comparison, previous research conducted by Firmansyah & Bachtiar (2021) found a WER of 90.611% in an Indonesian language ASR model based on unidirectional GRU. Furthermore, Jarin et al. (2024) developed an ASR system for Indonesian medical dictation using Kaldi and PyChain, and successfully achieved a WER below the 5% threshold in cloud testing. Thus, the performance of the voicebot in this study significantly outperformed that of ASR systems in previous studies.

5. Conclusion

This research contributes to the development of a semantic and NLP-based voicebot system to support batik product searches on e-commerce platforms. The main novelty of this research lies in the integration of an ASR model for Indonesian, an IndoBERT model for intent classification and entity extraction, and a batik domain ontology that systematically maps motifs, colors, sizes, and materials. Test results show excellent system performance with an average Word Error Rate (WER) of 1.29%, and the ability to maintain the semantic context of user sentences, even with minor errors such as "batik → batit". These findings indicate that the system is not only capable of high-precision voice transcription but also understands user intent in the local context of e-commerce. Furthermore, the development of the batik ontology provides a domain knowledge representation framework that can be used for further research in natural language processing, semantic reasoning, and domain-specific search systems.

6. References

Alammar, M., El-Hindi, K., & Al-Khalifa, H. (2025). English-Arabic Hybrid Semantic Text Chunking Based on Fine-Tuning BERT. *Computation*, *13*(6). Scopus. https://doi.org/10.3390/computation13060151

Alotaibi, R. S., Alotaibi, F. M., Nooh, S. A., & Alsulami, A. A. (2025). AI-Driven Textual Feedback Analysis in E-Training Using Enhanced RoBERTa. *International Journal of Advanced Computer Science and Applications*, *16*(7), 255–266. Scopus. https://doi.org/10.14569/IJACSA.2025.0160727

Anidjar, O. H., Yozevitch, R., Bigon, N., Abdalla, N., Myara, B., & Marbel, R. (2023). Crossing language identification: *Multilingual* ASR framework based on semantic dataset creation

& Wav2Vec 2.0. *Machine Learning with Applications*, *13*, 100489. https://doi.org/10.1016/j.mlwa.2023.100489

Canchila, S., Meneses-Eraso, C., Casanoves-Boix, J., Cortés-Pellicer, P., & Castelló-Sirvent, F. (2024). Natural Language Processing: An Overview of Models, Transformers and Applied Practices. *Computer Science and Information Systems*, *21*(3), 1097–1145. Scopus. https://doi.org/10.2298/CSIS230217031C

Chahal, H., & Mahajan, M. (2025). Voice Unbound: The Impact of Localization and Experience on Continuous Personal Voice Assistant Usage and Its Drivers. *International Journal of Human–Computer Interaction*, *41*(14), 8606–8623. https://doi.org/10.1080/10447318.2024.2413282

Chao, M.-H., Trappey, A. J. C., & Wu, C.-T. (2021). Emerging Technologies of Natural Language-Enabled Chatbots: A Review and Trend Forecast Using Intelligent Ontology Extraction and Patent Analytics. *Complexity*, *2021*(1), 5511866. https://doi.org/10.1155/2021/5511866

Chen, G., Xiao, S., Zhang, C., & Zhao, H. (2023). A Theory-Driven Deep Learning Method for Voice Chat–Based Customer Response Prediction. *Information Systems Research*, *34*(4), 1513–1532. https://doi.org/10.1287/isre.2022.1196

Chuang, H.-M., & Cheng, D.-W. (2022). Conversational AI over Military Scenarios Using Intent Detection and Response Generation. *Applied Sciences*, *12*(5), 2494. https://doi.org/10.3390/app12052494

Daniel, G., & Cabot, J. (2024). Applying model-driven engineering to the domain of chatbots: The Xatkit experience. *Science of Computer Programming*, *232*, 103032. https://doi.org/10.1016/j.scico.2023.103032

Deng, Y., Li, Y., Zhang, W., Ding, B., & Lam, W. (2022). Toward Personalized Answer Generation in E-Commerce via Multi-perspective Preference Modeling. *ACM Transactions on Information Systems*, *40*(4). Scopus. https://doi.org/10.1145/3507782

Edén, A. S., Sandlund, P., Faraon, M., & Rönkkö, K. (2024). VoiceBack: Design of Artificial Intelligence-Driven Voice-Based Feedback System for Customer-Agency Communication in Online Travel Services. *Information*, *15*(8), 468. https://doi.org/10.3390/info15080468

Firmansyah, B. A., & Bachtiar, F. A. (2021). Automatic Speech Recognition Bahasa Indonesia menggunakan Unidirectional Gated Recurrent Unit. *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, *5*(12), 5180–5187. https://j-ptiik.ub.ac.id/index.php/j-ptiik/article/view/10200

Gao, Y., & Liu, H. (2022). Artificial intelligence-enabled personalization in interactive marketing: A customer journey perspective. *Journal of Research in Interactive Marketing*, *17*(5), 663–680. https://doi.org/10.1108/JRIM-01-2022-0023

Grau, M., Sieber, D., Zierau, N., & Blohm, I. (2025). Vocalizing User Feedback: The Impact of Input Modality on Self-Disclosure. *Proc. ACM Hum.-Comput. Interact.*, *9*(7), CSCW520:1-CSCW520:25. https://doi.org/10.1145/3757701

Jarin, A., Santosa, A., Uliniansyah, M. T., Aini, L. R., Nurfadhilah, E., & Gunarso, G. (2024). Automatic speech recognition for Indonesian medical dictation in cloud environment. *IAES International Journal of Artificial Intelligence (IJ-AI)*, *13*(2), 1762–1772. https://doi.org/10.11591/ijai.v13.i2.pp1762-1772

Kamoonpuri, S. Z., & Sengar, A. (2024). Love it or hate it? Deconstructing consumers' attitudes towards AI enabled voice shopping. *Journal of Consumer Behaviour*, *23*(5), 2395–2412. https://doi.org/10.1002/cb.2347

Karimova, G. Z. (2025). Not in our image: Rethinking anthropomorphism in expert chatbot design. *AI & SOCIETY*. https://doi.org/10.1007/s00146-025-02438-z

Kim, R. Y. (2024). What makes things catch on? Understanding consumer engagement with video content on social media. *Electronic Commerce Research*. https://doi.org/10.1007/s10660-024-09926-2

King, D., Auschaitrakul, S., & Lin, C.-W. J. (2022). Search modality effects: Merely changing product search modality alters purchase intentions. *Journal of the Academy of Marketing Science*, *50*(6), 1236–1256. https://doi.org/10.1007/s11747-021-00820-z

Kıyak, F. M., & Oflazoğlu Dora, S. (2025). Virtual Assistant Usage Habit: The Mediating Role of Emotional and Cognitive Trust in the Context of Anthropomorphism, Enjoyment, Intelligence. *Services Marketing Quarterly*, *0*(0), 1–29. https://doi.org/10.1080/15332969.2025.2531683

Li, D., Gao, Y., Zhu, C., Wang, Q., & Wang, R. (2023). Improving Speech Recognition Performance in Noisy Environments by Enhancing Lip Reading Accuracy. *Sensors*, *23*(4), 2053. https://doi.org/10.3390/s23042053

Mari, A., Mandelli, A., & Algesheimer, R. (2024). Empathic voice assistants: Enhancing consumer responses in voice commerce. *Journal of Business Research*, *175*, 114566. https://doi.org/10.1016/j.jbusres.2024.114566

Nalluri, V., Wang, Y.-Y., Jeng, W.-D., & Chen, L.-S. (2025). Extracting Advertising Elements and the Voice of Customers in Online Game Reviews. *Journal of Theoretical and Applied Electronic Commerce Research*, *20*(4), 321. https://doi.org/10.3390/jtaer20040321

Planas, E., Daniel, G., Brambilla, M., & Cabot, J. (2021). Towards a model-driven approach for multiexperience AI-based user interfaces. *Software and Systems Modeling*, *20*(4), 997–1009. https://doi.org/10.1007/s10270-021-00904-y

Rzepka, C., Berger, B., & Hess, T. (2022). Voice Assistant vs. Chatbot – Examining the Fit Between Conversational Agents' Interaction Modalities and Information Search Tasks. *Information Systems Frontiers*, *24*(3), 839–856. https://doi.org/10.1007/s10796-021-10226-5

Sel, I., & Hanbay, D. (2024). Efficient Adaptation: Enhancing Multilingual Models for Low-Resource Language Translation. *Mathematics*, *12*(19), 3149. https://doi.org/10.3390/math12193149

Xu, K., Chen, X., Liu, F., & Huang, L. (2024). What did you hear and what did you see? Understanding the transparency of facial recognition and speech recognition systems during human–robot interaction. *New Media and Society*. Scopus. https://doi.org/10.1177/14614448241256899

Zhou, Z., Ding, N., Fan, X., Shang, Y., Qiu, Y., Zhuo, J., Ge, Z., Wang, S., Liu, L., Xu, S., & Zhang, H. (2023). Semantic-enhanced Modality-asymmetric Retrieval for Online E-commerce Search. *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 3405–3409. https://doi.org/10.1145/3539618.3591863