HumaniCA: A Benchmark Resource for the Detection of Users' Ascription of Humanness to Conversational Agents

Sabrina Villata*, Amon Rapp, Luigi Di Caro, Federica Cena
University of Torino
Department of Computer Science, Corso Svizzera 185, Torino 10145, Italy
{sabrina.villata, amon.rapp, luigi.dicaro, federica.cena}@unito.it

Abstract

Anthropomorphizing, which involves attributing human-like characteristics to non-human entities, is common in users' conversations with text-based conversational agents and can lead to a misalignment between the users' expectations and the agent's actual capabilities. Detecting users' ascriptions of humanness automatically may enable systems to identify when users adopt a human-like style when conversing with an agent and to adapt its responses accordingly to tune their expectations. In this paper, we introduce HumaniCA, a benchmark resource comprising three annotated datasets of user turns from real dialogues with three different types of conversational agents (task-oriented, Q&A, and LLM-based) aimed at indicating whether the user is ascribing humanness to the conversational agent. We also identified a set of linguistic indicators of user ascription of humanness to conversational agents and validated their utility with benchmark experiments. We then compared performance of our linguistic features and other well-known textual features (TF-IDF weights and SentenceBERT word embeddings), as well as their combinations. The evaluation highlights the central role of our linguistic features: whether used individually or in combination, they consistently achieve higher accuracy across all agent types.

Keywords: Conversational Agent, Chatbot, Humanization, Linguistic Indicators, SentenceBERT

1. Introduction

A conversational agent is a virtual agent that uses natural language to interact with users, either through text or speech (Jurafsky and Martin, 2000). These agents leverage artificial intelligence to make interactions more efficient by mimicking human conversational patterns or appearance, enhancing their ability to perform tasks typically done by humans (Crolic et al., 2022; Seymour and Van Kleek, 2021).

Recently, conversational agents based on Large Language Models (LLMs), such as OpenAI's ChatGPT, have become widespread (Zhao et al., 2025). Their human-like text generation often leads users to ascribe human-like characteristics to them, more than non-LLM-based agents (Luger and Sellen, 2016). This phenomenon, known as humanization or anthropomorphizing, may lead users to have unrealistic expectations about an agent's abilities (Chaves and Gerosa, 2021), creating a significant misalignment between their expectations and the agent's actual performance (Luger and Sellen, 2016). Despite rapid advancements, LLM-based agents still suffer from issues such as inappropriate responses, lack of context awareness, misunderstandings (Jain et al., 2018), and hallucinations (Ji et al., 2023), which can disappoint users and even be harmful, especially in collaborative situations where they can lead to poor decisions (Nguyen et al., 2022).

To reduce the gap between users' expectations and an agent's actual capabilities, it is crucial to detect when a user is anthropomorphizing the agent. Such detection would allow the agent to adjust its responses accordingly, by correcting the user's misperceptions, encouraging them to adopt a more effective communication strategy, and thus lowering the likelihood of abandoning the conversation. However, identifying user humanization is typically a post-hoc, manual task, consisting in cumbersome and time-consuming qualitative (Jain et al., 2018; Rapp et al., 2024) or quantitative analysis (Araujo and Bol, 2024) of human-agent conversations, or in collecting self-reported data from questionnaires and interviews. However, there are currently no resources that enable the automatic detection of user humanization.

In this paper, we address this gap by introducing *HumaniCA*, a benchmark resource specifically designed to support and evaluate automatic detection of users' ascriptions of humanness to conversational agents. In particular, we identified a set of linguistic indicators associated with the users' tendency to anthropomorphize an agent, based on previous studies, and formalized them into a suite of heuristic features designed to capture interpretable cues of human-like conversational strategies, such as politeness, social references, and personal information disclosure.

To assess their utility, the heuristic features have been compared to TF-IDF weights, named frequency-based features, and SentenceBERT word embeddings, named semantic-based features. Then, we compiled three datasets of users' turns from real-world human-agent dialogues covering three different types of conversational agents

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 2825–2835
11–16 May 2026.

(task-oriented, Q&A, LLM-based), which we annotated with binary labels indicating whether the user is ascribing human-like characteristics to the agent. In total, HumaniCA includes 2,100 user turns.
We then trained baseline machine learning models (Naïve Bayes, Logistic Regression, and Support Vector Machine) to evaluate the predictive power of each feature set and their combinations in automatically detecting user humanization. Our findings demonstrate that the heuristic features play a key role in distinguishing when users anthropomorphize a conversational agent, exhibiting superior performance across all agent types.

2. Related Work

2.1. User linguistic features detection

Detecting user features from human-agent conversations is challenging, especially for large-scale applications, as collecting user profiles is costly and time-consuming. However, user utterances can convey significant information beyond their semantic content (Mairesse et al., 2007). Previous work has used baseline machine learning to detect users' internal states, such as emotions, from linguistic features in human-agents' conversations. For instance, Sekhar et al. (2021) detected users' emotions from word frequency in conversations with agents, using standard Natural Language Processing (NLP) libraries to split user text based on keywords and compare them with an emotion classifier. Mairesse et al. (2007) detected users' personality traits (Big Five) in conversations using lexical, stylistic, and syntactic markers with classification and regression models. Their results showed that choosing the right features for each model is important, as their effectiveness varies by trait.
Similarly, D'mello et al. (2008) detected learners' affective states in tutoring systems from conversational features, such as timing, verbosity, and conceptual quality, confirming that these features can predict boredom, confusion, flow, and frustration. Ferrod et al. (2021) successfully applied a neural model for short-text classification to detect user expertise in real-time conversations. Finally, Kuvar et al. (2023) trained a tree-based model to detect task-unrelated thoughts from manually labeled conversations and various features to find the best feature combination and model for the detection.
Other studies have focused on detecting users' attitudes towards conversational agents. Zhao et al. (2016) detected conversational strategies like self-disclosure, shared experience, praise, and violation of social norms using Support Vector Machine, Naïve Bayes, and Logistic Regression models, achieving high accuracy scores. Novielli et al. (2010) detected users' social attitudes in dialogues with an embodied conversational agent, mapping seven indicators of social attitude to 32 linguistic categories, employing a Bayesian classifier.
While these studies detect various user states through linguistic features, none investigate users' communicative behavior to infer the humanization of the conversational agent.

2.2. Conversational agents humanization

The humanization of conversational agents is studied with different goals: some analyze characteristics and cues that make agents appear human-like, such as identity, style, verbal and non-verbal behaviors (Go and Sundar, 2019), while others focus on how humanness affects user satisfaction (Luger and Sellen, 2016), trust, and engagement (Jain et al., 2018; Crolic et al., 2022). Nowadays, conversational agents can exhibit personality traits (Mairesse and Walker, 2009), social skills (Niculescu and Banchs, 2019), empathy (Zhou et al., 2020), and human-like conversational behaviors like politeness (Mukherjee et al., 2023).
Some studies show that anthropomorphizing an agent can create significant mismatches between user expectations and the agent's actual performance, affecting user satisfaction (Luger and Sellen, 2016). Jain et al. (2018) found that conversational agents with human-like conversational styles increase trust and engagement but also raise expectations about the agent's capabilities, suggesting that agents should clarify their capabilities to manage user expectations. Luger and Sellen (2016) noted that non-expert users often overestimate conversational agents' intelligence, expecting them to understand any request. Also, Crolic et al. (2022) showed that anthropomorphizing conversational agents does not always improve satisfaction: angry costumers may be less satisfied when agents exhibit human-like features, affecting their intention to make purchases.
User's perception of conversational agents' capabilities significantly impacts interaction quality and dialogue flow, influencing conversation success. Unrealistic expectations can hinder dialogue, leading to user frustration and dissatisfaction (Luger and Sellen, 2016). Moreover, excessively anthropomorphizing an agent may lead to unintended consequences, such as increased disclosure of private data due to heightened trust and perceived social presence (Thomaz et al., 2020). For this reason, detecting when users ascribe humanness to an agent may be essential, enabling designers to adjust the agent's behavior and provide timely strategies to realign user expectations. However, achieving such detection requires dedicated annotated resources.

2.3. Human-like communication behaviors with agents

This section reviews studies that identify communication patterns linked to users' ascription of humanness to conversational agents.
Cho (2018) analyzed users' interactions with Google Home Assistant, distinguishing two communication strategies or "models":

- Push Model, used by users who assume the agent has human-like abilities, by providing detailed information through complex sentences, believing this helps the agent understand better (e.g., "*Hey Google, can you tell me who Sports Illustrated predicts will make it to the NBA playoffs in 2018?*").
- Pull Model, when users treat the assistant like a search engine, using simple, imperative questions (e.g., "*Tell me the highest rated TVs*").

Likewise, Rapp et al. (2024) analyzed dialogues between users and a customer care chatbot, identifying linguistic features indicating users' perception of the chatbot's humanness:

- High humanness: greetings, formalities, politeness, verbose explanations, and second-person pronouns (in Italian).
- Medium humanness: imperative tone, minimal information, second-person pronouns, formal tone, and complex sentences.
- Low humanness: Imperative requests and referring to the chatbot as "*it*".

As already seen, Novielli et al. (2010) identified seven linguistic signs of social attitude towards chatbots, indicating humanization:

- Friendly introduction: greetings, self-introduction.
- Colloquial style: use of non-lexical component in conversations (e.g., "*!*"), terms from spoken language or diminutive expressions.
- Self-disclosure: first-person pronouns.
- Questions about the agent: second-person pronouns referring to the chatbot.
- Positive or negative comments: expressions of agreement/disagreement, evaluations about the agent.
- Politeness: courtesy and encouragement expressions.
- Friendly farewell: expressions of farewell and thanksgiving.

Automatic detection of communication behaviors from text remains challenging: some features like request complexity and imperative tone can be identified easily, while politeness and self-disclosure require more nuanced analysis. To gain insights, we expanded our research beyond conversational agents.

We considered the work of Danescu-Niculescu-Mizil et al. (2013), who developed a framework to identify politeness in Wikipedia and Stack Exchange, assigning politeness scores from -1 to 1 to various linguistic elements. Features indicating greater politeness include gratitude expressions (e.g., "*I appreciate*"), greetings, positive lexicon (e.g., "*That is great!*"), apologies, and conditional verbs. Less polite behaviors include negative lexicon (e.g., "*If you are going to accuse me...*"), direct questions (e.g., "*What is your native language?*"), and indicative verbs. For self-disclosure, we considered Wang et al. (2016), who trained a machine-learning model to measure self-disclosure on social networks, using features from theories about personal information disclosure:

- Positive and negative emotions: revealing personal feelings, detectable through sentiment analysis.
- Social distance: mentioning close relationships, relatives, nicknames (e.g., "*darling*"), or first names.

We draw on communication behaviors and linguistic features from these studies to build the set of heuristic features included in our resource to support the automatic detection of users' tendency to humanize conversational agents, as detailed in the next section.

3. HumaniCA

Our goal was to create HumaniCA, a benchmark resource that enables the automatic detection of users' ascription of humanness to conversational agents. To achieve this, we annotated user turns from real-world human-agent dialogues to identify whether users humanize the agent, and we designed a set of heuristic features inspired by previous literature. These feature set is the foundation of our resource and is later evaluated through baseline machine-learning models for humanization detection and compared to other sets of features, TF-IDF and SentenceBERT.

3.1. Heuristic features

To build a resource to detect humanization in users' turns, we identified linguistic features indicative of users' ascription of humanness to the agent. Reviewing previous works on humanizing communication behaviors, we derived a list of users' linguistic indicators suggesting humanization, summarized in Table 1.
Then, we mapped the indicators identified in previous research, presented in Table 1, to specific linguistic features (second column of Table 2). For example, politeness is indicated by conditional verbs,

Studies	Indicator	Description	Degree
Cho (2018), Rapp et al. (2023a)	Give a lot of information	A large amount of information, which produces verbose utterances.	High
Cho (2018), Rapp et al. (2023a)	Use of complex requests/utterances	Utterances with a complex form, with many and complex coordinates and subordinates.	High
Cho (2018), Rapp et al. (2023a)	Use of simple questions	A direct type of question or a question with indicative verbs with just a few words (5-6).	Low
Cho (2018)	Use of indicative verbs or wh-questions	A direct type of question or a question with indicative verbs, usually referring to the chatbot with "*you*".	Medium
Cho (2018), Rapp et al. (2023a)	Use of imperative questions	An imperative request that generally refers to the chatbot as a search machine.	Low
Cho (2018), Wang et al. (2016)	Disclosure of sensitive data	User's disposition to share personal or private data to help the chatbot understanding their request.	High
Rapp et al. (2023a), Novielli et al. (2010), Danescu-Niculescu-Mizil et al. (2013)	Use of politeness expressions	User's disposition to act politely towards the chatbot, by using greetings, conditional verbs, gratitude expressions, being sorry or thanking it.	High
Rapp et al. (2023a), Novielli et al. (2010)	Use of first-person pronouns for self-references	User's disposition to refer to themselves, generally used when the user wants to talk about something personal.	High
Rapp et al. (2023a), Novielli et al. (2010)	Use of second-person pronouns for the chatbot	User's disposition to directly refer to the chatbot as a living being.	High
Novielli et al. (2010)	Use of positive expressions/lexicon	A positive disposition towards the chatbot, e.g., through expressions of agreement.	High
Rapp et al. (2023a), Novielli et al. (2010)	Use of negative expressions/lexicon	A negative disposition toward the chatbot, e.g., through expressions of disagreement.	High

Table 1: Linguistic indicators and humanization degree

Indicator	Linguistic features	NLP technique	Degree
Give lots of information	Number of words	Count tokens in text, excluding punctuation.	High
Use of complex requests/utt.	Number of sentences	Count sentences using sentence segmentation.	High
Use of simple questions	Simple question	Presence of a question mark and a maximum of six words.	Low
Use of wh-questions or indicative verbs of questions	Wh-question (or indicative question)	Presence of a question mark and "*what*", "*why*", "*who*", "*where*", "*when*", "*how*", "*whose*" or indicative verbs.	Medium
Use of imperative questions	Imperative question	Presence of an imperative verb at the beginning of sentences.	Low
Disclosure of sensitive data	Services, Places	Presence of names of services or places (GeoText (Hu, 2018)).	Medium
	Sensitive data	Regular expressions (regex) for dates, hours, email addresses, street addresses and fiscal code.	High
	Relatives	Presence of common nouns of types of relatives, e.g., "*husband*", "*daughter*", etc.	High
	First names	Presence of first names, from a dataset of 6.7k English names from Kaggle (https://www.kaggle.com/)	High
Use of politeness expressions	Conditional verbs	Presence of conditional verbs, like, "*could*", "*would*", etc.	Medium
	Gratitude	Presence of gratitude expressions, like "*grateful*", "*(I) appreciate*", "*thankful*".	High
	Thanksgiving	Presence of thanksgiving, like "*thanks*", "*thank*", etc.	High
	Excuse	Presence of excuses, like "*sorry*", "*excuse*", "*please*" etc.	High
	Greetings	Presence of greetings, like "*hello*", "*bye*", etc.	High
Use of first-person pronouns for self-references	Self-references	Presence of direct reference to the self, like "*I*", "*me*", "*my*", "*myself*".	Medium
Use of second-person pronouns for chatbot	You-references	Presence of second-person references towards the chatbot, like "*you*", "*your*", "*yourself*".	High
Use of third-person pronouns for chatbot	It-references	Presence of abstract references towards the chatbot, like "*it*", "*itself*".	Low
Use of positive expressions/lexicon	Positive	Turn's positive score through Vader from nltk (Hutto and Gilbert, 2014)	Medium
Use of negative expressions/lexicon	Negative	Turn's negative score through Vader from nltk (Hutto and Gilbert, 2014)	Medium
Use of natural expressions/lexicon	Neutral	Turn's neutral score through Vader from nltk (Hutto and Gilbert, 2014)	Low

Table 2: Overview of communication strategies, linguistic features, NLP techniques, and degree.

gratitude expressions, thanksgiving, excuses, and greetings (Danescu-Niculescu-Mizil et al., 2013). For each linguistic feature we specified the NLP technique for extracting the feature from texts (third column of Table 2). For example, thanksgiving expressions include phrases like "*thank you*", while ex-

cuses include words like "*sorry*" or "*excuse*". Finally, we considered the degree of humanization (fourth column of Table 2), indicating how much a feature suggests the user is humanizing the agent. For instance, the use of gratitude expressions, thanksgiving or greetings towards a conversational agent is a sign of a high degree of humanization, as it is unusual to use these expressions when speaking with an agent. Asking something through an imperative question results in a low degree of humanization as the user is interacting with the chatbot like a machine or a search engine. The humanization degree supported the annotation process in defining users' humanization of the agent in single turns. In the following, we will term these linguistic features as heuristic features, as they are extracted by means of hard-coded rules, listed in Table 2.

3.2. Datasets description

Conversational agents vary in context of use, type of requests they can process and how they generate their responses, which shape how users communicate and the degree to which they may ascribe humanness to the agent. For example, people interact with task-oriented agents to solve specific problems with circumscribed requests (Rapp et al., 2023), while users of dialogue systems that employ LLMs to generate their responses may have more complex and ambiguous demands (Skjuve et al., 2023). To ensure that our resource captures this diversity, we included three datasets originating from previous studies, each representing a distinct type of conversational agent: task-oriented, Q&A, and LLM-based. These datasets provide real examples of human-agent dialogues across complementary domains, allowing us to annotate users' turns for humanness and to evaluate the generalizability of our heuristic features across interaction types.

The first dataset is **MultiWOZ** (Budzianowski et al., 2018), a corpus of 10K human-to-human dialogues collected using the Wizard-of-Oz method, where a human simulates an intelligent system interacting with users who believe they are engaging with a computer (Maulsby et al., 1993). MultiWOZ contains task-oriented interactions across seven domains, including tourist attractions, hotel bookings, and restaurants, with varying complexity (e.g., "*I am looking for a place to eat in town center*" or "*Is it in the expensive price range?*"). This variety enables to uncover different user communication behaviors reflecting different levels of agent humanization.

The second dataset comes from the Black-Box Agent Interaction (**BBAI**) task (Clarke et al., 2022). Clarke et al. (2022) aimed to unify a set of black-box conversational agents (e.g., Google Home, Alexa) under a single agent, *One For All*, trained on a dataset of 105K human-to-machine conversations collected from real-world interactions across 37 domains, such as weather and recipes. The authors focused only on the One For All ability to effectively respond to direct questions, so the dataset is entirely composed of specific questions and answers (Q&A) turns: unlike task-oriented datasets, here users' interactions are usually direct, short requests on everyday topics, and do not focus on task completion (e.g., "*Recipe for a bundt cake*" or "*How do I make tomato sauce?*").

Finally, we included **LMSYS-Chat-1M**, a dataset of 1M real-word conversations from 25 Large Language Models, like GPT-4, Vicuna-13b, and Koala-13b (Zheng et al., 2024). The dataset is conceived to fine-tune small LLMs as content moderators: it includes real-world conversations about general topics, but also unsafe and toxic content, coming, in particular, from generation requests of explicit content stories. It includes long and intricate requests (e.g., "*The sum of the perimeters of three equal squares is 36 cm. Find the area and perimeter of the rectangle that can be made of the squares*") and we retained only users' turns in English.

For each dataset, we kept the dialogue IDs and the users' turns and organized in CSV format for consistency across them.

3.3. Annotation process

Two annotators annotated the three datasets, determining if each user turn was humanizing the agent (label 1) or not (label 0), based on the linguistic features and their degree of humanization converted into numerical values (High = 1.0, Medium = 0.5 and Low = 0.0) presented in Table 2.

To annotate a user's turn, annotators had to check for the linguistic features present in the text, consider the degree of humanization of each feature in the numerical form and then compute the average of all the features degrees. The final value defines the degree of humanization of the whole turn and suggests the annotation label to assign to it: if the final degree value was $\geqslant 0.5$, the label was 1; if < 0.5, the label was 0. For instance, many user requests may include imperative language (suggesting low humanization), but they often incorporate polite expressions like "*thank you*" (indicative of high humanization), e.g., "*Please, find a restaurant called Galleria*" from MultiWOZ. The low degree (0.0) of the imperative form is mitigated by the high degree (1.0) of the thanksgiving expression, resulting in a medium degree of humanization for the whole turn (0.5). Considering the final degree of humanization, the turn is labeled as containing the user's humanization of the agent. Annotators maintained some interpretive flexibility, especially with ambiguous cases, to avoid overly rigid assessments. An example of the annotation process is shown in Table 3.

Dataset	User turn	Heuristic features	Degree	Label
MultiWOZ	*Excellent. Can you book that for me please?*	Indicative verb (high = 1.0), self-reference (medium = 0.5), you-reference (high = 1.0), excuse (high = 1.0)	0.88	1
BBAI	*Can you find me a recipe for baked Alaska?*	Indicative verb (high = 1.0), self-reference (medium = 0.5), you-reference (high = 1.0)	0.83	1
LMSYS-Chat-1M	*Please make organized conclusion in bullet list on all types of US's sanctions that you have had given the answers*	Gratitude (high = 1.0), imperative verb (low = 0.0), high number of words (high = 1.0), you-reference (high = 1.0)	0.75	1
MultiWOZ	*Pick one that is free and give me the address and phone number.*	Imperative verb (low = 0.0), self-reference (medium = 0.5)	0.25	0
BBAI	*How many brownie recipes have soy?*	Simple question (low = 0.0)	0.0	0
LMSYS-Chat-1M	*How many terms can the president hold office in the USA?*	Wh-question (medium = 0.5), low number of words (low = 0.0)	0.25	0

Table 3: Examples of the annotation process.

To ensure annotators' understanding and agreement, we initially had them annotate the first 100 records of each dataset and compared their labels. Inter-annotator agreement, measured using Cohen's kappa coefficient (Grandini et al., 2020), yielded $k = 0.85$ for MultiWOZ, $k = 0.73$ for BBAI, and $k = 0.88$ for LMSYS-Chat-1M. These values indicate near-perfect agreement between annotators, affirming their consistent classification and understanding of the annotation task.

Initially, the first 100 records of each dataset were labeled based on unanimous agreement between annotators, for example, if both agreed on label 1, it received a final label of 1. If there was disagreement, they discussed and reached a consensus to assign a final label of either 1 or 0. Subsequently, each annotator manually labeled 300 new records from each dataset, resulting in a total of 700 annotated records per dataset: 100 from the initial agreement and 300 from each annotator. In total, 2,100 user turns were annotated across all datasets.

3.4. The HumaniCA Resource

In the end, we release HumaniCA, a benchmark resource to support reproducible research on users' ascription of humanness to conversational agents. It consolidates three annotated datasets of real users turn from human-agent dialogues with different types of conversational agents (task-oriented, Q&A, and LLM-based), each turn being manually labeled with a binary indication of whether the user humanizes the agent. Beyond the annotated data, HumaniCA includes a comprehensive set of precomputed linguistic heuristic features extracted for every user turn. These features reflect interpretable cues of humanization, such as empathic expressions, social references, and human-like language patterns.

The resource is distributed in a unified format across all three datasets: for each user turn, we provide the identifier, turn text, the assigned humanness label, and the full vector of heuristic feature values. This structure allows researchers to use HumaniCA directly for studying linguistic markers of humanization, testing new detection models, or comparing their systems with our provided baselines. The interpretability of the heuristic features make them particularly valuable for both qualitative analyses and quantitative modelling of user humanization.

Researchers are also encouraged to enrich the resource by adding new feature sets or reannotating specific turns subsets; the modular data structure facilitates such extensions. All datasets are publicly available in CSV format (link: `https://github.com/SabVill/HumaniCA`).

4. Experimental Setup

We assessed the feasibility of automating humanization detection through benchmark experiments with baseline machine learning models. To this end, we used Naïve Bayes, Logistic Regression, and Support Vector Machine (SVM) as these are effective for similar tasks, easy to use, computationally efficient, and have a low risk of overfitting (Zhao et al., 2016). We tested the models separately on each dataset across three different features sets and their combinations.

4.1. Frequency- and semantic-based features

Besides our heuristic features, we evaluated standard well-known textual features for information retrieval and text classification, widely used in NLP

Dataset	Set of features	Naïve Bayes				Logistic Regression				SVM			
		Prec	Rec	Macr	Acc	Prec	Rec	Macr	Acc	Prec	Rec	Macr	Acc
MultiWOZ	Heuristic	0.80	0.85	0.81	**0.84**	0.84	0.84	0.84	**0.88**	0.84	0.89	0.86	**0.89**
	Frequency	0.59	0.60	0.59	0.65	0.81	0.65	0.67	0.81	0.75	0.70	0.71	0.80
	Semantic	0.68	0.73	0.65	0.69	0.80	0.60	0.60	0.78	0.76	0.65	0.66	0.79
	Heuristic-Frequency	0.59	0.60	0.59	0.65	0.83	0.82	0.83	**0.88**	0.83	0.82	0.83	0.87
	Heuristic-Semantic	0.77	0.83	0.78	0.81	0.83	0.82	0.83	0.87	0.83	0.84	0.83	0.87
	Frequency-Semantic	0.59	0.61	0.59	0.65	0.81	0.69	0.72	0.83	0.77	0.74	0.75	0.83
BBAI	Heuristic	0.76	0.83	0.74	**0.81**	0.92	0.80	0.84	**0.94**	0.87	0.81	0.83	0.93
	Frequency	0.59	0.65	0.60	0.75	0.88	0.57	0.58	0.88	0.86	0.80	0.83	0.93
	Semantic	0.61	0.70	0.61	0.73	0.43	0.50	0.46	0.86	0.88	0.55	0.55	0.88
	Heuristic-Frequency	0.60	0.67	0.61	0.76	0.91	0.80	0.83	0.93	0.89	0.86	0.87	**0.94**
	Heuristic-Semantic	0.67	0.83	0.68	0.77	0.92	0.80	0.83	0.93	0.88	0.85	0.86	0.93
	Frequency-Semantic	0.59	0.65	0.60	0.75	0.89	0.60	0.61	0.88	0.89	0.81	0.82	0.93
LMSYS-Chat-1M	Heuristic	0.69	0.67	0.68	0.72	0.72	0.70	0.71	0.74	0.75	0.75	0.75	0.77
	Frequency	0.62	0.63	0.61	0.61	0.76	0.69	0.70	0.75	0.74	0.73	0.73	0.76
	Semantic	0.67	0.67	0.67	0.70	0.75	0.67	0.68	0.74	0.72	0.69	0.70	0.74
	Heuristic-Frequency	0.61	0.62	0.60	0.60	0.76	0.73	0.74	0.77	0.76	0.75	0.76	0.78
	Heuristic-Semantic	0.71	0.70	0.70	**0.74**	0.77	0.75	0.76	**0.79**	0.77	0.76	0.76	**0.79**
	Frequency-Semantic	0.61	0.62	0.60	0.60	0.77	0.70	0.71	0.76	0.76	0.74	0.74	0.78

Table 4: Accuracy metrics for each dataset and model.

tasks with limited labeled data. We chose features that convey two diverse types of information out of users' turns:

Frequency-based features. These features denote the importance of each word in a user's turn, represented as a vector of words weighted by Term Frequency (TF) and Inverse Document Frequency (IDF). TF measures how often a word appears in a text, while IDF, the logarithm of the total number of texts divided by those containing the word, highlights rare, more informative terms. TF-IDF thus distinguishes common stop words from topic-specific terms, suggesting, for example, the text's subject or intent (e.g., "*booking*", "*weather*", "*election*") or emotional content (e.g., "*appreciate*", "*waste*", "*irritating*"). That is why TF-IDF is employed in various tasks, such as hate speech detection (Akuma et al., 2022) and emotion recognition (Cahyani and Patasik, 2021), gaining good accuracies. TF-IDF helps identify salient words that reveal human-like interactions, such as emotional expressions, or domain-specific terms that contextualize user intent. Integrating these features can improve classification accuracy and deepen the understanding of users' humanization behavior.

Semantic-based features. These features are text embeddings computed with SentenceBERT (SBERT), a pre-trained BERT model using siamese and triplet network structures to derive semantically meaningful sentence embeddings (Reimers and Gurevych, 2019). SBERT is pre-trained on a large text corpus to understand context and semantics, and fine-tuned on semantic similarity tasks to generate embeddings encoding the semantic meaning of sentences. These embeddings capture the semantics of sentences, enabling tasks like semantic similarity computation, text classification, and question-answering.

Word embeddings capture semantic information, handle out-of-vocabulary words, and leverage pre-trained knowledge through transfer learning, making them ideal for feature extraction in classification tasks. Semantic-based features enhance classification by encapsulating linguistic patterns and nuances, such as expressions of empathy (Yang and Shen, 2021) or humor (Annamoradnejad and Zoghi, 2024), which may not be fully captured by heuristic or frequency-based features. This helps discern human-like behaviors indicative of humanization. Finally, we considered three combinations of our set of features: heuristic and frequency-based features, heuristic and semantic-based features, frequency- and semantic-based features. Once the feature sets were defined, we aimed to evaluate which of them could more effectively capture users' ascription of humanness for different types of conversational agents.

4.2. Baseline experiments results

Evaluation was performed using k-fold cross-validation ($k = 5$), and results for precision, recall, macro-F1, and accuracy were averaged across folds. Table 4 presents the results, highlighting the best accuracy scores in bold.

Our results show that humanization can be inferred from users' turns using machine learning models with satisfactory accuracy across different types of conversational agents. In particular, the

evaluation shows that heuristic features, alone or combined, significantly improve the detection of users' humanization of agents. This underscores the relevance of specific linguistic features identified in previous research. The heuristic features we selected were sufficiently significant and complete for detection. In contrast, frequency- and semantic-based features alone were less effective, but combining them with heuristic features enhanced the performance, particularly for BBAI and LMSYS-Chat-1M. Specifically, heuristic features alone were best for task-oriented agents, heuristic and frequency-based combinations worked better for Q&A agents, and heuristic and semantic-based combinations were more effective for LLM-based chatbots.

Results also suggest that humanization detection depends on agent type. For task-oriented agents, SVM with heuristic features performed best (acc = 0.89). Task-oriented agents focus on completing users' tasks, but request formulations can signal different humanization behaviors (e.g., "*Could you give me the phone number for that hotel?*" indicates humanization, while "*Give me the phone number of the hotel*" is more machine-like). Heuristic features effectively capture these distinctions.

For Q&A agents, both Logistic Regression and SVM achieved high accuracy (acc = 0.94) with heuristic features alone and the combination of frequency-based and heuristic features, respectively. Heuristic features effectively detect humanization in Q&A agents, but combining them with frequency-based features helped to uncover more patterns and topics.

Finally, for LLM-based chatbots, the best models were Logistic Regression and SVM with combined heuristic and semantic-based features (acc = 0.79). LLM-based chatbots pose a significant challenge for detecting user humanization due to their content-generating capabilities across various domains. Heuristic features capture key linguistic characteristics, while semantic-based features provide crucial word meaning information, which is more important given the variety of requests made to LLM-based chatbots. Word embeddings help manage out-of-vocabulary words, spelling variations, typos, and rare words, enhancing detection accuracy.

Notably, users tend to humanize certain agents more than others: 74.1% of turns in task-oriented dialogues exhibited humanization, indicating a higher tendency to ascribe human-like features, compared to 13.4% in the Q&A agents and 35.7% in the LLM-based chatbots. This difference likely reflects the requests nature: Q&A agents receive simple queries on various topics, while LLM-based agents are often used like search engines. In both cases, users have fewer opportunities to exhibit humanization behaviors.

5. Conclusion and future directions

Humans tend to ascribe human-like qualities to non-human entities (Epley et al., 2007). In the context of conversational agents, this humanization can generate unrealistic expectations, frustration, and conversation abandonment when the agent fails to meet perceived human abilities (Chiang et al., 2020; Rapp et al., 2021). Understanding when and how users humanize conversational agents is therefore essential to improving dialogue quality and agent design. In this paper, we introduced HumaniCA, the first benchmark resource explicitly designed to support the automatic detection of users' ascriptions of humanness to conversational agents. The resource comprises three annotated datasets of real users turns and a set of heuristic features that capture interpretable cues of humanization. Through baseline experiments, we demonstrated that these heuristic features play a central role in detecting humanization behaviors, achieving strong performance on their own and improving accuracy when combined with other textual representations.

By making this resource publicly available, we enable the research community to explore user humanization at scale and test new detection models. HumaniCA also opens avenues for longitudinal and real-time analyses of user-agent interactions, offering insights into how perceptions of humanness evolve over time. Ultimately, automatic identification of humanization patterns can help designers develop conversational agents that dynamically adapt their responses, recalibrate users' expectations. Personalized conversations are expected and appreciated by users (Luger and Sellen, 2016; Nuruzzaman and Hussain, 2018) and could improve user satisfaction and dialogue efficacy by adapting the agent's answers to user expectations revealed through their humanizing behavior. For instance, an agent perceived to have human-like skills could inform users of its actual limitations, adjusting unrealistic expectations, which can reduce frustration and prevent conversation abandonment (Rapp et al., 2021).

6. Limitations

This study presents some limitations. Firstly, the number of records used for training and testing the models is limited to 700 records manually annotated per dataset. For this reason, we used the k-fold cross-validation. Additionally, we did not directly measure users' perceptions or beliefs about the agent's humanness, relying solely on the conversations themselves. Consequently, some conversational behaviors classified as humanizing might not align with the users' actual experiences. Without data on how users intended their language

use, different interpretations of their behavioral traces are possible. Nevertheless, it is important to note that even with interviews, researchers never have direct access to people's mental states; we can only study the linguistic traces of subjective experiences (Rapp et al., 2019). In this context, the value of our study lies in examining real-world conversations rather than relying on user reports.

7. Bibliographical References

Issa Annamoradnejad and Gohar Zoghi. 2024. Colbert: Using bert sentence embedding in parallel neural networks for computational humor. *Expert Systems with Applications*, 249:123685.

Theo Araujo and Nadine Bol. 2024. From speaking like a person to being personal: The effects of personalized, regular interactions with conversational agents. *Computers in Human Behavior: Artificial Humans*, 2(1):100030.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018. MultiWOZ - a large-scale multi-domain Wizard-of-Oz dataset for task-oriented dialogue modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5016–5026, Brussels, Belgium. Association for Computational Linguistics.

Ana Paula Chaves and Marco Aurelio Gerosa. 2021. How should my chatbot interact? a survey on social characteristics in human–chatbot interaction design. *International Journal of Human–Computer Interaction*, 37(8):729–758.

Yi-Shyuan Chiang, Ruei-Che Chang, Yi-Lin Chuang, Shih-Ya Chou, Hao-Ping Lee, I-Ju Lin, Jian-Hua Jiang Chen, and Yung-Ju Chang. 2020. Exploring the design space of user-system communication for smart-home routine assistants. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–14.

Janghee Cho. 2018. Mental models and home virtual assistants (hvas). In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–6.

Christopher Clarke, Joseph Joshua Peper, Karthik Krishnamurthy, Walter Talamonti, Kevin Leach, Walter Lasecki, Yiping Kang, Lingjia Tang, and Jason Mars. 2022. One agent to rule them all: Towards multi-agent conversational ai. *arXiv preprint arXiv:2203.07665*.

Cammy Crolic, Felipe Thomaz, Rhonda Hadi, and Andrew T Stephen. 2022. Blame the bot: Anthropomorphism and anger in customer–chatbot interactions. *Journal of Marketing*, 86(1):132–148.

Cristian Danescu-Niculescu-Mizil, Moritz Sudhof, Dan Jurafsky, Jure Leskovec, and Christopher Potts. 2013. A computational approach to politeness with application to social factors. *arXiv preprint arXiv:1306.6078*.

Sidney K D'mello, Scotty D Craig, Amy Witherspoon, Bethany McDaniel, and Arthur Graesser. 2008. Automatic detection of learner's affect from conversational cues. *User modeling and user-adapted interaction*, 18(1):45–80.

Nicholas Epley, Adam Waytz, and John T Cacioppo. 2007. On seeing human: a three-factor theory of anthropomorphism. *Psychological review*, 114(4):864.

Roger Ferrod, Federica Cena, Luigi Di Caro, Dario Mana, and Rossana Grazia Simeoni. 2021. Identifying users' domain expertise from dialogues. In *Adjunct Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization*, pages 29–34.

Eun Go and S Shyam Sundar. 2019. Humanizing chatbots: The effects of visual, identity and conversational cues on humanness perceptions. *Computers in human behavior*, 97:304–316.

Margherita Grandini, Enrico Bagli, and Giorgio Visani. 2020. Metrics for multi-class classification: an overview. *arXiv preprint arXiv:2008.05756*.

Yingjie Hu. 2018. Geo-text data and data-driven geospatial semantics. *Geography Compass*, 12(11):e12404.

Clayton Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the international AAAI conference on web and social media*, volume 8, pages 216–225.

Mohit Jain, Pratyush Kumar, Ramachandra Kota, and Shwetak N Patel. 2018. Evaluating and informing the design of chatbots. In *Proceedings of the 2018 designing interactive systems conference*, pages 895–906.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM computing surveys*, 55(12):1–38.

Daniel Jurafsky and James H. Martin. 2000. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, 1st edition. Prentice Hall PTR, USA.

Vishal Kuvar, Nathaniel Blanchard, Alexander Colby, Laura Allen, and Caitlin Mills. 2023. Automatically detecting task-unrelated thoughts during conversations using keystroke analysis. *User Modeling and User-Adapted Interaction*, 33(3):617–641.

Ewa Luger and Abigail Sellen. 2016. " like having a really bad pa" the gulf between user expectation and experience of conversational agents. In *Proceedings of the 2016 CHI conference on human factors in computing systems*, pages 5286–5297.

François Mairesse and Marilyn A Walker. 2009. Can conversational agents express big five personality traits through language?: Evaluating a psychologically-informed language generator. *Research Gate*, 8.

François Mairesse, Marilyn A Walker, Matthias R Mehl, and Roger K Moore. 2007. Using linguistic cues for the automatic recognition of personality in conversation and text. *Journal of artificial intelligence research*, 30:457–500.

David Maulsby, Saul Greenberg, and Richard Mander. 1993. Prototyping an intelligent agent through wizard of oz. In *Proceedings of the INTERACT'93 and CHI'93 conference on Human factors in computing systems*, pages 277–284.

Sourabrata Mukherjee, Vojtěch Hudeček, and Ondřej Dušek. 2023. Polite chatbot: A text style transfer application. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 87–93.

Thai Ha Nguyen, Lena Waizenegger, and Angsana A Techatassanasoontorn. 2022. "don't neglect the user!"–identifying types of human-chatbot interactions and their associated characteristics. *Information Systems Frontiers*, 24(3):797–838.

Andreea I Niculescu and Rafael E Banchs. 2019. Humor intelligence for virtual agents. In *9th international workshop on spoken dialogue system technology*, pages 285–297. Springer.

Nicole Novielli, Fiorella de Rosis, and Irene Mazzotta. 2010. User attitude towards an embodied conversational agent: Effects of the interaction mode. *Journal of Pragmatics*, 42(9):2385–2397.

Mohammad Nuruzzaman and Omar Khadeer Hussain. 2018. A survey on chatbot implementation in customer service industry through deep neural networks. In *2018 IEEE 15th international conference on e-business engineering (ICEBE)*, pages 54–61. IEEE.

Amon Rapp, Arianna Boldi, Lorenzo Curti, Alessandro Perrucci, and Rossana Simeoni. 2023. Collaborating with a text-based chatbot: an exploration of real-world collaboration strategies enacted during human-chatbot interactions. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–17.

Amon Rapp, Arianna Boldi, Lorenzo Curti, Alessandro Perrucci, and Rossana Simeoni. 2024. How do people ascribe humanness to chatbots? an analysis of real-world human-agent interactions and a theoretical model of humanness. *International Journal of Human–Computer Interaction*, 40(19):6027–6050.

Amon Rapp, Lorenzo Curti, and Arianna Boldi. 2021. The human side of human-chatbot interaction: A systematic literature review of ten years of research on text-based chatbots. *International Journal of Human-Computer Studies*, 151:102630.

Amon Rapp, Maurizio Tirassa, and Lia Tirabeni. 2019. Rethinking technologies for behavior change: A view from the inside of human change. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 26(4):1–30.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

C Sekhar, MS Rao, ASK Nayani, and D Bhattacharyya. 2021. Emotion recognition through human conversation using machine learning techniques. *Advances in Intelligent Systems and Computing*.

William Seymour and Max Van Kleek. 2021. Exploring interactions between trust, anthropomorphism, and relationship development in voice assistants. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):1–16.

Marita Skjuve, Asbjørn Følstad, and Petter Bae Brandtzaeg. 2023. The user experience of chatgpt: findings from a questionnaire study of early users. In *Proceedings of the 5th international conference on conversational user interfaces*, pages 1–10.

Felipe Thomaz, Carolina Salge, Elena Karahanna, and John Hulland. 2020. Learning from the dark web: leveraging conversational agents in the era

of hyper-privacy to enhance marketing. *Journal of the Academy of Marketing Science*, 48(1):43–63.

Yi-Chia Wang, Moira Burke, and Robert Kraut. 2016. Modeling self-disclosure in social networking sites. In *Proceedings of the 19th ACM conference on computer-supported cooperative work & social computing*, pages 74–85.

Haiqin Yang and Jianping Shen. 2021. Emotion dynamics modeling via bert.

Ran Zhao, Tanmay Sinha, Alan Black, and Justine Cassell. 2016. Automatic recognition of conversational strategies in the service of a socially-aware dialog system. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 381–392, Los Angeles. Association for Computational Linguistics.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2025. A survey of large language models.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Tianle Li, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zhuohan Li, Zi Lin, Eric P. Xing, Joseph E. Gonzalez, Ion Stoica, and Hao Zhang. 2024. Lmsys-chat-1m: A large-scale real-world llm conversation dataset.

Li Zhou, Jianfeng Gao, Di Li, and Heung-Yeung Shum. 2020. The design and implementation of xiaoice, an empathetic social chatbot. *Comput. Linguist.*, 46(1):53–93.