Measuring and Clustering Heterogeneous Chatbot Designs

Outlier results without "toy" chatbots

PABLO C. CAÑIZARES, Universidad Autónoma de Madrid, Spain
JOSE MARÍA LÓPEZ-MORALES, Universidad Autónoma de Madrid, Spain
SARA PÉREZ-SOLER, Universidad Autónoma de Madrid, Spain
ESTHER GUERRA, Universidad Autónoma de Madrid, Spain
JUAN DE LARA, Universidad Autónoma de Madrid, Spain

We report the results of the outlier analysis without "toy" chatbots in our dataset.

CCS Concepts: • **Software and its engineering** → **Software design engineering**; *Extra-functional properties*; • **Computing methodologies** → *Natural language processing*;

Additional Key Words and Phrases: chatbot design, metrics, clustering, quality assurance, model-driven engineering

ACM Reference Format:
Pablo C. Cañizares, Jose María López-Morales, Sara Pérez-Soler, Esther Guerra, and Juan de Lara. 2023. Measuring and Clustering Heterogeneous Chatbot Designs: Outlier results without "toy" chatbots. *ACM Trans. Softw. Eng. Methodol.* 1, 1, Article 1 (January 2023), 2 pages. https://doi.org/XXXXXXX.XXXXXXX

We report the results of the outlier analysis without "toy" chatbots in our dataset, which amount to 21 Rasa chatbots and 10 Dialogflow agents. Specifically, from the Dialogflow platform we have excluded the following chatbots: ChronoGG, defaults-chatfuel, defaults-manychat, dialogflow-quotes, fulfillent-importer, fulfillment-multi-locale, HHandoffDAgent, HumanHandoffDemonstrationAgent, stock-bot.

Regarding the Rasa platform, we have excluded: 02-lead-bot, 04-feedback-bot, 07- survey-bot, Baisc-Demo, concertbot, diagrams2ai, Email-WhatsApp- Integration-Chatbot, heroku-demo, juwolfrum, jwheat, matiasguerrero, moodbot, MyPython-master, pydata18, Rasa-Docker-Test, rasa-playground, sathsaraRasan-tha, test-bot, TestFirstRasaBot, twb-asessement, yassinelamarti.

Authors' addresses: Pablo C. Cañizares, Universidad Autónoma de Madrid, Madrid, Spain, Pablo.Cerro@uam.es; Jose María López-Morales, Universidad Autónoma de Madrid, Madrid, Spain, JoseMaria.LopezM@uam.es; Sara Pérez-Soler, Universidad Autónoma de Madrid, Madrid, Spain, Sara.PerezS@uam.es; Esther Guerra, Universidad Autónoma de Madrid, Madrid, Spain, Esther.Guerra@uam.es; Juan de Lara, Universidad Autónoma de Madrid, Madrid, Spain, Juan.deLara@uam.es.

1049-331X/2023/1-ART1 $15.00
https://doi.org/XXXXXXX.XXXXXXX

Table 1. Metric outliers per technology – dataset without "toy" chatbots.

			High outliers			Low values		
Type	**Metric**	**Median**	**Cutoff**	**Chatbot%**	**Sample chatbots (max. 2)**	**Cutoff**	**Chatbot%**	**Sample chatbots (max. 2)**
Dialogflow								
Global	INT	6	27.75	5.6%	iLearn (89), Car (77)	2	0%	
	ENT	1	7.5	9.3%	MysteryAnimal (37), googleChallenge (34)	0.33	48.59%	iLearn (0), insurance_Bot (0)
	FLOW	6	20.25	8.4%	iLearn (89), MysteryAnimal (62)	2	3.73%	Currency-Converter (2), Hotel-Booking (2)
	PATH	6	24	8.41%	Car (117), iLearn (89)	2	2.80%	Currency-Converter (2), Hotel-Booking (2)
	CNF	4	91	14.01%	Car (1606), iLearn (1599)	1.33	35.51%	airportagent (0), basic-slotfilling (0)
	SNT^{+}	4	27.5	0.93%	BikeShop (38)	1.33	45.79%	dialogflow-silly-name-maker (0), fulfillment-temperature-converter (0)
	SNT^{-}	17	62.5	2.80%	dialogflow-google-sign-in (100), dialogflow-ssml (100)	5.66	33.64%	Car (0), MysteryAnimal (0)
Intent	TPI	5.65	19.99	5.60%	Dining-Out (94.67), Hotel-Booking (50.67)	1.88	14.01%	hackathon-group-10 (0.33), in-my-seats-jovo (0.33)
	WPTP	2.46	6.49	1.86%	googleChallenge (9.60), Car (6.8)	0.82	1.86%	fulfillment-multi-locale (0.5), hackathon-group-10 (0.67)
	VPTP	0.44	1.56	0.85%	googleChallenge (1.76)	0.14	12.82%	libsample-advanced (0.03), dialogflow-webhook-boilerplate (0.06)
	PPTP	0.51	1.53	0.93%	Dining-Out (8.33), Hotel-Booking (5)	0.17	12.14%	airportagent (0), keijiban (0)
	WPO	5.79	19.67	2.80%	Education_Chatbot (22.89), googleChallenge (19.81)	1.93	31.77%	Car (0), fulfillment-temperature-converter (1.5)
	CPO	23.56	85.6	5.6%	Education-Chatbot (125.56), googleChallenge (105.16)	7.85	30.84%	Car (0), Date (0)
	VPOP	1.04	3.41	0.93%	fulfillment-telephony (3.56)	0.34	32.71%	libsample-advanced (0.1), HOTEL-BOOKING-AGENT2 (0.29)
	READ	4	15	3.73%	Education-Chatbot (19), googleChallenge (16)	1.33	31.77%	fulfillment-temperature-converter (0), Car (0)
Entity	LPE	1	13.75	9.34%	Dining-Out (1177.13), ekgsBot (359.86)	0.33	48.59%	iLearn (0), airportagent (0)
	SPL	0	6.25	1.86%	Formats (7.13), iotairblower (6.83)	0	51.40%	iLearn (0), airportagent (0)
	WL	0	18.46	1.86%	gordobbot (36), ekgsBot (20.61)	0	51.40%	iLearn (0), airportagent (0)
Flow	FACT	2	3.69	1.86%	HOTEL-BOOKING-AGENT2 (8), keijiban (4.71)	0.66	0%	-
	FPATH	1	1	17.75%	Dining-Out (3.5), HR-Bot (2.33)	0.33	0%	-
	CL	1	3.5	4.67%	enoreese (8), Food-Ordering-Chatbot (7)	0.33	0%	-
Rasa								
Global	INT	10	32	9.91%	covid-19-chatbot (143), identity-cloning-toolkit (114)	3.33	0%	-
	ENT	0	0	19.83%	Foodie-Rasa-Chatbot (9), insurance-en (4)	0	80.17%	covid-19-chatbot (0), identity-cloning-toolkit (0)
	FLOW	4	14.5	9.91%	identity-cloning-toolkit (102), small-talk-rasa-stack (86)	1.33	16.52%	covid-19-chatbot (1), dong5854 (1)
	PATH	8	32.5	9.91%	finbot-master (207), identity-cloning-toolkit (115)	2.66	4.95%	formoriginal (1), aniketbangar(2)
	CNF	14	127.5	14.04%	covid-19-chatbot (10532), identity-cloning-toolkit (3461)	4.66	19.00%	WeatherBot (0), RasaProject-Docker (0)
	SNT^{+}	15	40	0.82%	-	5	14.87%	09-news-api (0), Ali (0)
	SNT^{-}	8	40	1.65%	FAQ-RASA-NLU (51), trackncov19 (44)	2.66	31.4%	05-event-bot (0), 09-news-api (0)
Intent	TPI	9.7	38.74	9.09%	aniketbangar (235.17), sokkalingam (214.33)	3.23	4.13%	Tiara-A-Chatbot (1), vardhaman-freshers (2.29)
	WPTP	3.02	6.4	0%	-	1.04	0.82%	Tiara-A-Chatbot (0.88), vardhaman-freshers (1.11)
	VPTP	0.6	1.5	0%	-	0.2	4.13%	09-news-api (0), 05-event-bot (0.12)
	PPTP	0.29	1.5	0.820%	flight-booking (1.6), Foodie-Rasa-Chatbot (1.5)	0.08	31.4%	rasa-faq-bot (0), 0-smalltalk-bot (0)
	WPO	6.89	14.83	8.26%	Tiara-A-Chatbot (87.63), Data-Mining-Chatbot (73.37)	2.29	0%	-
	CPO	29.13	66.12	9.91%	Tiara-A-Chatbot (440.1), Data-Mining-Chatbot (382.9)	9.46	0%	-
	VPOP	1.24	2.26	4.95%	Data-Mining-Chatbot (3.52), FAQ-RASA-NLU (3.33)	0.14	0.82%	Chatbot-Banking (0.17), WeatherBot (0.2)
	READ	5	11.5	8.26%	Tiara-A-Chatbot (75), Data-Mining-Chatbot (62)	1.66	0%	-
Flow	FACT	1.14	1.89	9.91%	Data-Mining-Chatbot (3.92), -	0.38	0%	-
	FPATH	1.92	8.0	11.57%	dong5854 (57), rasa-workshop-pydata-berlin (37)	0.64	0%	-
	CL	5	15.5	6.61%	aniketbangar (211), covid-19-chatbot (73)	1.66	8.26%	01-smalltalk-bot (1), Camillads (1)

Table 2. Summary of problems detected in chatbots with high metric outliers or low metric values. Legend: n/a indicates an empty set of outliers; * indicates the percentage is taken on the whole dataset of chatbots.

Aspect	**Metric**	**Value**	**# Problems, Outlier%**				**Problem description**	**Problem type**
			Dialogflow		**Rasa**			
Size	INT	low	0	0%	0	0%	Incomplete/Toy chatbot	Incomplete design, Usability
		high	2	33.4%	5	55.6%	Redundant intents (when CNF high)	Re-design
Conversation	FLOW	low	0	0%	0	0%	Incomplete/Toy chatbot	Incomplete design, Usability
		high	0	0%	7	46.7%	Repeated or redundant flows	Design error, Re-design
	PATH	low	0	0%	0	0%	Incomplete/Toy chatbot	Incomplete design, Usability
		high	0	0%	16	76.2%	Repeated or redundant paths	Design error, Re-design
	FPATH	low	n/a	n/a	n/a	n/a	Incomplete/Toy chatbot	Incomplete design, Usability
		high	1	5.3%	15	71.5%	Repeated or redundant paths	Design error, Re-design
	CL	high	0	0%	1	9.1%	Long conversation (hard to complete)	Usability
		high	0	0%	9	81.9%	Error in conversation design	Design error
Outputs	VPOP	high	1	100%	2	25%	Missing punctuation signs	Usability
	CPO	high	0	0%	4	33.4%	Long responses (>280 chars)	Usability, Deployability
	WPO	high	0	0%	5	50%	Long responses (>50 words)	Usability, Comprehensibility
	READ	high	0	0%	5	50%	Long reading times (>30 secs)	Usability, Efficiency
Inputs	TPI	low	9	60%	3	60%	Intents poorly trained (≤4 phrases)	Usability
			11	9.4%*	40	29.2%*	Intents without training phrases	Incomplete design, Re-design
		high	2	25%	14	77.8%	Repeated or redundant phrases	Re-design
	WPTP	low	2	100%	1	100%	Bad quality of training set	Usability, Incomplete design
	CNF	high	7	35%	5	27.8%	Confusing intents	Usability
Vocabulary	LPE	low	3	100%	n/a	n/a	Ill-defined entities (when ENT>0)	Usability
	WL	high	1	50%	n/a	n/a	Bad use of entity literals	Design error