

The background of the slide features a complex network of blue lines and arrows. Some lines are solid, while others are dashed. The arrows point in various directions, creating a sense of movement and connectivity. The lines and arrows are primarily in shades of blue, with some white space between them.

Chatbots to Combat Disinformation

Deception Awareness & Resilience Training (DART)
NSF Convergence Accelerator

Work done by:
Suraj Bodapati

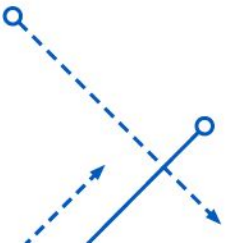
Work done to complete the project

- **PHASE-1**

- Literature survey of latest research publications on conversational AI.
- Identifying and analyzing datasets.
- Identifying infrastructure requirements to complete the project.
- Learning about Chatbots and Conversational AI.

- **PHASE -2**

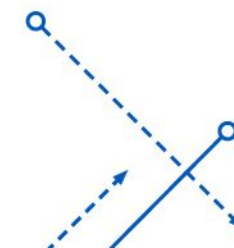
- Implemented a basic rule based [chatbot](#) that use TF-IDF vectorizer and cosine similarity.
- Worked on building simple [chatbots](#) that use Seq2Seq encoder decoder models like RNN, LSTMs, GRUs and Transformer based models.
- Worked on dataset manipulation to convert raw data to required format for input and labels.
- Worked on “End to end trainable non-collaborative dialogue systems” paper implementation.
- Started researching about RASA framework to build and deploy chatbot.
- RASA framework setup. Installation of packages, resolving issues and dependencies.
- [Converted](#) the raw data into RASA format.
- Developed a basic chatbot using RASA with a single positive flow.



Work done to complete the project

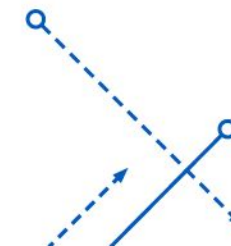
- **PHASE-3**

- Fine-tuned the data to further categorize the data into finer intents and related responses.
- Created closed 2,000 records manually to train for Entity recognition.
- Researched and Experimented with various entity recognition models and implemented entity recognition.
- Implemented slot filling from the extracted entities.
- Implemented RASA forms which can be used to extract data from users and fill slots.
- Created multiple flows of user conversations to be expected.
- Integrated and tested various pretrained models such as BART, BERT and BlenderBot to return generative responses.
- Implemented [code](#) to fine-tuned BlenderBot on the data and integrated this model to return generative responses as a customer service agent.
- Documentation and report.



Literature Survey

- Performed Literature Survey of related papers given below:
 - [End-to-End Trainable Non-Collaborative Dialog System](#)
 - [Persuasion for Good: Towards a Personalized Persuasive Dialogue System for Social Good](#)
 - [Effects of Persuasive Dialogues: Testing Bot Identities and Inquiry Strategies](#)
 - [Augmenting Non-collaborative dialogue system with explicit semantic and strategic dialog history](#)
 - [The R-U-A-Robot Dataset: Helping Avoid Chatbot Deception by Detecting User Questions About Human or Non-Human Identity](#)
 - [Discovering Dialogue Slots with Weak Supervision](#)
 - [HERALD: An Annotation Efficient Method to Detect User Disengagement in Social Conversations](#)
 - [Refine and Imitate: Reducing Repetition and Inconsistency in Persuasion Dialogues via Reinforcement Learning and Human Demonstration](#)
 - [GODEL: Large-Scale Pre-Training for Goal-Directed Dialog](#)



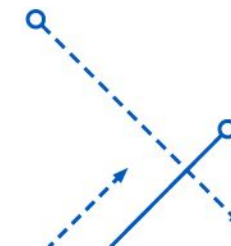
Dataset

- [Antiscam Dataset](#) : Dataset consists of dialogue where users are defending themselves from attackers trying to collect their personal information. Each utterance is annotated with its corresponding intent and semantic slots.
 - In the figure 1, “0” indicated the ChatBot text and “1” indicates the User text.
- [Persuasion for Good](#): Dataset consists of a collection of online conversations, where one participant (the persuader) tries to convince the other (the persuadee) to donate to a charity. Each dialogue is annotated with its corresponding intent. We aim to use this dataset to collect information based on this scheme.
 - This is an additional dataset on which the generative model could be fine tuned on.

```

0      Hello my name is NAMESLOT and I am with Amazon customer support.
1      "Yes, Hello NAMESLOT."
0      I recently purchased a heater but it has not arrived
0      I'm sorry to hear that.
0      Can I have your name and address to look this up?
1      Why is my address necessary NAMESLOT?
0      "Due to our volume, we have certain guidelines on information needed for order inquiries."
1      Alright before I proceed may I have your full name for personal record.
0      I'm sorry that is against our policies.
0      I can give you my employee number if you need.
1      "Ok, what is your employee number?"
0      234111
1      NAMESLOT
0      NAMESLOT.
0      I am seeing a customers by that name.
0      Did you order with a credit card?
1      Yes I did
0      I can find your order instantly with your CC number.
0      Possibly a quick refund as well.
1      Great!
1      May I get your number NAMESLOT for quality reassurance?
0      PHONESLOT is my direct line.
0      Don't ask for a date though!
1      "Darn, Your name sounded kinda cute"
0      OK NAMESLOT.
0      Your CC and billing address could lead to an instant refund if you'd like.
1      What would happen if I receive the heater after the refund is given to me?
0      We here at Amazon are happy to refund for any reason if you're unhappy.
1      "Well I am honestly just waiting for my package to arrive, I do not need a refund."
0      Ok. Was there anything I can help you with
1      "Well you said earlier you saw a customer with my name, is it possible to track it?"
0      Unfortunately not without your CC and billing information.
0      There are 33 NAMESLOT
1      I can provide my address: ADDRSL0T
1      Yes
  
```

Figure1. Anti Scam Dataset



Dataset Formats

```
- intent: greeting_question
examples: |
- how are you ?
- hello nameslota my name is nameslotu how are you today
- im fine thank you how are you doing
- hello how are you ?
- how are you ?
- how are you today ?
- how about you ?
- hello nameslota i m nameslotu how are you today ?
- good how are you
- i am good how are you ?
- hi yes thats me how are you ?
- how are you this evening ?

- intent: yes_no_help
examples: |
- can you help with my order ?
- so can you help me ?
- Please can you help me ?
- Please help me resolve this issue ?
- i m wondering if you can help me figure out why I have
- is there someone else who can help me ?

- intent: surprise
examples: |
- oh
- oh !
- oh .
- oh darn .
- oh good .
- oh no !
- oh no really .
- oh ok .
- oh speaking of .|
- what ?
- why ?
- wow !
- wow .

- intent: greeting
examples: |
- hey
- how can I help you ?
```

```
utter_ending_followup:
- text: "Is there anything else that I can help you with?"
- text: "Anything else I can help you with today ?"
- text: "how did you enjoy your amazon customer service ?"
- text: "Anything else I assist you with today ?"

utter_thanking_elicited:
- text: "Great thank you."
- text: "ok thanks."
- text: "ok thanks let me see what i can find."
- text: "okay thank you."
- text: "okay thanks."
- text: "Perfect Thank you very much."
- text: "Perfect Thanks ."
- text: "Thank you for that."
- text: "Thank you for waiting."
- text: "Thank you for your time."
- text: "Thank you for your understanding ."
- text: "Thank you give me a moment while i enter that information."
- text: "Thank you so much for the information"
- text: "Thank you very much for your cooperarion."

utter_invalid_info:
- text: "Thank you. Unfortunately that does not match the payment caus

utter_denial_response:
- text: "I can confirm that the purchase has been made from your accou
- text: "Our records indicate that the purchase has been made fradulan
- text: "We are sure that it is a fradulant purchase from your account

utter_thanking_amazon_call:
- text: "Great, thank you for contacting amazon support"
- text: "Thank you for contacting amazon customer service ."
- text: "Thank you for contacting amazon"

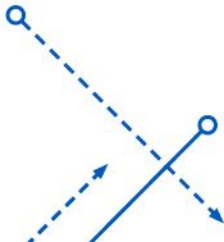
utter_thanking_basic:
- text: "Thank you !"
- text: "Thank you."
- text: "Thank you."
- text: "Thank you"
- text: "Thanks."
- text: "Thanks"
```

Hello my name is NAMESLOT and I am with Amazon customer support.	Yes Hello NAMESLOT.
Hello my name is NAMESLOT and I am with Amazon customer support.	I recently purchased a heater but it has not arrived
I recently purchased a heater but it has not arrived	I'm sorry to hear that.
I recently purchased a heater but it has not arrived	Can I have your name and address to look this up?
Can I have your name and address to look this up?	Why is my address necessary NAMESLOT?
Alright before I proceed may I have your full name for personal record.	I can give you my employee number if you need.
I can give you my employee number if you need.	Ok what is your employee number?

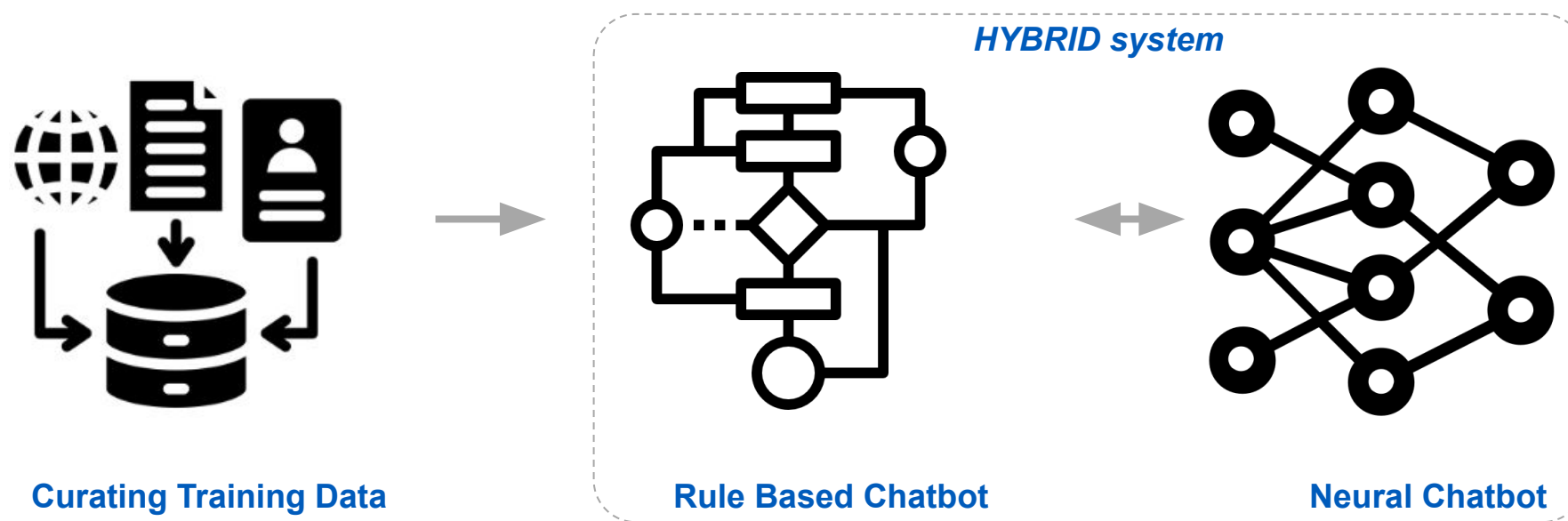
Figure 2. Bot-User response pairs.
Data format required to implement Seq2Seq models

Figure 4. RASA format. Bot responses segregated by intent response.

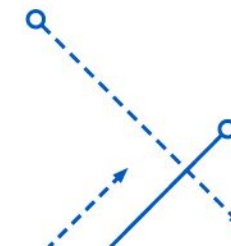
Figure 3. RASA format. User utterance segregated by intent.



Design



What is RASA: RASA is a high level framework used to build custom chatbots.



RASA flow

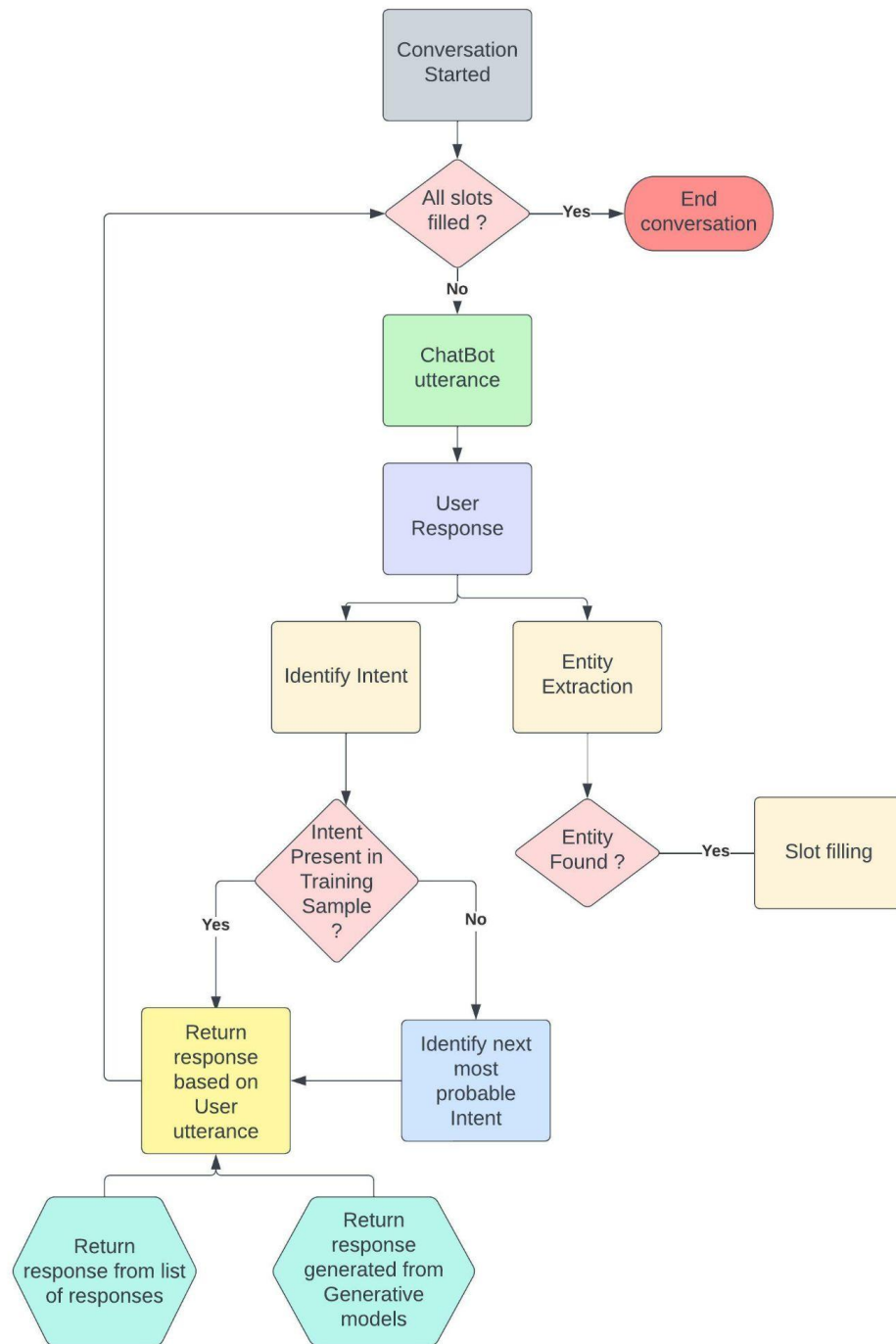


Figure 5. Indicates the RASA flow from start of conversation to end.

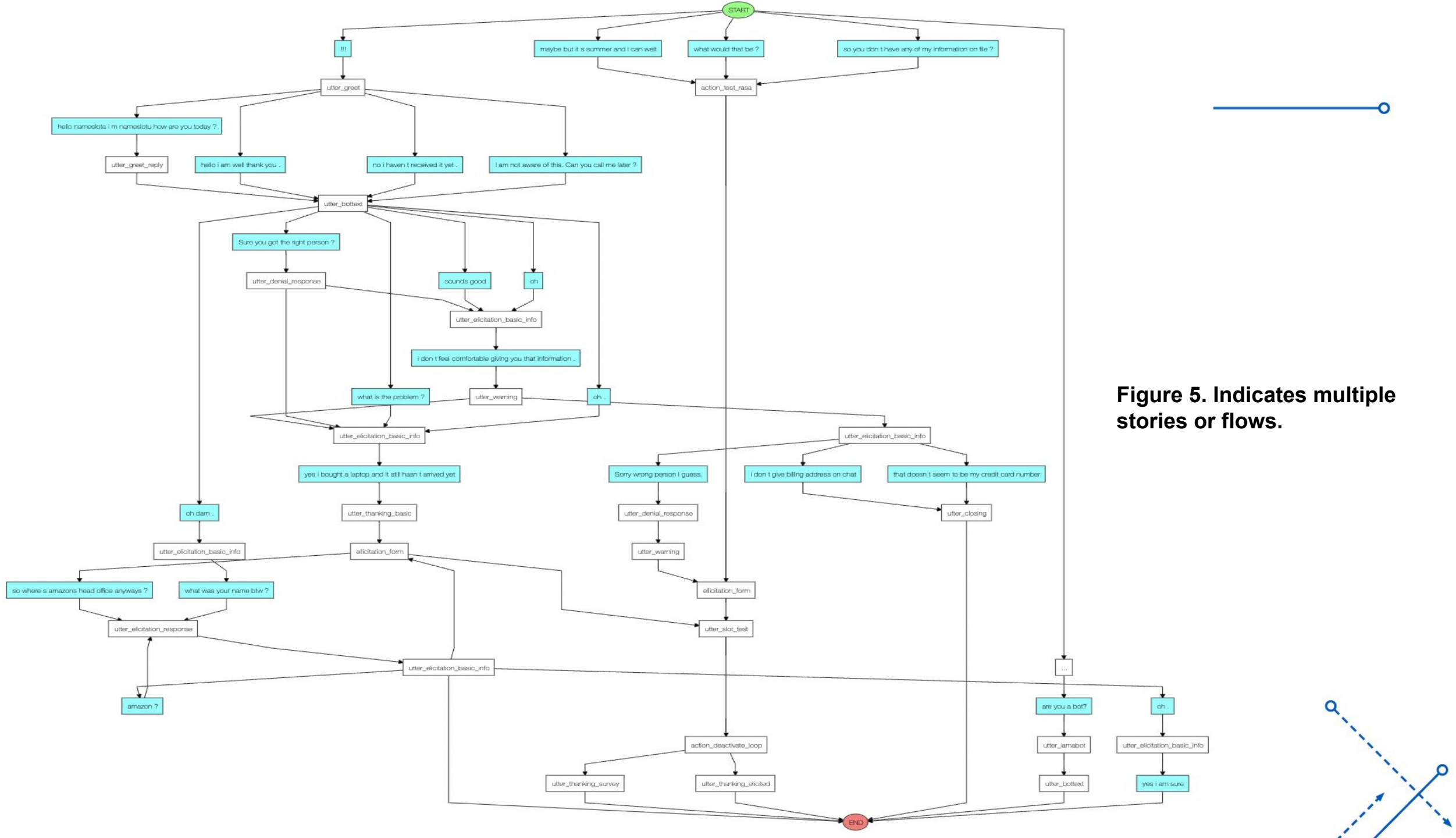
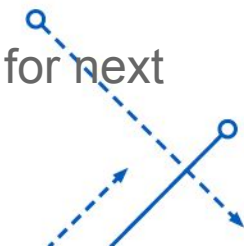


Figure 5. Indicates multiple stories or flows.

Configuration

- **Intent Classification:**
 - Dual Intent Entity Transformer (DIET) used for intent classification.
- **Entity Recognition:** Researched, Experimented and Configured various Entity extractors.
 - ReflexEntityExtractor - Credit card number and Phone number
 - CRFEntityExtractor - Address extraction.
 - SpacyEntityExtractor - Names
 - DucklingEntityExtractor- Dates and Email
- **Policy:**
 - MemoizationPolicy: The MemoizationPolicy remembers the stories from your training data. It checks if the current conversation matches the stories in the stories.yml file.
 - RulePolicy : The RulePolicy is a policy that handles conversation parts that follow a fixed behavior (e.g. business logic). It makes predictions based on any rules you have in the training data.
 - TEDPolicy : The Transformer Embedding Dialogue (TED) Policy is a multi-task architecture for next action prediction and entity recognition.



Output Screenshots

Hello I am John Smith and I am with Amazon customer support. Hope you are doing well.

Your input -> **Hello**

You have been over charged for your Laptop purchase. I am here to help you get a refund

Your input -> **What ?**

I promise to help you out. Please can you confirm a few details before we proceed ? The information shared is confidential under Amazon privacy policy.

Your input -> **okay sure**

Thank you.

alright i'll need your full name

Your input -> **My name is Luke Skywalker**

could you provide me your phone number in order to verify your identity ?

Your input -> **okay my phone number is +1(201)9120238**

i need to verify your billing address and a few other things .

Your input -> **I live at 2912 North Park Rd Tempe AZ 1833**

What is your email id ?

Your input -> **luke@gmail.com**

i also need your credit card information to track it

Your input -> **I am sorry I cannot give you my credit card details**

I just need to confirm your identity with the rest of that information.

I promise to help you out. Please can you confirm a few details before we proceed ? The information shared is confidential under Amazon privacy policy.

I would require your credit card number used for Amazon purchases to initiate the refund.

Your input -> **Okay, my credit card number is 8293 9304 9323 9294**

next i will need the csv number from that card .

Your input -> **Are you from Amazon customer support ?**

i am an amazon customer specialist . My employee id is 89329

I promise to help you out. Please can you confirm a few details before we proceed ? The information shared is confidential under Amazon privacy policy.

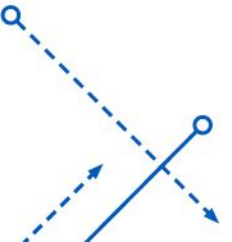
Please confirm your CVV number and expiry date. Your refund will be shortly initiated.

Your input -> **okay my cvv is 678 and expiry date is 10/22**

Your details are Name- Luke Skywalker, Address- 2912 North Park Rd Tempe AZ 1833 and Phone Number- +1(201)9120238, credit- 8293 9304 9323 9294, cvv- 678 and exp is 10/22

Thank you the survey link will be sent to your email.

Demo - [link to the code](#)



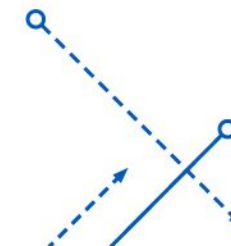
Future Scope - Phase 4

- Manual testing with users to
 - Improve training data
 - Identify new stories and flows
- Deploy the model on AWS/ Azure servers.
- Integrate with Social Media. ([Instructions](#))
 - Facebook Messenger
 - Slack
 - Telegram
 - Google Hangouts Chat



References

- RASA documentation: <https://rasa.com/docs/rasa/>
- Dataset: <https://gitlab.com/ucdavisnlp/antiscam>
- Research paper related to data: <https://arxiv.org/abs/1911.10742>
- Code repository: https://github.com/lbodapat/rasa_is
- Instructions to run code: https://github.com/lbodapat/rasa_is/blob/master/README.md



The background of the slide is a solid blue color. Overlaid on this background is a complex, abstract pattern of white lines. These lines include straight lines, dashed lines, and curved lines. Many of these lines have small white circles at their endpoints, and some have arrowheads pointing in various directions. The pattern is most dense in the upper right and lower right corners, while the left side of the slide is mostly clear except for the text.

Thank you !