

# "No Alexa, I don't want to book a flight; I want the book, 'The Flight': Improving Intent Detection for Virtual Assistants"

- CSI 5180 - Final Project

Presented by: Daniel Lobo (300319498) & Romario Vaz (300308477)



# Project Summary

- Intent detection:
  - Given a query, what does the user want?
- Dataset - CLINC150 - 150 intents (with 150 queries each) + OOS

Query	Intent
Move 100 dollars from my savings to my checking	Transfer
Tell me a joke please	Tell joke
How are my sports teams doing	Out-of-scope
Why can't you divide by zero	Out-of-scope


- Focus:
  - How do we improve detection of in-scope intents?
  - How do we improve detection of OOS intents?

# Methodology

## 1. Data Augmentation:

- Generate 100 additional queries per intent, using ChatGPT

DA

I want to extend my dataset of Virtual Assistant queries for particular intents. I have given an intent and some sample queries for that intent. Give me 100 additional queries that are phrased similar the sample queries. All the queries should be unique and use a wide range of vocabulary. 

intent: "who\_do\_you\_work\_for"

sample queries:

1. do you know who you report to
2. what is your boss's name
3. are you influenced by someone else
4. is there another company you work for
5. would you say you are working for me

# Methodology

## 2. Data Validation:

- Random sample (~20 per intent) from generated queries
- Manually verify correctness
- Re-generate problematic intent queries

Intent: "current\_location"

Original query: "check maps for my location"

ChatGPT generated query: "Which casino is closest to my current location?"

ChatGPT re-generated query: "What's my current geographic location?"

-----

Intent: "credit\_score"

Original query: "how good is my credit score"

ChatGPT generated query: "How does my credit score impact my loan application?"

ChatGPT re-generated query: "Can you help me find my credit score online?"

# Methodology

## 3. Baseline Reimplementation

- Using the BERT-based approach by Larson et al. [1]

## 4. RoBERTa [2]

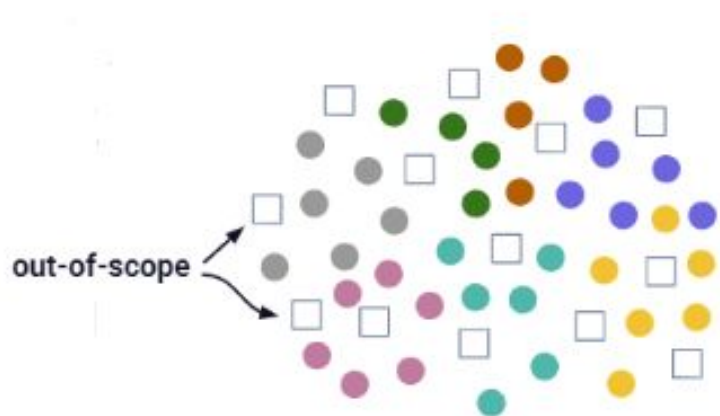
- Pre-trained on a much larger corpus of text than BERT.
- Uses dynamic masking and different pre-training tasks.
- Hypothesis:
  - RoBERTa's better generalization should improve in-scope accuracy as well as out-of-scope recall

# Methodology

## 5. Clustering:

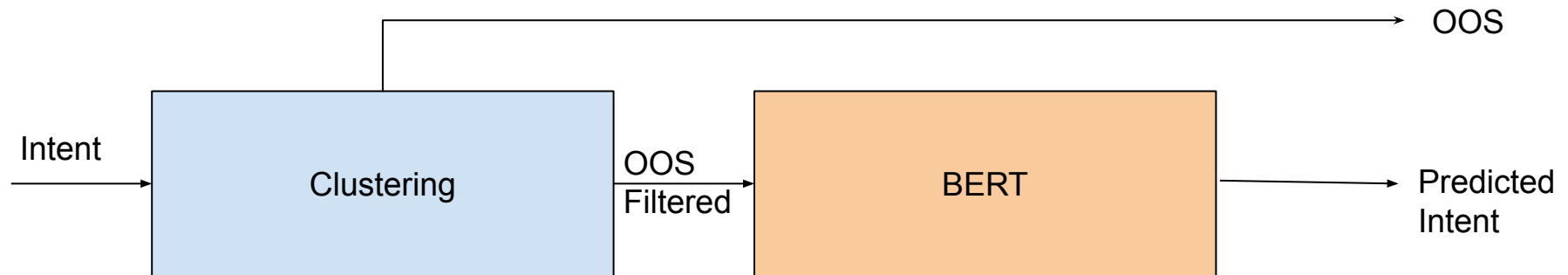
- Using K-Means with MPNET [3] Sentence Embeddings
- Hypothesis:
  - In-scope intents should form defined clusters
  - Out-of-scope intents are far from clusters

If distance to closest cluster (in higher dimensions)  $>$  threshold, then out-of-scope



# Methodology

## 6. Combined Approach:



# Activity Table

Activity	Why?	Deliverable	Responsible Member	Planned Time	Actual Time
Read Paper for the Dataset	Gather knowledge about the dataset	Summary of the article with important ideas highlighted	Both	1h	2h
Literature Review	Gather information about existing work	Review of literature on the dataset and related approaches	Both	5h	4h
Explore Dataset	Understand the structure of the data	Documentation of dataset characteristics and properties	Romario	2h	2h
Preprocessing data	Prepare data for training and evaluation	Cleaned and transformed dataset ready for training	Romario	3h	2h
Develop Baseline from paper	Establish a benchmark for comparison	Implementation of paper's approach on the dataset	Daniel	4h	3h
Generate Data for Augmentation	Increase dataset size	Augmented dataset for better performance	Daniel	3h	8h



# Activity Table

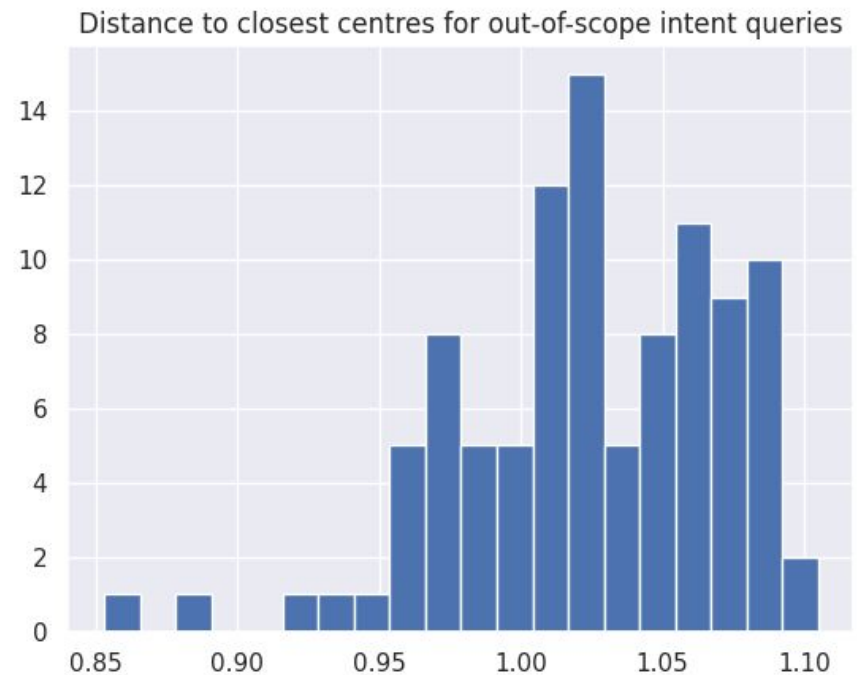
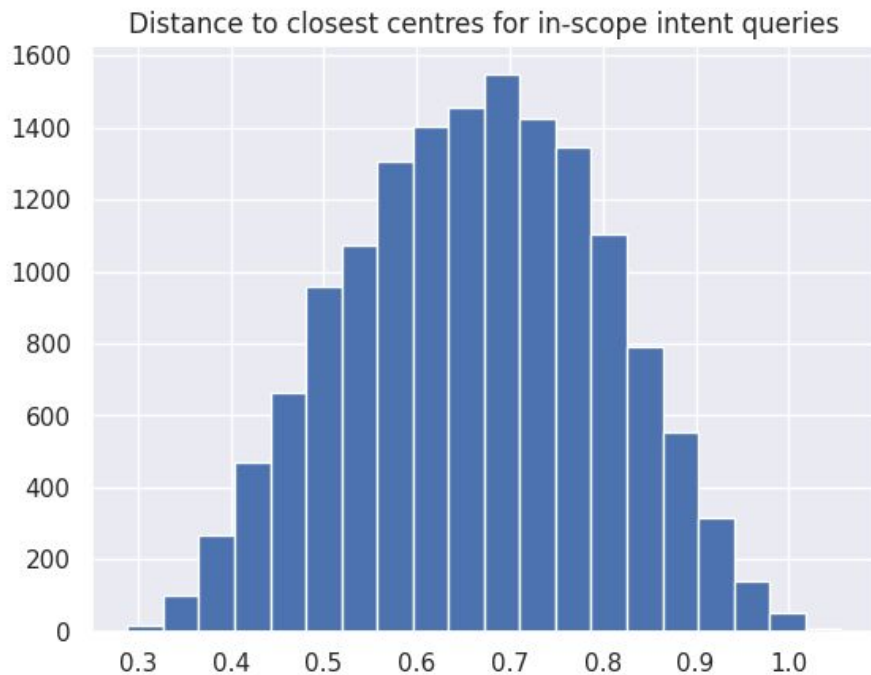
Activity	Why?	Deliverable	Responsible Member	Planned Time	Actual Time
Validate data and check for errors	Ensure data quality	Cleaned and validated dataset free of errors	Both	4h	4h
Build updated dataset and data splits	Prepare data for training and evaluation	Updated dataset ready for model training and evaluation	Romario	2h	1h
Re-implement baseline on new data	Evaluate baseline on new dataset	Performance of baseline approach on updated dataset	Daniel	1h	1h
Develop a new approach - A (clustering) + Analysis	Propose and evaluate a new approach	Implementation of clustering approach and analysis	Romario	6h	8h
Develop a new approach - B (RoBERTa) + Analysis	Propose and evaluate a new approach	Implementation of RoBERTa-based approach and analysis	Daniel	3h	3h
Comparative analysis	Compare performance of approaches	Comparison of performance of different approaches	Both	3h	3h

# Activity Table

Activity	Why?	Deliverable	Responsible Member	Planned Time	Actual Time
Combined Approach	Propose and evaluate a new approach	Implementation of Clustering + BERT	Both	3h	3h
Prepare Presentation	Document the project	Comprehensive presentation detailing the project and its results	Both	10h	6h
				<b>Total = 50h</b>	<b>Total = 50h</b>

# Results

Out-of-scope intents are far from clusters compared to in-scope intents ->  
Threshold of 0.95 selected based on graphs.



# Results

	Original Data		Augmented Data	
	In-Scope Accuracy	OOS Recall	In-Scope Accuracy	OOS Recall
Baseline BERT	0.97	0.46	0.97	0.48
RoBERTa	0.97	0.53	0.97	0.57
Clustering	N/A	0.77	N/A	0.74
Clustering + BERT	0.97	<b>0.77</b>	0.97	0.74

# Results

- Test Cases using the combined approach (Clustering + BERT)

Input Query	Predicted Intent	Actual Intent
i get the oil in my car changed quite frequently but i do not know how to do it	oil_change_how	oil_change_how
i get the oil in my car changed quite frequently but i want to do it myself	oil_change_when	oil_change_how
give me instructions to build a table	recipe	oos
can you guide me on how to build a table	oos	oos



# Results Discussion

- Data Augmentation was not very effective.
- More advanced models like RoBERTa did not help with in-scope accuracy, but were slightly more effective with OOS detection.
- OOS instances were farther away from cluster centres than in-scope instances.
- Unsupervised clustering was much more effective for OOS detection than supervised methods.
- A combination of clustering and BERT improves upon the baseline for OOS recall which results in better overall accuracy.

# Challenges

- Generating additional queries for augmenting the dataset was challenging.
- For some intent classes, ChatGPT did not understand the type of queries that should be generated, resulting in irrelevant or nonsensical queries.
- ChatGPT generated many queries with very different sentence structures than the original dataset query samples.
- Despite these challenges, generating additional data using ChatGPT was considerably faster than attempting to do it ourselves.

# What have we learned?

- Inherent subjectivity and ambiguity with manual verification.
- Large language models like ChatGPT are helpful for generating synthetic data, but are error-prone.
- Diminishing returns from dataset size increase
- Unsupervised approaches (like K-means) can help where supervised learning fails

# Conclusion

- Tackled a popular intent detection benchmark task
- Used novel (to the best of our knowledge) methods for:
  - ChatGPT-prompt engineering for data augmentation
  - Unsupervised clustering for OOS-detection
- Improved upon baseline metrics

## Resources and Links

- [1] S. Larson et al., 'An Evaluation Dataset for Intent Classification and Out-of-Scope Prediction', in Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 2019.
- [2] Y. Liu et al., 'RoBERTa: A Robustly Optimized BERT Pretraining Approach', CoRR, vol. abs/1907.11692, 2019.
- [3] K. Song, X. Tan, T. Qin, J. Lu, and T.-Y. Liu, 'MPNet: Masked and Permuted Pre-training for Language Understanding', arXiv [cs.CL]. 2020.

Code Repository:

<https://github.com/danlobo1999/csi5180-intent-classification/>