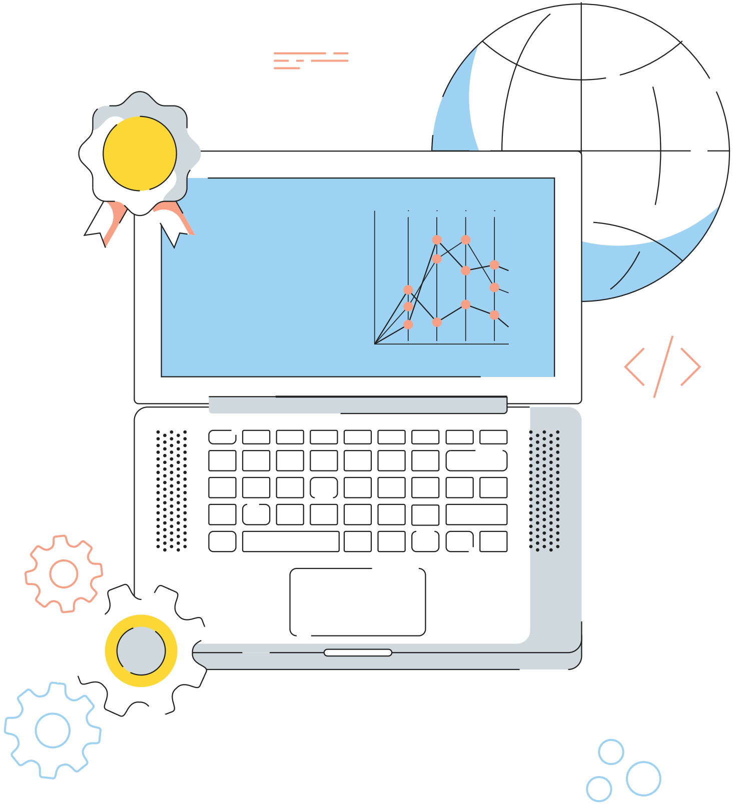
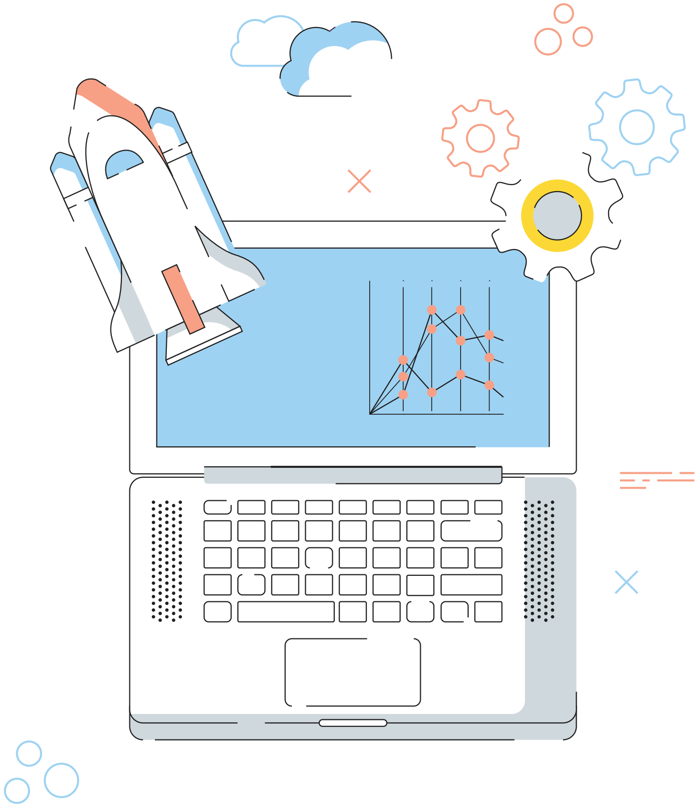
Sentiment Analysis on Human-to-Human Conversation



By Aden Lu, Garvin Chan, Jai Gollapudi



Background

- Based on our experience from service learning
  - Difficult to gauge the older adults' emotion
    - Monotone
    - Deadpan

- Existing research has shown that word choice can reflect a person's emotional wellbeing

- **Our Aim:** Gain a deeper understanding of the emotional states of older adults by analyzing their speech patterns, word choices, and conversational flow.

Data Overview

Dataset

EmoWOZ (MultiWOZ + ~~DialMAGE~~)

- Human-to-Human Conversations
- Manually annotated multi-domain task-oriented dialogue dataset (Booking, Reservations)

Statistics

- **Total Dialogues:** 10438
- **Total Sentences:** 71524 (65120 Unique)
- **Total Tokens:** 425933
- **Label Distribution:**

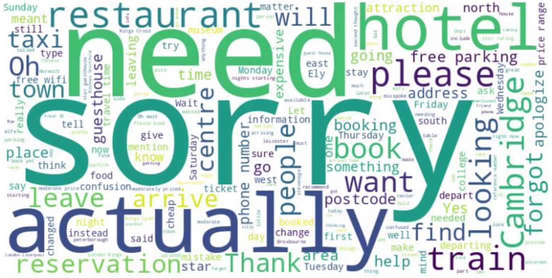
Distribution of Emotion Label in MultiWOZ:

	Label	Count	Percentage
0	Neutral	51426	71.9%
1	Fearful, Sad, Disappointed	381	0.53%
2	Dissatisfied, Disliking	914	1.28%
3	Apologetic	838	1.17%
4	Abusive	44	0.06%
5	Excited, Happy, Anticipating	860	1.2%
6	Satisfied, Liking	17061	23.85%

Word Cloud for Neutral Texts

Word Cloud for Satisfied, Liking Texts

Word Cloud for Apologetic Texts



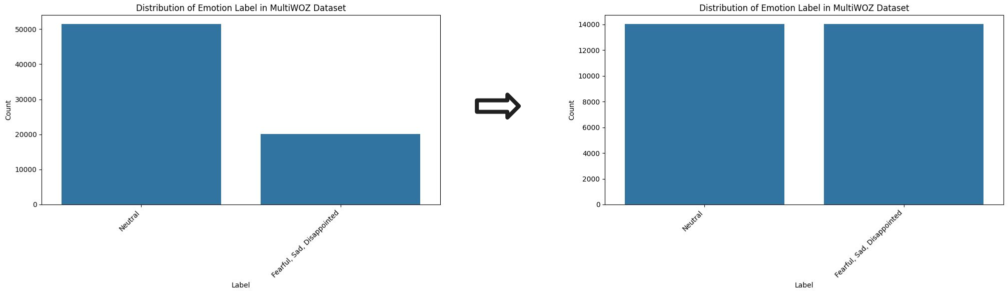
Data Pre-Processing

Problem:

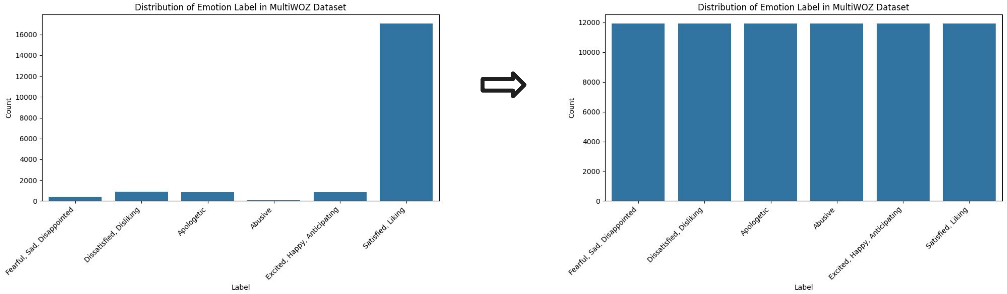
- Neutral and Satisfied, Liking make up 95% of the dataset
- Overfitting toward majority classes
- Back-translation, Synonym-replacement, etc (4+ hours on a T4 GPU) were insufficient to fully solve the disparity

Solution:

Sub-Model 1: Classifying as Neutral or not
Convert all labels other than neutral (‘0’) to non-neutral (‘1), upsample non-neutral to 50% and downsample neutral to 50%



Sub-Model 2: Classifying as one of the 6 emotions
Remove all labels that are neutral (‘0’) and upsample the minority classes to majority class (~16.5% each)



Model 1: Logistic Regression (Baseline)

Model	Feature Representation	Best Performing Representation	Best Hyperparameters (GridSearchCV)	Metrics (Weighted Avg)
Neutral Vs Non-Neutral	BoW, TF-IDF	BoW	C: 10 penalty': L2	**Accuracy:** 0.916 **Precision:** 0.92 **Recall:** 0.92 **F1:** 0.92
One of 6 emotions	BoW, TF-IDF	BoW	C: 10 penalty': L1	**Accuracy:** 0.889 **Precision:** 0.92 **Recall:** 0.89 **F1:** 0.90

Model 2: Feed Forward Neural Network

Model	Feature Representation	Best Performing Representation	Best Hyperparameters (Optuna)	Metrics (Weighted Avg)
Neutral Vs Non-Neutral	Word2Vec	Word2Vec (For now)	Epochs: 5+ LR: .001 Batch Size: 64 Dropout Rate: .5 Hidden Units: 128	**Accuracy:** 0.904 **Precision:** 0.91 **Recall:** 0.90 **F1:** 0.90
One of 6 emotions	Word2Vec	Word2Vec (For now)	Epochs: 5+ LR: .001 Batch Size: 64 Dropout Rate: .5 Hidden Unit: 128	**Accuracy:** 0.992 **Precision:** 0.99 **Recall:** 0.99 **F1:** 0.99

Model 3: BERT

Model	Feature Representation	Best Hyperparameters	Metrics (Weighted Avg)
All Emotions	None	**bert-based-uncased** Epochs: 3 Max Length: 128 Batch Size: 16 LR: 5e-5	**Accuracy:** 0.9251 **Loss:** .2374

Note: Haven't used hyperparameter tuning libraries to really fine-tune the values

Future Work

- Better data preprocessing
  - Back-translation (Requires more powerful GPU)
  - Other data augmentation methods/libraries (e.g. NLPAug, Synonym-replacement)

- Refine and tune neural network and BERT models
  - Run optuna and other hyperparameter libraries with a greater range of hyperparameter values

- Classify multiple emotions instead of one in a text
  - E.g. ‘I’m super excited that we went to this restaurant. I really enjoyed the food.’
  - Our current model might say this is either ‘Excited’ or ‘Satisfied’
  - In reality, this is around 50% ‘excited’ and 50% ‘satisfied’

- Text-only analysis currently
  - We want to capture and transcribe conversations with older adults to better understand their tone and conversation pattern