

# Capstone Project

Goh Hong Aik

# Problem Statement

Suppose you are a new hire at XYZ company and it's your first day at work, after a routine orientation with HR, your boss summons you to his office.

Hey! We just launched a new product and management wants to know what our customers think. We have about 10,000 reviews - can you analyse them and let me have the results by next week? Our usual guy is down with COVID...



Err... okay

# Problem Statement

Thinking that this is a basic unsupervised learning task with the well-established LDA and NMF, you get down to work, only to realise that these don't work very well for multi-label classification. Further, there seems to be nothing much you can do further to improve the model's accuracy.

Luckily, you remember your lessons from DSI and decide to combine supervised and unsupervised methods to build a model that can perform decently well.

# Problem Statement

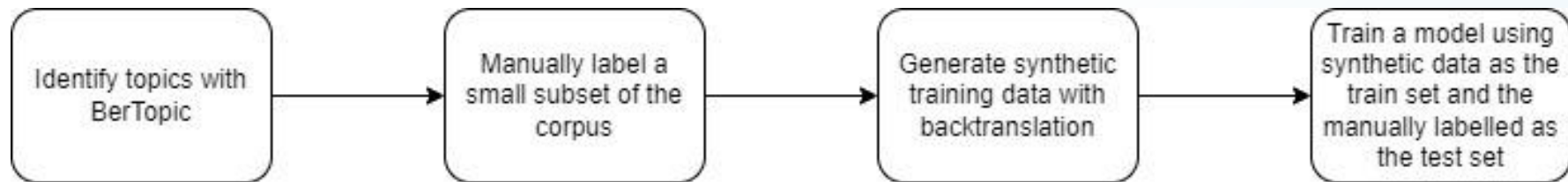
Intended outcome:

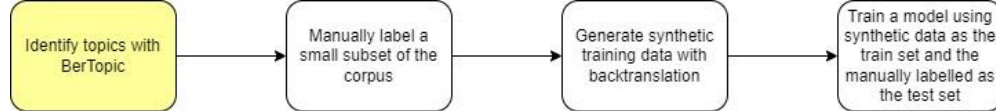
| Text                                      | Price | Flavor | Packaging |
|-------------------------------------------|-------|--------|-----------|
| I think it's delicious and affordable!    | 1     | 1      | 0         |
| The design could be more well thought out | 0     | 0      | 1         |

Data:

~10,000 documents in corpus

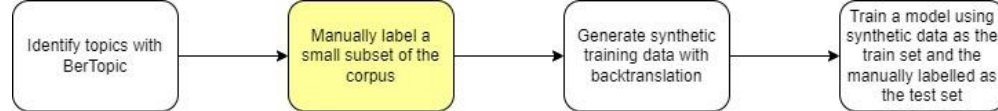
# Pipeline Overview





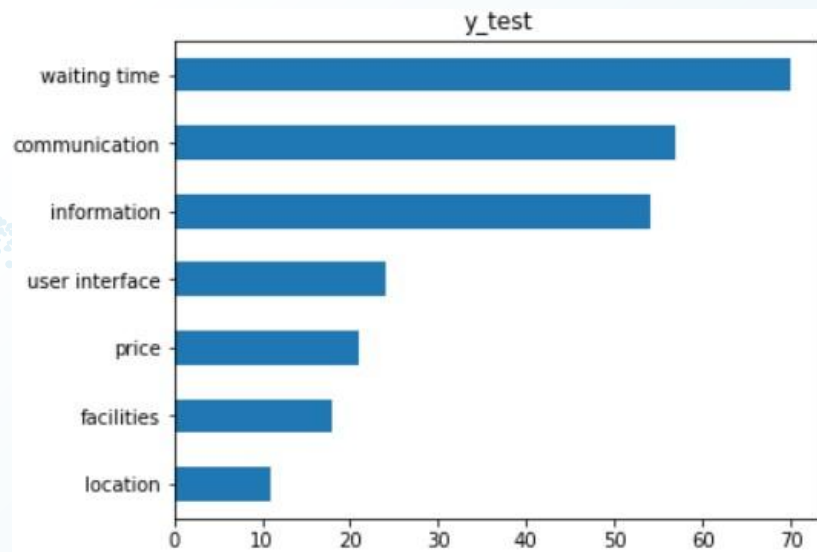
# BerTopic

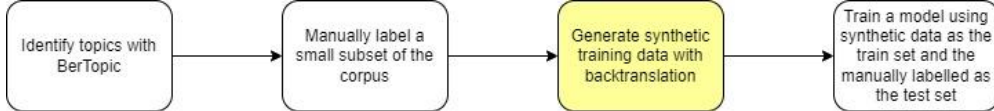
- Uses BERT sentence embeddings to vectorize text, unlike LDA and NMF which uses statistical methods
- Stochastic process, therefore results are not reproducible
- Does not do well in multi-label problems; insufficient as standalone topic modelling tool
- Produced  $> 100$  topics, condensed to 7 topics



# Manual Labelling

- Important to guide downstream modelling
- Labelled ~200 records, with each topic having at least 10 rows
- Act as test dataset





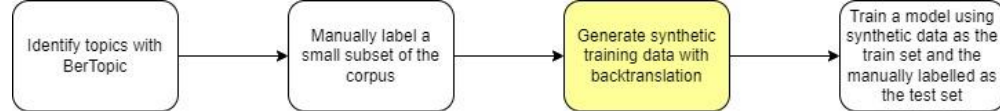
# Data Augmentation



The chicken was  
delicious

The poultry was  
tasty

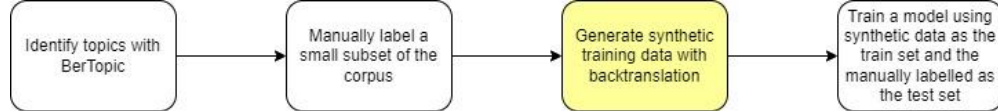




# BackTranslation

- Translates text into another language and back





# BackTranslation

- “Softer” way of altering text
- Leaks information into training data
- Better to use languages that are more different from English

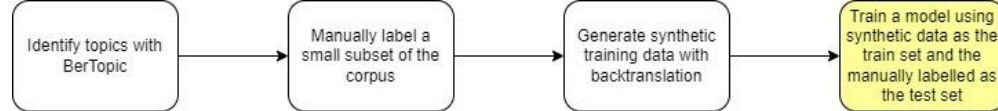
The language used by partners is very friendly and polite and easy to understand, the connection is also smooth

Backtranslation,  
ContextualWordEmbs,  
Sometimes

His words are very friendly, polite, and understandable, and my manner proceeds smooth.

My partner's words sound very friendly and easy, easy will understand, and smooth to connect.

His words are very friendly, polite, and understandable, and his communication is smooth.

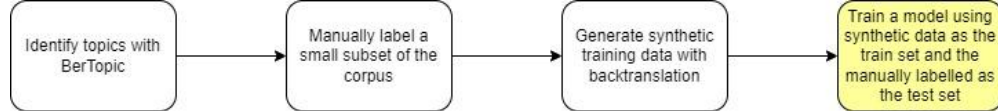


# Model Training

- Baseline model → Zero Shot Classification

```
{'sequence': '\n\nThe language used by partners is very friendly and polite and easy to understand, the connection is also smooth\n',  
'labels': ['communication',  
          'behavior',  
          'competency',  
          'information',  
          'facilities',  
          'user interface',  
          'location',  
          'price',  
          'response time',  
          'waiting time'],  
'scores': [0.9803360104560852,  
           0.7886410355567932,  
           0.754747211933136,  
           0.7030088901519775,  
           0.6719058156013489,  
           0.6212607622146606,  
           0.3871709108352661,  
           0.33242109417915344,  
           0.3129715621471405,  
           0.13848033547401428]}
```

**Weighted F1 score = 54%**



# Model Training

- Multi-label classification → OneVsRestClassifier
- SVC as default classifier due to superior performance

| Expt | Vectorizer                 | Weighted F1-score (test set) |
|------|----------------------------|------------------------------|
| 1    | Word2Vec                   | 88%                          |
| 2    | BERT                       | 80%                          |
| 3    | BERT with PCA              | 92%                          |
| 4    | BERT (fine-tuned)          | 72%                          |
| 5    | BERT (fine-tuned) with PCA | 90%                          |

- Much more whitespace to finetune compared to LDA/NMF/BerTopic alone!

# Further Evaluation (Holdout set)

| Expt | Vectorizer                 | Weighted F1-score (holdout set) |
|------|----------------------------|---------------------------------|
| 1    | Word2Vec                   | 63%                             |
| 3    | BERT with PCA              | 54%                             |
| 5    | BERT (fine-tuned) with PCA | 35%                             |

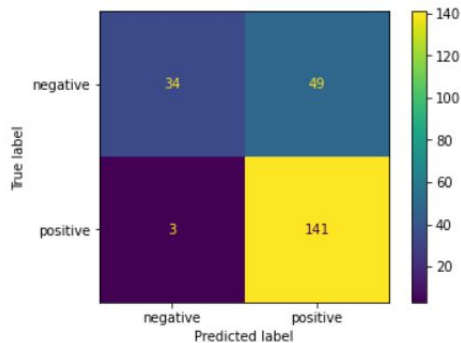
## Limitations

- Volatile changes in %
- Language is subjective; could have multiple interpretations how a text should be labelled

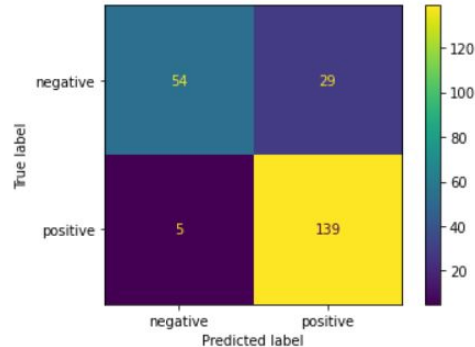
# Further Evaluation (Inspecting “Others”)

| sequence                                | w2v                          | bert_untuned              | bert_tuned   |
|-----------------------------------------|------------------------------|---------------------------|--------------|
| my problem is now 6 weeks old and       | others                       | information               | waiting time |
| as to the staff it really varies per br | others                       | others                    | others       |
| there was clearly a language barrier    | others                       | information, waiting time | waiting time |
| the space was small and in the mid      | others                       | waiting time, information | waiting time |
| i think that the call agents can be m   | others                       | others                    | waiting time |
| I had an excellent experience with      | information, user interface  | information               | others       |
| The person at the counter mention       | information, price           | others                    | others       |
| Very good way of explaining what        | information                  | others                    | others       |
| Repair notification process: becaus     | waiting time                 | others                    | others       |
| I have a [REDACTED] Smart TV Wh         | information                  | others                    | others       |
| I took my phone in after a update a     | information                  | others                    | waiting time |
| For me I am satisfy with the service    | price                        | others                    | others       |
| I walked into Harvey Norman with        | information                  | others                    | waiting time |
| When I first called the retail shop t   | communication, price         | others                    | waiting time |
| Because from the original, the sma      | information, user interface, | others                    | information  |

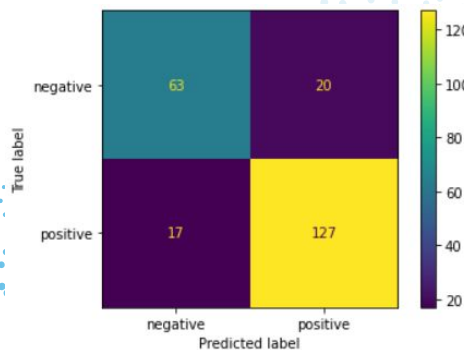
# Sentiment Analysis



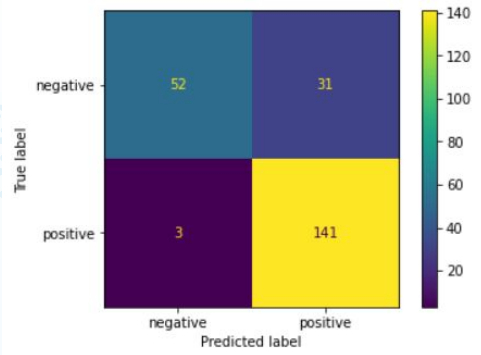
cardiffnlp/twitter-roberta-base-sentiment (F1-weighted = 74%)



facebook/bart-large-mnli (F1-weighted = 84%)



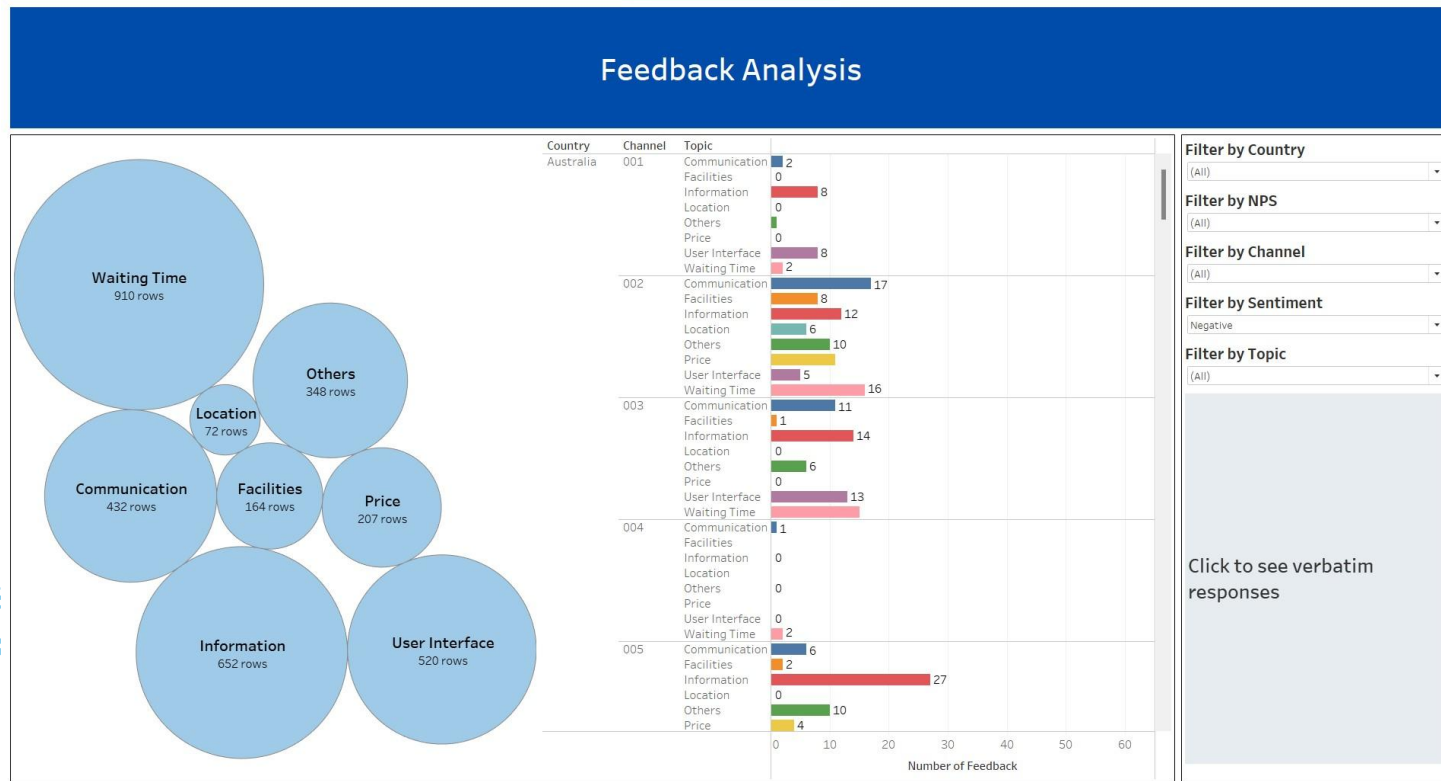
distilbert-base-uncased-finetuned-sst-2-english (F1-weighted = 84%)



Majority Vote (F1-weighted = 84%)



# Presentation





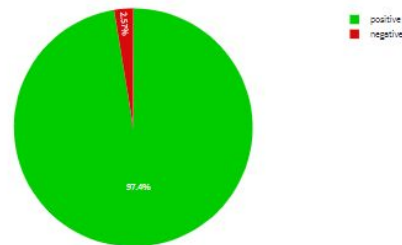
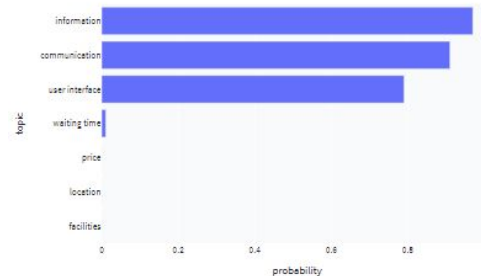
# Deployment

## Text Classification for Service Feedback

Type your text here

The website was user friendly and the agent provided good solutions

Click for predictions!



Or... Upload a csv file if you have a file instead.

Download sample file here

Please upload a csv file with only 1 column of texts.

# Conclusion

- Even with advancement of technology, topic modelling is notoriously hard to gauge model effectiveness without sufficient labelled data.
- With this pipeline, despite its theoretical flaws, the model performed decently.
- Also provided much whitespace to fine-tune to the context of the task, which is typically limited in unsupervised learning.

# Future Work

- Test out the pipeline on labelled dataset with multi-labels

# Thanks!

[LinkedIn](#) | [Email](#) | [GitHub](#) | [App Deployment](#)

