# TripAdvisor: Multiclass Classification on Unlabeled User Reviews
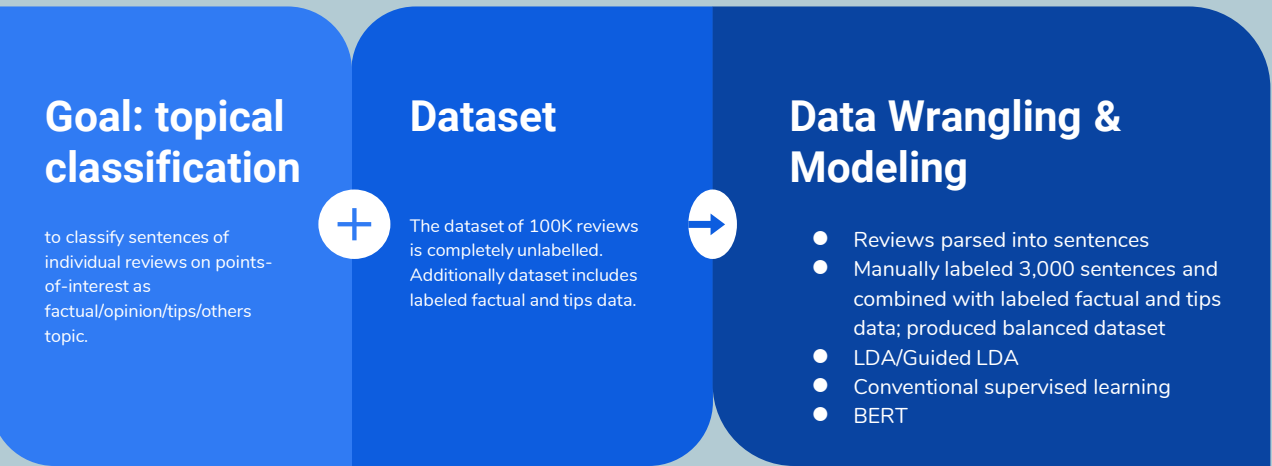
Yucong John Hu, Isaac Heeseok Joo, Zexi Ye

Center for Data Science *at* New York University

## Abstract

In association with mentors from **TripAdvisor** (NASDAQ: TRIP), we tested several machine learning models on unlabeled, raw online user reviews on points of interest (POI) to produce a satisfactory accuracy result on 4-class topical classification.

For the final output, we not only predicted a single label class, but also the distribution of probability over the 4 classes to provide more granular data and to allow levels of thresholding.

**Goal: topical classification**
to classify sentences of individual reviews on points-of-interest as factual/opinion/tips/others topic.

**Dataset**
The dataset of 100K reviews is completely unlabelled. Additionally dataset includes labeled factual and tips data.

**Data Wrangling & Modeling**
- Reviews parsed into sentences
- Manually labeled 3,000 sentences and combined with labeled factual and tips data; produced balanced dataset
- LDA/Guided LDA
- Conventional supervised learning
- BERT

## Data Exploration

- 100,000 rows of unlabeled user reviews on 100 points of interest
- 3,000 reviews manually labeled into categories of 'Facts', 'Emotion', 'Tips', or 'Others'


Number for each label class
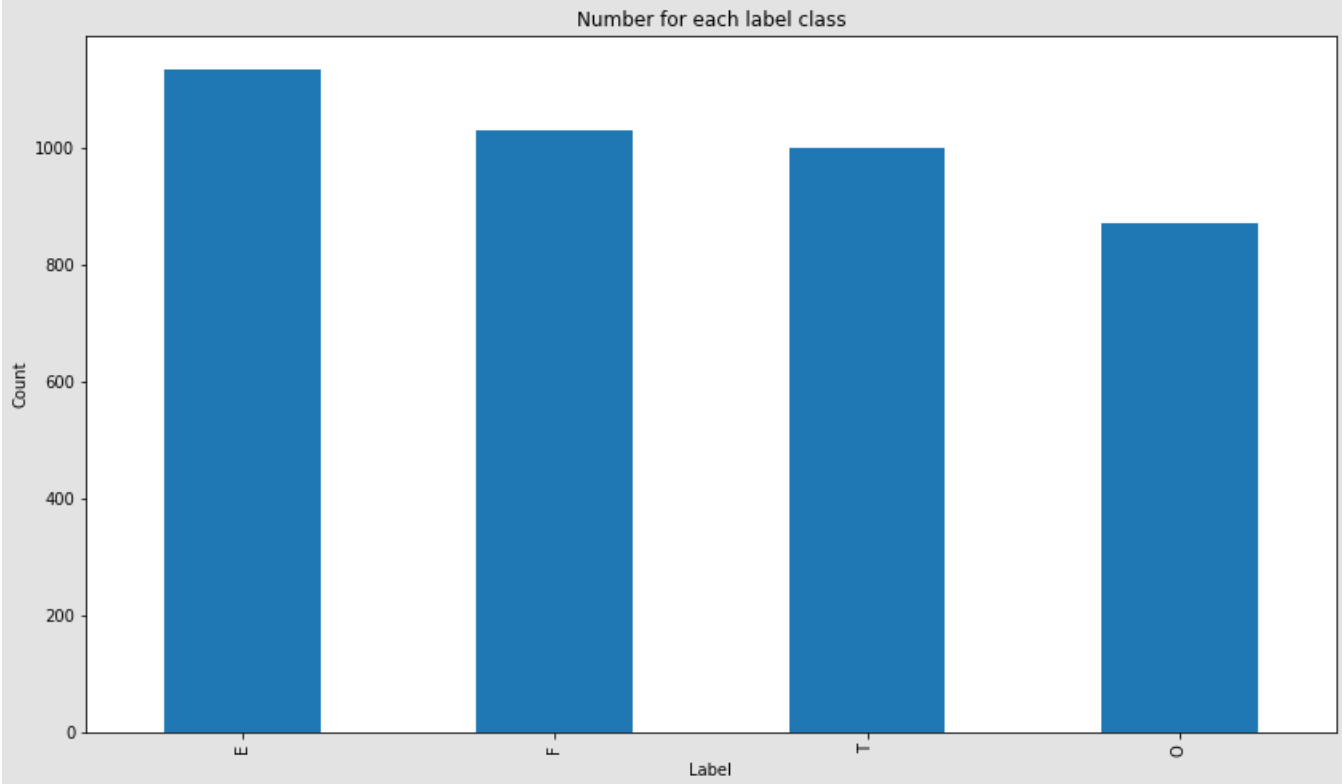
- With additional labeled data from TripAdvisor, we ended up with 4,100 reviews to work with
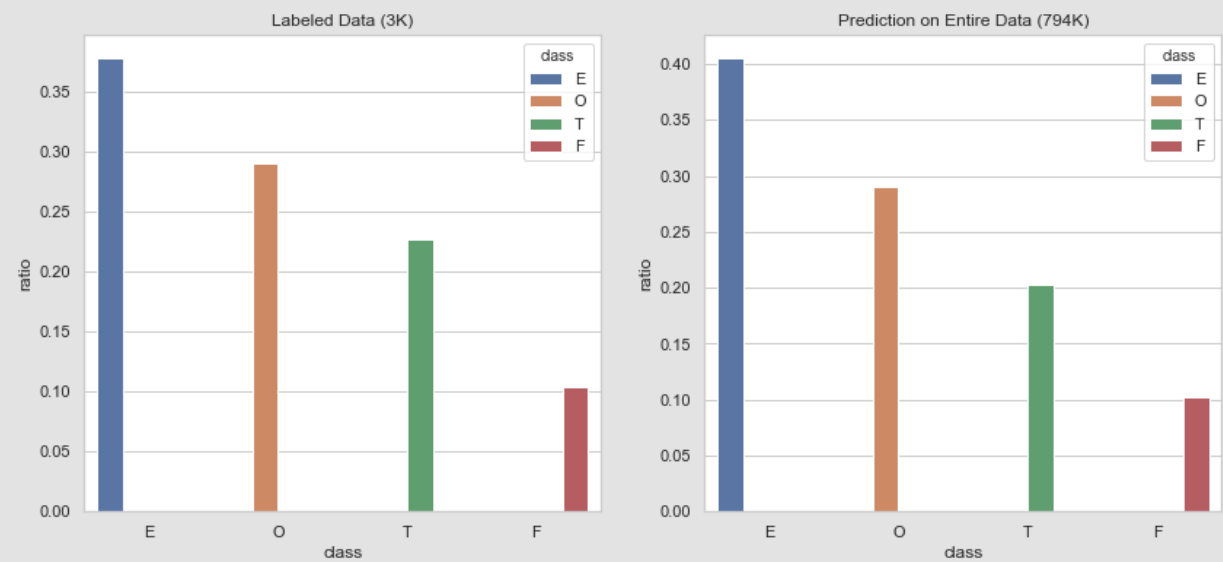


## LDA and Guided LDA

- LDA - Unsupervised topic modelling to categorize each review
- Regular LDA produced vague and ambiguous topic words that were not related to our target categories

- Guided LDA is semi-supervised modelling that nudges the model into desired categories by providing seed words
- Qualitatively, Guided LDA was able to produce more reasonable related words to each topic

| Categories | Seed Words | Resulting Words |
|---|---|---|
| Fact | history, tall, long, old, year, meter, wide | ride, park, visit, people, experience, year, make |
| Emotion | interesting, fantastic, fun, disgusting, beautiful, terrible | view, great, place,. beautiful, well, must, like, good |
| Tip | aware, recommend, tip, advice, suggest | get, ticket, go, time, day, line, wait, queue, buy, tour |

- Ambiguous words were still produced, so these words were grouped together to make another seed category 'Others', but it only lowered the testing score

Best test accuracy: **25.50%**

## Supervised Models

We select four conventional supervised learning models using 5-fold CV as baselines to see how they compare to BERT:
- Random forest
- Linear SVC
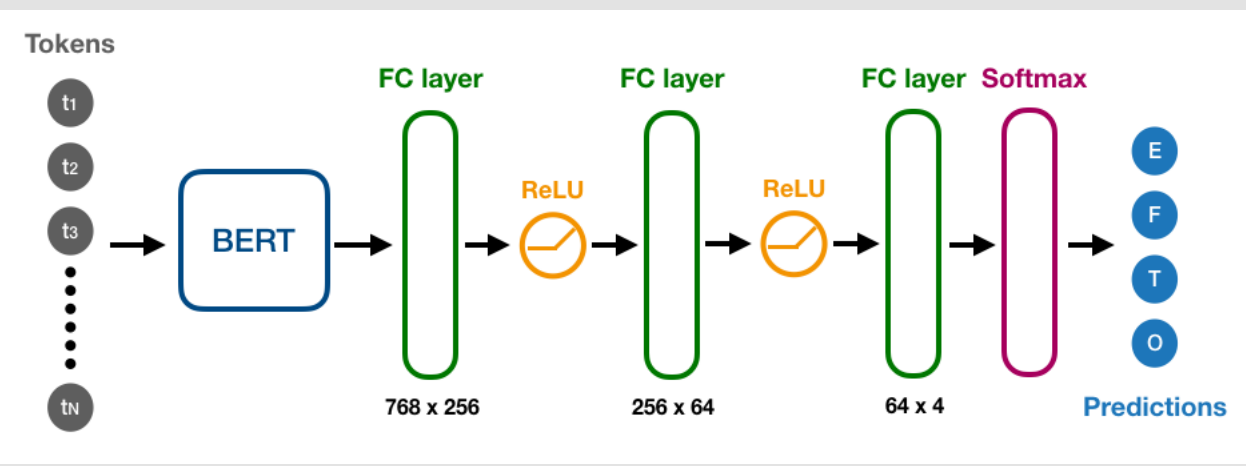- Multinomial NB
- Logistic regression

Best test **accuracy**: **64%**

Motivation: Using Amazon mechanical turk/manual labeling in industry to save engineering time v.s. computationally more expensive models
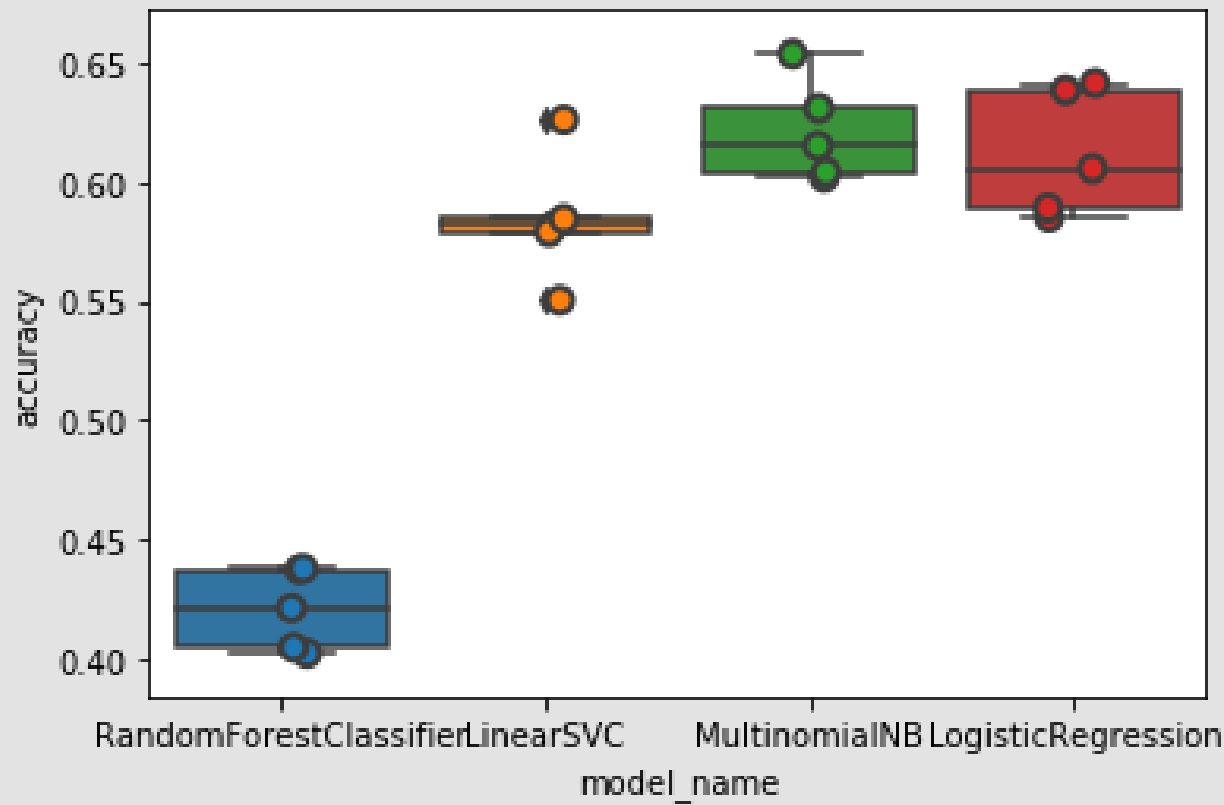


## BERT

- Key components
  - Self-attention
  - Transformer
  - Pre-trained weights

- Architecture
  - BERT feature extractor
  - Feed-forward network



- Training schemes
  - Supervised learning
  - Semi-supervised learning using Masked Language Modeling (MLM)

- Evaluation
  - Best test accuracy: **77.48%**

| | Training Accuracy on BERT Models | | | |
|---|---|---|---|---|
| | Supervised Frozen | Supervised Unfrozen | MLM Fine-tuned Frozen | MLM Fine-tuned Unfrozen |
| 1-layer classifier | 0.6559 | 0.7079 | 0.7203 | 0.7673 |
| 3-layer classifier | 0.7079 | 0.7475 | 0.7327 | 0.7748 |

- Examples of predictions

| Sentence | Prediction | Truth |
|---|---|---|
| In most cases the presidents only vaguely resembled their live (or previously alive) counterparts. | Fact | Fact |
| On the night tour it is still day light at the start, and they open up more sections for your to explore. | Tips | Fact |

- Ablation Studies
  - Weights re-randomization for identifying importance of each layer
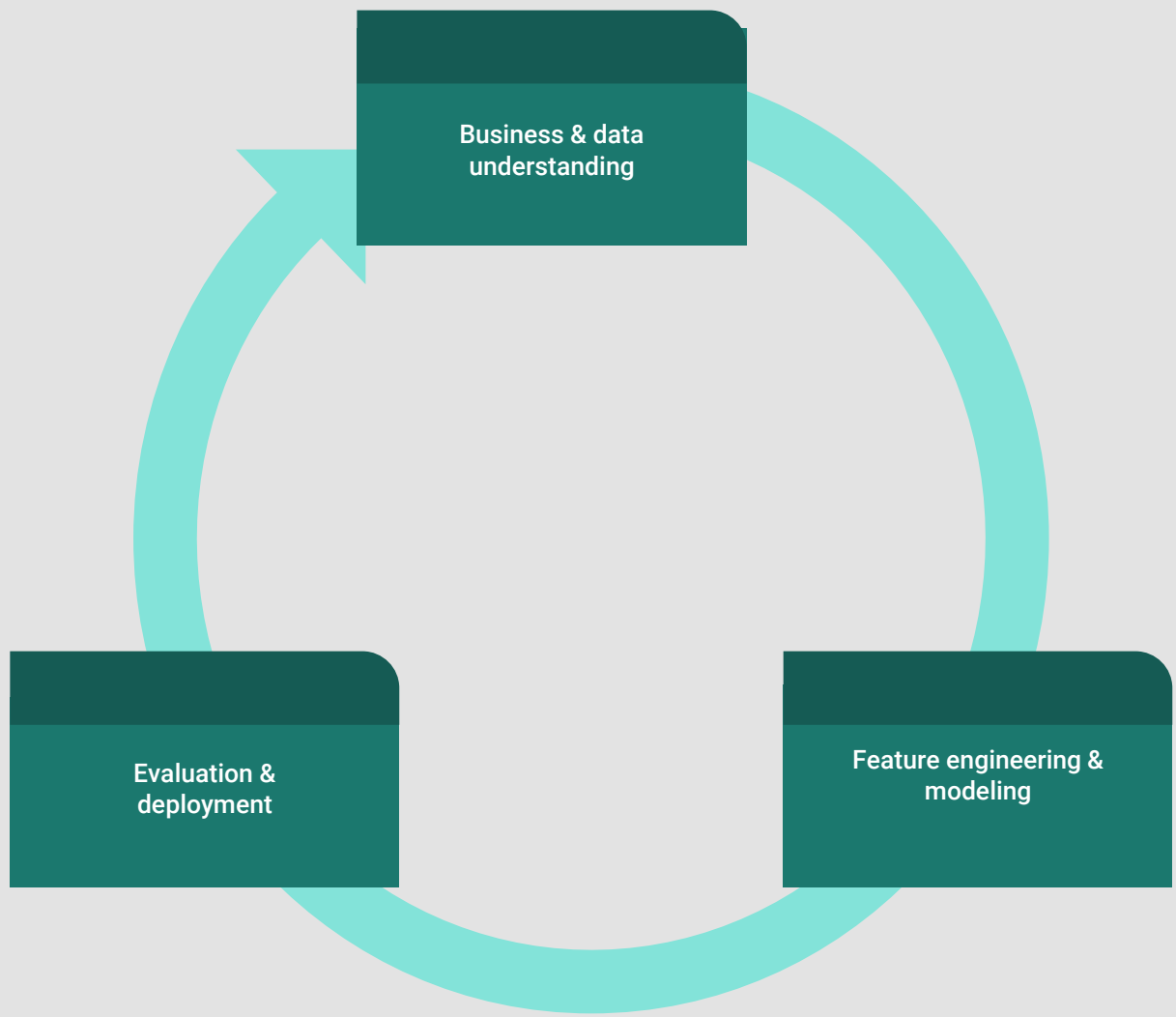  - Error compounds throughout forward-propagation
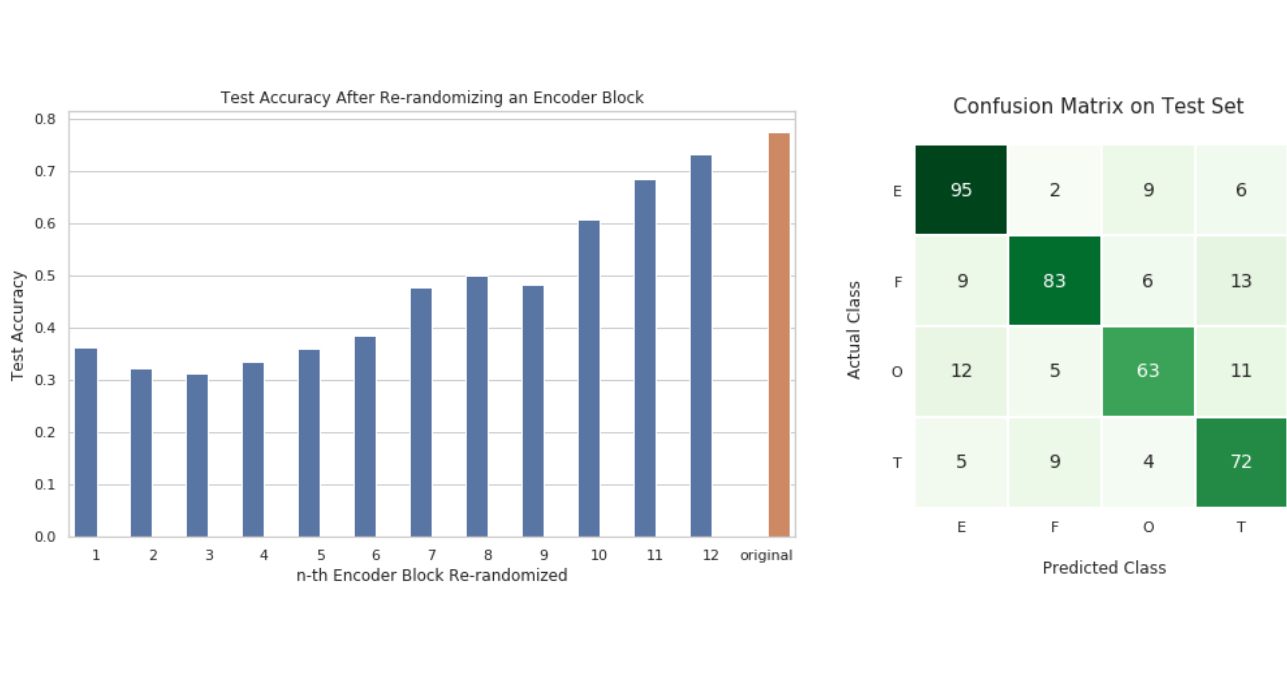


## Conclusion

- Our best model (BERT) achieves superior result, and produces a probability distribution that is relevant to the business goal
- LDA could provide higher accuracy if the parts of speech is known since 'Tips' often had imperative verbs

| | Deep Learning | Test accuracy | Algorithms |
|---|---|---|---|---|
| 1 | Supervised learning | ✗ | 64% | Naive Bayes; Logistic Regression; Support Vector Machine (SVM); Random Forests |
| 2 | Unsupervised learning | ✗ | 25.50% | Latent Dirichlet Allocation (LDA) & Guided LDA |
| 3 | Semi-supervised learning | ✓ | 77.48% | Bidirectional Encoder Representations for Transformers (BERT) |

## Recommendations

- closely align model output to business goal
- expand training dataset size (manual labeling / feedback from model)
- continuously retrain model



Business & data understanding

Feature engineering & modeling

Evaluation & deployment

## Acknowledgements