Recommending Smartphones Based on User Preferences

Introduction to Machine Learning Final Project

Emily Song, Aijia Xia, Tianji Li

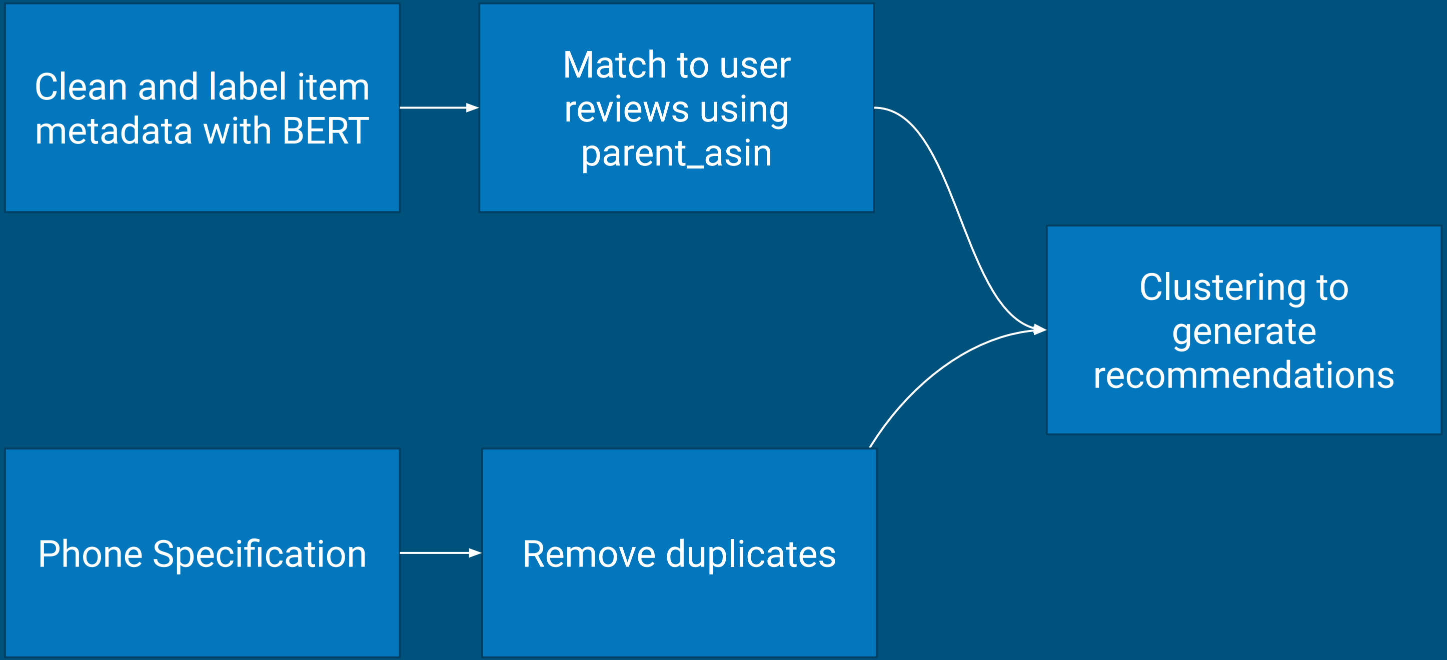
Problem Statement

- Huge global smartphone market -> vast number of available options -> a challenge to make informed decisions
- Goal: create a model that recommends similar smartphones that a user will most likely enjoy based on their existing smartphone

Datasets

- Phones 2024: Phone Listings from GSMArena.com
  - Comprehensive collection of phone information and specifications
  - Columns: phone_brand, phone_model, price_USD, storage, ram, dispaly_type, etc.
- Amazon Reviews 2023 - McAuley Lab
  - Subsets: user reviews and item metadata
  - Major product categories, focus on Cell_Phones_and_Accessories
  - User review columns: rating, review titles, review text, etc.
  - Item metadata columns: item title, features, description, price, etc.

Workflow



Clean and label item metadata with BERT

Match to user reviews using parent_asin

Clustering to generate recommendations

Phone Specification

Remove duplicates

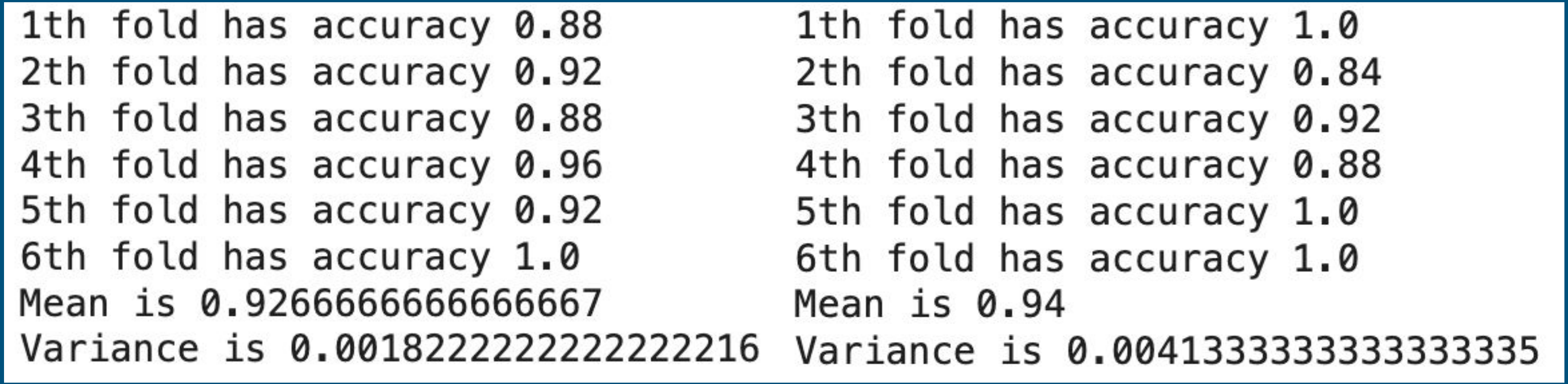
Data labeling using BERT

Amazon Review: Data Cleaning and Labeling

- 7271 items in item metadata containing both phones and phone accessories (case, charger, …)
- Need to classify phones from phone accessories
- Use BERT-based Model to conduct text classification based on product title
- Manually label 150 entries, train BERT Model, and apply to the rest of data for classification

BERT vs. RoBERTa

K-Fold Cross Validation Result: (125 train, 25 test)



BERT

```
1th fold has accuracy 0.88
2th fold has accuracy 0.92
3th fold has accuracy 0.88
4th fold has accuracy 0.96
5th fold has accuracy 0.92
6th fold has accuracy 1.0
Mean is 0.9266666666666667
Variance is 0.0018222222222222216
```

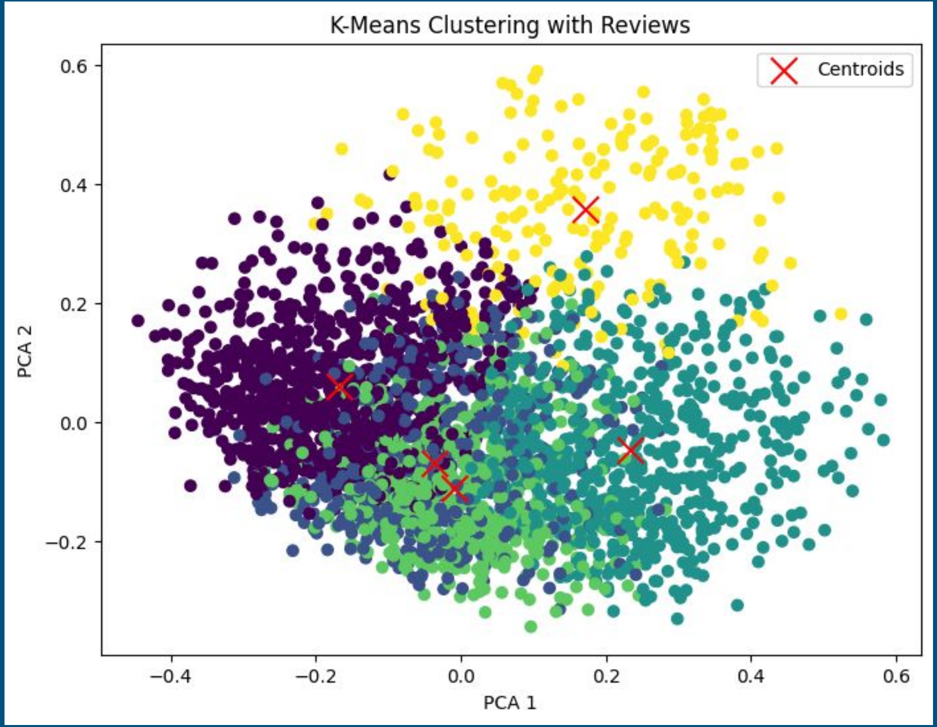
RoBERTa

```
1th fold has accuracy 1.0
2th fold has accuracy 0.84
3th fold has accuracy 0.92
4th fold has accuracy 0.88
5th fold has accuracy 1.0
6th fold has accuracy 1.0
Mean is 0.94
Variance is 0.004133333333333335
```

User Recommendation

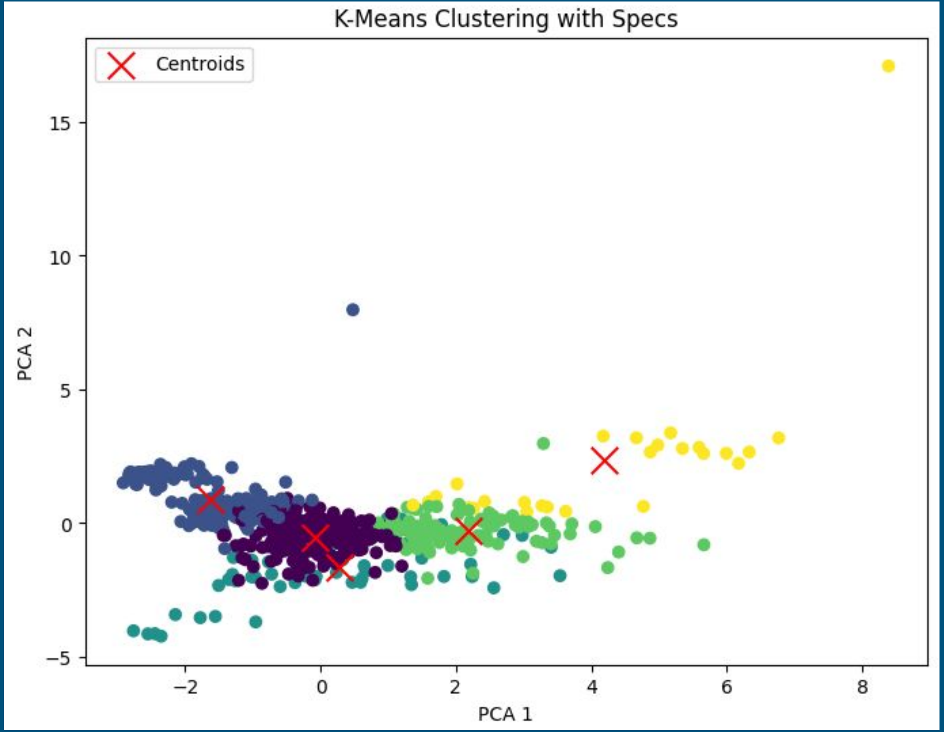
Clustering with Reviews

- Data cleaning + sentence embeddings
  - sentence transformer
- K-means + PCA
- Top n nearest neighbors in the same cluster



Clustering with Specs

- Feature selection
- One-hot encoding
- K-means + PCA
- Top n nearest neighbors in the same cluster

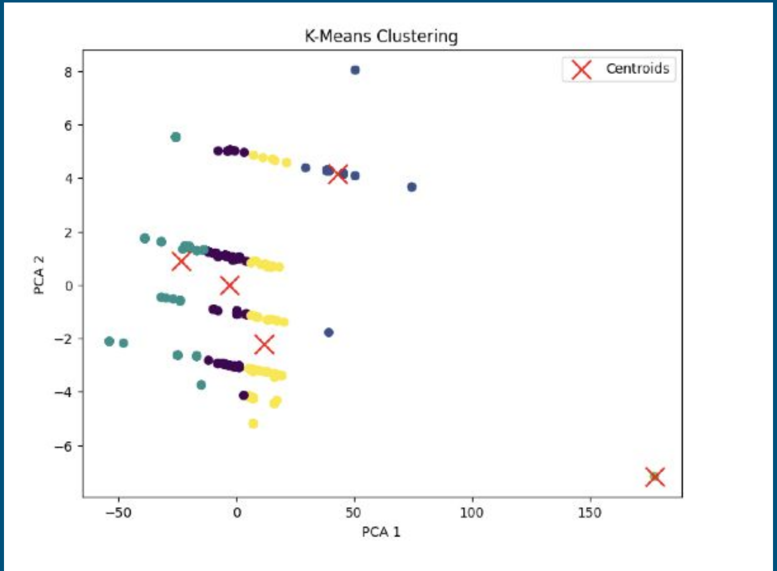


Use Cases

	Input	Recommendations
Review	Samsung Galaxy S22 5G	Samsung Galaxy A32 5G LG V30 Alcatel One Touch Fierce 2 Google Pixel 3 Samsung Galaxy S6
Specifications	Samsung Galaxy A23 5G	Samsung Galaxy A23 5G Samsung Galaxy A23 Nokia G60 Samsung Galaxy A23 Samsung Galaxy A13

Further Studies & Discussion

- Matched datasets - match reviews to phone specs
- Small dataset of only 438
- Linear formation
- Tight cluster
- Need more data or robust matching techniques



References

[1] McAuley-Lab/Amazon-Reviews-2023 · Datasets at Hugging Face — huggingface.co. https://huggingface.co/datasets/McAuley-Lab/Amazon-Reviews-2023. [Accessed 18-12-2024].

[2] Phones 2024 — kaggle.com. https://www.kaggle.com/datasets/jakubkhalponiak/phones-2024/data. [Accessed 18-12-2024].

[3] Topic: US smartphone market — statista.com. https://www.statista.com/topics/2711/us-smartphone-market/#topicOverview. [Accessed 18-12-2024].

[4] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, volume 1, page 2. Minneapolis, Minnesota, 2019.

[5] Yinhan Liu. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:*1907.11692, 364, 2019.

Thank you!