Introduction to Machine Learning Final Project Report: Recommending Smartphones Based on User Preferences

Emily Song, Aijia Xia, Tianji Li

December 18, 2024

Introduction

Our goal is to create a model that recommends similar smartphones that a user will most likely enjoy based on their existing smartphone.

The smartphone market in the United States is one of the largest in the world, with smartphone sales forecasted to reach almost 110 billion U.S. dollars in 2024 [3]. With the huge market and diverse choices, there is a significant need to meet customer expectations with appropriate smartphone recommendations, thereby increasing customer satisfaction and potential sales.

Data

We used the Phones 2024 data set[2], which is licensed as an open dataset. This data set includes various relevant features such as phone brand, model, price, storage, display type, RAM, chipset, battery, CPU, GPU, operating system, colors, and more.

The Phones 2024 data is preprocessed into a csv file, and we further process it by combining the same model of phones from different sellers that are listed as different entries to ensure there are no duplicate models.

A useful supplement to this data is Amazon Reviews 2023[1], which is open to use for research purposes and contains 571.54M Amazon user reviews for a wide range of products. It hosts abundant features such as user reviews (ratings, text, helpfulness votes, etc.), item metadata (descriptions, price, raw image, etc.), and links (*user-item / bought together* graphs), enabling studies into user behavior and consumer-product interactions. Specifically, we look at the user reviews and item metadata of the Cell_Phones_and_Accessories category.

Data Selection using BERT

Due to Amazon classification, the majority of entries in Cell_Phones_and_Accessories category are not smartphones but electronic accessories like cases, chargers, and holders. Since the original data set has not labeled whether a product is a smartphone or not and manually labeling for all data sets is burdensome, we use the methodology of first training and testing model on a small manually labeled dataset, then apply the model to label all the data. Following this methodology, we decide to use Bidirectional encoder representations from transformers (BERT)[4], a powerful self-supervised model that studies latent representation of tokens in texts, for both its strong performance in various

natural language processing tasks and its strong ability to transfer learning on specific tasks by fine tuning with a small train set. For the purpose of our task, we append a classification layer on BERT for text classification.

For the purposes of reducing the dimension of data (the original dataset has 220000+ entries), we take a cut of the price at $50 and only consider smartphones with available price information that are priced $50 and upwards.

In order to train BERT, we first manually labeled 150 data randomly selected from the dataset, where 0 is "smartphone", and 1 is "not smartphone". To decrease possible human error, we conduct this by independently labeling these 150 data, and then compare and cross-check for all the differences, then reach the final labeling.

To test the validity of the BERT model, we conduct a K-Fold Cross Validation by splitting train and test to 125 and 25(6 folds in total). We have run the program several times, and the results show that the average accuracy for BERT in these folds is around 93%(highest 100%, lowest 88%), with variance around 0.001, which shows a pretty good and consistent performance.

We have also conducted the same procedure for RoBERTa[5], which is a BERT-based model with some modifications in pre-train objective and hyperparameters. The results show RoBERTa has an average accuracy of around 94%(highest 100%, lowest 84%), and a variance of around 0.004. Although the performance for RoBERTa is slightly better than BERT, we decide to stick with BERT for its smaller variance and larger minimum accuracy across the folders. We then used the trained BERT model to label all the 7000+ data in the data set.

	BERT	**RoBERTa**
1st Fold	88%	100%
2nd Fold	92%	84%
3rd Fold	88%	92%
4th Fold	96%	88%
5th Fold	92%	100%
6th Fold	100%	100%
Mean	92.6%	94%
Variance	0.18%	0.41%

Table 1: Comparison of BERT and RoBERTa Performances

Model

Given the large size and nature of our data, we have selected clustering to generate recommendations. Clustering algorithm is a fundamental unsupervised machine learning technique that groups similar data points based on their inherent characteristics. The unsupervised nature makes clustering suitable for the thousands of phones and reviews, and clustering with embedded review texts allows for the extraction of underlying themes or sentiments that can contribute to higher similarity and thus more informed recommendations.

Clustering of Phone Models with Reviews

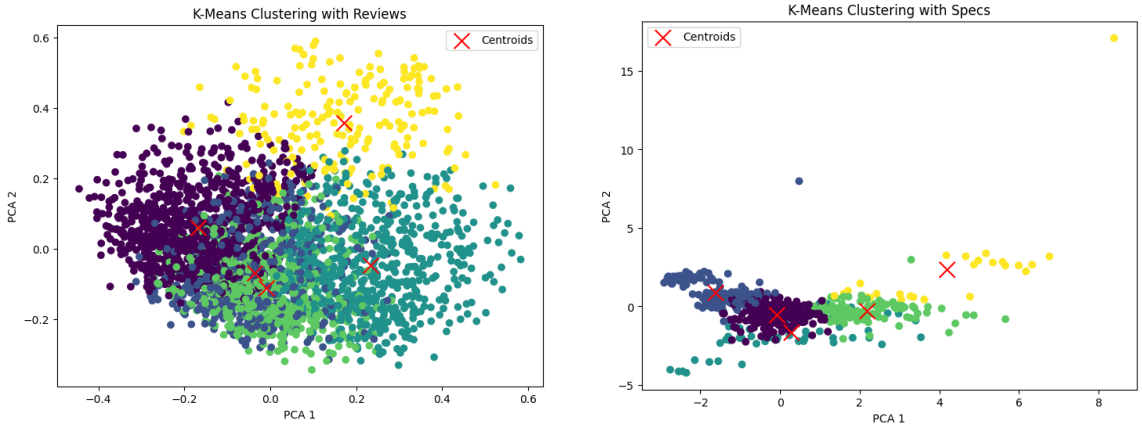
With the processed Amazon Reviews 2023 dataset, We concatenate the individual reviews for each product and remove symbols and bracketed texts. We choose to embed the resulting review texts with a pretrained sentence transformer model as we expect sentence embeddings to be essential for embedding reviews. Out of the available pretrained sentence transformers, we select "all-MiniLM-L6-v2" as it has relatively satisfying performance (58.80) with reasonable size (80 MB)[6]. Given sentence embeddings, we used k-means clustering and PCA (Principal Component Analysis) to reduce the dimensions for quick visualizations. Since varying the size of k does not reflect in silhouette scores and Davies-Bouldin scores, we decide to use k=5 based on the visualizations. Given the review embeddings and clusters, we return the top n nearest neighbors in the same cluster of a given product, where n is specified. Thus, we could recommend phone products simply based on a past/preferred choice of the user.

Clustering of Phone Models with Specifications

With the processed dataset Phones 2024, we remove features that we deem as uninformative or repetitive. We normalize the numerical features with the standard scaler and use one-hot encoding to transform categorical features to numerical ones. We again use k-means clustering and PCA on the processed features. Varying k also does not change the silhouette scores and Davies-Bouldin scores, so k = 5 is chosen for consistency. The same function is used here to find the top n nearest neighbors in the same cluster, generating a recommendation based on phone product features.

Results

The plots of clustering are as follows:

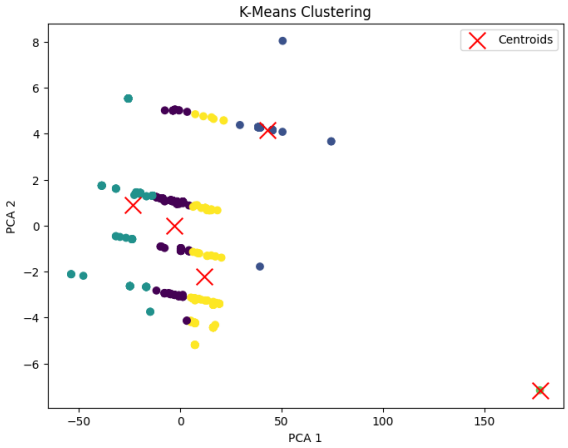


We have tested the two recommendation methods with different specified phone products. Generally, the recommender based on reviews give suggestions with higher variance in models. For example, given input "Samsung Galaxy S22 Plus 5G", the top 5 recommendations would be: "Samsung Galaxy A32 (5G), LG V30 Silver for Verizon, Alcatel One Touch Fierce 2, Smartphone, Pixel Phone

3, Samsung Galaxy S6". For the recommender based on phone features, given the input "Samsung Galaxy A23 5G", it would recommend: "Samsung Galaxy A23 5G, Samsung Galaxy A23, Nokia G60, Samsung Galaxy A23, Samsung Galaxy A13". So we infer that the recommender based on reviews might give more innovative suggestions due to its user-oriented property.

Discussion

A natural next step is to cross-relate the two data sets and analyze reviews for phone models with available specifications. To achieve this, we matched the products labeled as "smartphone" by BERT with phone models in the Phones 2024 dataset by considering model names and storage and comparing with product title. The resulting matched data set has 438 phone models with their specifications and Amazon reviews. Clustering algorithm is then run considering both specifications and review tokens. For dimensionality considerations we selected the most relevant features in specifications only. The result is as follows:



Clearly, performance is hindered by the significantly smaller data size, and it appears that data points are aligned in linear formation and clustered vertically with respect to PCA 1. This likely indicate the algorithm picked up a strong relevant feature targeting highly standardized features, which can be useful in determining most desirable smartphone traits. Further studies can be carried out with more available matched data.

Conclusion

Using two large-scale real-life datasets, we have demonstrated the feasibility of machine learning techniques in solving problems relevant to everyday tasks like smartphone recommendation. We utilized both BERT for data processing and labeling and clustering for feature correlation and similarity extraction to develop robust recommendation systems based on smartphone features and past reviews. The methods described can be generalized to similar complex real-life scenarios involving a large amount of text-based data.

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.

[6] SBERT.net. Pretrained Models &x2014; Sentence Transformers documentation — sbert.net. https://www.sbert.net/docs/sentence_transformer/pretrained_models.html. [Accessed 19-12-2024].