



Walmart Inc.



The University of Texas at Austin
Texas McCombs
MS Business Analytics
McCombs School of Business



Product Item Structure

Junhong Xu, Palakh Gupta, Sichun Zuo, Shivank Sood
Prof. Tejwansh Anand, Prof. Daniel Mitchell



The University of Texas at Austin
Texas McCombs
MS Business Analytics
McCombs School of Business

Agenda



Business intent



Exploratory Data Analysis



Approach & Methodology



Operational Execution



Conclusion



Next Steps

The Problem

Walmart.com

Moosejaw

BONOBOS

hayneedle

jet

65+ Online Banners



Consistent Experience Across Online Banners

Goal: Develop one consistent product categorisation using Walmart.com



Better Visibility of Product Structure

Goal: Form a space in which the relative positions of items represent their semantic similarity.

Business Value

- The final product can be integrated to the internal text analytics platform at Walmart
- Cleaning and Standardizing Categories for future product addition

Product	Walmart Category	Shop by Brand / Electronics by Brand / HP / HP Desktops
Refurbished HP 290-p0043w Slim Celeron G4900 3.1GHz 4GB RAM 500GB HDD Win 10 Home Black	Standardized Category	Electronics/Computers/Desktop

- Reduced manual labelling (categorization) hours for new products and products on subsidiary websites
- Reduce computation time during Product Quote Analysis by the Audit Team

Agenda



Business intent



Exploratory Data Analysis



Approach & Methodology



Operational Execution



Conclusion



Next Steps

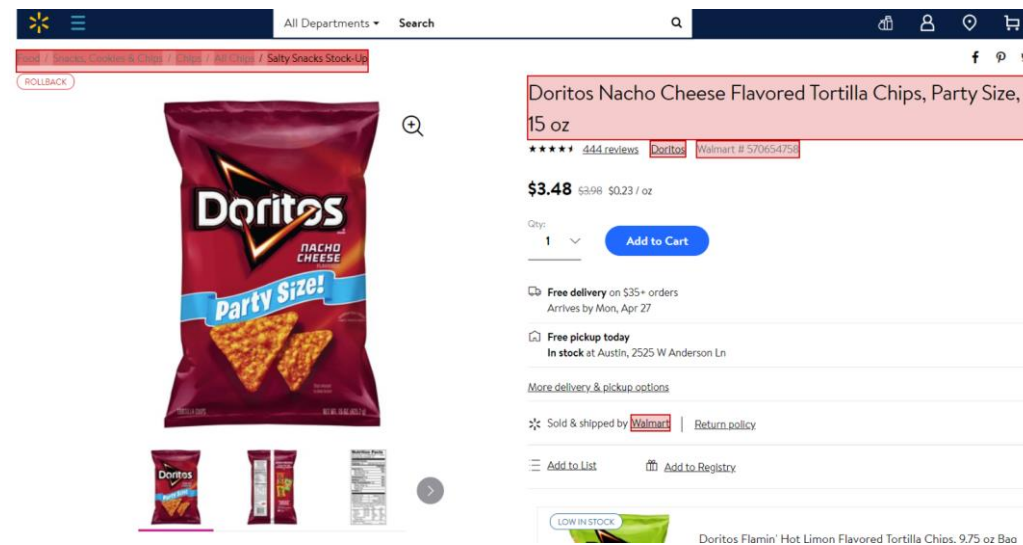
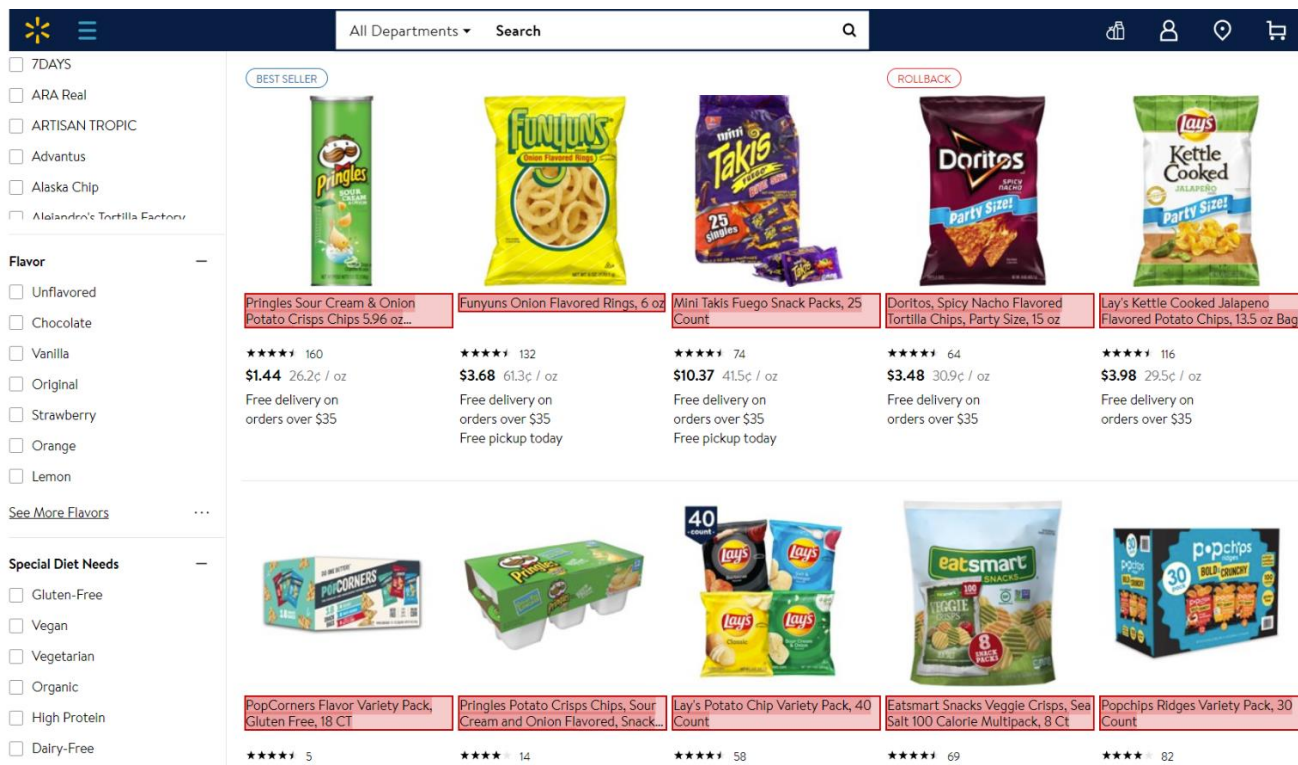
The Data

43M+
Products on
Walmart.com

16,895
Products
Scraped

3M+
Words Text Data
Product Name
Product Description

5
Base Categories
20
Sub Categories
112
Sub-Sub Categories



About This Item

We aim to show you accurate product information. Manufacturers, suppliers and others provide what you see here, and we have not verified it. [See our disclaimer](#)

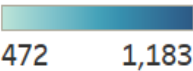
DORITOS isn't just a chip. It's fuel for disruption - our flavors ignite adventure and inspire action. With every crunch, we aim to redefine culture and support those who are boldly themselves. Are you ready? If so, crunch on.

- Delicious spicy nacho flavor
- Large bag is perfect for sharing
- Crunchy straight from the bag or after scooping up dip or salsa

Products in Each Category

Number of Products in Each Category

Main Category	Subcategory	
Beauty	Hair Care	640
	Makeup	637
	Nail Care	598
	Skin Care	643
Electronics	Computers	797
	Home Audio & Theater	479
	Printer & Supplies	957
	TV	472
Food	Beverages	963
	Coffee	967
	Meal Solutions, Grains & Pasta	740
	Snacks, Cookies & Chips	777
Garden & Patio	Grills & Outdoor Cooking	1,090
	Outdoor Heating	934
	Outdoor Power Equipment	1,183
	Patio Furniture	749
Home, Furnitures & Appliances	Appliances	1,093
	Furnitures	1,093
	Home Decor	1,046
	Kitchen & Dining	1,037



SubCategory	Sub Sub Category	
Beverages	Energy Drinks/All Energy Drinks	187
	Juices	196
	Milk	200
	Soda Pop	191
	Water	189
		963

Sum of Number of Products broken down by Main Category and Subcategory. Color shows sum of Number of Products. The marks are labeled by sum of Number of Products.

Agenda



Business intent



Exploratory Data Analysis



Approach & Methodology



Operational Execution



Conclusion



Next Steps

Feature Engineering

Three Word Embedding Technology:

- 

*fast*Text

-  – Universal Sentence Encoder

Feature Engineering

Data Input:

- **Product name**

Doritos Nacho Cheese Flavored Tortilla Chips,
Party Size, 15 oz

★★★★★ [444 reviews](#) [Doritos](#) Walmart # 570654758

\$3.48 ~~\$3.98~~ \$0.23 / oz

- **Product Description**

About This Item

We aim to show you accurate product information. Manufacturers, suppliers and others provide what you see here, and we have not verified it. [See our disclaimer](#)

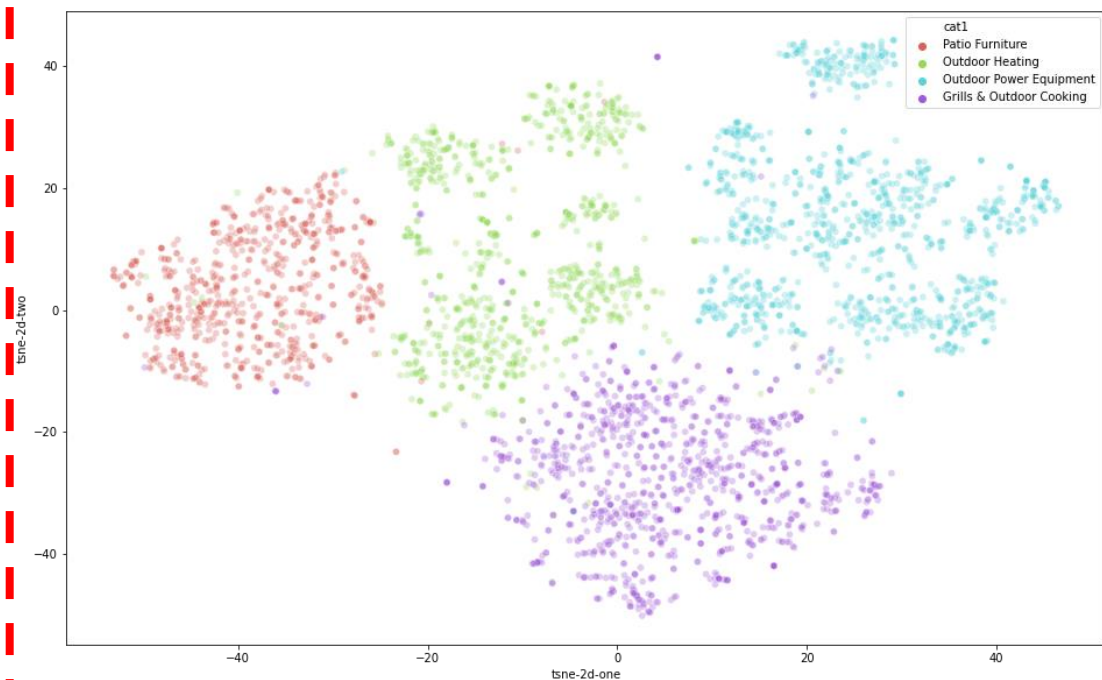
DORITOS isn't just a chip. It's fuel for disruption - our flavors ignite adventure and inspire action. With every crunch, we aim to redefine culture and support those who are boldly themselves. Are you ready? If so, crunch on.

- Delicious nacho cheese flavor
- Large bag is perfect for sharing
- Crunchy straight from the bag or after scooping up dip or salsa

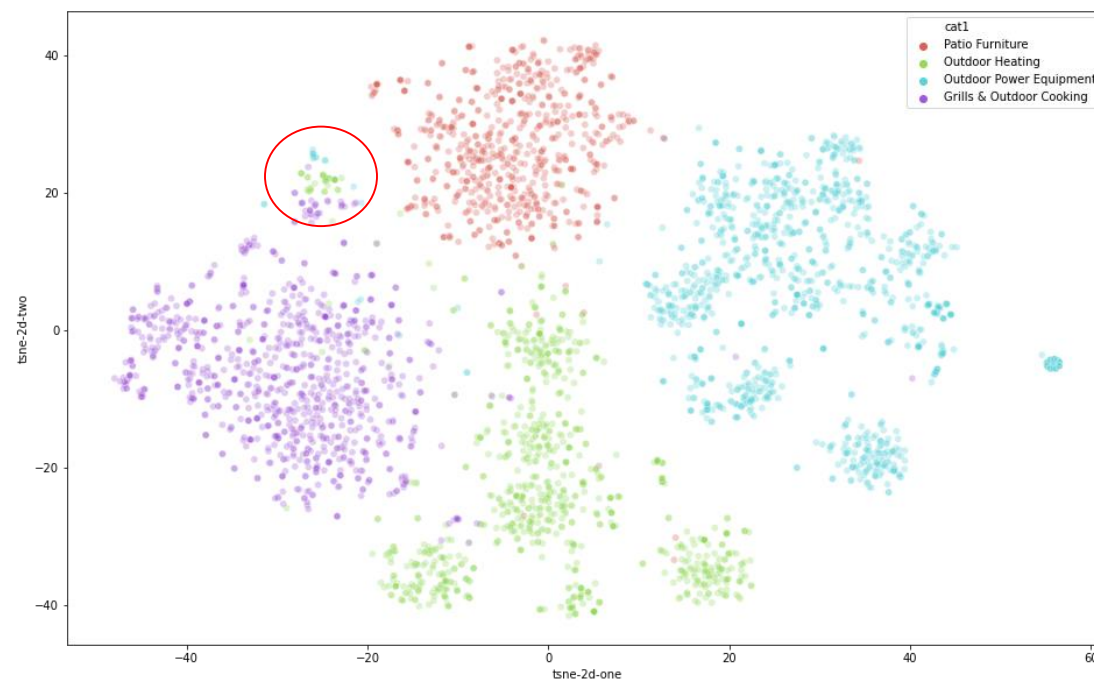


Feature Engineering

USE + Product Name

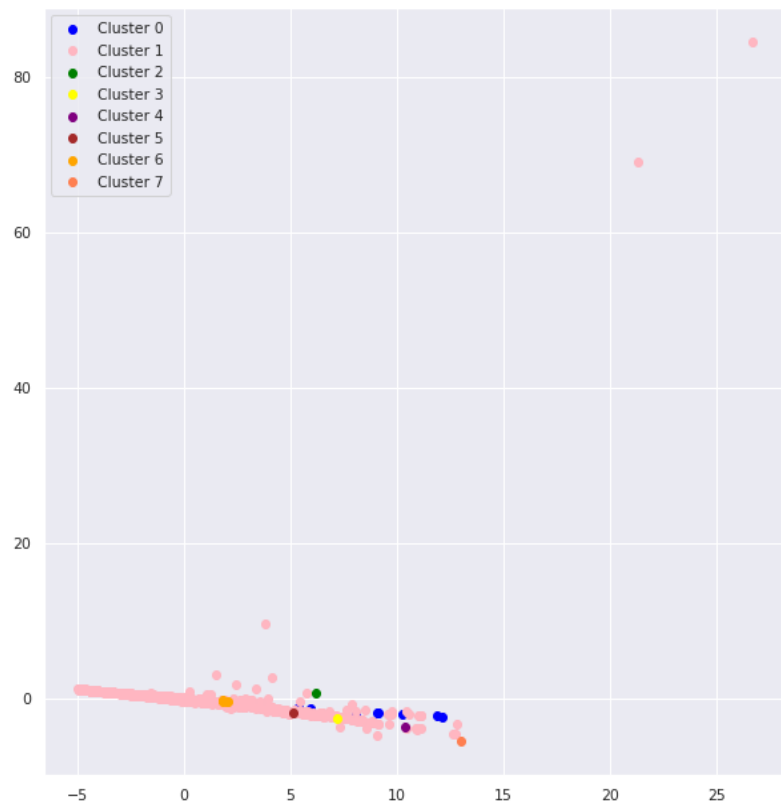


USE + Product Description



Clustering Analysis for Best Embedding Model

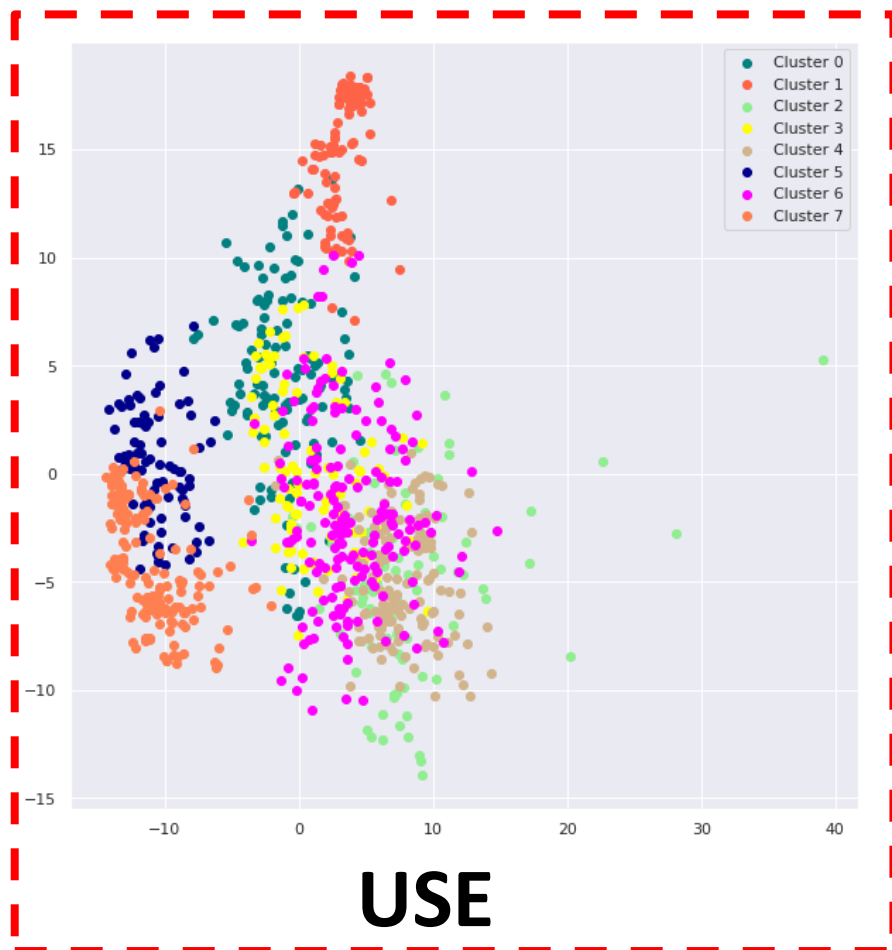
Gaussian Mixture Modelling (GMM)



tf-idf

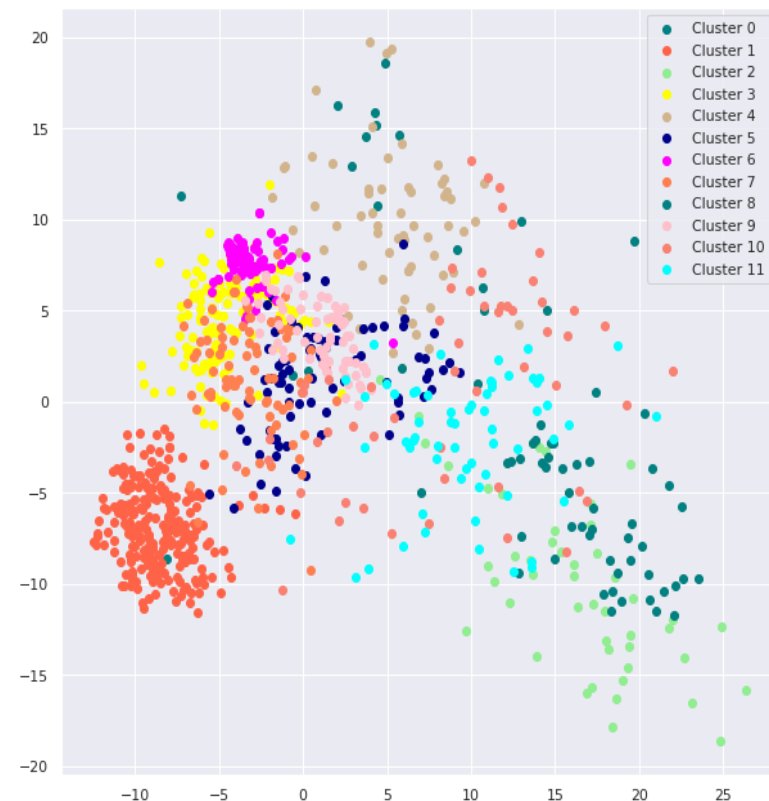
0.000039

Silhouette Score



USE

0.2086

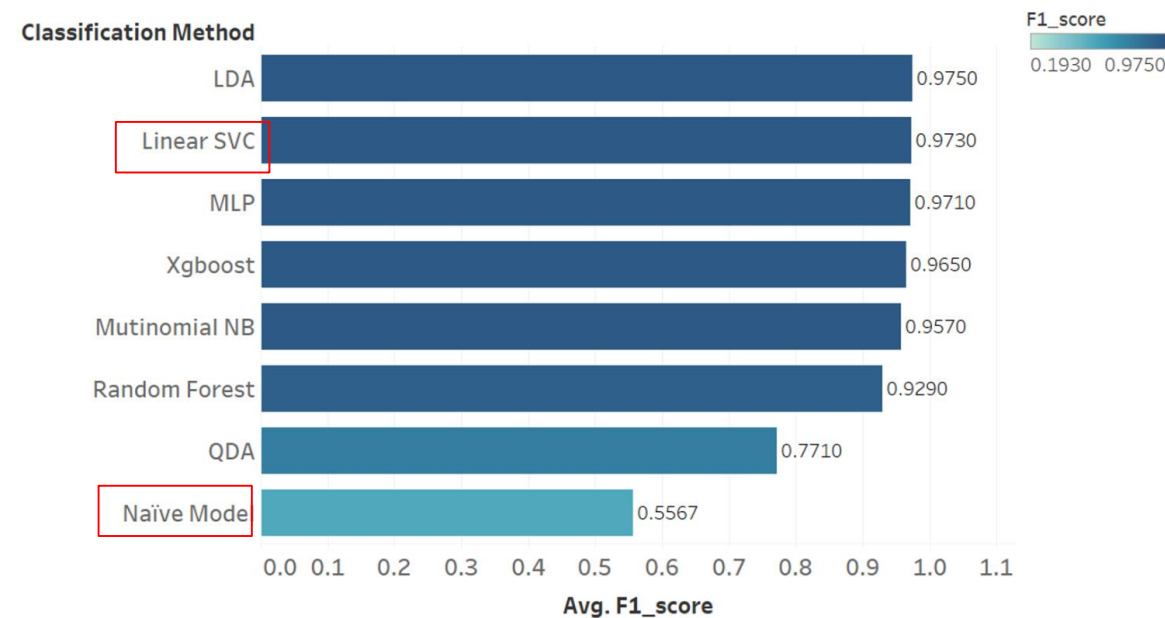


FastText

0.251260

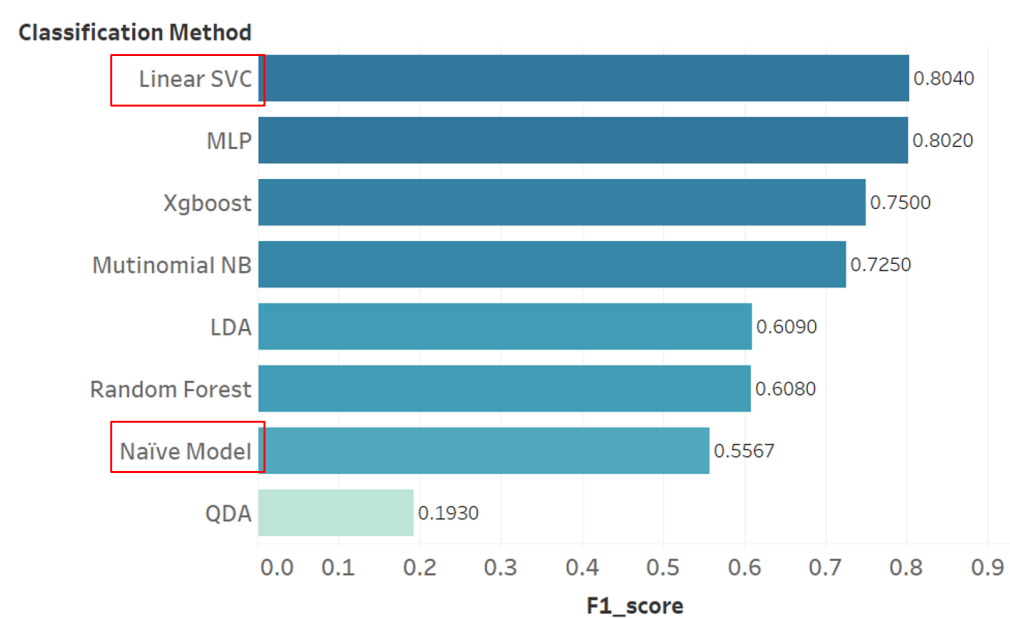
Predictive Modeling: Model Selection -- Classification

Main Category Weighted Average



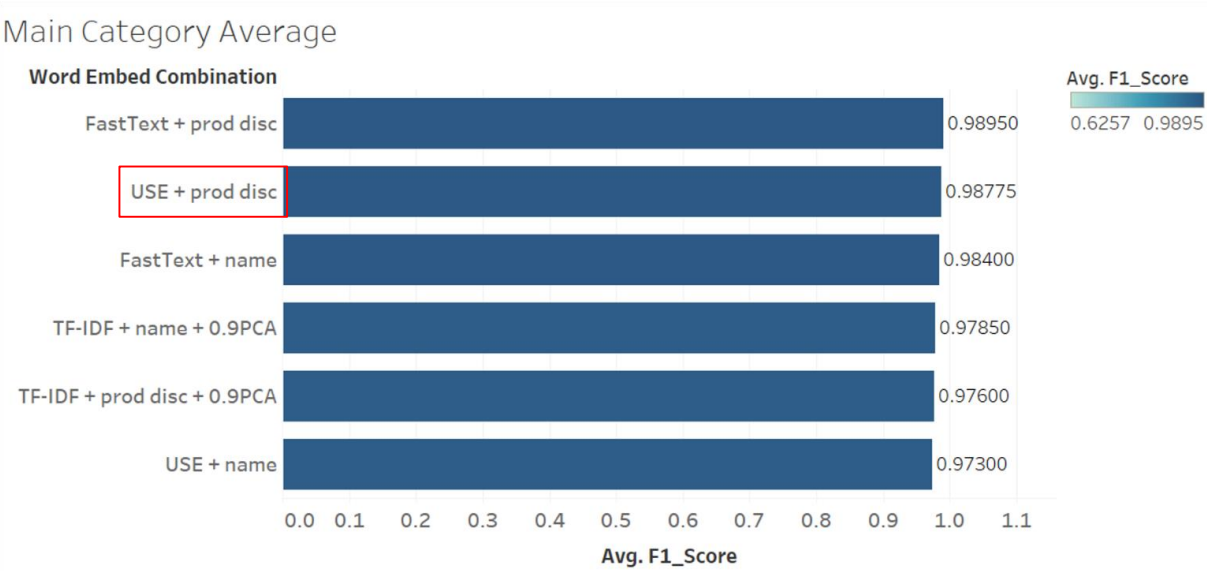
Average of F1_score for each Classification Method. Color shows sum of F1_score. The marks are labeled by sum of F1_score. The data is filtered on Main Sub and Category Name. The Main Sub filter keeps main category. The Category Name filter keeps weighted average.

Subcategory

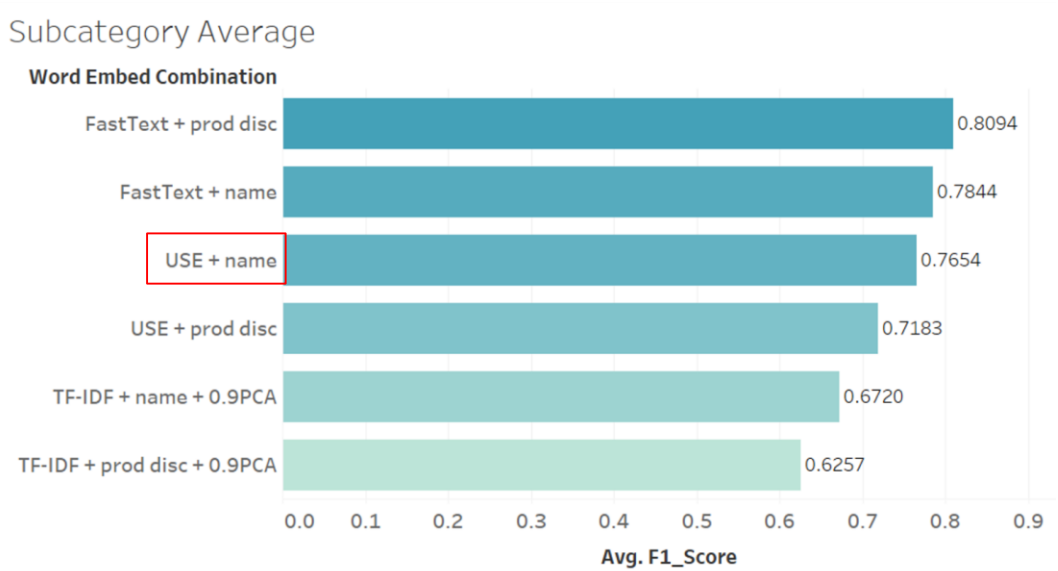


Sum of F1_score for each Classification Method. Color shows sum of F1_score. The marks are labeled by sum of F1_score. The data is filtered on Main Sub and Category Name. The Main Sub filter keeps subcategory. The Category Name filter keeps weighted average.

Predictive Modeling: Model Selection -- Word Embed

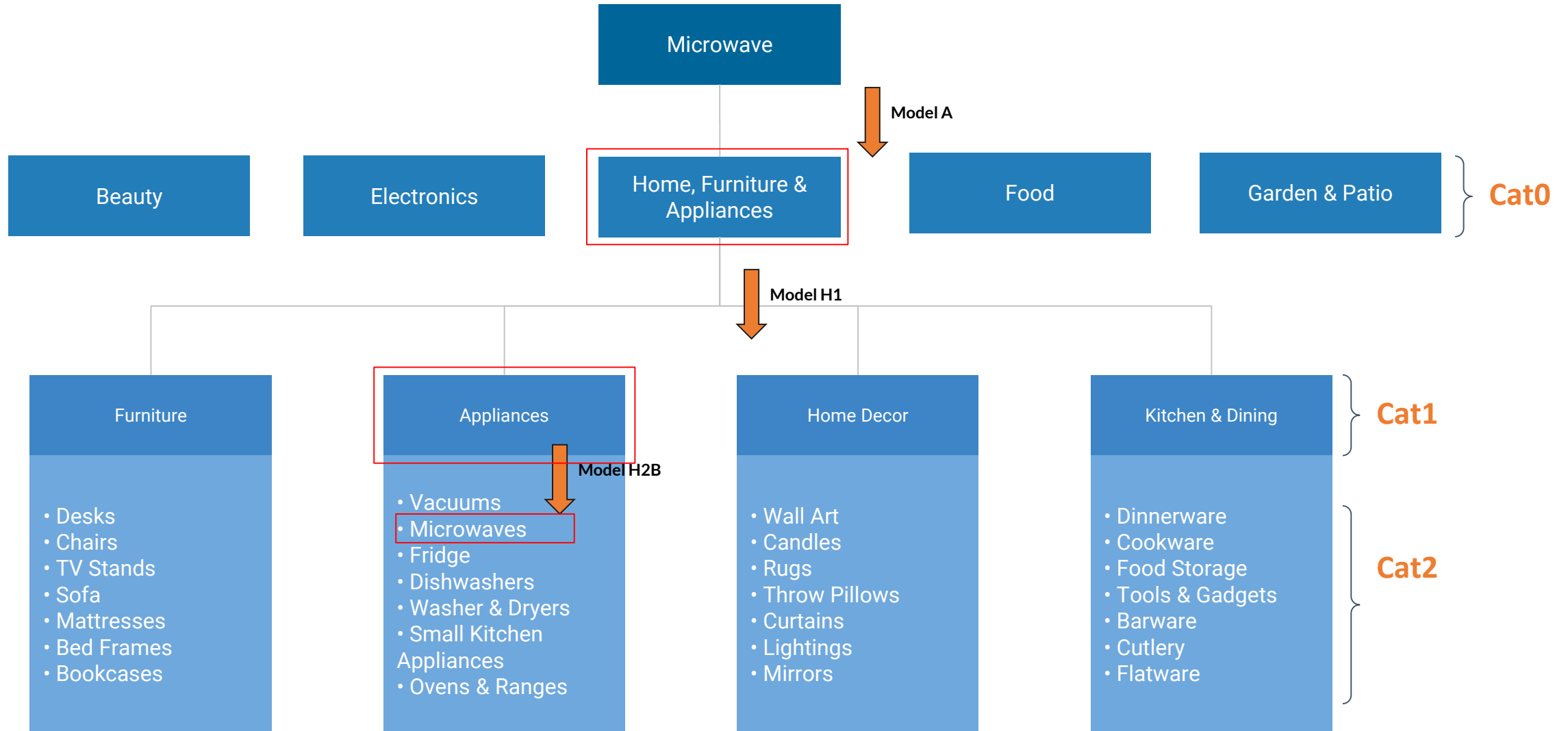


Average of F1_Score for each Word Embed Combination. Color shows average of F1_Score. The marks are labeled by average of F1_Score. The data is filtered on Main Sub and Category Name. The Main Sub filter keeps main. The Category Name filter keeps grills & outdoor cooking, outdoor heating, outdoor power equipment and patio furniture.

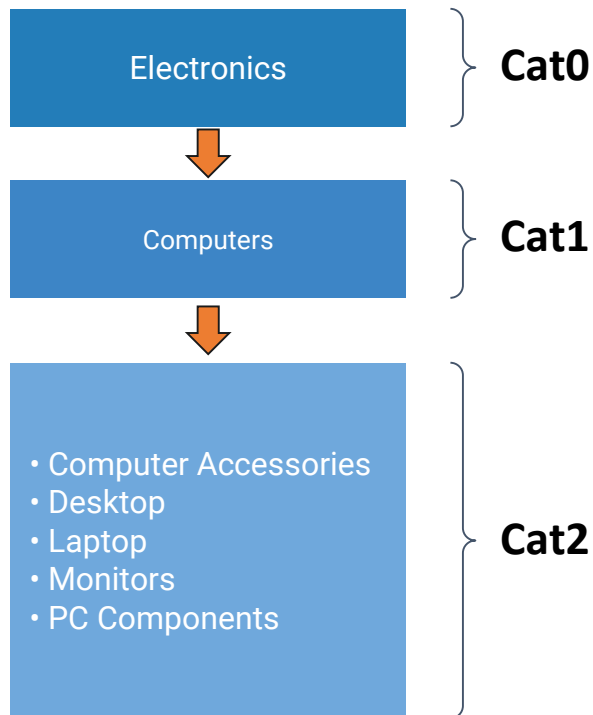


Average of F1_Score for each Word Embed Combination. Color shows average of F1_Score. The marks are labeled by average of F1_Score. The data is filtered on Category Name and Main Sub. The Category Name filter keeps 11 of 11 members. The Main Sub filter keeps subcategory.

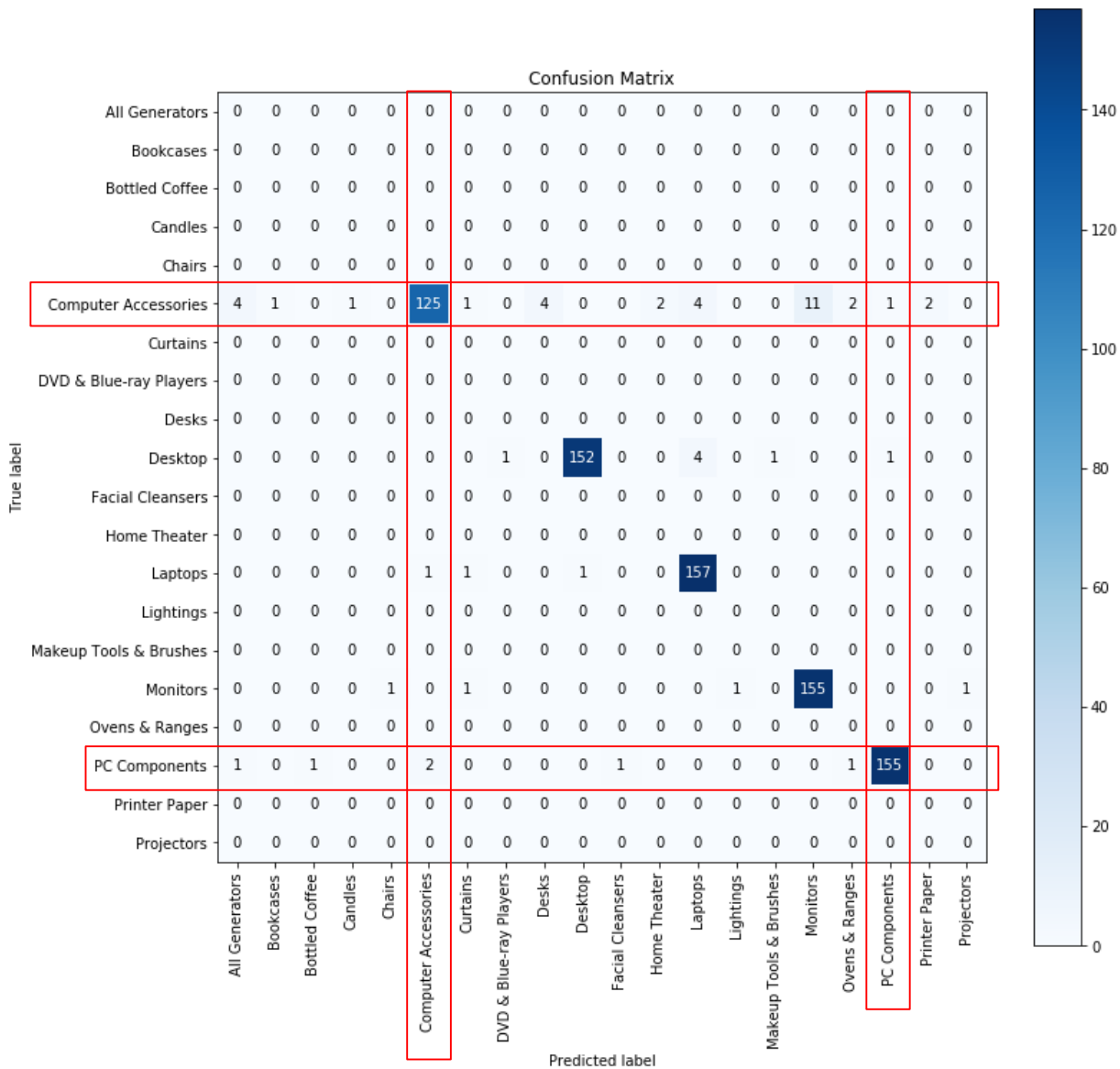
Predictive Modeling: Hierarchical Structure



Predictive Modeling: Results - Electronics(Category 2)



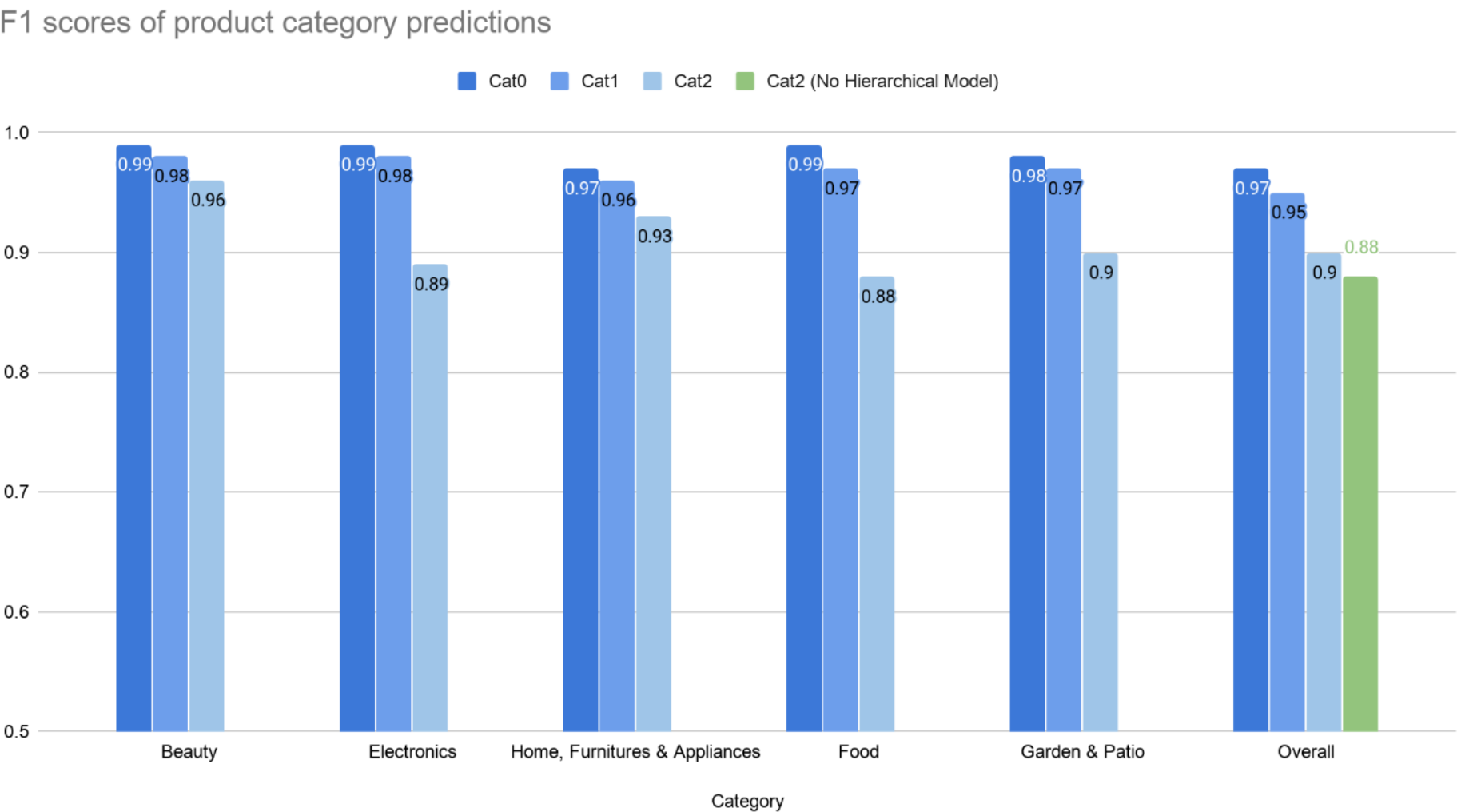
- All products predicted under the “**Computer**” category
- Accuracy: 0.91
- F1 Score: 0.95



Predicted label

Predictive Modeling: Results

- Accuracy: 0.898
- F1 Score: 0.897



Agenda



Business intent



Exploratory Data Analysis



Approach & Methodology



Operational Execution



Conclusion



Next Steps

Product Path Generator



FIXBODY Nail Clipper Set – Black Stainless Steel Fingernails & Toenails Clippers & Nail File Sharp Nail Cutter with Leather Case, Set of 3 (Straight & Curved)



Doritos Nacho Cheese Flavored Tortilla Chips, 1 oz (Pack of 40)



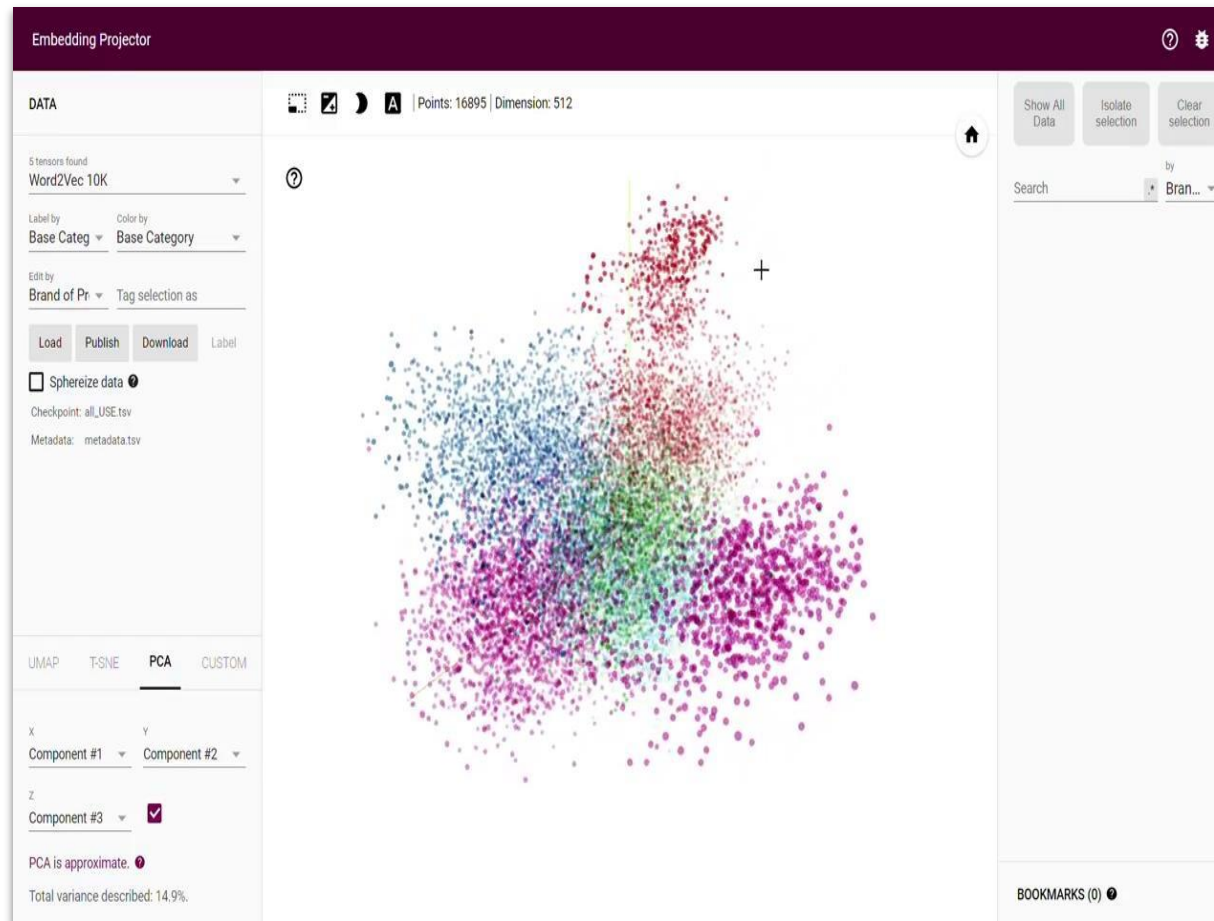
Dell XPS 13 9380 13.3" Notebook - 1920 X 1080 - Core I5-8265U - 8GB RAM - 256GB SSD - Platinum Silver, Carbon Black

```
df_test = encoder(['FIXBODY Nail Clipper Set - Black Stainless Steel Fingernails & Toenails Clippers & Nail File Sharp Nail Cutter with Leather Case, Set of 3 (Straight & Curved)',  
                  'Doritos Nacho Cheese Flavored Tortilla Chips, 1 oz (Pack of 40)',  
                  'Dell XPS 13 9380 13.3" Notebook - 1920 X 1080 - Core I5-8265U - 8GB RAM - 256GB SSD - Platinum Silver, Carbon Black'])
```

```
0      Beauty/Nail Care/Manicure Tools  
1      Food/Snacks, Cookies & Chips/Chips  
2      Electronics/Computers/Laptops  
dtype: object
```

Similar(Semantically) Product Finder

TensorFlow Embedding Projector
(Word Embedding Method: Universal Sentence Encoder , Dimension Reduction: PCA)



Agenda



Business intent



Exploratory Data Analysis



Approach & Methodology



Operational Execution & Change Management

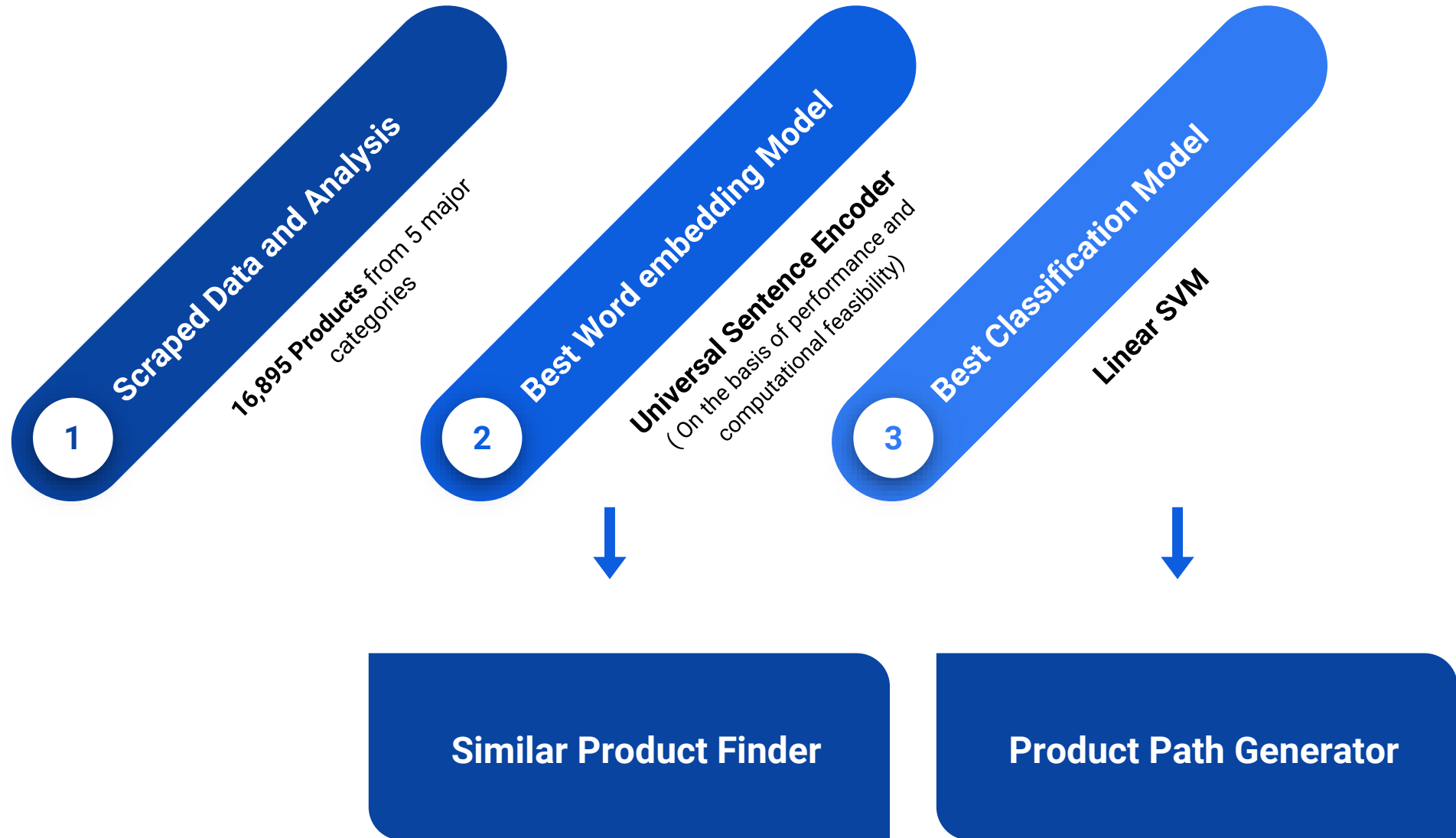


Conclusion



Next Steps

Let's recap



Agenda



Business intent



Exploratory Data Analysis



Approach & Methodology



Operational Execution



Conclusion



Next Steps

Recommendations to Walmart

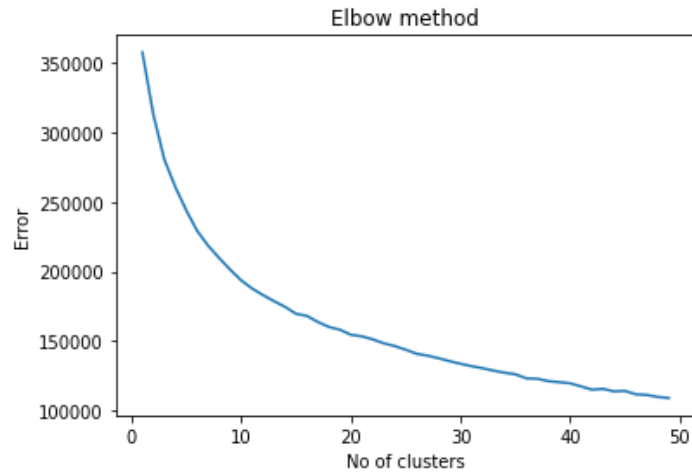
- Cross-Site availability of data (MooseJaw, Bonobos)
- Standardizing categories will improve Similarity Product Finder's accuracy
- Cat3 level of classification in our data

Clustering ----> Topic Modeling ----> Classification (Topic Memory Networks)

Cat3 level of Categorization

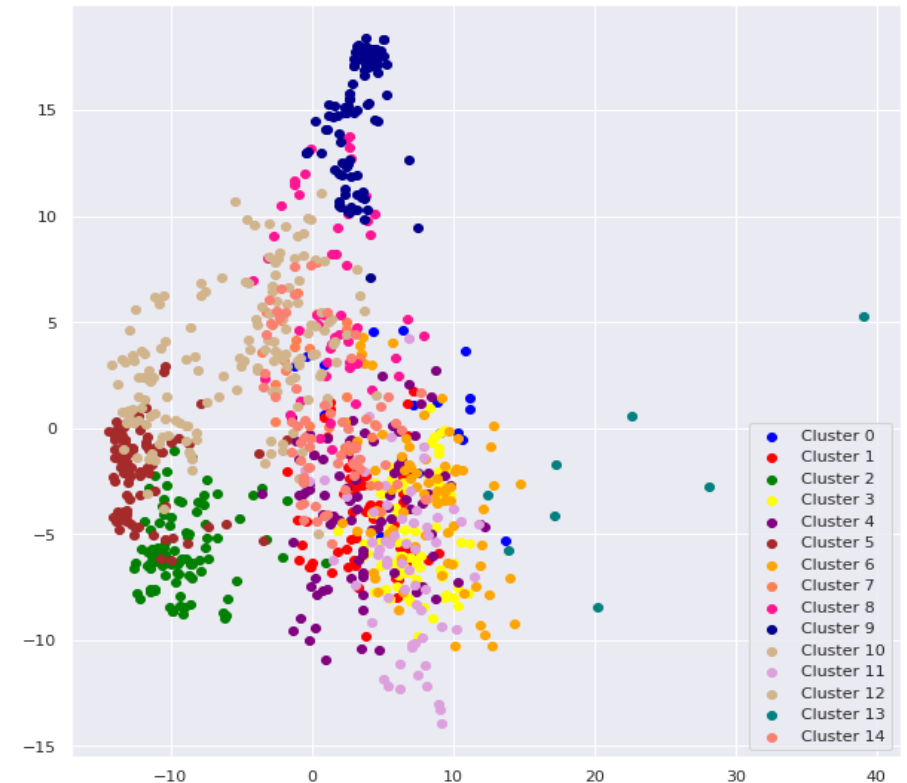
Separability in data after the deepest level of categorization validates possibility of further classification

k-means Clustering



k = 15

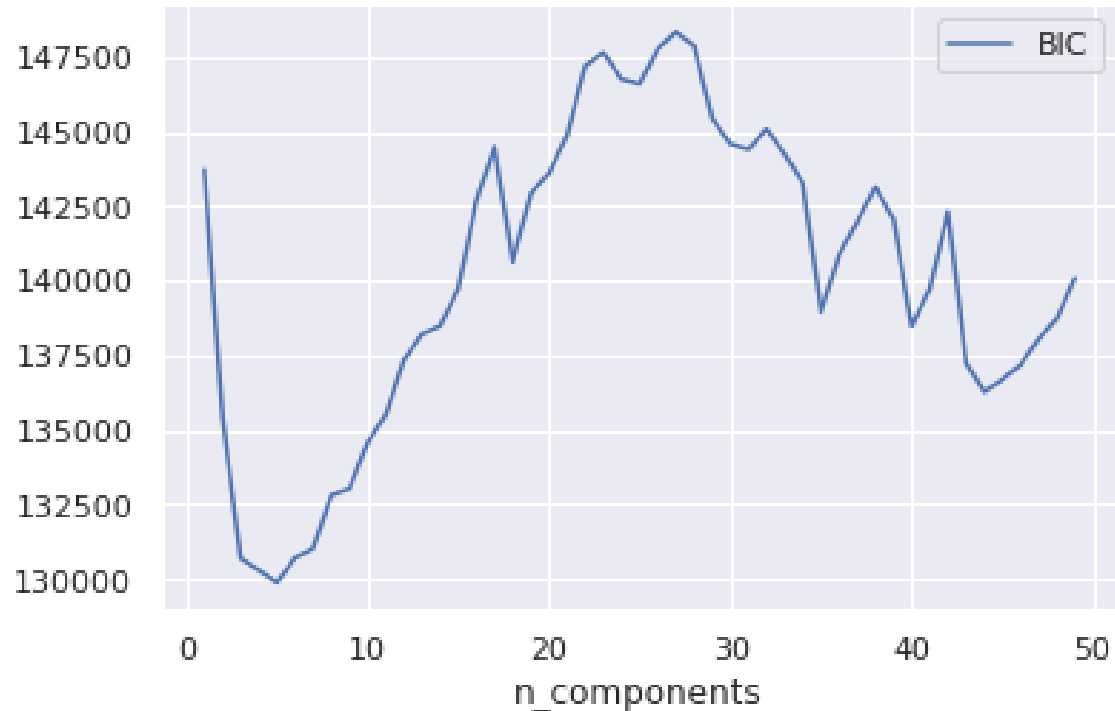
Silhouette Score = 0.1934



Topic Modeling Results are not well defined due to multiple overlaps in features

Gaussian Mixture Modelling (GMM)

Hyper Parameter Tuning for ideal number of clusters



Clusters	Silhouette Score
8	0.208617
41	0.206045
32	0.201018
28	0.199282

Silhouette Score ->

k = 8

BIC value -> k = 6- 8

Therefore k = 8

Topic Modeling on the Clusters using LDA (Latent Dirichlet allocation)



Cluster 0	LCD Screen , Tower Desktops
Cluster 1	Dual Laptops, VGA Specs
Cluster 2	Apple , iMac , Silver, Dents, Old
Cluster 3	Intel , Professional, LCD , Pavillion
Cluster 4	Gaming , RTX , Black, Omen
Cluster 5	Monitor , Refurbished, Core
Cluster 6	Pavilion , Dual, Refurbished
Cluster 7	Refurbished , Windows Pro



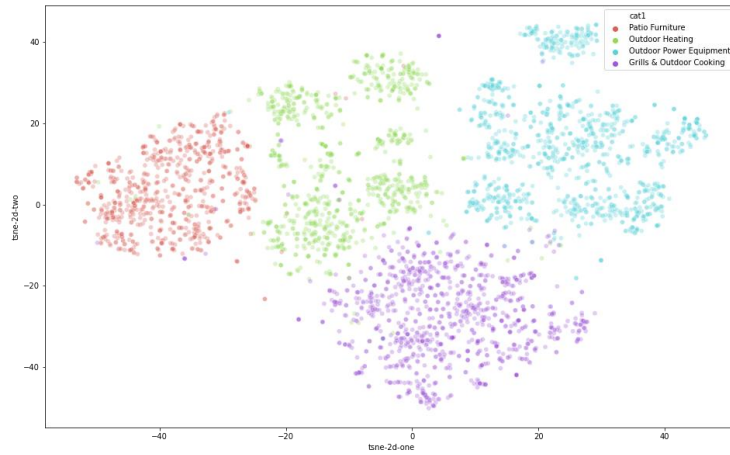
Thank you



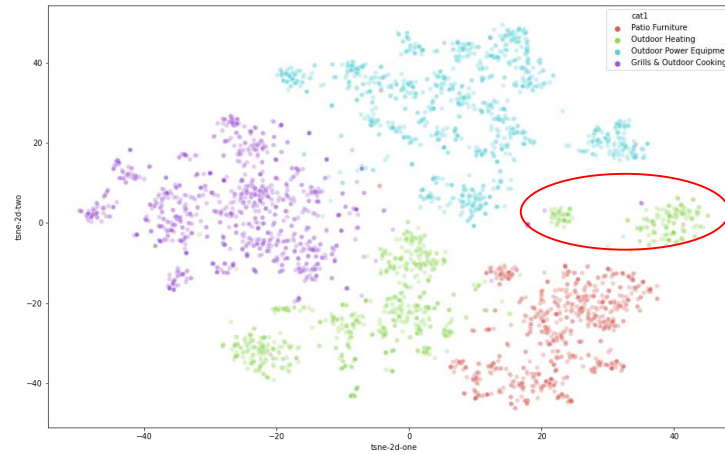
The University of Texas at Austin
Texas McCombs
MS Business Analytics
McCombs School of Business

Appendix

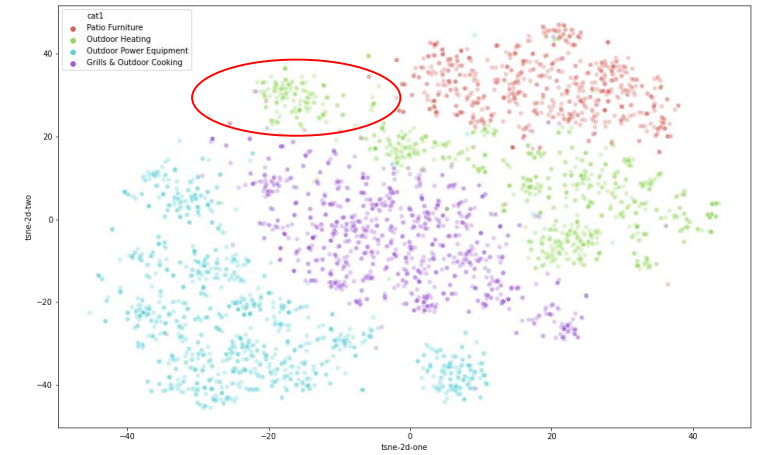
Feature Engineering



USE + Product Name



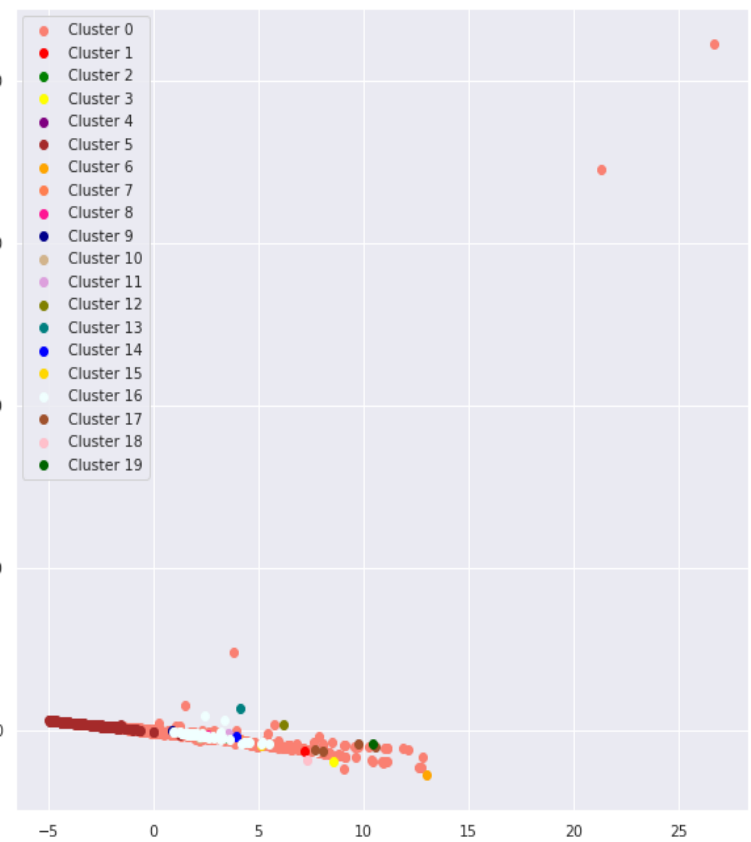
FastText + Product Name



Tf_idf + Product Name

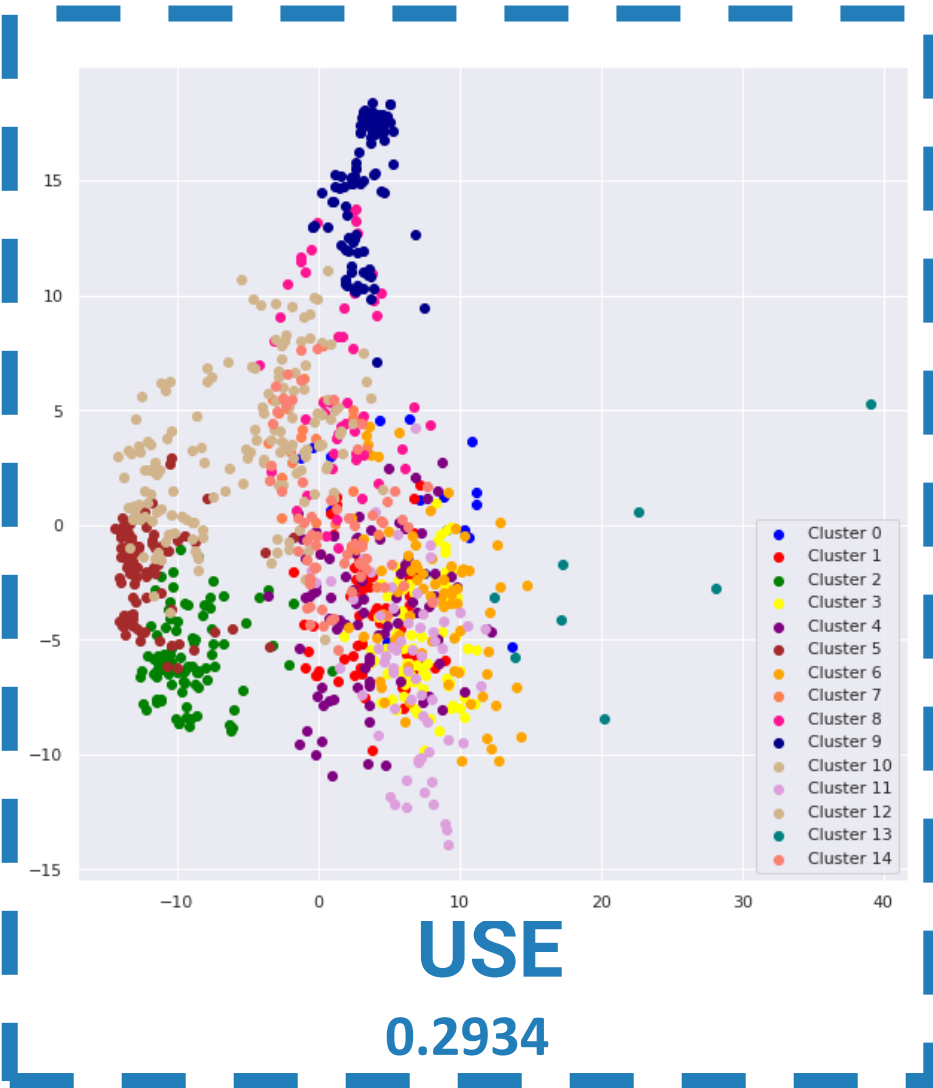
Clustering Analysis for Best Embedding Model

k-
means



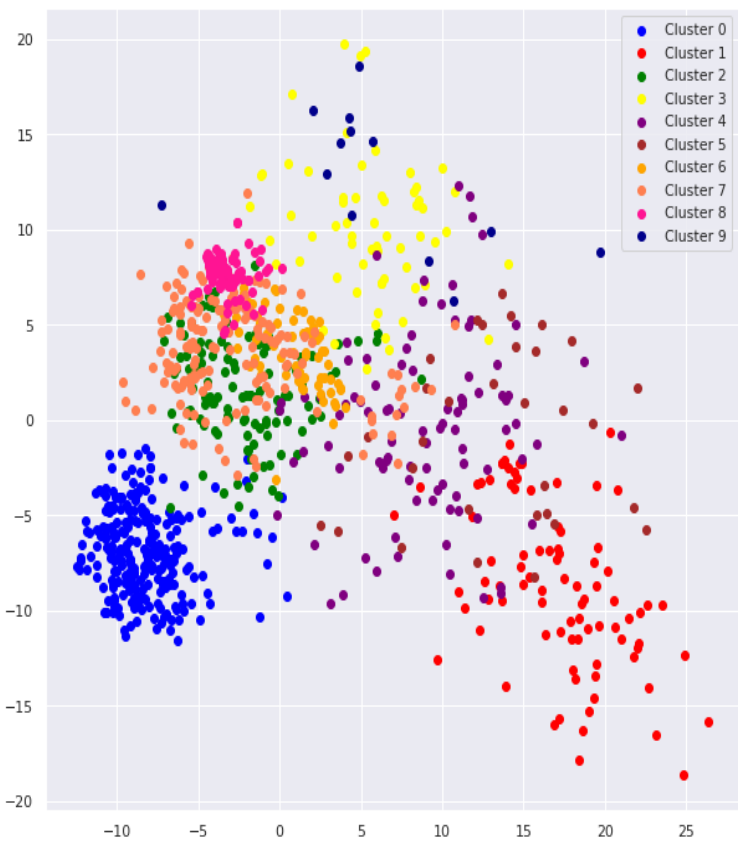
tf-idf

Silhouette Score 0.00751



USE

0.2934

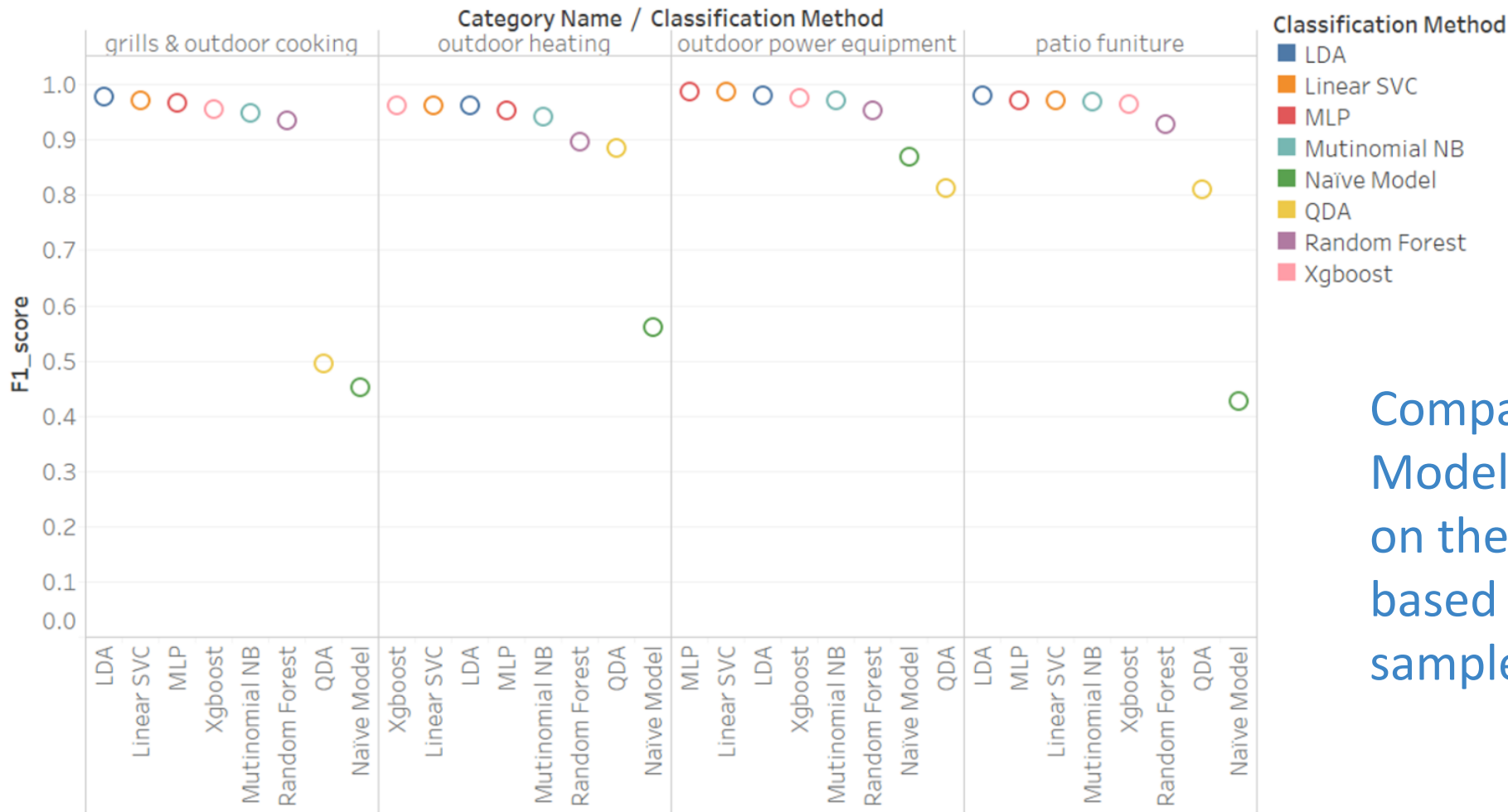


FastText

0.2644

Predictive Modeling: Model Selection - Classification

Main Category Seperate

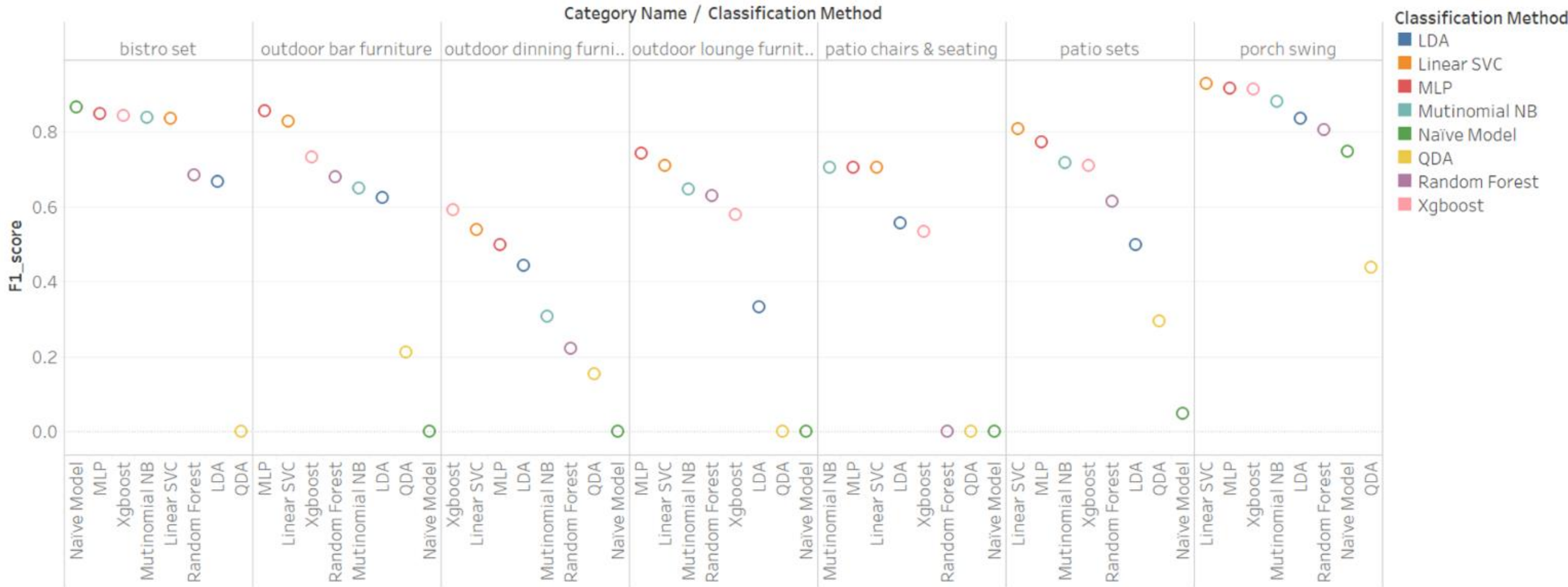


Compare Classification Models by running them on the same USE result based on the same garden sample data

Sum of F1_score for each Classification Method broken down by Category Name. Color shows details about Classification Method. The data is filtered on Main Sub, which keeps main category. The view is filtered on Category Name, which keeps grills & outdoor cooking, outdoor heating, outdoor power equipment and patio funiture.

Predictive Modeling: Model Selection - Classification

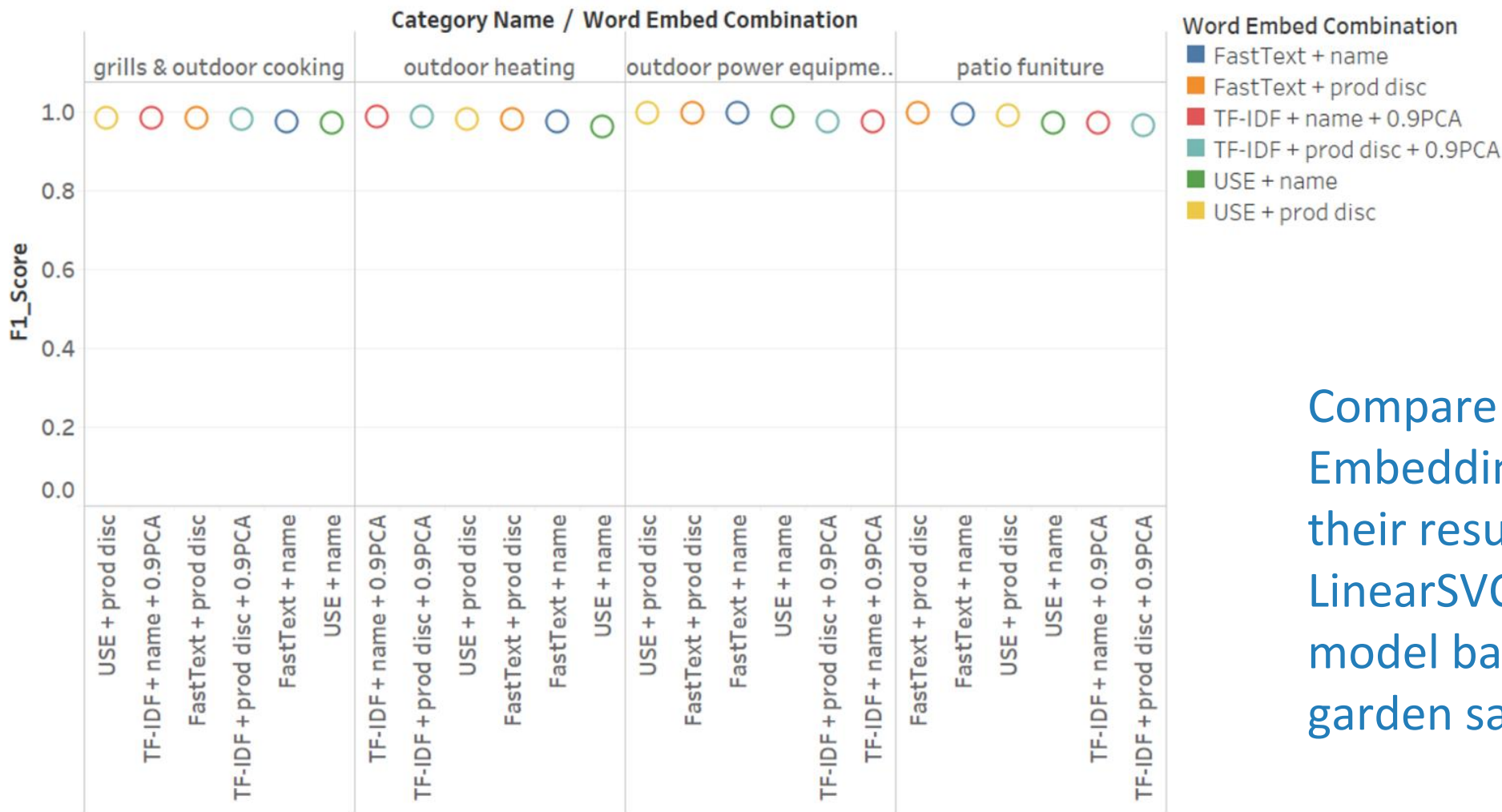
Subcategory Seperate



Sum of F1_score for each Classification Method broken down by Category Name. Color shows details about Classification Method. The data is filtered on Main Sub, which keeps subcategory. The view is filtered on Category Name, which has multiple members selected.

Predictive Modeling: Model Selection -- Word Embed

Main Category Seperate

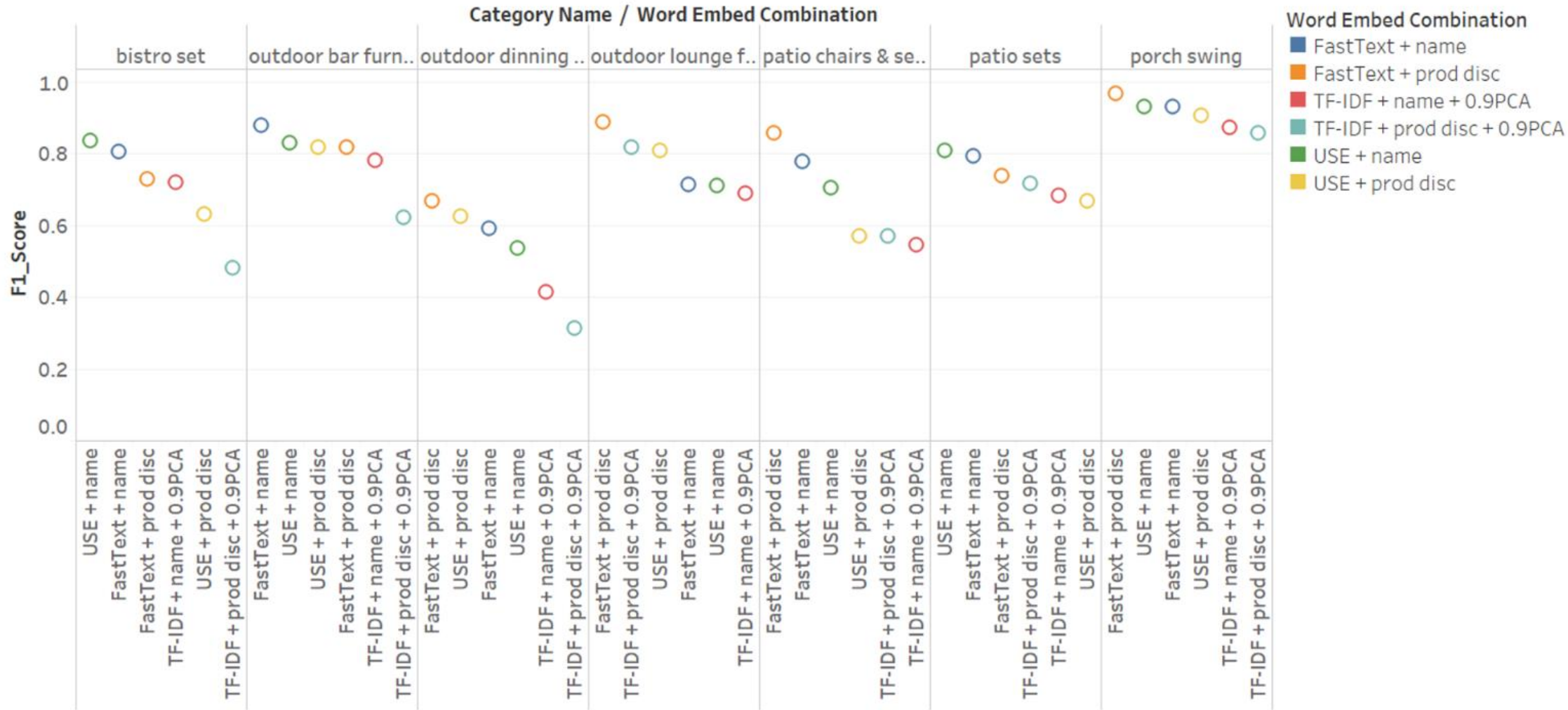


Compare different Word Embedding tech by using their results on the same LinearSVC classification model based on the same garden sample data

Sum of F1_Score for each Word Embed Combination broken down by Category Name. Color shows details about Word Embed Combination. The data is filtered on Main Sub, which keeps main.

Predictive Modeling: Model Selection -- Word Embed

Subcategory Seperate



Sum of F1_Score for each Word Embed Combination broken down by Category Name. Color shows details about Word Embed Combination. The data is filtered on Main Sub, which keeps subcategory.