E-COMMERCE PRODUCT CATEGORIZATION

*Welcome to the "**Ecommerce Product Categorization**" hackathon, hosted by KnowledgeHut.This exciting competition challenges you to apply your data science, machine learning & NLP skills to a real-world problem in the retail sector.*

LOAD DATASET

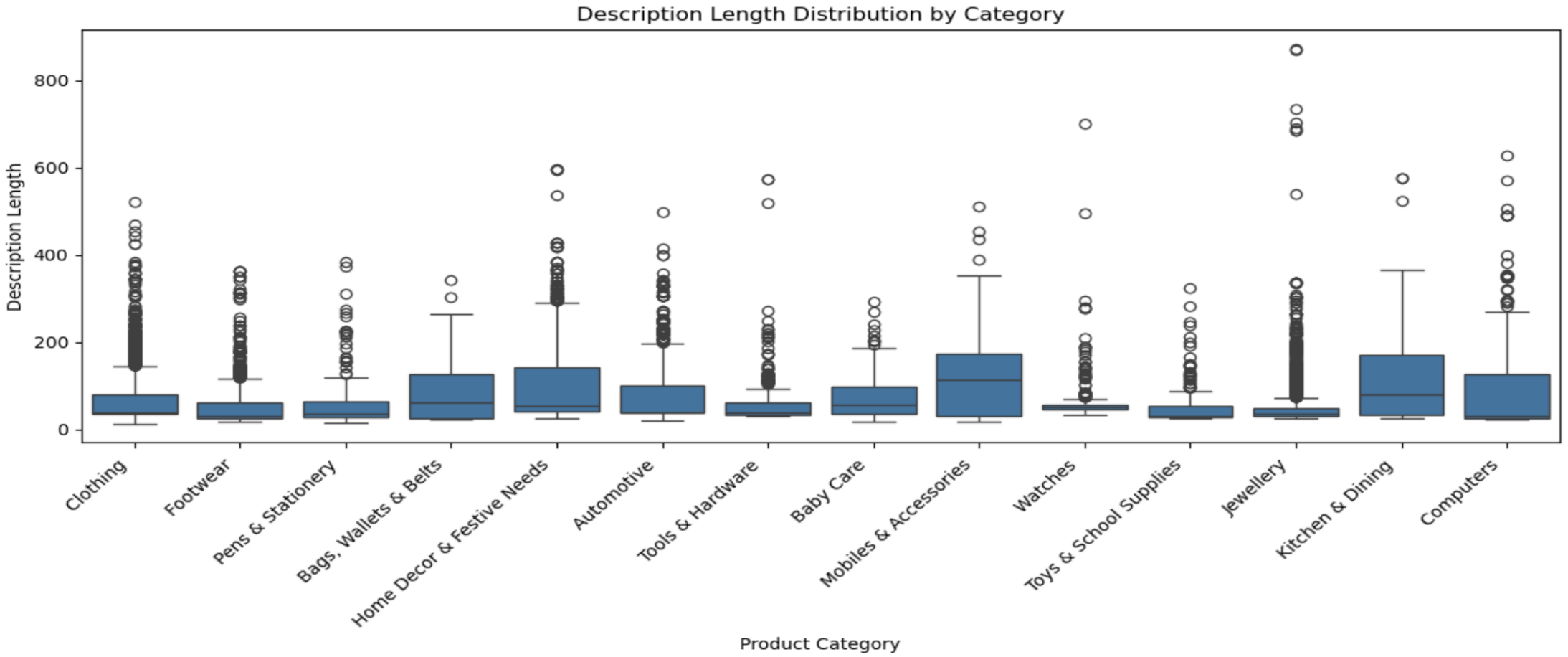
[66]:
```
product_data = pd.read_csv("train_product_data.csv")
```

[67]:
```
product_data.head()
```

[67]:

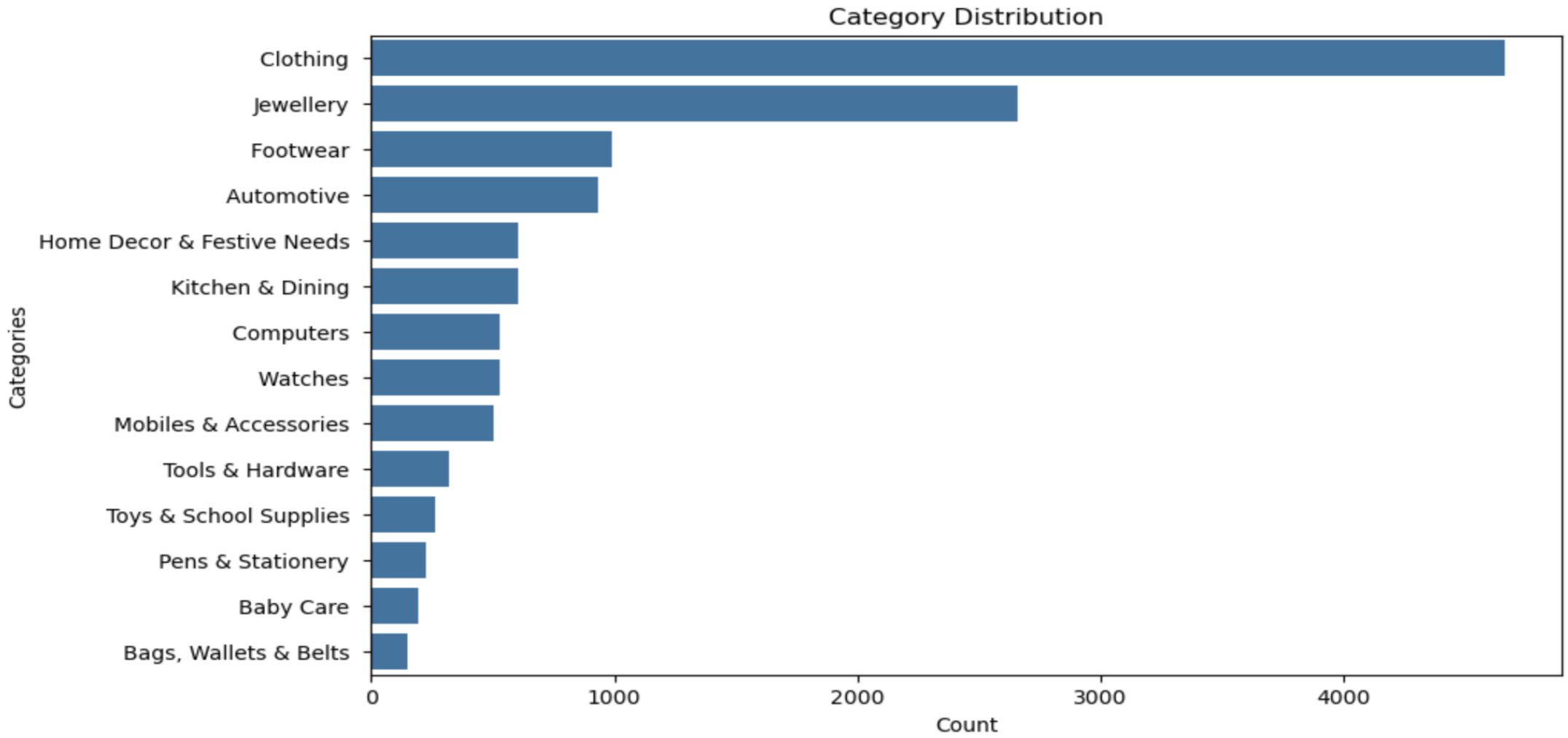
	uniq_id	crawl_timestamp	product_url	product_name	product_category_tree	pid	retail_price
0	c2d766ca982eca8304150849735ffef9	2016-03-25 22:59:23 +0000	http://www.flipkart.com/alisha-solid-women-s-c...	Alisha Solid Women's Cycling Shorts	Clothing	SRTEH2FF9KEDEFGF	999.0
1	f449ec65dcbc041b6ae5e6a32717d01b	2016-03-25 22:59:23 +0000	http://www.flipkart.com/aw-bellies/p/itmeh4grg...	AW Bellies	Footwear	SHOEH4GRSUBJGZXE	999.0
2	0973b37acd0c664e3de26e97e5571454	2016-03-25 22:59:23 +0000	http://www.flipkart.com/alisha-solid-women-s-c...	Alisha Solid Women's Cycling Shorts	Clothing	SRTEH2F6HUZMQ6SJ	699.0
3	ce5a6818f7707e2cb61fdcdbba61f5ad	2016-03-25 22:59:23 +0000	http://www.flipkart.com/alisha-solid-women-s-c...	Alisha Solid Women's Cycling Shorts	Clothing	SRTEH2FVVKRBAXHB	1199.0
4	29c8d290caa451f97b1c32df64477a2c	2016-03-25 22:59:23 +0000	http://www.flipkart.com/dilli-bazaaar-bellies-...	dilli bazaaar Bellies, Corporate Casuals, Casuals	Footwear	SHOEH3DZBFR88SCK	699.0

DESCRIPTION LENGTH DISTRIBUTION BY CATEGORY



Descriptions for products in the "Mobiles & Accessories" category tend to be the longest, followed by "Kitchens & Dining" and "Computers." Descriptions for products in the "Clothing" ,"watches" and "Footwear" categories tend to be on the shorter side.

CATEGORY DISTRIBUTION



Dataset contains more categories of “Clothing” followed by "Jewellery" and "Footwear", this implies that customers are more interested in buying this products probably female customers

WORD DISTRIBUTION AFTER NORMALISATION

Wordcloud with ECommerce stopwords

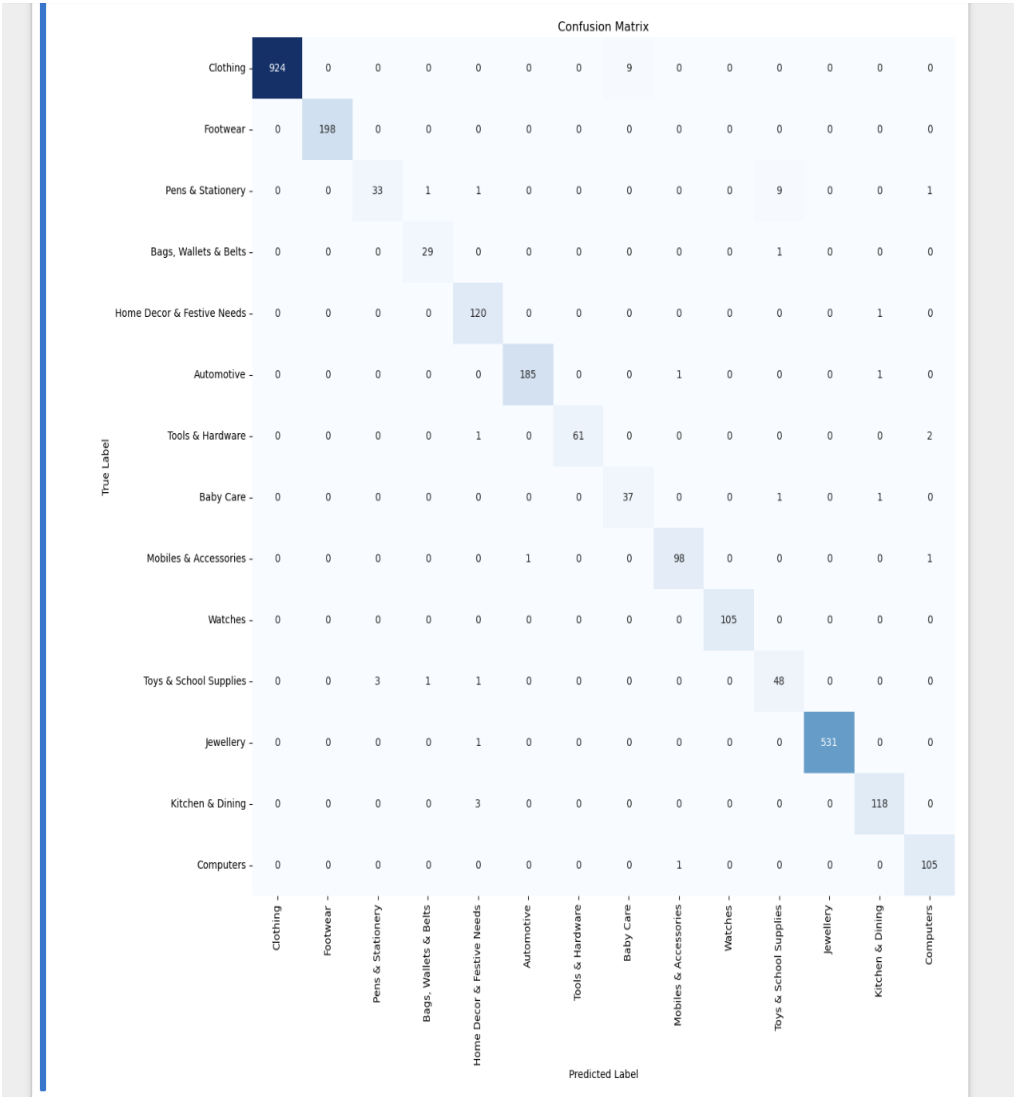
Wordcloud with General stopwords

APPLY MODELS ON TF-IDF VECTORIZATION

	Classifier	Training accuracy	Validation accuracy
5	SGD Classifier	0.992404	0.983675
6	Ridge Classifier	0.994873	0.983675
3	Linear SVM	0.992974	0.982916
7	XGBoost	0.999335	0.972665
4	Random Forest	0.999335	0.959757
2	Decision Tree	0.999335	0.952164
0	MultinomialNB	0.915875	0.896735
1	KNN Classifier	0.896791	0.892938
8	AdaBoost	0.584125	0.583523

Ridge Classifiers performed best of all

HYPERPARAMETER TUNING ON BEST PERFOMED



Classification report for training set

	precision	recall	f1-score	support
Automotive	1.00	1.00	1.00	748
Baby Care	0.87	1.00	0.93	156
Bags, Wallets & Belts	1.00	1.00	1.00	122
Clothing	1.00	0.99	1.00	3730
Computers	1.00	1.00	1.00	423
Footwear	1.00	1.00	1.00	790
Home Decor & Festive Needs	1.00	1.00	1.00	485
Jewellery	1.00	1.00	1.00	2126
Kitchen & Dining	1.00	1.00	1.00	485
Mobiles & Accessories	1.00	1.00	1.00	401
Pens & Stationery	0.99	0.94	0.97	179
Tools & Hardware	1.00	1.00	1.00	257
Toys & School Supplies	0.95	1.00	0.97	209
Watches	1.00	1.00	1.00	421
accuracy			1.00	10532
macro avg	0.99	0.99	0.99	10532
weighted avg	1.00	1.00	1.00	10532

Classification report for test set

	precision	recall	f1-score	support
Automotive	0.99	0.99	0.99	187
Baby Care	0.80	0.95	0.87	39
Bags, Wallets & Belts	0.94	0.97	0.95	30
Clothing	1.00	0.99	1.00	933
Computers	0.96	0.99	0.98	106
Footwear	1.00	1.00	1.00	198
Home Decor & Festive Needs	0.94	0.99	0.97	121
Jewellery	1.00	1.00	1.00	532
Kitchen & Dining	0.98	0.98	0.98	121
Mobiles & Accessories	0.98	0.98	0.98	100
Pens & Stationery	0.92	0.73	0.81	45
Tools & Hardware	1.00	0.95	0.98	64
Toys & School Supplies	0.81	0.91	0.86	53
Watches	1.00	1.00	1.00	105
accuracy			0.98	2634
macro avg	0.95	0.96	0.95	2634
weighted avg	0.98	0.98	0.98	2634

E-COMMERCE PRODUCT CATEGORIZATION PERFORMANCE

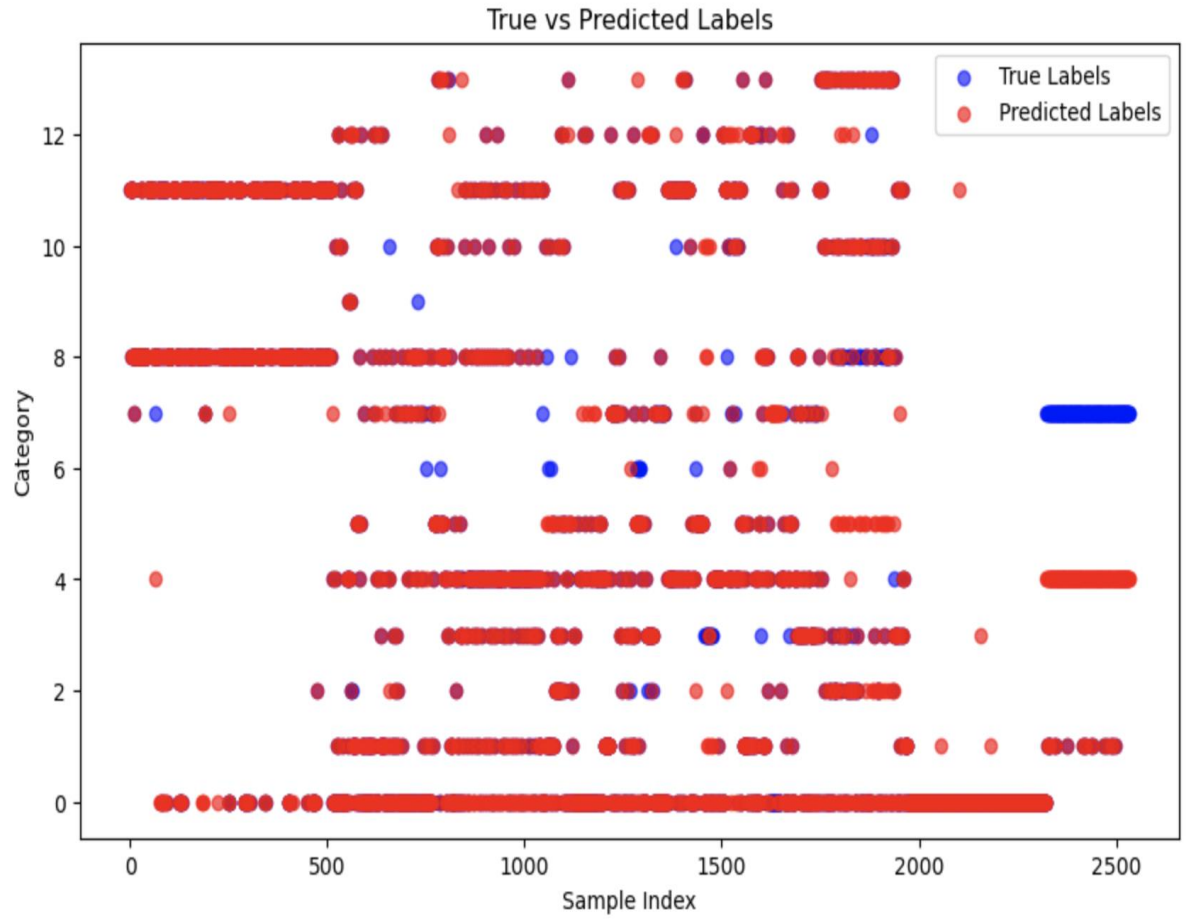
Training Set:

- **Overall Performance:** The model achieved an accuracy of 99% on the training set.
- **Precision and Recall:** Most categories exhibit high precision and recall values, indicating effective categorization.
- **Performance Balance:** Both macro-average and weighted-average F1-scores are high, suggesting balanced performance across categories.

Test set:

- **Overall Performance:** The model achieved an accuracy of 98% on the test set.
- **Precision and Recall:** While precision and recall remain generally high, some categories display slight variations.
- **Imbalance:** Certain categories, like "Pens & Stationery" and "Toys & School Supplies," exhibit lower precision and recall, possibly due to inherent complexities or class imbalance.

PREDICTION ON TEST DATA



Conclusion:

The model achieved a high accuracy of **85.83%**, indicating strong performance in predicting the categories of the given samples. The scatter plot analysis further confirms this observation, revealing a substantial overlap between the **true** and **predicted labels** for most indices, signifying consistent and accurate predictions.

THANK YOU