Project Report
Product Title Classification

Course: Information Retrieval (IR)

Course Teacher: Dr. Muhammad Rafi

Group:

- Muhammad Qasim Alias Haseeb (21K-4889)
- Sunny (21K-4562)
- Bhavish Kumar (21K-3450)

Introduction

The research paper titled "Product Title vs. Text Classification" investigates various methods for classifying product titles into predefined categories using SVM (Support Vector Machine) models. The study focuses on enhancing the accuracy of classification through different feature extraction techniques and SVM configurations.

Objective

The primary objective of this research paper is to develop a robust machine learning model that can accurately classify product titles into specific categories. This involves experimenting with various feature extraction methods and SVM models to determine the optimal approach for this task.

Data Preprocessing

- **Text Normalization**: The product titles are converted to lowercase to ensure uniformity.
- **Tokenization**: The titles are tokenized into individual words or phrases.
- **Label Encoding**: Categories are encoded as integers using label encoding to convert categorical labels into a numerical format suitable for machine learning algorithms.
- **TF-IDF Vectorization**: The text data is transformed into numerical vectors using TF-IDF, which assigns weights to words based on their frequency and importance within the dataset. The paper specifically compares unigrams (single words) and bigrams (pairs of consecutive words) for this purpose.

Feature Extraction Techniques

- **Unigrams:** Each word in the product title is considered a separate feature. This is the simplest form of text representation.
- **Bigrams:** Pairs of consecutive words are treated as features. This method captures some context that unigrams might miss.
- **Degree-2 Polynomial Mapping (poly2):** This method expands the feature space by including squared terms and interaction terms of the original features. The resulting feature vector for a title includes the original unigrams, the squares of these unigrams, and the products of each pair of unigrams.

SVM Models

- **One-vs-One (OvO):** Constructs a binary classifier for every pair of classes. For n, n(n−1)/2 classifiers are trained.
- **One-vs-Rest (OvR):** Constructs a binary classifier for each class against all other classes. For n, n classifiers are trained.
- **Crammer & Singer:** A multiclass SVM method that solves a single optimization problem for all classes simultaneously, unlike OvO and OvR which decompose the problem into multiple binary classifications.

Comparison of Methods

The paper evaluates the performance of different SVM models and feature extraction techniques using the Relative Error (REbaseline) metric. The key findings include:

- **Bigram Features:** Incorporating bigrams significantly improves classification performance compared to using unigrams alone.
- **Polynomial Features (poly2):** This method shows the best results by reducing the number of classification errors to around 70% of the baseline error.

SVM Models:

- **One-vs-One (OvO) and One-vs-Rest (OvR):** Both models show competitive performance, but OvO tends to be more computationally intensive due to the larger number of classifiers.
- **Crammer & Singer:** This method generally outperforms the others in terms of both accuracy and computational efficiency. It requires solving a single optimization problem, making it more efficient for large datasets.

Results

The research shows that advanced feature extraction methods like bigrams and polynomial mappings can greatly enhance the accuracy of product title classification. Specific findings include:

- **Bigram + Unigram:** Using a combination of bigrams and unigrams results in 11,799,345 features.
- **Polynomial Mapping (poly2):** Expanding the feature set using polynomial mappings results in 41,689,205 features. This method provides slightly better results across all three SVM strategies compared to bigram + unigram.

Performance Metrics:

- **REbaseline:** The best results show a reduction in the number of errors to approximately 70% of the baseline error.

Implementation Details

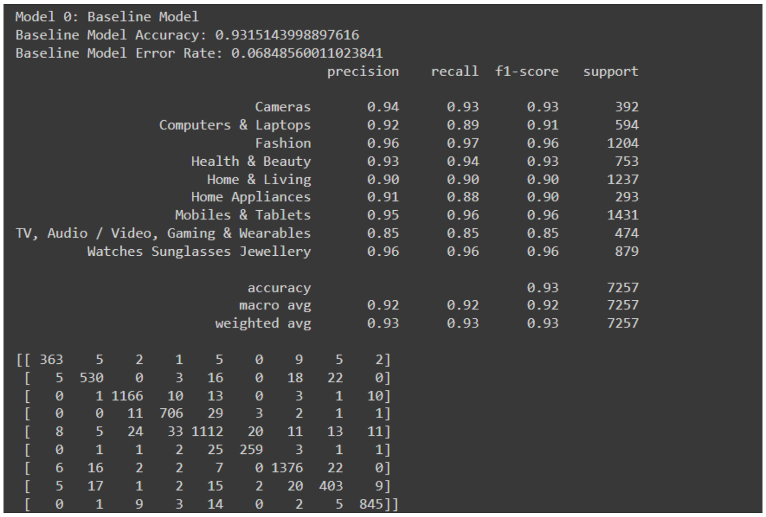
The paper emphasizes the importance of efficient implementation techniques to handle the large number of features generated by bigram and polynomial mappings. Key implementation details include:

- **Normalization:** Each instance is normalized to have unit length to ensure that no single feature dominates the others.
- **Hashing Technique:** To efficiently manage the vast feature space, a hashing technique is employed to remove unnecessary features that never have non-zero values in the training set.

Conclusion

The research paper concludes that using sophisticated feature extraction techniques such as bigrams and polynomial mappings can significantly improve the accuracy of product title classification. Among the SVM models, the Crammer & Singer method is recommended for its superior performance and computational efficiency. This study highlights the importance of feature engineering in text classification tasks and provides valuable insights into the implementation of effective SVM models for multi-class classification.

Snapshots

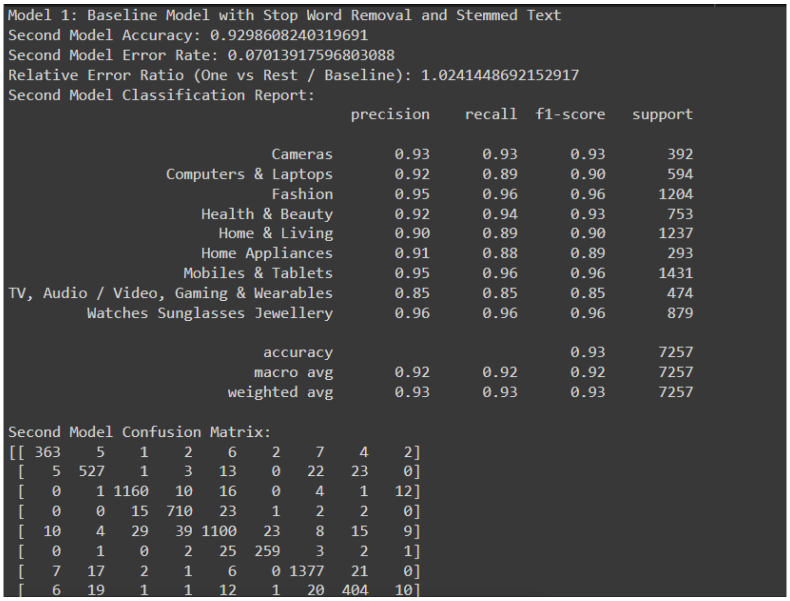


```
Model 0: Baseline Model
Baseline Model Accuracy: 0.9315143998897616
Baseline Model Error Rate: 0.06848560011023841
                                       precision    recall  f1-score   support

                              Cameras       0.94      0.93      0.93       392
                  Computers & Laptops       0.92      0.89      0.91       594
                              Fashion       0.96      0.97      0.96      1204
                      Health & Beauty       0.93      0.94      0.93       753
                        Home & Living       0.90      0.90      0.90      1237
                      Home Appliances       0.91      0.88      0.90       293
                    Mobiles & Tablets       0.95      0.96      0.96      1431
TV, Audio / Video, Gaming & Wearables       0.85      0.85      0.85       474
         Watches Sunglasses Jewellery       0.96      0.96      0.96       879

                             accuracy                           0.93      7257
                            macro avg       0.92      0.92      0.92      7257
                         weighted avg       0.93      0.93      0.93      7257

[[ 363    5    2    1    5    0    9    5    2]
 [   5  530    0    3   16    0   18   22    0]
 [   0    1 1166   10   13    0    3    1   10]
 [   0    0   11  706   29    3    2    1    1]
 [   8    5   24   33 1112   20   11   13   11]
 [   0    1    1    2   25  259    3    1    1]
 [   6   16    2    2    7    0 1376   22    0]
 [   5   17    1    2   15    2   20  403    9]
 [   0    1    9    3   14    0    2    5  845]]
```

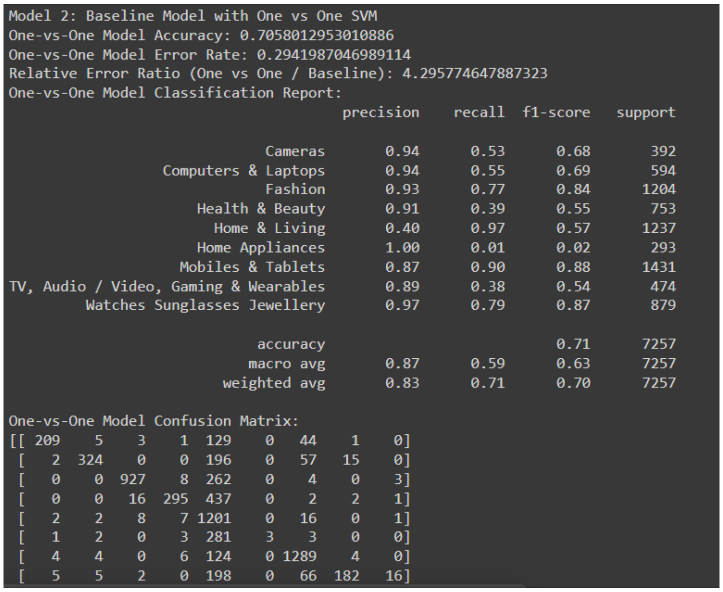


```
Model 1: Baseline Model with Stop Word Removal and Stemmed Text
Second Model Accuracy: 0.9298608240319691
Second Model Error Rate: 0.07013917596803088
Relative Error Ratio (One vs Rest / Baseline): 1.0241448692152917
Second Model Classification Report:
                                       precision    recall  f1-score   support

                              Cameras       0.93      0.93      0.93       392
                  Computers & Laptops       0.92      0.89      0.90       594
                              Fashion       0.95      0.96      0.96      1204
                      Health & Beauty       0.92      0.94      0.93       753
                        Home & Living       0.90      0.89      0.90      1237
                      Home Appliances       0.91      0.88      0.89       293
                    Mobiles & Tablets       0.95      0.96      0.96      1431
TV, Audio / Video, Gaming & Wearables       0.85      0.85      0.85       474
         Watches Sunglasses Jewellery       0.96      0.96      0.96       879

                             accuracy                           0.93      7257
                            macro avg       0.92      0.92      0.92      7257
                         weighted avg       0.93      0.93      0.93      7257

Second Model Confusion Matrix:
[[ 363    5    1    2    6    2    7    4    2]
 [   5  527    1    3   13    0   22   23    0]
 [   0    1 1160   10   16    0    4    1   12]
 [   0    0   15  710   23    1    2    2    0]
 [  10    4   29   39 1100   23    8   15    9]
 [   0    1    0    2   25  259    3    2    1]
 [   7   17    2    1    6    0 1377   21    0]
 [   6   19    1    1   12    1   20  404   10]
```

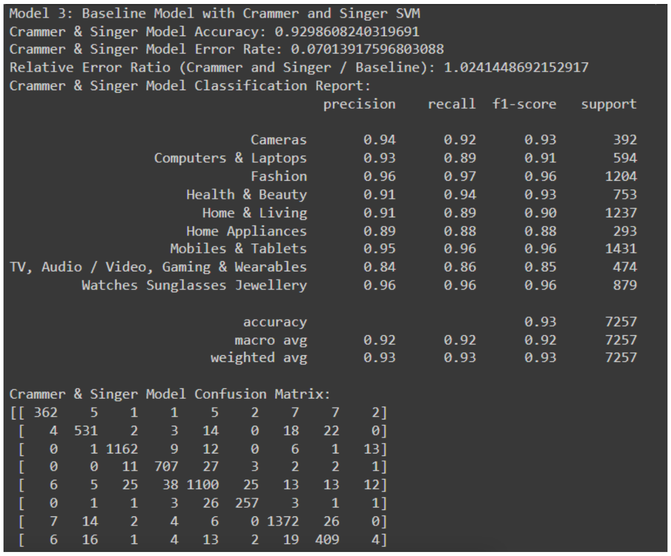


```
Model 2: Baseline Model with One vs One SVM
One-vs-One Model Accuracy: 0.7058012953010886
One-vs-One Model Error Rate: 0.2941987046989114
Relative Error Ratio (One vs One / Baseline): 4.295774647887323
One-vs-One Model Classification Report:
                                       precision    recall  f1-score   support

                              Cameras       0.94      0.53      0.68       392
                  Computers & Laptops       0.94      0.55      0.69       594
                              Fashion       0.93      0.77      0.84      1204
                      Health & Beauty       0.91      0.39      0.55       753
                        Home & Living       0.40      0.97      0.57      1237
                      Home Appliances       1.00      0.01      0.02       293
                    Mobiles & Tablets       0.87      0.90      0.88      1431
TV, Audio / Video, Gaming & Wearables       0.89      0.38      0.54       474
         Watches Sunglasses Jewellery       0.97      0.79      0.87       879

                             accuracy                           0.71      7257
                            macro avg       0.87      0.59      0.63      7257
                         weighted avg       0.83      0.71      0.70      7257

One-vs-One Model Confusion Matrix:
[[ 209    5    3    1  129    0   44    1    0]
 [   2  324    0    0  196    0   57   15    0]
 [   0    0  927    8  262    0    4    0    3]
 [   0    0   16  295  437    0    2    2    1]
 [   2    2    8    7 1201    0   16    0    1]
 [   1    2    0    3  281    3    3    0    0]
 [   4    4    0    6  124    0 1289    4    0]
 [   5    5    2    0  198    0   66  182   16]
```

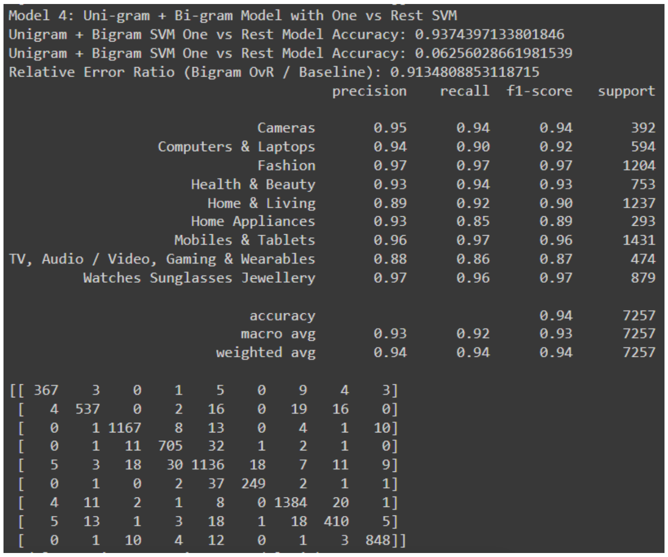


```
Model 3: Baseline Model with Crammer and Singer SVM
Crammer & Singer Model Accuracy: 0.9298608240319691
Crammer & Singer Model Error Rate: 0.07013917596803088
Relative Error Ratio (Crammer and Singer / Baseline): 1.0241448692152917
Crammer & Singer Model Classification Report:
                                       precision    recall  f1-score   support

                             Cameras       0.94      0.92      0.93       392
                 Computers & Laptops       0.93      0.89      0.91       594
                             Fashion       0.96      0.97      0.96      1204
                     Health & Beauty       0.91      0.94      0.93       753
                       Home & Living       0.91      0.89      0.90      1237
                     Home Appliances       0.89      0.88      0.88       293
                   Mobiles & Tablets       0.95      0.96      0.96      1431
TV, Audio / Video, Gaming & Wearables       0.84      0.86      0.85       474
         Watches Sunglasses Jewellery       0.96      0.96      0.96       879

                            accuracy                           0.93      7257
                           macro avg       0.92      0.92      0.92      7257
                        weighted avg       0.93      0.93      0.93      7257

Crammer & Singer Model Confusion Matrix:
[[ 362    5    1    1    5    2    7    7    2]
 [   4  531    2    3   14    0   18   22    0]
 [   0    1 1162    9   12    0    6    1   13]
 [   0    0   11  707   27    3    2    2    1]
 [   6    5   25   38 1100   25   13   13   12]
 [   0    1    1    3   26  257    3    1    1]
 [   7   14    2    4    6    0 1372   26    0]
 [   6   16    1    4   13    2   19  409    4]
```

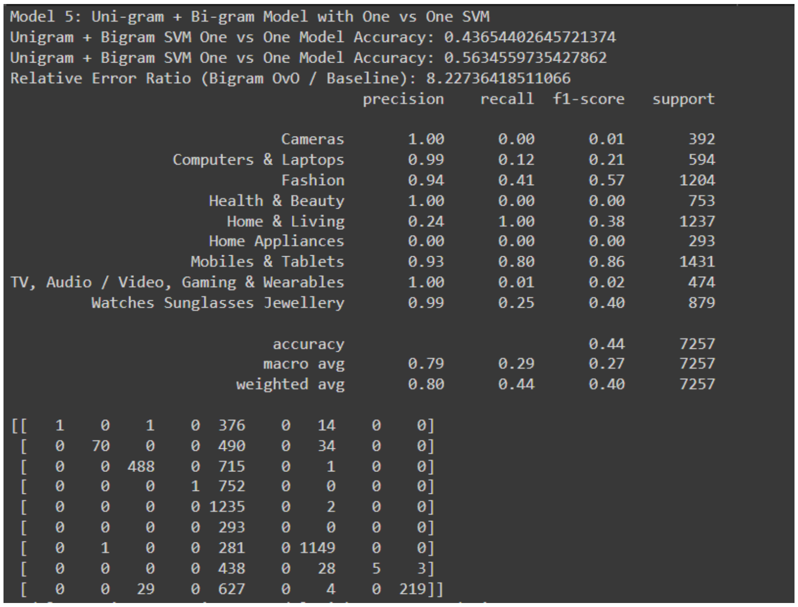


```
Model 4: Uni-gram + Bi-gram Model with One vs Rest SVM
Unigram + Bigram SVM One vs Rest Model Accuracy: 0.9374397133801846
Unigram + Bigram SVM One vs Rest Model Accuracy: 0.06256028661981539
Relative Error Ratio (Bigram OvR / Baseline): 0.9134808853118715
                                       precision    recall  f1-score   support

                             Cameras       0.95      0.94      0.94       392
                 Computers & Laptops       0.94      0.90      0.92       594
                             Fashion       0.97      0.97      0.97      1204
                     Health & Beauty       0.93      0.94      0.93       753
                       Home & Living       0.89      0.92      0.90      1237
                     Home Appliances       0.93      0.85      0.89       293
                   Mobiles & Tablets       0.96      0.97      0.96      1431
TV, Audio / Video, Gaming & Wearables       0.88      0.86      0.87       474
         Watches Sunglasses Jewellery       0.97      0.96      0.97       879

                            accuracy                           0.94      7257
                           macro avg       0.93      0.92      0.93      7257
                        weighted avg       0.94      0.94      0.94      7257

[[ 367    3    0    1    5    0    9    4    3]
 [   4  537    0    2   16    0   19   16    0]
 [   0    1 1167    8   13    0    4    1   10]
 [   0    1   11  705   32    1    2    1    0]
 [   5    3   18   30 1136   18    7   11    9]
 [   0    1    0    2   37  249    2    1    1]
 [   4   11    2    1    8    0 1384   20    1]
 [   5   13    1    3   18    1   18  410    5]
 [   0    1   10    4   12    0    1    3  848]]
```

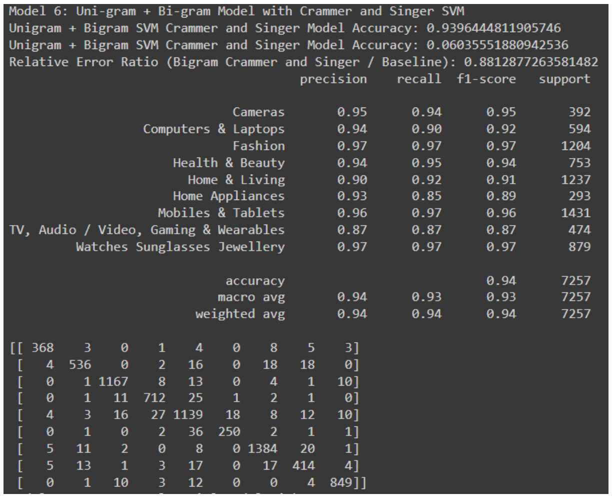


```
Model 5: Uni-gram + Bi-gram Model with One vs One SVM
Unigram + Bigram SVM One vs One Model Accuracy: 0.43654402645721374
Unigram + Bigram SVM One vs One Model Accuracy: 0.5634559735427862
Relative Error Ratio (Bigram OvO / Baseline): 8.22736418511066
                                       precision    recall  f1-score   support

                             Cameras       1.00      0.00      0.01       392
                 Computers & Laptops       0.99      0.12      0.21       594
                             Fashion       0.94      0.41      0.57      1204
                     Health & Beauty       1.00      0.00      0.00       753
                       Home & Living       0.24      1.00      0.38      1237
                     Home Appliances       0.00      0.00      0.00       293
                   Mobiles & Tablets       0.93      0.80      0.86      1431
TV, Audio / Video, Gaming & Wearables       1.00      0.01      0.02       474
         Watches Sunglasses Jewellery       0.99      0.25      0.40       879

                            accuracy                           0.44      7257
                           macro avg       0.79      0.29      0.27      7257
                        weighted avg       0.80      0.44      0.40      7257

[[   1    0    1    0  376    0   14    0    0]
 [   0   70    0    0  490    0   34    0    0]
 [   0    0  488    0  715    0    1    0    0]
 [   0    0    0    1  752    0    0    0    0]
 [   0    0    0    0 1235    0    2    0    0]
 [   0    0    0    0  293    0    0    0    0]
 [   0    1    0    0  281    0 1149    0    0]
 [   0    0    0    0  438    0   28    5    3]
 [   0    0   29    0  627    0    4    0  219]]
```

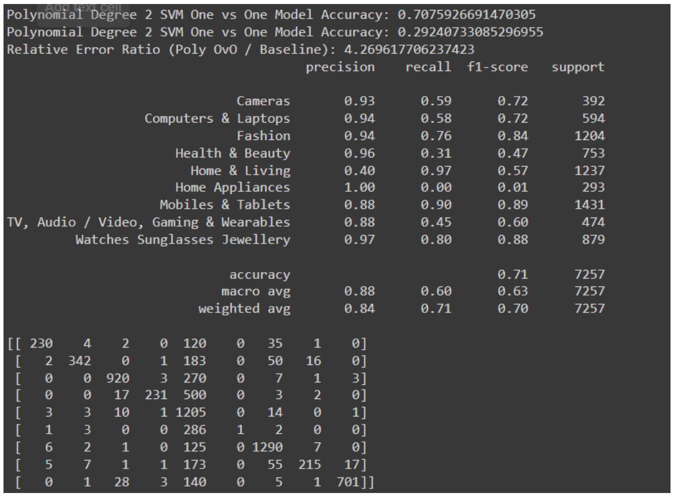


```
Model 6: Uni-gram + Bi-gram Model with Crammer and Singer SVM
Unigram + Bigram SVM Crammer and Singer Model Accuracy: 0.9396444811905746
Unigram + Bigram SVM Crammer and Singer Model Accuracy: 0.06035551880942536
Relative Error Ratio (Bigram Crammer and Singer / Baseline): 0.8812877263581482
                                       precision    recall  f1-score   support

                              Cameras       0.95      0.94      0.95       392
                  Computers & Laptops       0.94      0.90      0.92       594
                              Fashion       0.97      0.97      0.97      1204
                      Health & Beauty       0.94      0.95      0.94       753
                        Home & Living       0.90      0.92      0.91      1237
                      Home Appliances       0.93      0.85      0.89       293
                    Mobiles & Tablets       0.96      0.97      0.96      1431
TV, Audio / Video, Gaming & Wearables       0.87      0.87      0.87       474
          Watches Sunglasses Jewellery       0.97      0.97      0.97       879

                             accuracy                           0.94      7257
                            macro avg       0.94      0.93      0.93      7257
                         weighted avg       0.94      0.94      0.94      7257

[[ 368    3    0    1    4    0    8    5    3]
 [   4  536    0    2   16    0   18   18    0]
 [   0    1 1167    8   13    0    4    1   10]
 [   0    1   11  712   25    1    2    1    0]
 [   4    3   16   27 1139   18    8   12   10]
 [   0    1    0    2   36  250    2    1    1]
 [   5   11    2    0    8    0 1384   20    1]
 [   5   13    1    3   17    0   17  414    4]
 [   0    1   10    3   12    0    0    4  849]]
```

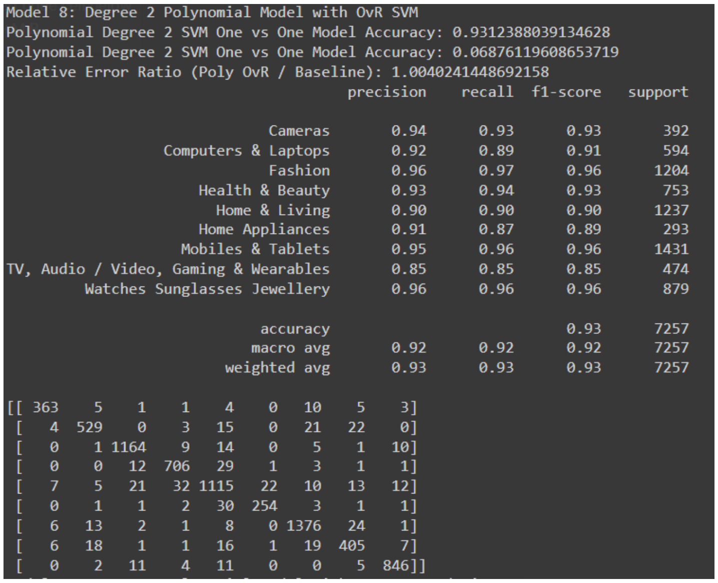


```
Polynomial Degree 2 SVM One vs One Model Accuracy: 0.7075926691470305
Polynomial Degree 2 SVM One vs One Model Accuracy: 0.29240733085296955
Relative Error Ratio (Poly OvO / Baseline): 4.269617706237423
                                       precision    recall  f1-score   support

                              Cameras       0.93      0.59      0.72       392
                  Computers & Laptops       0.94      0.58      0.72       594
                              Fashion       0.94      0.76      0.84      1204
                      Health & Beauty       0.96      0.31      0.47       753
                        Home & Living       0.40      0.97      0.57      1237
                      Home Appliances       1.00      0.00      0.01       293
                    Mobiles & Tablets       0.88      0.90      0.89      1431
TV, Audio / Video, Gaming & Wearables       0.88      0.45      0.60       474
          Watches Sunglasses Jewellery       0.97      0.80      0.88       879

                             accuracy                           0.71      7257
                            macro avg       0.88      0.60      0.63      7257
                         weighted avg       0.84      0.71      0.70      7257

[[ 230    4    2    0  120    0   35    1    0]
 [   2  342    0    1  183    0   50   16    0]
 [   0    0  920    3  270    0    7    1    3]
 [   0    0   17  231  500    0    3    2    0]
 [   3    3   10    1 1205    0   14    0    1]
 [   1    3    0    0  286    1    2    0    0]
 [   6    2    1    0  125    0 1290    7    0]
 [   5    7    1    1  173    0   55  215   17]
 [   0    1   28    3  140    0    5    1  701]]
```

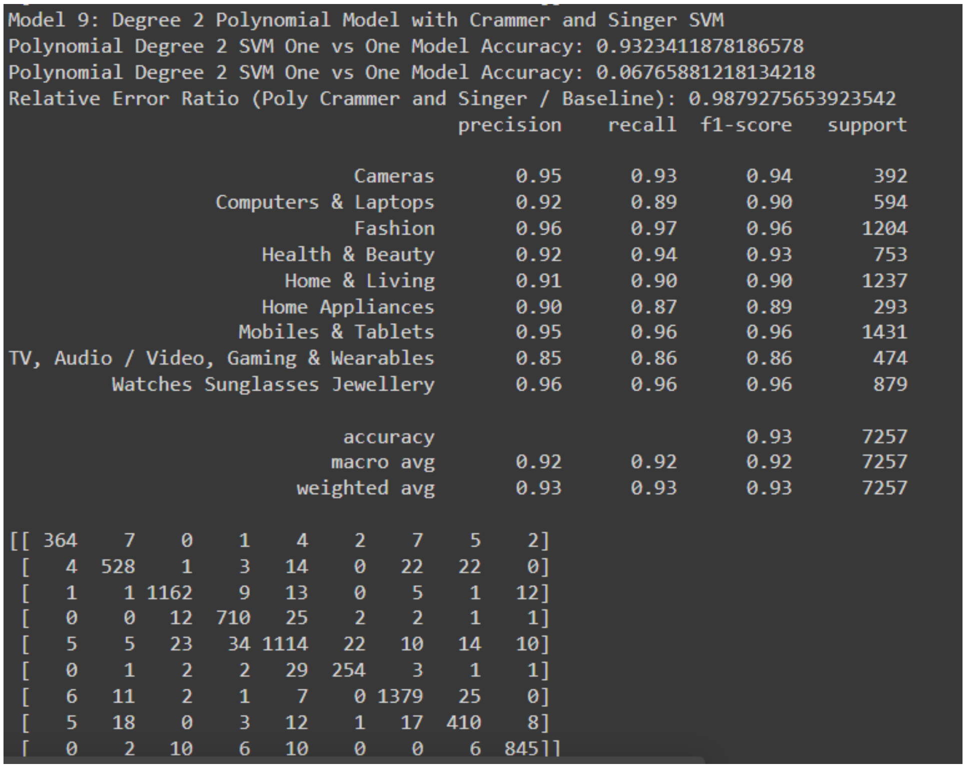


```
Model 8: Degree 2 Polynomial Model with OvR SVM
Polynomial Degree 2 SVM One vs One Model Accuracy: 0.9312388039134628
Polynomial Degree 2 SVM One vs One Model Accuracy: 0.06876119608653719
Relative Error Ratio (Poly OvR / Baseline): 1.0040241448692158
                                       precision    recall  f1-score   support

                              Cameras       0.94      0.93      0.93       392
                  Computers & Laptops       0.92      0.89      0.91       594
                              Fashion       0.96      0.97      0.96      1204
                      Health & Beauty       0.93      0.94      0.93       753
                        Home & Living       0.90      0.90      0.90      1237
                      Home Appliances       0.91      0.87      0.89       293
                    Mobiles & Tablets       0.95      0.96      0.96      1431
TV, Audio / Video, Gaming & Wearables       0.85      0.85      0.85       474
          Watches Sunglasses Jewellery       0.96      0.96      0.96       879

                             accuracy                           0.93      7257
                            macro avg       0.92      0.92      0.92      7257
                         weighted avg       0.93      0.93      0.93      7257

[[ 363    5    1    1    4    0   10    5    3]
 [   4  529    0    3   15    0   21   22    0]
 [   0    1 1164    9   14    0    5    1   10]
 [   0    0   12  706   29    1    3    1    1]
 [   7    5   21   32 1115   22   10   13   12]
 [   0    1    1    2   30  254    3    1    1]
 [   6   13    2    1    8    0 1376   24    1]
 [   6   18    1    1   16    1   19  405    7]
 [   0    2   11    4   11    0    0    5  846]]
```

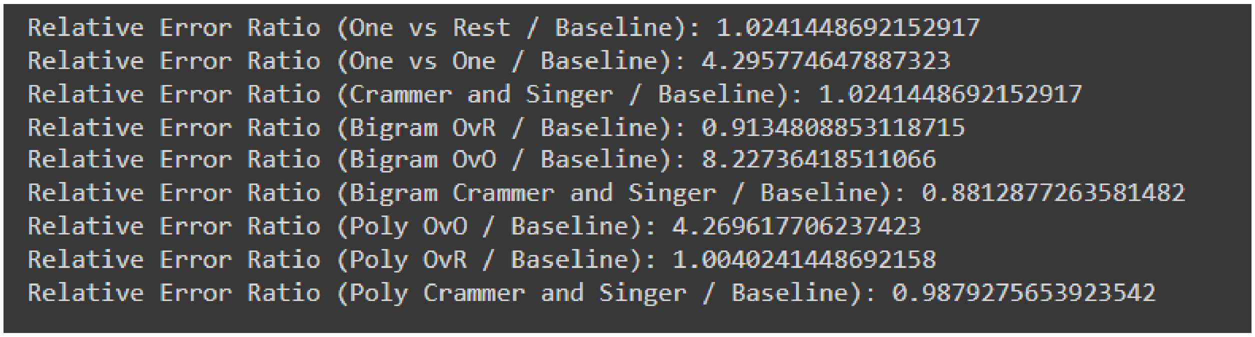


```
Model 9: Degree 2 Polynomial Model with Crammer and Singer SVM
Polynomial Degree 2 SVM One vs One Model Accuracy: 0.9323411878186578
Polynomial Degree 2 SVM One vs One Model Accuracy: 0.06765881218134218
Relative Error Ratio (Poly Crammer and Singer / Baseline): 0.9879275653923542
                                       precision    recall  f1-score   support

                              Cameras       0.95      0.93      0.94       392
                  Computers & Laptops       0.92      0.89      0.90       594
                              Fashion       0.96      0.97      0.96      1204
                      Health & Beauty       0.92      0.94      0.93       753
                        Home & Living       0.91      0.90      0.90      1237
                      Home Appliances       0.90      0.87      0.89       293
                    Mobiles & Tablets       0.95      0.96      0.96      1431
TV, Audio / Video, Gaming & Wearables       0.85      0.86      0.86       474
         Watches Sunglasses Jewellery       0.96      0.96      0.96       879

                             accuracy                           0.93      7257
                            macro avg       0.92      0.92      0.92      7257
                         weighted avg       0.93      0.93      0.93      7257

[[ 364    7    0    1    4    2    7    5    2]
 [   4  528    1    3   14    0   22   22    0]
 [   1    1 1162    9   13    0    5    1   12]
 [   0    0   12  710   25    2    2    1    1]
 [   5    5   23   34 1114   22   10   14   10]
 [   0    1    2    2   29  254    3    1    1]
 [   6   11    2    1    7    0 1379   25    0]
 [   5   18    0    3   12    1   17  410    8]
 [   0    2   10    6   10    0    0    6  845]]
```



```
Relative Error Ratio (One vs Rest / Baseline): 1.0241448692152917
Relative Error Ratio (One vs One / Baseline): 4.295774647887323
Relative Error Ratio (Crammer and Singer / Baseline): 1.0241448692152917
Relative Error Ratio (Bigram OvR / Baseline): 0.9134808853118715
Relative Error Ratio (Bigram OvO / Baseline): 8.22736418511066
Relative Error Ratio (Bigram Crammer and Singer / Baseline): 0.8812877263581482
Relative Error Ratio (Poly OvO / Baseline): 4.269617706237423
Relative Error Ratio (Poly OvR / Baseline): 1.0040241448692158
Relative Error Ratio (Poly Crammer and Singer / Baseline): 0.9879275653923542
```