Experimenting with a multiclass classification model

Sachin Patel (22265660)
Department of Computing, Dublin City University
Dublin, Ireland
sachin.patel3@mail.dcu.ie

***Abstract*— In this paper, we propose the development of a multiclass classification machine learning model for finding the top category id, bottom category id, and primary colour id of a product listed on the Etsy e-commerce website. The goal is to create an efficient and accurate model that can predict all three target categories simultaneously or using separate models. The model will be trained on a dataset containing product titles, tags, types, and craft types as features. The paper will discuss the preprocessing of the text data, including feature extraction and transformation using techniques such as CountVectorizer and TfidfTransformer. The Support Vector Machine (SVM) algorithm will be used for the classification task, with a linear kernel and a regularization parameter of 1.0. The paper will present the results of the model's performance in terms of F1-score and classification report on a test dataset. The proposed model has the potential to significantly improve the accuracy and efficiency of product categorization on the Etsy e-commerce website, leading to better user experience and increased customer satisfaction.**

Keywords—machine learning, classification, transformer, support vector machine, Etsy, regularization, model development, text classification, multi-class classification model

I. INTRODUCTION

In the ever-growing e-commerce sector, improving customer satisfaction requires effective and fast product classification. Online retailers selling a variety of handcrafted articles, clothing, home and furniture and one-of-a-kind items, like Etsy, must be able to classify things properly into relevant categories like top category, bottom category, and primary colour.

Machine learning algorithms have recently demonstrated promising outcomes in a variety of classification applications, including multiclass classification. Multiclass classification is particularly pertinent for the task of categorising objects on Etsy because it entails predicting numerous target categories at once. By automatically classifying things into their top category, bottom category, and primary colour, a successful multiclass classification model can help Etsy create a robust and well-organized product catalogue.

II. RELATED WORK

Multi-class classification is particularly relevant in the context of categorizing products on e-commerce websites, such as Etsy, where predicting multiple target categories simultaneously is necessary. In this literature review, we review several papers that discuss different approaches and techniques for multi-class classification, with a focus on improving the accuracy and efficiency of product categorization.

One common approach for text-based multi-class classification is to use term weighting schemes to represent text data effectively. The paper "Inverse-Category-Frequency based Supervised Term Weighting Schemes for Text Categorization" (Wang and Zhang, no date) proposes an inverse-category-frequency-based term weighting scheme for text categorization. The authors show that their proposed method outperforms traditional term weighting schemes in multi-class classification tasks, including improving accuracy and reducing computational overhead.

Another paper "Multiclass Classification Methods: A Review" (Alsafy, Aydam and Mutlag, 2018) provides an overview of various multiclass classification methods, including decision tree-based, distance-based, ensemble-based, and transformation-based methods. The authors review the strengths and weaknesses of different methods and discuss their applications in various domains.

Feature engineering is another critical aspect of improving multi-class classification performance. The paper "Improving Quality of the Multiclass SVM Classification Based on Feature Engineering" (Klyueva, 2019) proposes a feature engineering method for multi-class SVM classification. The authors introduce a method that uses feature engineering techniques, such as feature selection and feature extraction, to enhance the quality of the SVM classifier, leading to improved classification accuracy.

In addition to traditional machine learning algorithms, deep learning techniques, such as neural networks, have also been employed for multi-class classification tasks. The paper "Policy Text Classification Algorithm Based on BERT" (Yu, Deng and Bu, 2022) proposes a policy text classification algorithm based on the BERT (Bidirectional Encoder Representations from Transformers) model. The authors show that their proposed algorithm achieves high accuracy in policy text classification tasks, demonstrating the effectiveness of deep learning techniques for multi-class classification.

Furthermore, some studies have focused on specific domains or applications of multi-class classification. For example, the paper "Albanian News Category Predictor System using Multinomial Naïve Bayes and Logistic Regression Algorithms" (Shkurti and Kabashi, 2021) presents a news category prediction system for Albanian news articles using multinomial Naïve Bayes and logistic regression algorithms. The authors discuss the performance of their proposed system in predicting news categories and highlight the potential applications of their approach in the context of news categorization.

Finally, some studies have investigated the impact of imputation techniques on multi-class classification performance. For instance, the paper "Data Imputation Techniques: An Empirical Study using Chronic Kidney Disease and Life Expectancy Datasets" (Reddy Sankepally, Kosaraju and Mallikharjuna Rao, 2022) conducts an empirical study on different data imputation techniques and their effects on multi-class classification performance using chronic

kidney disease and life expectancy datasets. The authors provide insights into the importance of data imputation

III. DATASET

More than 5 million active sellers on Etsy's marketplace have approximately 100 million current listings for sale. A portion of those items are given as training data. The Etsy dataset was provided by the company itself and it contains a zip files consisting of two file types (i) Parquet (ii) TFRecords Both these files types have train and test folder consisting of train and test files. The parquet files consist only textual data whereas TFRecords files have images in it. The overall size of the dataset was 14Gb zipped.

Data Understanding

- The training dataset contains 21 columns whereas, the test dataset contains 15 columns.
- Total records in training dataset 245K and in test dataset has 27K records.
- The following product attributes which were to be predicted had no missing values in them.
- Top Category ID or Top Category Text has 15 unique classes, Bottom Category ID or Bottom Category ID has 2782 classes and Color ID or color Text has 20 unique classes.
- The remaining 14 out of 15 columns have missing values of different proportions and it was observed that these were Missing at random, and some were Missing not at random.
- The data in color text column is highly skewed.

IV. DATA PREPARATION

The data preparation stage specially deals with the handling of the missing data, analysing outliers, data cleaning, data transformation, data integration, data encoding, data splitting, data preprocessing.

A. Handling missing values

Initially, the title and description column was checked for missing values and it was observed that 0.4% data was missing along with missing in other columns too for these 940 records. So, these records can be dropped as it will not impact much as almost all the columns have missing data for these records. The 'type' attribute also has 0.5% missing records and the distribution of data in it is highly imbalanced. All the imputation techniques to impute the missing records were appropriately chosen considering the computational capacity of the available system. For the "type" mode would be appropriate choice but that will also create bias against the minority class. So, a different approach to impute the data from the description column was followed. As description of a product contains all the related product information and for the type of shipping or delivery method is physical or download, a text matching using regex can be applied to check which record has "download" or "no physical" keywords and impute the same in type column. The all three occasions, holiday and craft type attributes have more than 60% data missing. Therefore, a similar approach was utilized to impute values from description column to these columns. For the rest of the columns which contain more than 90% missing data it can be dropped as it will be difficult to impute values with computational constraints. Basic cleaning of special characters, new lines and stop words were carried out during this phase.

B. Feature Engineering

One of the three attributes to predict is the primary color of the product. Every product in the dataset has some primary color and it is mentioned either in the title of the product or in the product tags or in the description of the product. There are two methods to predict the primary color of the product:

1. Use the TFRecords training data files to build a model using a transfer learning approach and then detect the primary color in the image.
2. Another method is to extract the color names from the title, tags, description of the products using pattern matching regular expression.

According to the second method, a new feature "extracted color" was created by extracting the color names by pattern matching and it contains a list of all the colors available in the title, tags, description for a particular record. Now the aim is to extract only one color from the list of extracted colors for each record and another feature called reduced color was created. While extracting the color from list most frequent color was preferred and if there is single occurrence of all the color in the list then the color name from title was preferred as it was observed that the primary color was mentioned in the product title or in the product tags.

There were two specific observations: The color name "burgundy" was present in title, tags, description of a product was labeled as red color in the color text column. This implies that burgundy which is like red color was also considered as red color. Another observation was that the craft type attribute contained "jewelry / jewellry" and gold is a metal in this context not a color. But in the unique color list there is gold color which should not be counted when the craft type is jewelry. These two points were taken in consideration while extracting color name and creating a new feature column.

V. MODEL DEVELOPMENT & EVALUATION

1. For Top Category ID

a) Baseline Model – Multinomial Naïve Bayes: A simple multinomial Naïve Bayes model which can handle discrete features such as word occurences or frequencies in text data and is efficient and fast was considered to build a simple model which will act as a baseline model for further model development. The MNB was trained on the title of the product to predict "top category id" and it performed good with the validation F1 score of 63%.

b) Stochastic Gradient Descent Classifier (SGDClassifier): The SGDClassifier model was trained on title for predicting "top category id" but this time with three different learning rates with other parameters as penalty as 12 max iteration 10 and the model performed pretty well outperforming baseline MNB model with validation F1 score of 81.64% at 1e-05 learning rate.

c) Support Vector Machine (SVM): The SVM with a linear kernel is well suited for handling high dimensional data as it can separate data points in a linearly seprarable manner

without complex non-linear transformations. It can handle large datasets with many training samples and also has good generalization performance and robustness to noise. This model was trained on title, tags, type, craft_type for predicting "top category id" and it performed pretty well with validation F1 score of 86.27% with regularization hyperparameter C=1.0 which can be tuned and evaluate on different values of C.

2. For Bottom Category ID

a) Baseline Model – Multinomial Naïve Bayes: NVB for bottom category was trained on title and after evaluation the validation F1 score was 47%

b) Stochastic Gradient Descent Classifier: SGD was trained on title, type, craft type at three different learning rates and the model outperformed baseline model with validation F1 score of 50% at 1e-05 learning rate.

c) SVM with sigmoid kernel: SVM model was trained on title, type, craft type and since there are more than 2500+ classes in the bottom category, the machines avialble on open source like google colab and Kaggle are unable to handle in terms of memory capacity and therefore relevant metrics is not available for this model. The sigmoid or Radial Bias function are the kernel types which could have effective for bottom category id.

3. For Color ID

a) *Logistic Regression*: A simple logistic regression model was trained on the featured engineered column "reduced color" and title column for predicting color id and the model performed decent with the validation F1 score of 52.28% which provides a baseline model for predicting color id.

b) Multinomial Logistic Regression (Softmax Regression): The multinomial Logistic Regression model, which was trained on reduced color, title, tags for predicting color id performed same as that of logistic regression without SoftMax.

c) SVM with Linear Kernel: The SVM model, which was trained on reduced color, title, tags, type, craft type for predicting color id outperformed Logistic Regression model with validation F1 score of 55.83% which clearly shows improvement over the logistic regression model.

TABLE I. MODEL EVALUATION METRICS

Target Column	Model Evaluation Results		
	Model Name	***Accuracy***	***F1 score***
Top cat. id	Naïve Bayes	0.6698	0.6391
Top cat. id	SGDClassifier	0.8204	0.8164
Top cat. id	SVM (Linear Kernel)	0.8623	0.8627
Bottom cat id	Naïve Bayes	0.5021	0.4756
Bottom cat id	SGDClassifier	0.5187	0.5002
Color id	Logistic Regression	0.5281	0.5228
Color id	Multinomial Logistic Regression (Softmax)	0.5272	0.5180
Color id	SVM (Linear kernel)	0.5570	0.5583

Fig. 1. Model Evaluation Metrics

Figure 1. shows the different types of machine learning model implemented on the Etsy dataset. F1 score as metric is used as it is the appropriate metric for multiclass classification problems.

CONCLUSION

Thus, experimenting with a multiclass classification model can provide valuable insights and help improve the accuracy of predictions. With careful consideration of data preparation, feature engineering, model selection, and hyperparameter tuning, a well-designed multiclass classification model can offer a powerful tool for solving complex classification problems. Overall, the findings demonstrate the importance of data preparation and feature engineering for accurate prediction of product attributes in e-commerce datasets.

Project Links : Etsy Notebook

REFERENCES

[1] Alsafy, B.M., Aydam, Z.M. and Mutlag, W.K. (2018) 'Multiclass Classification Methods: A Review', 5(3).

[2] Klyueva, I. (2019) 'Improving Quality of the Multiclass SVM Classification Based on the Feature Engineering', in *2019 1st International Conference on Control Systems, Mathematical Modelling, Automation and Energy Efficiency (SUMMA)*. *2019 1st International Conference on Control Systems, Mathematical Modelling, Automation and Energy Efficiency (SUMMA)*, pp. 491–494. Available at: https://doi.org/10.1109/SUMMA48161.2019.8947599.

[3] Reddy Sankepally, S., Kosaraju, N. and Mallikharjuna Rao, K. (2022) 'Data Imputation Techniques: An Empirical Study using Chronic Kidney Disease and Life Expectancy Datasets', in *2022 International Conference on Innovative Trends in Information Technology (ICITIIT)*. *2022 International Conference on Innovative Trends in Information Technology (ICITIIT)*, pp. 1–7. Available at: https://doi.org/10.1109/ICITIIT54346.2022.9744211.

[4] Shkurti, L. and Kabashi, F. (2021) 'Albanian News Category Predictor System using a Multinomial Naïve Bayes and Logistic Regression Algorithms', in *2021 5th International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT)*. *2021 5th International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT)*, pp. 642–647. Available at: https://doi.org/10.1109/ISMSIT52890.2021.9604602.

[5] Wang, D. and Zhang, H. (no date) 'Inverse-Category-Frequency based Supervised Term Weighting Schemes for Text Categorization'.

[6] Yu, B., Deng, C. and Bu, L. (2022) 'Policy Text Classification Algorithm Based on Bert', in *2022 11th International Conference of Information and Communication Technology (ICTech))*. *2022 11th International Conference of Information and Communication Technology (ICTech))*, pp. 488–491. Available at: https://doi.org/10.1109/ICTech55460.2022.00103.