Category and Colour Prediction using Multiclass Classification on Etsy Dataset.

Agnideep Mukherjee
Student id: 22260149
agnideep.mukherjee2@mail.dcu.ie

***Abstract*-- This research article examines the category and colour trends discovered in a dataset of items from the popular online marketplace Etsy. Using a machine learning approach, I investigate the capacity to predict the category and colour of an item based on its title and image. Our findings imply that, whereas category prediction is a relatively reliable task, colour prediction is more difficult due to the subjective character of colour perception. Overall, our study sheds light on the potential of machine learning techniques for analysing large-scale e-commerce data and provides insights into the ways in which consumers interact with and perceive different product categories and colours.**

***Index terms*- Random Forest, K-Nearest Neighbour (KNN), Bag-of-Words, Natural Language Processing (NLP)**

I. INTRODUCTION

Etsy is a global marketplace for one-of-a-kind and original products. It is the home of a vast array of exceptional and unusual products, including handcrafted rarities and vintage treasures. Etsy runs two-sided online marketplaces that link 7.7 million sellers and approximately 100 million passionate, creative shoppers globally. [1]

The study of massive datasets in a variety of fields, including e-commerce, has been transformed by the use of machine learning techniques. Exploring the potential of machine learning for understanding consumer behaviour and preferences has grown more crucial as online marketplaces like Etsy have proliferated. Using a dataset made up of item titles and photographs, the purpose of this study is to evaluate the capacity to predict the categories and colours of things sold on Etsy. Similar tasks, such as detecting garment features from photographs, have been investigated in the past, but few studies have concentrated on predicting categories and colours for a wide variety of objects. As a result, this work significantly advances our knowledge of e-commerce data analysis.

The colour detection of the real-world object has always been a daunting task because of the illumination, angles, reflection, shadow etc. Numerous methods, such as the Feature Context and Normalized RGB Histogram (NRH), have been suggested to address these issues. However, NRH has its own limitations. To address this A. Tariq et al have leveraged Faster RCNN, a type of deep learning model, to classify and recognize the colours of vehicles. [2]

Predicting the top and bottom categories of products is important for e-commerce websites as it improves the user experience, helps in organizing products, improves SEO provides insights into customer preferences and aids inventory management. [3]

II. RELATED WORKS

Product category matching is the process of determining what branches correspond to a particular item. For e-commerce systems like online marketplaces, which have a vast inventory of different articles, it is very important to identify and place the items under the correct category for ease of navigation on the website.[4] Feature engineering is used in traditional target-dependent sentiment classification to optimize classifier performance. However, this takes a significant amount of time, resources, and domain knowledge, and it is prone to error. There have been numerous sentiment lexicons established, but they are domain-specific and require significant human work.[6]

Once a text vector has been constructed using vectorization techniques such as Bag-of-Words (BOW), Term Frequency-Inverse Document Frequency (Tf-Idf), Fast Text, and others, various state-of-the-art classification models can be used to predict the classes based on the word vectors. Vishan Kumar Gupta et al. have used classification methodologies like Support Vector Machine (SVM), Random Forrest, Decision Tree, and Multinomial Logistic Regression. [7].

ULMFiT, ELMo, and BERT are a few examples of pre-trained language models that have recently shown an extraordinary aptitude for text's semantic expressiveness.[6] X et al. have used Bidirectional Transformer (BERT) to pre-train a language model based on a large corpus. It's used to express a sentence or a set of phrases as a series of tokens, either individually or together. [6]

Computer vision is important in e-commerce websites because it allows for the automation of visual activities like product classification, image tagging, and visual

search. Computer vision algorithms can assist e-commerce websites improve the user experience, boosting product discovery, and raising sales conversion rates by analysing product photos. Furthermore, computer vision can help e-commerce companies with fraud detection, quality control, and inventory management. [8] [5].

Colour detection on a large e-commerce website might be difficult due to cluttered backdrops, bad lighting conditions, the usage of poor cameras, and so on. K. Aryafar et al. overcame this difficulty by employing object detection techniques on the Etsy dataset. They generated an image mask by detecting edges and outlines and then utilized it to detect colour. [5]

III. METHODOLOGY

The colour and category prediction has been performed in separate flows. The prediction for the bottom and top category id has been described in Figure 1.

The first and most time-consuming stage was data cleaning and preparation. The Null values in the columns: *type, room, craft_type, recipient, material*, etc. have been omitted because the information in those columns is not used in the feature matrix. However, the rows with missing titles, descriptions, or tags were truncated because they made up 22.5% of the overall dataset, leaving the balance of the dataset with 210574 data rows to work with.

The content from the title, description, and tags has been combined into a single column that provides all of the textual information needed to generate the feature matrix. The text was then cleaned by lowercasing it and deleting any unnecessary characters and symbols. To improve readability, all stopwords have been eliminated. The ready-made set of stopwords was taken from NLTK.corpus. To further enhance the text, stemming has been used to reduce words to their base or root form.

To create the matrix of features Bag of Words has been used, where the max feature attribute has been set to 600 in order to increase training performance and eliminate terms with a low frequency that do not contribute to prediction accuracy.

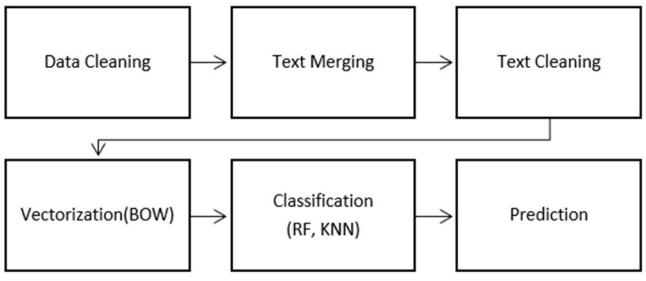

Figure 1: Prediction workflow for Top & Bottom Category Prediction

From Figure 2, we can observe that the dataset is heavily biased for the category_id=8, 6 & 5. To address this issue I have used oversampling by SMOTE. The SMOTE algorithm calculates the nearest neighbours of the fault samples using the KNN approach and synthesizes fresh minority class samples using random linear interpolation.[9]

The balanced dataset has been divided into test and training sets in an 80:20 ratio. The training set has been used for training classification models like K-NN (K nearest neighbours) and Random Forest. The top category.

First, I applied the Random Forest Model. First, I assigned each review to one of ten categories using Random Forest's categorization. By randomly picking data samples and feature groups, this approach builds a large number of decision trees. The trees are then trained on a subset of the data, and their individual forecasts are combined to create a final prediction. Random forests are well-suited for NLP applications because they can manage a large number of features, which is common in textual data. I designed the decision trees with entropy as the criterion, which results in better-balanced trees with less overfitting. Unfortunately for Bottom category prediction, since the number of the multiclassification is mammoth (>2000) classes, training it with Random Forest is consuming up all the RAM memory, thereby crashing the Kernel.

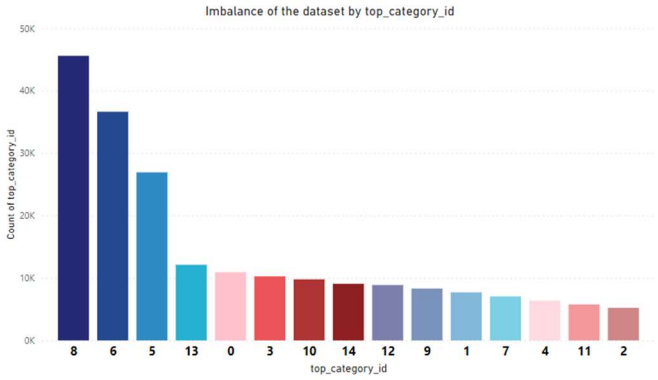

Figure 2: Distribution of Etsy products by top_category_id

Secondly, I tested the K-Nearest Neighbours (KNN) clustering method. The KNN technique is a supervised learning methodology that predicts a new data point's categorization based on how close it is to existing data points in a feature space. To calculate the distance between two locations, I used the Minkowski distance, which is a generalization of the Manhattan and Euclidian distances.

For the classification of Images by their colour I have chosen a different workflow as mentioned in Figure 3. I start the flow by storing the images in a local path to optimise the usage of our limited computational power. I have stored 1000 images per colour category as my

training dataset, and 100 images per colour as the test dataset. Subsequently, I started the cleaning procedure, in this step, I removed all the images below 8kb in size and truncated the images which do not fall under the following extensions: *'jpeg', 'jpg', 'bmp' & 'png'*.

The next step was critical. Because pixel values in photographs generally vary from 0 to 255, where 0 indicates black and 255 represents white, the pixels in the photos were divided by 255. Because the input characteristics are more consistent and have a smaller range, scaling the pixel values to be in the range of 0 to 1 allows the model to learn better.

Post that the image had been trained by ResNet50. It is a deep convolutional neural network which is pre-trained with 50 layers. I have used ImageGenerator to augment versions of the training data in real time during the training process. This helps to increase the diversity of the training data and improve the model's ability to generalize to new data. A batch size of 54 and an epoch of 20 has been set for the training.

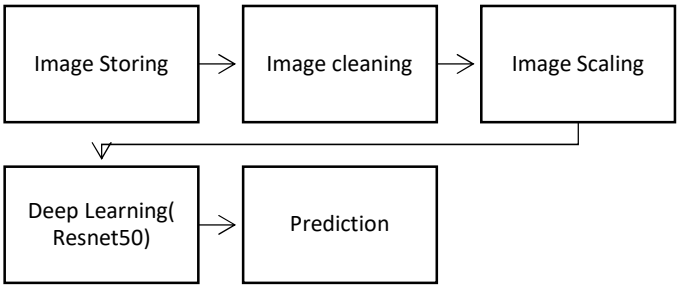

Figure 3: Prediction workflow for Colour Prediction

IV. EVALUATION

I considered Accuracy, Precession, Recall and F1 score as our metrics for evaluation. At first, I had predicted the Top category. We can clearly see from Table 1, that Random Forest has outperformed the KNN. It yielded around 88.6% whereas KNN has been able to reach 67.8%.

Metric	*Random Forest*	*KNN*
Accuracy	0.886	0.678
Precession	0.885	0.723
Recall	0.886	0.678
F1 Score	0.886	0.663

Table 1: Top Category Prediction Scores

Due to computing power limits, classification methods such as Decision Tree and Random Forest failed with a 'memory error' for Bottom Category Prediction. However, I used KNN to train the model, which performed well and provided an accuracy of 97%. The other metrics of evaluation are mentioned in Table 2.

Metric	*Random Forest*	*KNN*
Accuracy	--	0.973
Precession	--	0.977
Recall	--	0.974
F1 Score	--	0.974

Table 2: Bottom Category Prediction Scores

For colour category prediction I have tried ResNet50 as the deep learning algorithm. The results of the metric can be seen in Table 3.

Metric	*ResNet50*
Accuracy	*0.287*
Precession	*0.281*
Recall	*0.273*
F1 Score	*0.281*

Table 3: Colour Id Prediction Scores

V. CONCLUSION

In this project, we aimed to predict the top and bottom categories of products on Etsy.com, as well as the colour of the products using machine learning algorithms. We used Random Forest and K-Nearest Neighbour classification algorithms to predict the top and bottom categories, achieving an accuracy of over 85% for both the category. However, for the colour prediction, the accuracy was poor.

For future scope, I would suggest Experimenting with other deep learning models: ResNet50 is just one of many deep learning models that can be used for image classification. It would be interesting to try out other models like Inception, VGG, or MobileNet to see if they improve the accuracy of the colour classification. If we can detect the object from an image using another Machine Learning model and pick the colour from the object mask, then it might make more accurate predictions. Also, The Random Forest and KNN models used in this project were not fine-tuned. Fine-tuning these models could lead to better results.

VI. REFERENCES

[1] About Etsy," Etsy. https://www.etsy.com/ie/about?ref=ftr

[2] A. Tariq, M. Z. Khan, and M. U. Ghani Khan, "Real Time Vehicle Detection and Colour Recognition using tuned Features of Faster-RCNN," IEEE Xplore, Apr. 01, 2021. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9425106 (accessed Apr. 15, 2023).

[3] C. Corbiere, H. Ben-Younes, A. Rame, and C. Ollion, "Leveraging Weakly Annotated Data for Fashion Image Retrieval and Label Prediction,"

2017 IEEE International Conference on Computer Vision Workshops (ICCVW), Oct. 2017, doi: https://doi.org/10.1109/iccvw.2017.266.

[4] R. Peeters and C. Bizer, “Dual-objective fine-tuning of BERT for entity matching,” Proceedings of the VLDB Endowment, vol. 14, no. 10, pp. 1913–1921, Jun. 2021, doi: https://doi.org/10.14778/3467861.3467878.

[5] K. Aryafar, C. Lynch, and J. Attenberg, “Exploring User Behaviour on Etsy through Dominant Colors,” IEEE Xplore, Aug. 01, 2014. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6976966 (accessed Apr. 15, 2023).

[6] Z. Gao, A. Feng, X. Song, and X. Wu, “Target-Dependent Sentiment Classification With BERT,” IEEE Access, vol. 7, pp. 154290–154299, 2019, doi: https://doi.org/10.1109/access.2019.2946594.

[7] V. K. Gupta, A. Gupta, D. Kumar, and A. Sardana, “Prediction of COVID-19 confirmed, death, and cured cases in India using random forest model,” Big Data Mining and Analytics, vol. 4, no. 2, pp. 116–123, Jun. 2021, doi: https://doi.org/10.26599/bdma.2020.9020016.

[8] X. Meng, J. Lin, and Y. Ding, “An extended HOG model: SCHOG for human hand detection,” IEEE Xplore, May 01, 2012. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6223584 (accessed Apr. 16, 2023).