Categories Classification Depending on the Textual Context

Abstract – **Buying and selling goods online have grown over the years. The technical term for this is E-commerce. This allows businesses to reach a large group of audience and offer customers the convenience of shopping from anywhere, at any time, using a variety of devices. Etsy is an e-commerce company which focuses on online shopping for handmade craft supplies and vintage items. The growth in this domain led to the building of some systems which recommend to customers products depending on their likings and searches. The dataset for the same problem was shared by the company which contained the required columns for textual analysis. Compare to other analysis models the base model was found to be a better choice. A comparison of the models like Logistic Regression, MultinomialNB, Decision Tree, and SVM was done for getting a better model for the classification of the features.**

I. INTRODUCTION

E-Commerce refers to buying and selling items/products over the Internet. Over the years there has been great growth in online shopping since it is easy for customers to buy stuff and get it delivered to their doorstep. This has given big and small businesses have got large audiences for their product not only in their country but also globally. Businesses have reached globally due to E-commerce and social media. Online shopping has got a boost in recent years due to its convenience, as consumers can browse and purchase any product from the comfort of their homes. Etsy is one of the best E-commerce companies which has grown over the years. Etsy mainly focuses on handmade or vintage items and craft supplies. These things mainly include jewelry, bags, clothing, home décor, furniture, toys etc.

This research tried to find a way in which depending on the textual columns is it possible to find the top, bottom and color id for that product. The aim of this research is to develop a model that can forecast demand for the product category which will help the company to manage its inventory with ease. For the same, the data was provided by the company. Various models were built to accurately forecast the requirements of the company.

II. PREVIOUS WORK

[1] This research compared different models and tried to tell which algorithm is better for text classification. This research describes how the BBC news text classification model was built using machine learning models such as logistic regression, random forest, and KNN. The logistic regression model with TFIDF vectorizer was found to be most accurate than the other two models. The accuracy for Logistic regression was 97% followed by Random forest at 93% and then KNN at 92%.

[2] The authors in this research used supervised machine learning approaches for pre-processing and feature selection. They also compared the performance of several machine learning models such as Naïve Bayes, SVM, and Logistic Regression. The results showed that the SVM outperformed the other models. They used precision, recall and f1 score as their matrices.

The growing use of textual data has made it important to develop mechanisms for automatic text classification. The researchers in [3] talked about the important steps for the classification of text which includes data collection, text pre-processing, feature extraction, and text classification methods. The experimental results suggested that incorporating relation information into the classification process may significantly enhance the quality of the underlying results.

Hence depending on the studies of [1] and [2], the models like Logistic Regression, SVM, Decision Tree and MultinomialNB were shortlisted for the classification of the categories. Also cleaning and regularising the data for modelling was learnt through it.

III. Data

The data used for the analysis was provided by the company. This data contained parquet files for training and testing. The only difference between training and testing files was the number of columns in them. Testing data had three fewer columns than training data.

The data contained a total of 21 columns and 245485 rows. This data had mixed data types. The columns in the dataset were product_id, title, description, tags, type, room, craft_type, recipient, material, occasion, Holiday, art_subject, style, shape, pattern, bottom_category_id, bottom_category_text, top_category_id, top_category_text, color_id, color_text. The below table shows more information about the data which was provided.

Columns	Count	DataType
product_id	245485	int64
title	244545	object
description	244545	object
tags	210575	object
type	244211	object
room	8727	object
craft_type	32520	object
recipient	13753	object
material	20876	object
occasion	53229	object
holiday	41019	object
art_subject	2773	object
style	17032	object
shape	2358	object
pattern	10678	object
bottom_category_id	245485	int64
bottom_category_text	245485	object
top_category_id	245485	int64
top_category_text	245485	object
color_id	245485	int64
color_text.	245485	object

IV. Methodology

Data cleaning:

The data which was received needed some necessary cleaning to be done for better computation.

- The columns such as room, craft_type, recipient, material, occasion, holiday, art_subject, style, shape, and pattern were removed since they were having more than 50% missing values. Also, they were not adding any important value to the analysis.
- The column 'type' was also dropped since it was not giving too much insight into the analysis of the data.
- The remaining columns like tags, titles and descriptions were used for further analysis.

For the cleaning of those columns, there was the use of NLP.

- Special characters, digits, URLs, extra spaces etc were removed from the data which helped in focusing on the words only. Standardisation of the data was achieved by doing this.
- The stopwords such as 'the', 'an', 'is', 'a' etc were removed for focusing on the significant words in the data.
- Technique such as lemmatization was used to get the base word of the words in the text. The reason for using lemmatization was to normalize the text columns which helps in improving the performance and reducing vocabulary size for easy computation of the data. Lemmatization positively affects the model by increasing its accuracy.

All the above-stated cleaning steps were done on the columns' title, description, and tags which were found more proficient for the classifications of the ids.

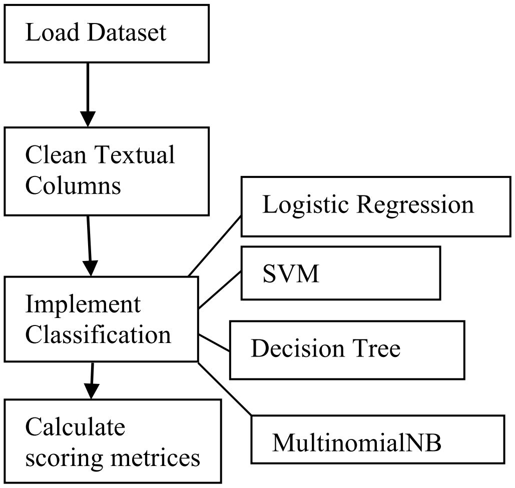
FEATURE ENGINEERING:

After cleaning columns, they were combined and a new column was created named 'combined' which had all the cleaned text and the title, description and tags separated by spaces in it.

This column contained all cleaned text. A CSV of cleaned data was created for backup.

MODEL BUILDING:

The below flow chart shows the exact flow of the work.



The flow chart shows four models used to find which one is the best algorithm for the classification of the ids. Logistic regression is used as the base model for our classification.

For better performance of the models, batching of data was done. It means the large dataset was divided into smaller chunks of data for processing. This helped for efficient processing of the data. This also helped to reduce memory usage and improve processing speed. Before fitting the data into the model, the textual data was first vectorized using CountVectorizer to convert the texts into the matrix of tokens. It was used to convert the textual data into a format that is more suitable for analysis and machine learning. The reason behind choosing CountVectorizer over TFIDF was that it is simpler than TFIDF and it preserves the order of the words which helps in classification.

After the vectorization of the data, the data were fitted into the specified models for classification. Every model stated was run for the top_category_id, bottom_category_id, and color_id. The accuracy, f1 score and recall were seen better on the classification of the top_category_id compared to the other two features. The reason that top_category_id was efficiently classified was that there were only 13 classes in that column which was very less compared to the other two. The color_id had the lowest score is the color in the combined text and the actual color in the color column was different in many cases which led to less score. The bottom_category_id was having better scores than the color_id but was not as good as the top_category_id scores, reason for that is that there are more than 100 classes in the bottom category to classify which leads to less score. The table below will give more specific information about the models used for classifying top_category_id.

Model Name	**Accuracy**	**F1 Score**	**Recall**	
Logistic Regression	0.99	0.99	0.99	Train
	0.89	0.89	0.89	Validation
MultiNomialNB	0.88	0.87	0.87	Train
	0.78	0.79	0.79	Validation
SVM	0.93	0.93	0.93	Train
	0.83	0.83	0.83	Validation
Decision Tree	0.99	0.99	0.99	Train
	0.74	0.74	0.74	Validation

Table 2

Similar results were found for bottom_category_id and color_id.

As we can see above there is a gap of approximately 0.10 between the training and the validation dataset, this is because the data which was used in the model was hard to sample since there was a little imbalance in it. For resolving this sampling of data is the solution which we can use to reduce the gap between the validation and training. The Random Forest model was also used on the same data, but due to its requirement for heavy computation and imbalance dataset, it took too much time to run and gave accuracy and other scores of 0. Random Forest's less interpretability was also the reason for getting 0 in the scores, so it was dropped from the models.

As far as research is concerned the baseline model was found to be more precise than the other models which were used.

V. FUTURE WORK

Currently, there are three models for predicting three different features, in future, we can optimise the data and create a single model to predict all

three features which are required. Sample the data in a proper way which will help for better accuracy and lower the gaps between train and validation.
If the cleaning and sampling of data are done more precisely, then training and validation gaps may reduce.

VI. CONCLUSION

In this research, we found that the base model Logistic Regression Model was the best for the classification of the data. The reason for this is that it is simpler and easy to compute compared to other algorithms. From this, we can say that even the base model can give better results than the complex models if the given data is used in the right way. This also gives the importance of cleaning the data for analysis of the data. We achieved better results for top_category_id prediction. The lowest scores were for the color_id reason being the color given in the title, description and tags were not matching the color which were put. This may occur due to human error or some other problem. Overall for all the features Logistic Regression have given better results than other models.

References

[1] Shah, K., Patel, H., Sanghvi, D. et al. A Comparative Analysis of Logistic Regression, Random Forest and KNN Models for the Text Classification. Augment Hum Res 5, 12 (2020). https://doi.org/10.1007/s41133-020-00032-0

[2] Xiaoyu Luo,Efficient English text classification using selected Machine Learning Techniques, Alexandria Engineering Journal, Volume 60, Issue 3, 2021, Pages 3401-3409, ISSN 1110-0168,https://doi.org 10.1016/j.aej.2021.02.009.

[3] Kadhim, A.I. Survey on supervised machine learning techniques for automatic text classification. Artif Intell Rev 52, 273–292 (2019). https://doi.org/10.1007/s10462-018-09677-1

[4] https://www.etsy.com/ie

[5] Qian Li, Hao Peng, Jianxin Li, Congying Xia, Renyu Yang, Lichao Sun, Philip S. Yu, and Lifang He. 2022. A Survey on Text Classification: From Traditional to Deep Learning. ACM Trans. Intell. Syst. Technol. 13, 2, Article 31 (April 2022), 41 pages. https://doi.org/10.1145/3495162

[6] Haihua Chen, Lei Wu, Jiangping Chen, Wei Lu, Junhua Ding, A comparative study of automated legal text classification using random forests and deep learning, Information Processing & Management, Volume 59, Issue 2, 2022, 102798, ISSN 0306-4573, https://doi.org/10.1016/j.ipm.2021.102798.