Deep Learning Framework for Multi-Output Product Categorization

Abstract—**This paper presents a multifaceted machine learning project aiming to predict four categorical variables using text data. Utilizing a rich dataset from an e-commerce platform, we employ various models ranging from Naïve Bayes to advanced neural networks, exploring their performance in multi-output classification tasks.**

I. Introduction

In the dynamic realm of e-commerce, product categorization plays a pivotal role in shaping the consumer shopping experience and enhancing operational efficiency. It enables customers to easily locate desired products amidst the vast array of items available online. Accurate categorization improves search functionality and recommendation systems, directly impacting customer satisfaction and retention. It also plays a crucial role in influencing the discoverability of products, as well-organized categories are essential for effective search engine optimization (SEO), which attracts organic traffic to the site.

However, with the ever-growing volume of products entering the digital market daily, manual categorization becomes impractical and prone to inconsistencies. This challenge has spurred the development of automated categorization systems employing various machine learning techniques which not only enhances efficiency and scalability but also offer the adaptability to handle the dynamic and ever-expanding inventory characteristic of e-commerce platforms.

In this study, we design a machine learning pipeline for Etsy; the global marketplace for unique and creative goods, to predict product categories including top and bottom categories as well as primary and secondary colors. An LSTM model was incorporated in a MultiOutput framework, with text data using Pre-trained GloVe embeddings. PCA visualization is used to reveal the content features' reduction and connection to product categories.

II. Literature Review

In e-commerce, product categorization and color identification is a significant factor that improves the user experience. More recently, machine learning and natural language processing (NLP) has been utilized to automate these processes.

Text Classification and Feature Extraction: Textual data analysis constitutes the foundation of product categorization, and it is mainly based on product titles, descriptions, tags, etc. Conventional machine learning models such as Naive Bayes, Support Vector Machines (SVM), and Random Forests have been extensively used for text classification. To some extent, the neural network, a type of neural network that has been applied in text data recently, can capture extremely complex patterns but at the same time, it requires large datasets and a lot of computational resources (Zhang et al., 2015). Feature extraction techniques as TF-IDF (Term Frequency-Inverse Document Frequency) are popular approaches aimed at converting text into a numerical format for machine learning algorithms to process. TF-IDF measures the density of each word in the document against its representativeness across all documents to produce a balanced weight measure (Ramos et al., 2003).

Multi-Output Classification: In case of top and bottom category IDs (e.g. multiple labels per instance) then multi-output classification approaches are usable. Random Forests, which are an ensemble learning method, are preferred in many cases because of their robustness and ability to deal with multiple outputs in a simultaneous manner. They do such by building many decision trees during training and outputting the majority class in a classification task (Breiman, 2001).

Dimensionality Reduction: High-dimensional data, which is one of the essential features of text data applications, is transformed into a lower-dimensional space through the utilization of the PCA (Principal Component Analysis) method. PCA decreases the dimension by identifying the principal components that explain the biggest variance in the dataset. Furthermore, this not only helps reduce the computational load, but it also helps in visualizing data patterns more effectively (Jolliffe, 2002).

Embeddings and Pre-trained Models: The progress of computing resources has received widespread studies on the capability of pre-trained models such as Word2Vec, GloVe, and lately BERT (Bidirectional Encoder Representations from Transformers) to capture the meanings of text. These models can be tuned on specific datasets, so that deeper embeddings that could capture linguistic and contextual nuances better (Devlin et al., 2018) can learn.

Comparative Studies: The comparisons may assess the efficiency or effectiveness of distinct models in different configurations. For instance, experiments comparing pre-trained embeddings with task-specific feature extraction could show the relationship between generalization and task specific-tuning. Furthermore, visualizing embeddings or reduced feature space can illustrate how well the model clusters similar items, it can give empirical grounds to believe that the model is capable to create meaningful presentation.

On this basis of research, an advanced machine learning models that aims at improving both backend classification efficiency and frontend user satisfaction may be deployed in the ecommerce platforms such as Etsy.

III. Dataset Description and exploration

The dataset used is a proprietary dataset provided by Etsy; a global marketplace for unique and creative goods. Etsy's online marketplace collectively connects approximately 100 million passionate and creative buyers and 7.7 million sellers around the world.

The dataset provided contains 229,624 unique products and 26 features with features including product title, description, and image encoded in bytes. The corresponding product category is included in two features; a top category and bottom category. There are 15 possible top categories a product can belong to and 2,609 bottom categories which contain hierarchical information about the products top and sub-categories. In addition, there are labels for the color of a product contained in 2 features (primary and secondary color).

The dataset is extremely imbalanced with the maximum number of products in a category being 54,600 and the minimum 5,744. The bottom category, primary color, and secondary color features are also imbalanced. The distributions of the length of text contained in the title reveal that most titles are relatively short, with a significant number falling below 100 characters. There is a sharp peak, indicating that there is a common title length in the dataset and a long tail stretching towards 300 characters but very few titles reach this length. The width of the images is highly concentrated around a specific value (about 200 pixels). However, the image height distribution is more varied than the width distribution with noticeable secondary peaks and a prominent peak around 400 pixels.

IV. Experimental Results and Analysis

A. Baseline Model

For Text classification, a Naive Bayes classifier is used as a benchmark to compare the performance of other models. The title feature was vectorized using TF-IDF and served as input to the model. Results were evaluated using the F1 score (micro-averaged), which is an accepted metric for multi-label classification and imbalanced datasets. The baseline model achieved F1-scores of 64%, 51%, 24% and 15% in predicting the top category, bottom category, primary color, and secondary color respectively.

B. ***Dimensionality Reduction and Visualization***

For visualization purposes, PCA was implemented to the TF-IDF features reducing the data to two principal components. This reduction approach enabled us to plot the data points with colors representing different top category IDs and displaying clusters between similar products (Uddin et al., 2021). The PCA plot revealed that the TF-IDF features captured meaningful patterns that are related to the product top categories, as the products clustered clearly in the reduced space.

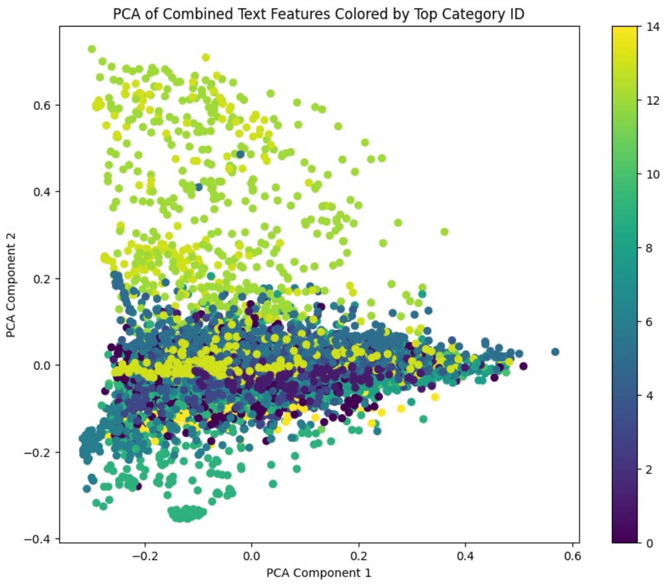

Fig. 1. PCA Plot

C. Data Preprocessing

The title and description was pre-processed to remove stopwords, HTML entities, URLs, punctuations and digits. Other natural language processing techniques including tokenization and lemmatization were also employed to prepare the data for further analysis.

To capture the semantic relationships within the text data, we employed pre-trained GloVe embeddings, which transform text into dense vectors. Each product's title was converted into a vectorized format, allowing us to perform more complex numerical analysis. Words in the corpus not covered by GloVe were initialized to vectors containing all zeros.

D. Model Training:

To categorize products, we experimented with various machine learning models, including Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs). Pretrained GloVe embeddings as well as training our own embeddings were used with our trained embeddings providing better performance in all cases. The LSTM network also outperformed the CNN network.

E. Model Evaluation:

Using a split of training, validation, and test sets, we rigorously evaluated our models' performance using accuracy, loss and F1 scores. The LSTM model was further optimized by applying hyperparameter tuning techniques using libraries like Keras Tuner. This step was crucial in refining the model to achieve the best possible performance. We used a multi-output classifier framework to predict multiple attributes simultaneously.

The tuning process resulted in a model giving best performance that incorporates an input layer designed to handle sequences of 200 elements (corresponding to the title length as most titles are below 200 in length), each with 100 features, processed through an embedding layer that converts these features into a dense 100-dimensional space. The network features three LSTM layers, each followed by a dropout layer to help in mitigating overfitting by randomly disabling neurons during training. The LSTM layers, with 320, 256, and 192 units respectively, are engineered to capture both short-term and long-term dependencies within the data. The architecture concludes with four dense layers, representing the four outputs to be predicted. The model contains a total of 9,267,626 parameters, with 1,988,326 trainable, suggesting a substantial capacity to learn from large and complex datasets.

For computational efficiency, the model was re-trained for 10 epochs using all available data (training and validation sets combined) and evaluated on the test set. The resulting F1 scores for the 4 labels were 83%, 47%, 46%, and 27% respectively.

V. Conclusion

The methodology employed in this study addresses the complexities of product categorization in a large-scale commercial dataset. Through the use of advanced NLP techniques and neural network architectures, we were able to extract meaningful insights from the data. The models demonstrated reasonable accuracy, with the LSTM showing particular promise in handling the multi-faceted nature of textual data.

However, the challenges of handling diverse and sometimes sparse data were evident, particularly in the variability of model performance across different product categories. Future work should focus on enhancing the robustness of the models through more sophisticated natural language processing techniques and exploring deeper neural network architectures. Additionally, incorporating more contextual and image data into the models could potentially improve their predictive power and offer more granular insights into product categorization.

This research sets a foundation for further studies that could incorporate real-time data analysis for dynamic category prediction in e-commerce platforms.

VI. references

Artama, M., Sukajaya, I. N., & Indrawan, G. (2020, April). Classification of official letters using TF-IDF method. In *Journal of Physics: Conference Series* (Vol. 1516, No. 1, p. 012001). IOP Publishing.

Breiman, L. (2001). Random Forests. Machine Learning, 45(1), 5-32.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv preprint arXiv:1810.04805.

Jolliffe, I. T. (2002). Principal Component Analysis. Springer Series in Statistics.

Le, Q. V., & Mikolov, T. (2014). Distributed Representations of Sentences and Documents. In International Conference on Machine Learning (pp. 1188-1196).

Nakano, F. K., Pliakos, K., & Vens, C. (2022). Deep tree-ensembles for multi-output prediction. *Pattern Recognition*, *121*, 108211.

Ramos, J. (2003). Using TF-IDF to Determine Word Relevance in Document Queries. In Proceedings of the First Instructional Conference on Machine Learning.

Uddin, M. P., Mamun, M. A., & Hossain, M. A. (2021). PCA-based feature reduction for hyperspectral remote sensing image classification. *IETE Technical Review*, *38*(4), 377-396.

Zhang, X., Zhao, J., & LeCun, Y. (2015). Character-level convolutional networks for text classification. In Advances in neural information processing systems (pp. 649-657).