NLP Classification task for Shopping Products

Faisal Rasheed Khan vb02734@umbc.edu
Maitri Mistry mmistry3@umbc.edu
Tarun Varma tarunv2@umbc.edu
Rohith Reddy rohithm4@umbc.edu

Abstract

With the growth of online shopping since the pandemic, it is essential to keep momentum going in the e-commerce industry. One of the main reasons people prefer e-commerce platforms for market research and purchase is the ease of the process and variety of options that are available online. Product search and browsing is a crucial component that contributes to making shopping easier and consequently improving sales. The categorization of products must be done properly to assist in product search. For effective organization of products on e-commerce websites, each product is usually assigned a category. A product is typically represented by several features like title, description, and image and is identified by a category label. In this work, we will focus on the prediction of product categories based on their titles and textual descriptions. Our main focus will be on the data processing and feature extraction process. We shall explore different kinds of feature extraction methods such as TF-IDF, and word embeddings to extract semantic information that will guide the classification task. Using these features, we plan to evaluate different classification models. Various combinations of features and classification models will be explored and compared to find a good product categorizing model.

1 Description and Motivation

In this project, the problem we will be handling is the multi-class classification for an E-commerce text dataset. This dataset contains 12k rows wherein each row has category labels and descriptions. Overall, there are 4 different product type labels. We will investigate different embeddings, feature extraction methods and their influence on the outcome and accuracy of classification models. We are also interested in finding ways to combine them for better performance.

Over recent years, we have witnessed a surge in online shopping. On such E-commerce platforms, we have noticed that sometimes products may not be shown to the user due to wrongful categorization. This could cause issues for both sellers and buyers as well. This motivated us to work on developing an efficient and accurate product categorizing model using advanced methods in natural language processing.

2 Proposed solution

In order to train the data on the classifier, we will convert the text into useful features using various word embeddings - where the embeddings capture the appropriate relationships required. We plan on working with embeddings given by BERT, Word2Vec. We also intend to use methods like TF-IDF, and Bag-of-Words. With the help of these embeddings, we will categorize products using different models such as SVM, Random Forest, and Logistic Regression. We will analyze the different approaches of the embeddings and the different models specified above.

3 How our solution fits into previous work

While prior research has addressed classification tasks in e-commerce, there is noticeably less focus on fully exploring the interplay between various embeddings and models. Our work aims to contribute to that aspect by delving deeper into the selection and effectiveness of embeddings, specifically investigating BERT, Word2Vec, and other methods. Feature engineering, specifically embeddings, is a primary focus area since the textual information from the product description in the E-commerce dataset is pivotal for product classification. Having read through multiple publications on this topic, we developed a methodology employing best practices in the field of Natural language Processing and Machine learning for e-commerce product classification. This focus on understanding the synergy between embeddings and models could offer valuable insights for future developers

and researchers working on e-commerce categorization tasks. Essentially, our proposed solution builds upon the foundations established by prior work in e-commerce product categorization.

4 Experimentation

Our paper aims to offer a comprehensive understanding of the intricate relationship between different embeddings and models for e-commerce product categorization. Based on the detailed literature review conducted, we selected embedding techniques and models for our study. Our approach incorporates a variety of feature engineering methods, including contextual and word embeddings such as BERT and Word2Vec, an importance-based statistical method like TF-IDF, and a frequency-based method like CountVectorizer. The generated features are then used as input for classification models like Random Forest, Support Vector Classifier, and Logistic Regression. We utilized metrics such as accuracy and F1 scores to assess and compare the performance of these embeddings and models. Our code implementation uses several libraries, including transformers, gensim, nltk, sklearn, cupy, cuml, pandas, and numpy.

5 Methodology

The dataset comprises over 12,000 rows, each containing product descriptions of an e-commerce product and corresponding categories. These labels fall into four categories: books, household, electronics, and clothing & accessories. Despite not being perfectly balanced, the dataset exhibits no severe imbalance, with household being the most common label (38%) and clothing & accessories the least (17%). Data cleaning is applied to address indexing issues. Any row with null values in the product description column or its associated label is dropped. The models are trained on 80% of the dataset and rest 20% is used to test.Accuracy and F1 scores are considered the primary metrics for evaluation.

BERT and Word2Vec embeddings are generated from the product descriptions, and the resulting vectors are appended to the data frame. TF-IDF and Countvectorizer cannot be directly added to the data frame due to the need for separate transformations on the training and test splits. For Word2Vec, where each row contains a list of arrays, computational constraints necessitate calculating the mean of each array instead of flattening them.

Applying TF-IDF and Countvectorizer directly to the DataFrame leads to a substantial increase in dimensionality, resulting in over 40,000 columns. To address this issue and manage the high-dimensional feature space more efficiently, we employ incremental PCA (Principal Component Analysis) on a GPU.

Three machine learning models, namely Random Forest Classifier, SVM with a linear kernel, and Logistic Regression, are selected for training. The training is conducted using the GPU-accelerated methods provided by the CUML library. Default hyperparameters are utilized for these models, with the maximum number of iterations set to 500 for logistic regression.

Using trained models, accuracy and F1 scores for all model-embedding pairs are systematically stored in lists. These metrics serve as the basis for evaluating and comparing model performance. The results are visualized through graphs, offering a good understanding of how different embeddings impact the results of each model.

6 Analysis of result

The BERT embeddings perform well across all models, with Logistic Regression achieving the best accuracy and F1 score (93.1% accuracy and 93.1% F1). This shows that the extensive contextual information captured by BERT embeddings considerably contributes to the classification task's effectiveness. Notably, BERT embeddings with Random Forest and SVM models show excellent accuracy and F1 scores, demonstrating BERT's adaptability across model architectures.

Word2Vec embeddings present a lower performance compared to BERT. Among the Word2Vec embeddings, the Random Forest model has the highest accuracy (60.5%) and F1 score (57.6%), highlighting the importance of model selection. The limitations in processing Word2Vec embedding arrays to useful representations might have impacted the model's ability to capture complex relationships within the data.

TF-IDF embeddings consistently perform well across all three models, with SVM achieving the highest accuracy (94.1%) and F1 score (94.1%). The ability of TF-IDF to capture the importance of terms in the corpus appears to greatly contribute to model effectiveness. The Logistic Regression model also performs well, demonstrating the ability of TF-IDF embeddings in this classification prob-

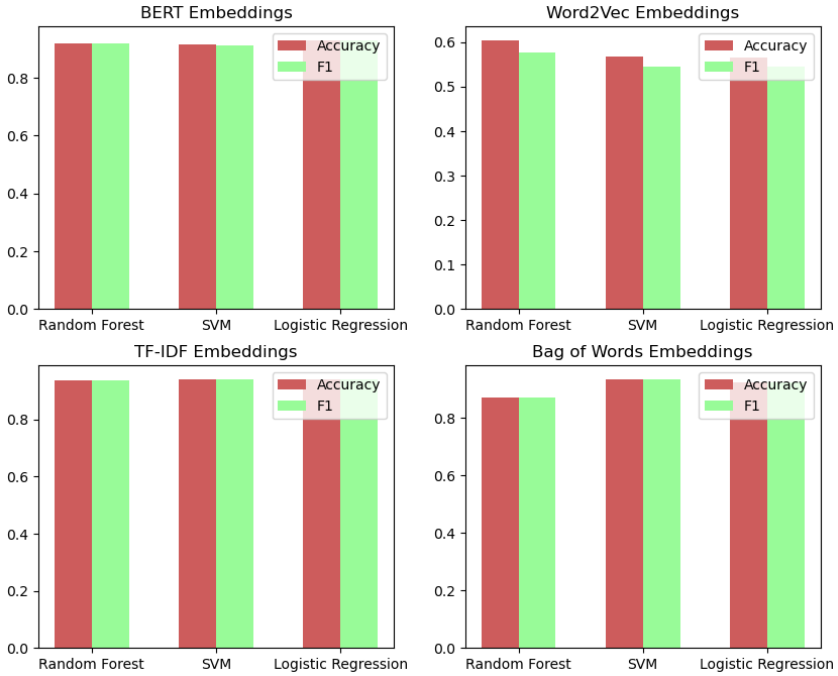

Figure 1: Accuracy and F1 scores for Different Embeddings-Model pairs

lem across different architectures.

Bag of Words embeddings show comparable results, with SVM surpassing other models in both accuracy (93.7%) and F1 score (93.7%). The simple nature of Bag of Words representations, which capture word frequencies without regard for context, appears to correspond well with the properties of the SVM model.

The embedding-model pair choices have a considerable impact on classification results. BERT embeddings consistently outperform Word2Vec embeddings in the current e-commerce data, emphasizing the value of using contextual information. The robustness of TF-IDF embeddings across multiple models emphasizes their efficiency in capturing word significance. While competitive, Bag of Words embeddings lag behind BERT and TF-IDF embeddings.

Logistic Regression consistently performs well across all embeddings, indicating its adaptability and effectiveness in leveraging different types of semantic information. SVM, especially with TF-IDF and Bag of Words embeddings, demonstrates strong discriminatory power, aligning with its suitability for text classification tasks. While effective, Random Forest performs significantly worse than SVM and Logistic Regression, presumably because of its ensemble nature.

7 Limitations

Bert embeddings for real-time inferences: Because of the computational intensity, the use of BERT embeddings in our approach is a major limitation. Creating BERT embeddings for user text inputs during inference requires a considerable amount of computational power. This presents a challenge in situations where real-time responses are critical, particularly in e-commerce platforms with frequently changing product listings, potentially hindering the app's capacity to deliver timely predictions. Furthermore, when the volume of data or the number of concurrent users grows, the model should be scaled up. Scaling the system to manage the increasing load becomes more difficult, necessitating a larger infrastructure. As a result, while BERT embeddings provide rich language representation, their computing requirements pose practical limits in real-time and scalable deployment situations, particularly in dynamic e-commerce contexts.

Limitations of taking only text as input: The current approach prioritizes text-based embeddings, while not considering information contained in product images. By relying just on textual input, the model might miss visual cues and details that are sometimes critical in efficiently recognizing and categorizing objects. There is a clear scope to integrate image-based features to remove this constraint and increase the model's understanding

Embeddings	Random Forest	SVM	Logistic Regression
BERT	0.9191	0.9139	0.9310
Word2Vec	0.6049	0.5680	0.5660
TF-IDF	0.9357	0.9413	0.9393
Bag of Words	0.8707	0.9365	0.9254

Table 1: Accuracy for Different Embeddings and Models

Embedding	Random Forest	SVM	Logistic Regression
BERT	0.9189	0.9139	0.9309
Word2Vec	0.5761	0.5446	0.5442
TF-IDF	0.9357	0.9412	0.9392
Bag of Words	0.8712	0.9365	0.9254

Table 2: F1 Score for Different Embeddings and Models

of products. Incorporating visual data, alongside textual data, may improve the model's overall performance and robustness in e-commerce categorization tasks.

Complexity in handling Word2Vec embeddings: The outputs of Word2Vec embeddings are lists of arrays of shape 100. Unfortunately, there was no method to efficiently process this array to a trainable format with a better computational complexity than O(n_2). This becomes problematic with large datasets, rendering the process computationally intensive and impractical even when utilizing GPU resources. As a solution, we chose to compute the mean of each array, resulting in a more comprehensible representation that could be used effectively for training.

Handling change in output labels: The models become outdated and less accurate as product categories grow, or get reorganized in the e-commerce platform. It results in failure to categorize things into newly created or changed categories, leading to a drop in performance. To deal with this, the model has to be re-trained on updated datasets on a regular basis. Methods like incremental learning, or online learning could help us overcome this limitation.

Lack of transparency: A limitation of the models used in the project is their interpretability and lack of transparency, especially in models like BERT and Random Forest. Both theories make understanding the underlying decision-making processes difficult. These models' complex connections may be difficult to understand and figure out why various products are classified into specific groups. This constraint reduces the transparency of the classification process, which may have an influence on user trust and makes it difficult to provide explanations for model predictions in real-world applications.

8 Future Work

While evaluating the models, we sensed some short-term and long-term improvements required to help improve the model. One of the suggested improvements is fine-tuning, where we tune the pre-trained models using the dataset to improve the performance for our specific needs. We could also perform hyperparameter tuning to improve the model and its working. This was not something we were able to do with our current computational budget. Another possible improvement is using ensemble methods, where we could use the predictions from multiple models and combine those to improve the performance compared to just an individual model. For long-term improvements, we need to consider continuous improvements by maintaining the model and training it with the latest data with a larger variety of labels. The model should be able to perform well outside of the domain of E-Commerce and in a wider field. We could also utilize a feedback model to improve accuracy. Possible scaling of the model to include different category labels would require training again with new data, including new labels. Multimodal Learning could be incorporated into the model to handle modalities such as images using computer vision techniques to further enhance the model. The short-term improvements are focused on refining the existing models and techniques, while the long-term involves continuous learning, scalability with more labels, and multimodal learning.

9 Acknowledgement

The authors would like to acknowledge the use of the E-commerce Text Dataset (Version - 2) (Gautam, 2019) in this project.

References

Ali Cevahir and Koji Murakami. 2016. Large-scale multi-class and hierarchical product categorization for an e-commerce giant. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 525–535.

Lei Chen and Hirokazu Miyake. 2021. Label-guided learning for item categorization in e-commerce. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Papers*, pages 296–303.

Ying Ding, M. Korotkiy, B. Omelayenko, V. Kartseva, V. Zykov, Michel Klein, and E. Schulten. 2002. Goldenbullet: Automated classification of product data in e-commerce.

Python Software Foundation. 2023. Python documentation. `https://docs.python.org/3/`. Accessed: December 20, 2023.

Gautam. 2019. E-commerce text dataset (version - 2).

Vivek Gupta, Harish Karnick, Ashendra Bansal, and Pradhuman Jhala. 2016. Product classification in E-commerce using distributional semantics. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 536–546, Osaka, Japan. The COLING 2016 Organizing Committee.

Amit Mandelbaum and Adi Shalev. 2016. Word embeddings and their use in sentence classification tasks. *arXiv preprint arXiv:1610.08229*.

Wenhu Yu, Zhiqiang Sun, Haifeng Liu, Zhipeng Li, and Zhitong Zheng. 2018. Multi-level deep learning based e-commerce product categorization. In *The SIGIR 2018 Workshop On eCommerce co-located with the 41st International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2018), Ann Arbor, Michigan, USA, July 12, 2018*, volume 2319 of *CEUR Workshop Proceedings*. CEUR-WS.org.