|  |
| --- |
| **NLP Classification task for Shopping Products** |
|  |
| **Faisal Rasheed Khan** **Maitri Mistry Tarun Varma Rohith Reddy** |
| vb02734@umbc.edu [mmistry3@umbc.edu](mailto:mmistry3@umbc.edu) [tarunv2@umbc.edu](mailto:tarunv2@umbc.edu) 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.

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 and their influence on the outcome and accuracy of classification models. We are also interested in finding ways to combine the embeddings 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.

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, TF-IDF, and also performed using 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.

How the proposed solution fits with previous research done

While we did go through many papers that attempted to perform classification on E-commerce data, none of them went deeper into their reasoning for choosing a specific embedding. As we believe classification of products in an E-commerce domain is a highly paramount task in modern society, we wish to have a clear understanding of exactly which embeddings work best with which models and why. We believe this project would give future developers and researchers further knowledge which they can leverage if they were to build a model that pertained to E-commerce.

How solution fits previous problem

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.

project involves 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.

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 involved addressing indexing issues. And any row with null values in product description or its associated label, are dropped . Accuracy and F1 scores are considered as 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.

Description and analysis of results

The evaluation of various embedding-model combinations in the classification task reveals nuanced insights into their respective performances. BERT embeddings, when integrated with the Random Forest model, exhibit a commendable accuracy of 89.11% and an F1 score of 88.88%. This highlights the robust feature representation capabilities of BERT in conjunction with the ensemble-based classification.

Similarly, BERT embeddings paired with the SVM model demonstrate even stronger performance, achieving an accuracy of 93.56% and an F1 score of 93.52%. This outcome underscores the adaptability of BERT to capture intricate relationships within the data, making it particularly effective for support vector machine classification.

Moreover, the synergy between BERT embeddings and the Logistic Regression model results in the highest accuracy and F1 score among the evaluated combinations, reaching 94.55% and 94.52%, respectively. The linear nature of Logistic Regression seems to align well with the information captured by BERT embeddings, leading to superior classification outcomes.

Conversely, Word2Vec embeddings present challenges when paired with both Random Forest and SVM models, yielding lower accuracy and F1 scores. This suggests limitations in capturing semantic information and nuanced relationships within the data compared to BERT embeddings.

On the other hand, TF-IDF embeddings exhibit competitive performance, particularly when combined with the SVM model, achieving an accuracy of 92.57% and an F1 score of 92.52%. This indicates the efficacy of the traditional TF-IDF weighting scheme in capturing important features for SVM-based classification.

In a broader context, BERT consistently outperforms Word2Vec and TF-IDF embeddings across all models, highlighting its effectiveness in capturing contextual information. The results also emphasize the compatibility of BERT with various classification models, showcasing its versatility in different contexts.

While BERT proves dominant in terms of accuracy and F1 score, challenges lie in the interpretability of its predictions. Future work could focus on enhancing model explainability to better understand the reasoning behind BERT's classifications.

Moreover, considerations for data augmentation and hyperparameter tuning could contribute to addressing potential biases and optimizing model performance. Exploring additional embeddings or pre-trained models may also offer insights into the impact of embedding diversity on classification outcomes, paving the way for more comprehensive and robust models in the realm of NLP classification tasks.

Analysis of limitations of work

1. 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 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.

1. 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 of products. Incorporating visual data, alongside textual data, may improve the model's overall performance and robustness in e-commerce categorization tasks.
2. 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.
3. 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.
4. 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.

Potential follow-up work for short and long-term improvements

We have trained and evaluated our model on the shopping dataset to classify the labels. It is a multi classification task. We have used different kinds of embeddings such as Bert, Word2Vec, TF-IDF and also used bag-of-words on the dataset. The potential improvements for the models are always there. While evaluating the models, we sensed some improvements need to be done to help the model train on the shopping dataset both for short term and long term works. The model we trained is on a dataset which has a label and lengthy product description. Actually training on a lengthy description of a label is good, because we have a variety of words to be seen and trained, but it is a challenging task to train as it requires a good amount of time to run. Based on the word embeddings used we train models using Random Forest, Support Vector Machines and logistic regression, these models work good but can be improved further. One of the improvements which can be done is fine-tuning, where we tune the pretrained model using the dataset for good performance. Fine-tuning using pretrained models actually helps in improving the performance for our specific needs. We could also perform hyperparameter tuning to improve the model and work fine for a particular type of dataset. This was not something we were able to do with our current computational budget.

Another method of improvement is using ensemble methods where we could use the predictions from many models and combine those to improve the performance compared to just one single individual model. For the long term improvements, we need to consider the 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 context of just E-Commerce, and into a wider field. We could also have a feedback model to improve accuracy. Large scale up of the model to include different category labels and that requires training again with new data with the new labels.This is scaling the model to add more categories. Multimodal Learning could be incorporated into the model which handles modalities such as images using computer vision techniques to further improve the model both in terms of accuracy and scalability of not just handling the text data but also with the images. The short-term improvements are focussed on refining the existing models and techniques, while the long-term involves continuous learning, scalability with more labels and multimodel learnings.

References

Lei Chen and Hirokazu Miyake. 2021. [Label-Guided Learning for Item Categorization in e-Commerce](https://aclanthology.org/2021.naacl-industry.37). 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, Online. Association for Computational Linguistics.

Ali Cevahir and Koji Murakami. 2016. [Large-scale Multi-class and Hierarchical Product Categorization for an E-commerce Giant](https://aclanthology.org/C16-1051). In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 525–535, Osaka, Japan. The COLING 2016 Organizing Committee.

Gautam. (2019). E commerce text dataset (version - 2) [Data set]. Zenodo. <https://doi.org/10.5281/zenodo.3355823>

Ding, Ying & Korotkiy, M. & Omelayenko, B. & Kartseva, V. & Zykov, V. & Klein, Michel & Schulten, E.. (2002). GoldenBullet: Automated Classification of Product Data in E-commerce.

Vivek Gupta, Harish Karnick, Ashendra Bansal, and Pradhuman Jhala. 2016. [Product Classification in E-Commerce using Distributional Semantics](https://aclanthology.org/C16-1052). In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 536–546, Osaka, Japan. The COLING 2016 Organizing Committee.

Wenhu Yu, Zhiqiang Sun, Haifeng Liu, Zhipeng Li, Zhitong Zheng: Multi-level Deep Learning based e-Commerce Product Categorization. eCOM@SIGIR 2018

[ecom18DC\_paper\_2.pdf (ceur-ws.org)](https://ceur-ws.org/Vol-2319/ecom18DC_paper_2.pdf)