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 the 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 the 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 the 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 50k+ 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.

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, GLUE, Glove, Word2Vec, TF-IDF, and our custom-defined embeddings. With the help of these embeddings, we will categorize products using different models such as BERT, SVM, and Random Forest. We will analyze the different approaches of the embeddings and the different models specified above.

3 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.

4 Blocks experienced or anticipate experiencing.

1. The data set contains product title and description, using both of them as input will be challenging as the product descriptions are lengthy (5-10 sentences per product).
2. We are uncertain of the baseline model to be used.
3. We are unsure if the size of the chosen dataset is sufficient to train, validate and test the model.
4. Defining the embeddings and using these kinds of embeddings to get the appropriate results will be a challenge.

References

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, Online. Association for Computational Linguistics.

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, 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