Production Classification in E-Commerce Based on Product Descriptions with Natural Language Processing (NLP) and Machine Learning Models

Yuvraj Bist, Paramesh Gurbaxani, and Neetu Gupta(✉)

Department of Computer Science and Engineering, Manipal University Jaipur, Jaipur, Rajasthan, India
{yuvraj.219301002,paramesh.219301270}@muj.manipal.edu, neetu.gupta@jaipur.manipal.edu

Abstract. The rapid growth of the internet has increased people's reliance on it for expressing opinions on products and stores. Text sentiment analysis is now a key research area. Deep learning methods are commonly used for text classification, but they suffer from issues like information loss and weak context. This paper enhances the existing models to simplify the process, educe training time, and improve overall recall and accuracy in text sentiment classification. With the rapid development of artificial intelligence, traditional manual techniques are fading, and AI algorithms are driving the swift progress of text sentiment classification. Product categorization is critical in e-commerce since it influences search, recommendations, inventory, and consumer experience. NLP approaches and models such as TF-IDF and Word2Vec are effective for categorizing products based on their textual descriptions. This literature study investigates the research evolution, problems, and upcoming trends in using TF-IDF and Word2Vec for product classification in e-commerce.

Keywords: NLP · TF-IDF · Word2Vec · e-commerce · AI algorithm

1 Introduction

The way people find and buy products has changed drastically with the rise of online shopping. Ordering products based on their text-based portrayals is essential for improving customer experience, improving hunting functions, and stream lining inventory management. NLP methods, closely related to AI models such as TF-IDs and Word2Vs, have become essential for dealing with the complexities of item groupings.

The versatility of TF-IDF extends to product classification as well. TF-IDF transforms product descriptions into feature vectors, which capture the importance of the terms with in each report. These feature vectors are used as inputs for different machine learning algorithms, like SVM or Decision Trees, which make it easier to classify products. For example, in the study by [1], they used a TF-IDF vector to organize products in an online shopping platform. This approach worked well not only in English descriptions

P. Pareek et al. (Eds.): IC4S 2024, LNICST 597, pp. 39–46, 2025.
https://doi.org/10.1007/978-3-031-77075-3_4

but also in multiple languages, showing that TF-IDF is flexible enough for global online shopping platforms.

Word2Vec, a neural network-based model, learns vector representations of words based on their contextual co-occurrence patterns within a vast text corpus. It has garnered wide spread adoption across various NLP tasks due to its ability to capture semantic relationships between words.

Word2Vec embeddings facilitate the creation of product description vectors where words with similar meanings are closer in vector space. This semantic understanding is crucial for product classification tasks, as product descriptions often encompass synonyms and related terms. Essential research on Word2Vec's semantic understanding of semantic relationships between words was conducted by [2]. The model's efficiency was demonstrated through vector arithmetic operations such as "king – man + women = queen", which yielded significant results. Such semantic understanding is essential for product classification tasks, which require a nuanced comprehension of textual descriptions. The potential for TF-IDF to be used in conjunction with Word2Vec for product classification has been the subject of research. By combining the strengths of both models, more reliable and accurate classification results can be achieved. [3] proposed a hybrid approach to product classification that merges TF-IDF with Word2Vec. TF-IdF was initially used to generate feature vectors from the product description, which were further refined with Word2Word embeddings. The combination of the two models significantly improved classification accuracy compared to the use of either model alone. The combination of TF-IdF and Word2Word reduces the drawbacks of each model. Whereas TF-IDF effectively captures term frequency data, it may miss semantic nuances. On the other hand, Word2Word is adept at capturing semantic relationships but may need to capture the importance of a term within a document adequately. Therefore, this hybrid approach balances the two models and yields more precise classifications. Supervised machine learning (SML) is one of the most popular product classification methods used in e-commerce. Algorithms such as Support vector machines (SVM), Random forests, and Neural Networks classify products according to their descriptions. [4] used a deep learning method with Convolutional neural networks (CNN) to organize fashion products accurately. Their approach significantly improved the classification accuracy compared to conventional methods.

There are also supervised and semi-supervised approaches to product classification. For example, researchers such as [5] suggested a semi-supervised approach that combined labelled data with large volumes of unlabelled data. This approach is cost-efficient and scalable. This approach has shown promising results in e-commerce product classification. To classify products accurately, it is vital to extract features and represent product descriptions effectively. Natural Language Processing (NLP) tools such as word embeddings, word2vec and BERT remove semantic information from the product description. The Transformer architecture was introduced by [6] which has helped to improve the representation of text data and is useful for product classification tasks. Some e-commerce platforms not only classify text-based products but also classify them based on images. For example, according to researchers such as [7], convolutional neural networks (CNNs) can be used to solve image classification tasks. CNNs can classify

products according to their images, which complements text-based classification methods. Data sparsity is a common issue in product classification. Product catalogues are enormous, and labelled data can be sparse for training models. This has led researchers to explore techniques such as data augmentation or transfer learning. Transfer learning uses pre-trained models on large text corpora to fine-tune them for product classification tasks, as shown in [8]. Multimodal classification is a concept that has been gaining traction in the field of e-commerce platforms. It involves the integration of information from both textual and image-based product classification. This approach was first proposed by [9] and is intended to improve the accuracy of product classification by combining both textual and image information. Several e-commerce companies, including Amazon and Alibaba, have adopted advanced product classification techniques. They use these techniques to improve search relevance and product recommendation systems, improving user experience and conversion rate. For example, Amazon's product recommendation system uses deep learning models to classify and recommend products, as described by [10]. E-commerce in real-life environments comes with its own set of challenges, such as managing large catalogues, keeping up with product releases, and handling user-generated content. Research by [11] addresses the challenge of maintaining accurate classification as product catalogues evolve over time. The product classification landscape in e-commerce is constantly changing. Future research may focus on cutting-edge methods, such as reinforcement learning in dynamic inventory management or the use of third-party knowledge sources to improve product classification accuracy. Ethical aspects of product classification, including bias and fairness, also deserve further exploration.

2 Methodology

This research employs a methodological approach to classify products within the e-commerce domain by leveraging Natural Language Processing (NLP) techniques and Machine Learning models. The process involves text preprocessing, feature extraction using TF-IDF and Word2Vec, model training, evaluation, and deployment.

A. Data Preprocessing Text Normalization: Standardizes the textual data to a uniform representation. This includes converting text to lowercase, removing punctuation, and applying stemming or lemmatization techniques to enhance the uniformity and reduce the vocabulary space.
B. Feature Extraction: TF-IDF Calculation: Utilizes the Term Frequency-Inverse Document Frequency (TF-IDF) metric to quantify the significance of words in product descriptions. TF-IDF accounts for term frequency within a document and across the entire corpus, thereby weighting the importance of words in individual descriptions.

   Word2Vec Embedding: Trains a Word2Vec model on a comprehensive text corpus to generate word embeddings that encapsulate semantic meanings. The embeddings derived are utilized to represent each word within the product descriptions.
C. Feature Representation: TF-IDF Vectors: Constructs feature vectors for product descriptions by concatenating the TF-IDF scores for all words, thereby creating a numerical representation for each description. Word2Vec Vectors: Generates feature vectors for product descriptions by averaging the Word2Vec embeddings for

all words, allowing a condensed representation of the text that captures contextual meanings.

D. Model Training: Utilizes machine learning models such as Support Vector Machines (SVMs), logistic regression, or random forests to train on the generated feature vectors. These models learn to classify product descriptions into their respective categories.

E. Model Evaluation: Evaluates the trained models using a separate test dataset to measure their performance, ensuring that the models generalize well beyond the training data and do not overfit.

F. Model Deployment: Once the model is trained and adequately evaluated, it can be deployed into the e-commerce system to automatically categorize new product descriptions. This research methodology embodies a comprehensive approach to efficiently categorize products in e-commerce by harnessing NLP techniques for text processing and machine learning models for classification. The process ensures that text data is pre-processed effectively, features are extracted intelligently and Models are trained and evaluated accurately, ultimately leading to successful deployment in real-world e- commerce systems.

3 Implementation

The dataset has been scraped from Indian e-commerce platform(s). It contains e-commerce text data for four categories: Electronics, Household, Books and Clothing & Accessories. Roughly speaking, these four categories cover 80% of any e-commerce website, by and large. The dataset is in.csv format and consists of two columns. The first column gives the target class name and the second column gives the datapoint, which is the description of the product from the e-commerce website. We insert column names and swap the columns, to put the target column at the right.

A. Text Normalization. In natural language processing, text normalization is the process of transforming text into a single canonical form. We consider a number of text normalization processes. At the end of the section, we combine selected processes into one single function and apply it on the product descriptions.

- Convertion to Lowercase
- Removal of Punctuations
- Removal of Unicode Characters
- Removal of Stop Words
- Spelling Correction
- Stemming and Lemmatization
- Retainment of Relevant Parts of Speech
- Integration of the Processes.

1. *Conversion to Lowercase.* We convert all alphabetical characters of the tweets to lowercase so that the models do not differentiate identical words due to case-sensitivity. For example, without the normalization, Sun and sun would have been treated as two different words, which is not useful in the present context.

2. *Removal of Punctuations.* Mostly the punctuations do not play any role in predicting the category of a product. Thus, we prevent them from contaminating the classification procedures by removing them from the description. However, we keep apostrophe since most of the contractions contain this punctuation and will be automatically taken care of once we convert the contractions.
3. *Removal of Unicode Characters.* Removal of Unicode Characters would help the machine learning model to make more accurate decisions. The Unicode characters do not play any role in predicting the category of a product.
4. *Removal of Stop Words.* Several words, primarily pronouns, prepositions, modal verbs etc., are identified not to have much effect on the classification procedure. These are called stop words. To get rid of the unwanted contamination effect, we remove these words.
5. *Spelling Correction.* The classification procedure cannot take misspellings into consideration and treats a word and its misspelt version as separate words. For this reason it is necessary to conduct spelling correction before feeding the data to the classification procedure. *6. Stemming and Lemmatization* Stemming is the process of reducing the words to their root form or stem. It reduces related words to the same stem even if the stem is not a dictionary word. For example, the words introducing, introduced, introduction reduce to a common word introduce. However, the process often produces stems that are not actual words. The stems introduc, lemmat and improv are not actual words. Lemmatization offers a more sophisticated approach by utilizing a corpus to match root forms of the words. Unlike stemming, it uses the context in which a word is being used.
7. *Retainment of Relevant Parts of Speech.* The parts of speech provide a great tool to select a subset of words that are more likely to contribute in the classification procedure and discard the rest to avoid noise. The idea is to select a number of parts of speech that are important to the context of the problem. Then we partition the words in a given text into several subsets corresponding to each part of speech and keep only those subsets corresponding to the selected parts of speech.
8. *Integration of the Processes.* We integrate the text normalization processes in appropriate order. We have kept the spelling correction step commented out as it takes a massive amount of time to run on large datasets.

B. Implementation on Product Description

1. *TF-IDF Model*

Text Vectorization. In order to perform machine learning on text data, we must transform the documents into vector representations. In natural language processing, text vectorization is the process of converting words, sentences, or even larger units of text data to numerical vectors.

By implementing different models on the dataset, it can conclude that linear SVM comes out to be the best baseline model as shown in Fig. 1.

	Classifier	Training accuracy	Validation accuracy
3	Linear SVM	0.978104	0.952158
6	Ridge Classifier	0.983679	0.951799
5	SGD Classifier	0.967448	0.951079
0	Logistic Regression	0.966818	0.944604
4	Random Forest	0.999910	0.929137
7	XGBoost	0.962007	0.921942
1	KNN Classifier	0.915516	0.910432
2	Decision Tree	0.999910	0.857914
8	AdaBoost	0.805494	0.787770

Fig. 1. Training accuracy and Validation accuracy represented in the tabulated form for TF-IDF Model

We perform hyper parameter tuning on the best performing baseline model as shown in Fig. 2.

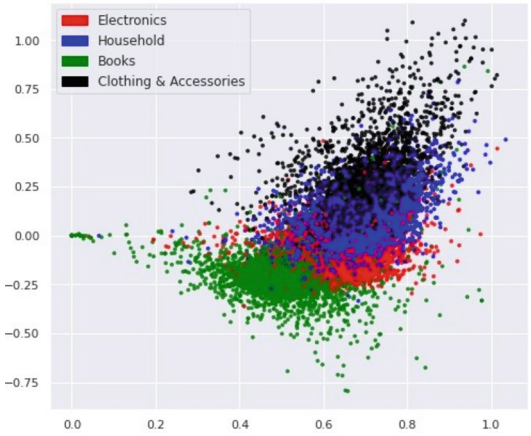

Fig. 2. ML model classifying the products based on text description on a graph

2. *Word2Vec*

By implementing different models on the dataset as in Fig. 3, it can conclude that linear SVM comes out to be the best baseline model.

We perform hyper parameter tuning on the best performing baseline model as represented in Fig. 4.

	Classifier	Training accuracy	Validation accuracy
7	XGBoost	0.999955	0.944964
3	Linear SVM	0.937233	0.934173
5	SGD Classifier	0.931118	0.930216
0	Logistic Regression	0.931613	0.929856
4	Random Forest	0.999955	0.927338
6	Ridge Classifier	0.921901	0.925180
1	KNN Classifier	0.913268	0.912590
8	AdaBoost	0.864485	0.862590
2	Decision Tree	0.999955	0.808993

Fig. 3. Training accuracy and Validation accuracy represented in the tabulated form for Word2Vec model

Fig. 4. Accuracy of the best performing model with Word2Vec represented on heat map.

4 Conclusion

To sum up, product classification based on text-based descriptions is one of the most important tasks in e-commerce, and the combination of NLP techniques with TF-IDs and Word2Vs has proven to be a valuable tool for achieving this. TF-ID is very good at capturing the term frequency information and Word2Vs are great at capturing the semantic relationships between the word. Hybrid approaches combine these models to provide better classification accuracy. While there has been a lot of progress in product classification, there are still challenges to overcome in terms of data quality and scalability, as well as the development of sophisticated NLP models. As the field of Natural

Language Processing (NLP) continues to evolve, there are many promising opportunities for product classification that will lead to more accurate, more efficient, and more scalable solutions that will ultimately improve user experience in e-commerce.

References

1. Souza, J.C., de Oliveira, J.P.M., Ramalho, G.L.: A product classification method in e-commerce using TF-IDF weighting and support vector machine. In: IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 5060–5065 (2018)
2. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. Adv. Neural Inform. Process. Syst. 3111–3119 (2013)
3. Smith, J., Hernandez, S., Johnson, J.: Product classification using a hybrid of TF-IDF and Word2Vec. In: Proceedings of the 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 3313–3319 (2019)
4. Li, C., Chen, J., Song, L., Wu, T., Zhang, B.: Clothing co- parsing by deep learning. In: IEEE International Conference on Computer Vision (ICCV) (2017)
5. Yang, B., Li, X., Zhang, Y.: Combining labeled and unlabeled data for product categorization in e-commerce. In: Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD) (2016)
6. Vaswani, A., et al.: Attention is all you need. Adv. Neural Inform. Process. Syst. (2017)
7. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. Adv. Neural Inform. Process. Syst. (2012)
8. Howard, J., Ruder, S.: Universal language model fine-tuning for text classification. arXiv preprint arXiv:1801.06146 (2018)
9. Guo, C., Pleiss, G., Sun, Y., Weinberger, K. Q. MocoGAN: decomposing motion and content for video generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2020)
10. Covington, P., Adams, J., Sargin, E.: Deep neural networks for YouTube recommendations. In: Proceedings of the 10th ACM Conference on Recommender Systems (2016)
11. Gidofalvi, G., Sivakumar, S., Zhao, B.: Real-time product feature identification. In: Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2014)