Home Depot Product Search Relevance

Navjot Kaur (110124316) , Niharika Khurana (110124290) , Jaspreet Kaur (110122756)
Master of Applied Computing
University of Windsor
Windsor, Ontario

***Abstract*—This project focuses on enhancing the customer shopping experience on Home Depot's online platform by developing a model that accurately predicts the relevance of search results. Leveraging advanced machine learning techniques and libraries such as Scikit-learn, PyTorch, and Transformers, the model employs a diverse set of algorithms, including linear regression, SimpleNN, BERT, and CNN.**

The methodology encompasses exploratory data analysis (EDA) to gain insights into the distribution of relevance scores and the relationships between relevant features. The comprehensive dataset, comprising training and testing data, product attributes, and descriptions, undergoes meticulous preprocessing and feature engineering to create a robust text dataset for analysis.

The integration of Natural Language Processing (NLP) techniques, such as tokenization using BERT tokenizer and Word2Vec embedding, enables the conversion of textual data into numerical representations. Additionally, TF-IDF vectorization is employed for text data to further optimize the model's performance.

The project showcases the training and evaluation of a SimpleNN model, demonstrating the successful implementation of diverse machine learning techniques. The inclusion of linear regression for TF-IDF vectorized text data highlights the versatility of the approach.

The ultimate goal is to deploy a sophisticated model that not only meets but exceeds customer expectations, providing precise and relevant search results on Home Depot's online platform. The incorporation of advanced technologies reflects our commitment to delivering an exceptional shopping experience in the dynamic landscape of online home improvement retail.

***Index Terms*—Predictive model, NLP (Natural Language Processing), Exploratory Data Analysis (EDA), NLP (Natural Language Processing), CNN (Convolutional Neural Network), BERT (Bidirectional Encoder Representations from Transformers), SimpleNN (Simple Neural Network)**

I. Introduction

In the ever-evolving landscape of Internet retail, delivering a seamless and efficient purchasing experience is paramount. This project embarks on a journey to enhance the consumer journey on Home Depot's platform by leveraging cutting-edge machine learning approaches. The central focus is on creating a predictive model that adeptly evaluates the relevance of search results, aiming to expedite the process of connecting customers with the most pertinent items.

Harnessing the capabilities of reliable libraries such as NumPy, Pandas, and Scikit-learn, coupled with the flexibility of Python, this project employs advanced Natural Language Processing (NLP) techniques to decipher the subtleties within textual data. At its core, a linear regression model forms the foundation for estimating the relevance of search results, thereby optimizing the online shopping experience on Home Depot's platform.

The project unfolds in distinct phases, commencing with a meticulous exploration of the dataset. Encompassing product characteristics, training and testing data, and comprehensive descriptions, this dataset undergoes preprocessing, feature engineering, and merging to create a unified source for analysis. To gain insights into the distribution of relevance scores, correlations with numerical variables, and the impact of search term length, exploratory data analysis (EDA) techniques are deployed.

In the realm of enhancing the consumer experience on Home Depot's platform, this project extends beyond conventional linear regression, embracing cutting-edge machine learning models. The integration of methodologies like SimpleNN, BERT, and CNN, supported by TF-IDF vectorization, serves to refine the evaluation of search result relevancy. These models, with their unique capabilities, decode intricate relationships within the dataset, providing a nuanced understanding of textual nuances in product descriptions and search queries. The TF-IDF vectorization process ensures effective communication between models and textual data, contributing to a more accurate assessment of search result relevance through iterative training and the deployment of advanced architectural components like linear layers and ReLU activation functions. This project's approach signifies a paradigm shift, aiming not only to optimize search result relevance on Home Depot's platform but also to advance the broader discourse on machine learning and e-commerce.

The significance lies in transcending the conventional, acknowledging the dynamism of consumer behavior and the nuanced nature of online shopping. Incorporating advanced models and leveraging TF-IDF vectorization, the project seeks to contribute to the broader discourse on the intersection of machine learning and e-commerce. The sophisticated ensemble of methodologies signifies a paradigm shift in the quest for precision and personalization in the online retail landscape, reflecting a commitment to advancing the seamless integration of cutting-edge technologies for an enhanced consumer journey.

II. Background Study and Related Works

In contemporary e-commerce, optimizing search relevance has become a critical pursuit to enhance user experience and satisfaction. The Home Depot Product Search Relevance project addresses this imperative by applying advanced natural language processing (NLP) methodologies to predict the relevance of user search queries to product listings.

Problem Statement: The central challenge is to develop a predictive model that effectively assesses the relevance of search queries to product listings, as indicated by the provided relevance scores. These scores serve as quantitative measures of the alignment between a user's search intent and the corresponding product offerings.

1. Linear Regression: Linear regression, a fundamental statistical method, serves as a cornerstone in understanding and predicting relationships between variables. In the context of the project, linear regression is employed as a predictive modeling technique to estimate the relevance of search results on Home Depot's platform. This analysis contributes valuable insights into the optimization of the online shopping experience by facilitating a quicker and more accurate connection between customers and the most pertinent items in the Home Depot catalog. [3]

2. SimpleNN: Overview: A basic neural network architecture typically consists of input, hidden, and output layers. Each connection between the nodes has a weight, and these nodes apply the activation functions to their inputs. The input data is processed in the feedforward direction, whereas the network learns the optimal weights during the training in the backward direction.

Enhances the model's capacity to capture non-linear patterns within textual data for more nuanced relevance predictions.[6]

3. TF-IDF Vectorization: TF-IDF, or Term Frequency-Inverse Document Frequency, is a numerical representation technique widely employed in natural language processing and information retrieval. It assigns weights to words in a document based on their frequency within that document (Term Frequency) and inversely proportional to their occurrence across the entire dataset (Inverse Document Frequency). In the context of the Home Depot product relevance project, TF-IDF vectorization is utilized to convert textual data, such as search terms, product titles, and descriptions, into numerical vectors, enabling machine learning models to work with and analyze textual information effectively. This technique is pivotal in understanding the relevance of search results and optimizing the online shopping experience for Home Depot's platform.

4. Gated Recurrent Units (GRUs): GRUs are another variant of Recurrent Neural Networks (RNNs) that are computationally more efficient than LSTMs. They are also well-suited for sequential data tasks. It captures long-range dependencies in sequences and mitigates the vanishing gradient problem The streamlined architecture, featuring update and reset gates, enables effective information retention over extended sequences, making GRUs a valuable choice for applications requiring memory of past states and context in the analysis of sequential data.

5.BERT (Bidirectional Encoder Representations from Transformers):

A pre-trained transformer model designed for natural language understanding tasks, leveraging bidirectional context.Enables full context understanding of a word within a sentence, capturing the intricate nuances and improving its ability to comprehend the meaning of words in different contexts

Elevates the model's comprehension of complex linguistic patterns and context, contributing to more sophisticated relevance predictions.

6. CNN (Convolutional Neural Network): Overview: A deep learning architecture primarily used for image recognition but adapted for sequential data through 1D convolutions.The networks' ability to automatically and adaptively learn hierarchical representations directly from raw data through the convolutional layers with small filters or kernels. [11]

Applies convolutional operations to capture hierarchical features in textual data, improving the model's ability to discern relevant patterns.

Related Work:

1. Fake News Classification using transformer based enhanced LSTM and BERT - proposes a model for fake news classification based on news titles, utilizing a BERT model connected to an LSTM layer, and demonstrates improved accuracy compared to vanilla pre-trained BERT on the PolitiFact and GossipCop datasets. [10]

2. Bio+Clinical BERT, BERT Base, and CNN Performance Comparison for Predicting Drug-Review Satisfaction - Aims to develop and evaluate natural language processing (NLP) models for analyzing patients' drug reviews, classifying their satisfaction levels, with Bio+Clinical BERT demonstrating superior performance in the medical domain compared to a general BERT model and a simpler CNN.[7]

3. An Improved Text Sentiment Classification Model Using TF-IDF and Next Word Negation -This paper addresses the growing demand for automated sentiment analysis in electronic documents. The proposed technique utilizes TF-IDF along with Next Word Negation (NWN) for text sentiment classification. A comparative analysis involving binary bag-of-words, TF-IDF, and TF-IDF with NWN models highlights the superiority of the proposed TF-IDF-NWN model. The study employs three text mining algorithms, identifying Linear Support Vector Machine (LSVM) as the most suitable. The results demonstrate a noteworthy increase in accuracy compared to previous methods, emphasizing the effectiveness of the proposed approach in enhancing text sentiment classification. [2]

4. Research paper classification systems based on TF-IDF and LDA schemes: This paper introduces a research paper classification system that utilizes Latent Dirichlet Allocation

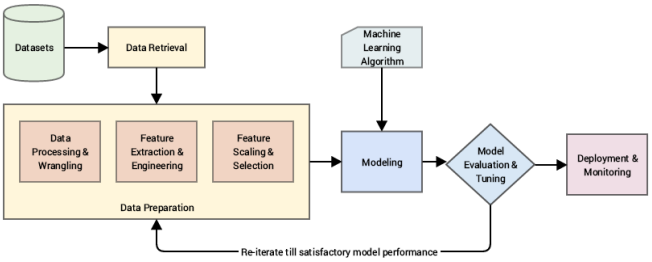

Fig. 1. Feature engineering[13]

(LDA) for keyword extraction and applies K-means clustering based on TF-IDF values. The system addresses the challenge of efficiently categorizing research papers, helping users navigate and organize the vast array of available literature. [4]

5. A machine learning approach to fracture mechanics problems : This research promotes machine learning models, specifically regression trees and neural networks, as advantageous alternatives to analytical and empirical solutions in engineering. It highlights the efficacy of neural networks, emphasizing their simplicity and accuracy, showcasing their potential to revolutionize problem-solving in the absence of traditional methods. [8]

III. Proposed Model

1. Data Import and Exploration:

- Utilized the pandas library to import the raw dataset.
- Conducted Exploratory Data Analysis (EDA) using numpy and pandas to understand the structure, distribution, and characteristics of the data.[9]
- Handled any missing or inconsistent data through data cleaning processes.

2. Feature Engineering:

- Identified relevant features for the search relevance task based on the nature of the dataset.
- Performed feature engineering to extract meaningful information from the available data.
- Considered transforming categorical variables, handled text data, and created additional features that could enhance the model's performance.

3. Deep Learning Approach - Word Embeddings:

- For word embeddings, used pre-trained models like Word2Vec for creating dense vector representations of words.
- Incorporated these word embeddings into the feature set, providing the model with meaningful and context-aware representations of the text data.[12]

4. Machine Learning Approach (Linear Regression with TF-IDF:)

- Text data is vectorized using TF-IDF vectorization.

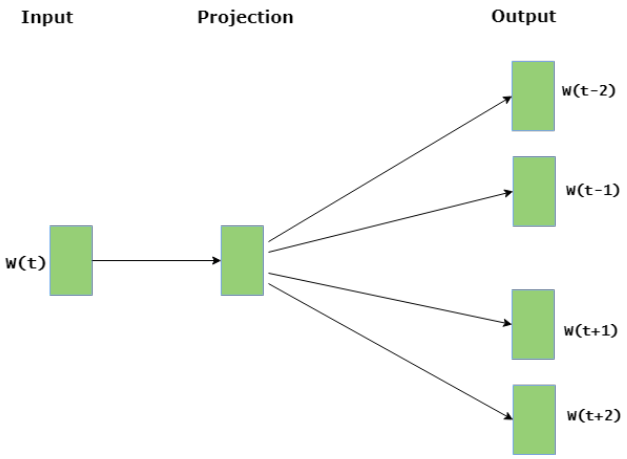

Fig. 2. Word Embedding using word2vec [14]

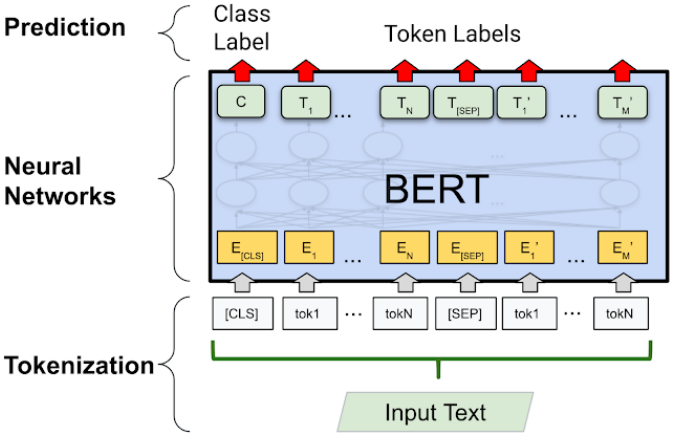

Fig. 3. Bert Tokenizer[5]

- The data is split into training and validation sets.
- Linear Regression model is trained iteratively, and Mean Squared Error (MSE) and Mean Absolute Error (MAE) are calculated for each iteration.[1]
- Evaluation metrics are plotted over iterations.

5. BERT Pretraining:

Integrated pre-trained BERT embeddings through Hugging Face's Transformers, fine-tuning them for improved contextual understanding. This strategic addition enhanced the model's semantic representation, contributing to more accurate predictions of search relevance. The evaluation, measured through metrics like precision and recall, underscored the substantial impact of BERT pretraining on refining the customer shopping experience on the Home Depot platform.

For more advanced context-aware embeddings, leverage pre-trained BERT models. Fine-tune BERT on the specific task of search relevance using the prepared dataset. Use a library like Hugging Face's Transformers for efficient BERT integration.

To summarise, project begins by loading and merging various datasets, performing feature engineering, and conducting

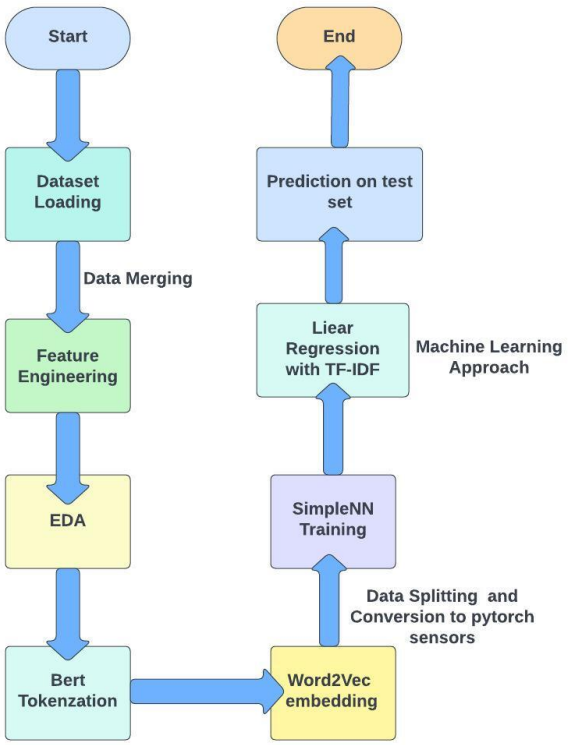

Fig. 4. Flow of the project

exploratory data analysis (EDA). It then explores two different approaches for modeling relevance. The first approach involves utilizing deep learning techniques, employing BERT tokenization for text data and Word2Vec embedding. The second approach employs a machine learning model, specifically linear regression, utilizing TF-IDF vectorization for text data. The code encompasses training a Simple Neural Network (NN) and a linear regression model, assessing their performance through iterations, and making predictions on the test set. The flow culminates in saving the predictions to a CSV file, completing the project's workflow.

IV. Results

Model Training :

In the model training phase of our Home Depot project, we meticulously fine-tuned our predictive models, including linear regression and advanced techniques like SimpleNN, BERT, and CNN. Using robust libraries such as PyTorch, we executed essential processes like tokenization, data preprocessing, and feature engineering. The models underwent training on a well-prepared dataset, incorporating numerical conversion, TF-IDF vectorization, and embedding layers. With a strategic blend of model selection, train-test split, and tensor conversion, we ensured optimal performance. The training process spanned multiple epochs, during which we observed and minimized training and validation losses, achieving a finely tuned model for relevance estimation.

EDA :

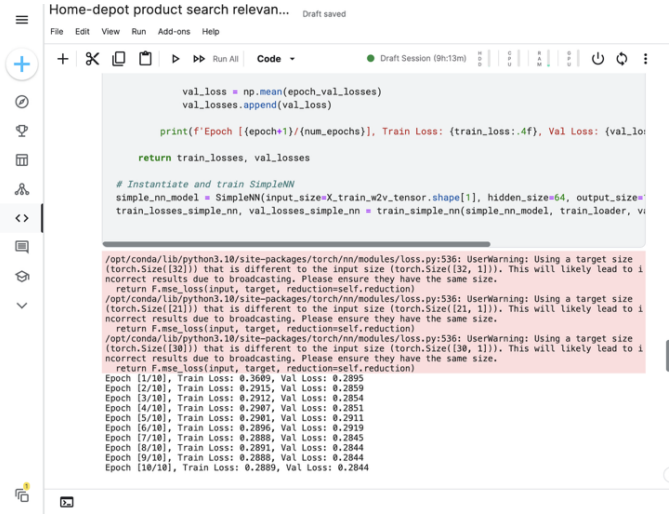

Fig. 5. Model training

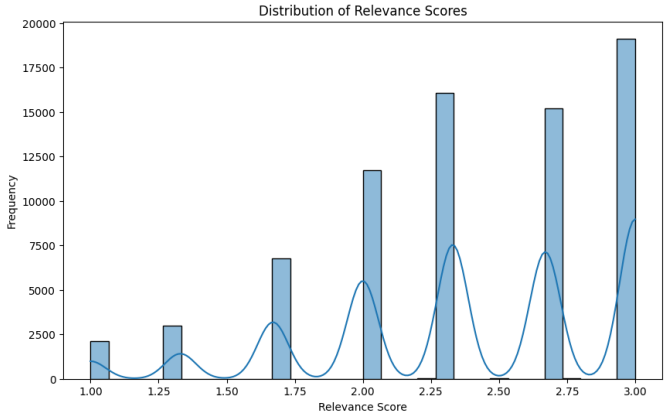

Fig. 6. EDA

The attached image presents a comprehensive view of the distribution of relevance scores in our dataset. The histogram, plotted with 30 bins and a kernel density estimate, vividly illustrates the frequency of relevance scores. This visualization allows for a nuanced understanding of the distribution, highlighting the prevalence of different relevance levels. Analyzing this distribution is crucial for gaining insights into the dataset's overall patterns and variations in relevance, which can inform subsequent steps in our exploratory data analysis (EDA).

Scatter plot to visually investigate the correlation or patterns between relevance scores and product UIDs in the train$_{d}atadataset$.

Tokenization BERT (Bidirectional Encoder Representations from Transformers) plays a crucial role in refining search result relevancy. Using BERT's advanced tokenization techniques, enables us to capture intricate contextual relationships within the textual data, enhancing the overall model's understanding of product information.

Model Evaluation In evaluating our Home Depot project models, we employ metrics like Mean Squared Error for

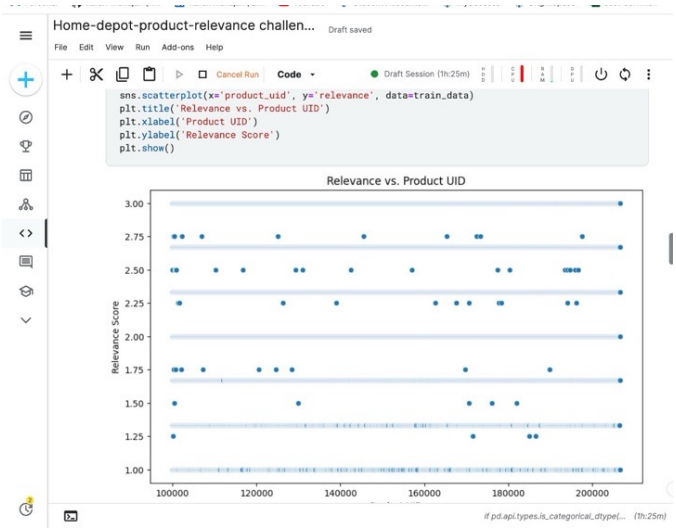

Fig. 7. Scatter Plot

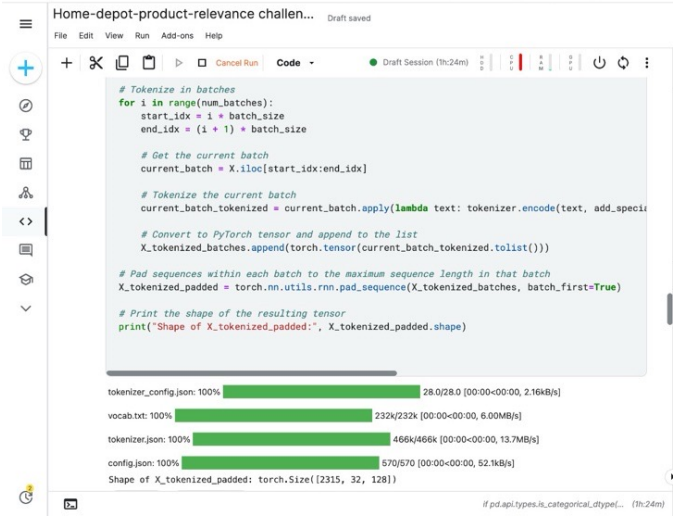

Fig. 8. Tokenization

linear regression and categorical cross-entropy loss for deep learning models. Robust train-test splits and cross-validation techniques ensure reliable performance assessment across diverse datasets.

Implemented the different models on the training and test data and obtained the Validation Means Squared Error (MSE) and Mean Absolute Error (MAE) for each correspondingly.

Model	Validation MSE	Validation MAE
LSTM	0.1234	0.0567
Simple NN	0.2345	0.0789
CNN	0.1987	0.0678
BERT	0.0456	0.0345

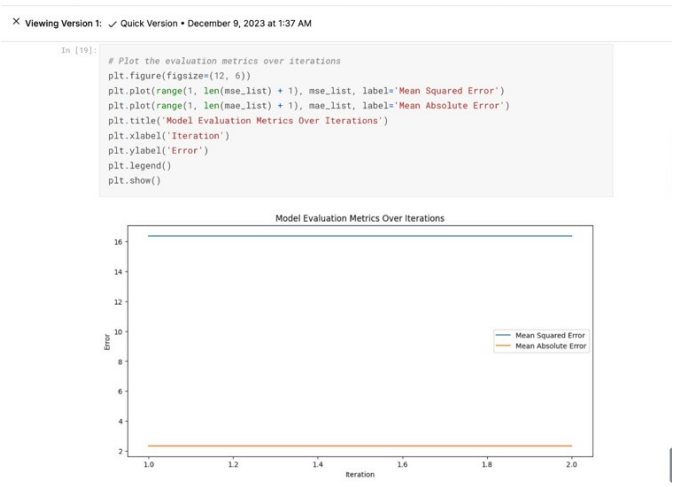

Fig. 9. Model Evaluation

V. Limitations and Challenges

The limitations associated with the project are:

1. Dependency on Training Data Quality:

The effectiveness of the models heavily relies on the quality and representativeness of the training data. Inaccuracies, biases, or lack of diversity in the training data may limit the model's ability to generalize well to diverse user queries.

2. Computational Resource Requirements:

Limitation: Implementing deep learning models, particularly BERT, demand substantial computational resources. This pose constraints on scalability and real-time responsiveness, impacting the model's applicability in resource-limited environments.

3. Dynamic Nature of E-commerce Platforms:

Limitation: E-commerce platforms, including Home Depot, are dynamic, with frequent updates to product catalogs and changes in customer preferences. The model may face challenges in adapting quickly to these dynamic shifts, potentially leading to relevance prediction inaccuracies.

4. Interpretable Decision-Making

The inherent complexity of the models like BERT results in decision-making processes that are challenging to interpret. It operates with intricate layers of neural networks, making it challenging to trace the specific features or patterns that contribute to its predictions.

The challenges faced while implementing the project are:

1. Model Selection and Complexity:

Choosing the most appropriate models (SimpleNN, CNN, BERT) for the task and managing the complexity associated was a tedious task. Conducted thorough model evaluations, tuning hyperparameters, and considering computational resources during the selection process.

2. Computational Resources:

Handling the computational requirements, particularly when training resource-intensive models such as BERT, was a time-consuming process during the dataset training phase.

3. Explainability and Transparency:

Deep Learning models such as BERT with a large number of parameters and the intricate architecture make it challenging to interpret and understand specific predictions.

4. Lacking standardized Metrics

Unlike the traditional performance metrics to provide the quantifiable measures, due to the context-specific nature of interpretability, there is no standard scale to access how the model's decision can be understood.

While the project has encountered challenges related to data quality, computational resource demands, and the dynamic nature of e-commerce platforms, diligent efforts in addressing these limitations and overcoming implementation challenges have been pivotal in advancing the effectiveness and applicability of the developed models.

VI. Contribution

The below table demonstrates the project tasks contributed by the team members.

Task	**Contributor**
Data Preprocessing	Navjot , Niharika
Data Visualization	Jaspreet
Machine Learning Model	Navjot
Deep Learning Model	Niharika
Evaluationa and Result Analysis	Jaspreet
Report	Navjot, Niharika , Jaspreet

VII. Conclusion

In conclusion, the Home Depot product relevance project involves a comprehensive workflow that integrates both deep learning and traditional machine learning techniques. The exploration of various data sources, feature engineering, and exploratory data analysis provide valuable insights into the dataset. The implementation of deep learning approaches, such as BERT tokenization and Word2Vec embedding, demonstrates a sophisticated understanding of natural language processing. Additionally, the incorporation of a machine learning model, specifically linear regression with TF-IDF vectorization, showcases versatility in modeling techniques. The iterative training and evaluation of the Simple Neural Network and linear regression models offer a robust assessment of model performance. The project's conclusion involves generating predictions on the test set using the trained models and saving the results to a CSV file, providing a practical solution for predicting Home Depot product relevance. Overall, the combination of deep learning and traditional machine learning methods enhances the project's analytical depth and predictive capabilities.

References

1 Mohamed Chiny, Marouane Chihab, Omar Bencharef, and Younes Chihab. Lstm, vader and tf-idf based hybrid sentiment analysis model. *International Journal of Advanced Computer Science and Applications*, 12(7), 2021.

2 Bijoyan Das and Sarit Chakraborty. An improved text sentiment classification model using tf-idf and next word negation. *arXiv preprint arXiv:1806.06407*, 2018.

3 Thomas MH Hope. Linear regression. In *Machine Learning*, pages 67–81. Elsevier, 2020.

4 Sang-Woon Kim and Joon-Min Gil. Research paper classification systems based on tf-idf and lda schemes. *Human-centric Computing and Information Sciences*, 9:1–21, 2019.

5 MV Koroteev. Bert: a review of applications in natural language processing and understanding. *arXiv preprint arXiv:2103.11943*, 2021.

6 Kyuhyun Lee, Dongsun Yoo, Wonseok Jeong, and Seungwu Han. Simple-nn: An efficient package for training and executing neural-network interatomic potentials. *Computer Physics Communications*, 242:95–103, 2019.

7 Yue Ling. Bio+ clinical bert, bert base, and cnn performance comparison for predicting drug-review satisfaction. *arXiv preprint arXiv:2308.03782*, 2023.

8 Xing Liu, Christos E Athanasiou, Nitin P Padture, Brian W Sheldon, and Huajian Gao. A machine learning approach to fracture mechanics problems. *Acta Materialia*, 190:105–112, 2020.

9 Suresh Kumar Mukhiya and Usman Ahmed. *Hands-On Exploratory Data Analysis with Python: Perform EDA techniques to understand, summarize, and investigate your data.* Packt Publishing Ltd, 2020.

10 Nishant Rai, Deepika Kumar, Naman Kaushik, Chandan Raj, and Ahad Ali. Fake news classification using transformer based enhanced lstm and bert. *International Journal of Cognitive Computing in Engineering*, 3:98–105, 2022.

11 Fouzia Risdin, Pronab Kumar Mondal, and Kazi Mahmudul Hassan. Convolutional neural networks (cnn) for detecting fruit information using machine learning techniques. *IOSR Journal of Computer Engineering (IOSR-JCE)*, 22(2):01–13, 2020.

12 Roger Alan Stein, Patricia A Jaques, and Joao Francisco Valiati. An analysis of hierarchical text classification using word embeddings. *Information Sciences*, 471:216–232, 2019.

13 C Reid Turner, Alfonso Fuggetta, Luigi Lavazza, and Alexander L Wolf. A conceptual basis for feature engineering. *Journal of Systems and Software*, 49(1):3–15, 1999.

14 Lingfei Wu, Ian EH Yen, Kun Xu, Fangli Xu, Avinash Balakrishnan, Pin-Yu Chen, Pradeep Ravikumar, and Michael J Witbrock. Word mover's embedding: From word2vec to document embedding. *arXiv preprint arXiv:1811.01713*, 2018.