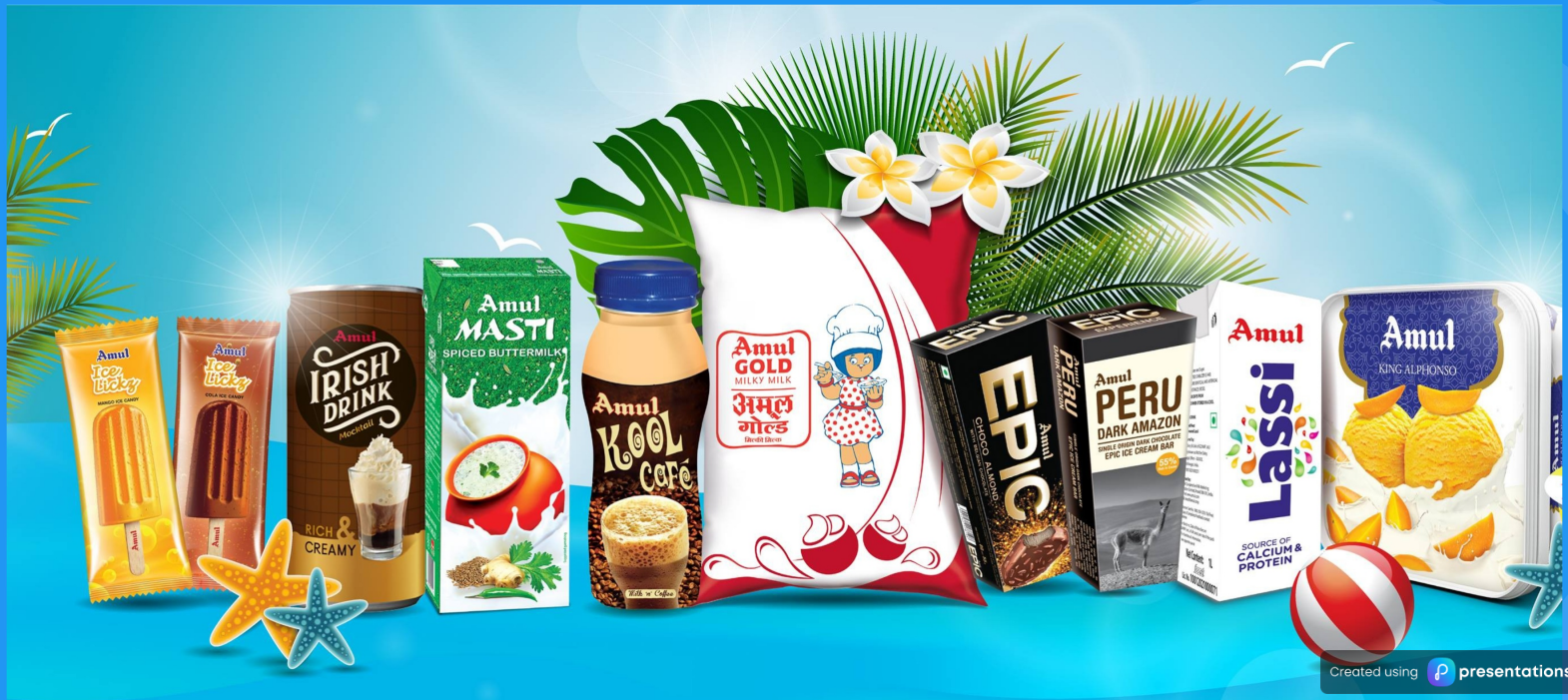
Maximizing Sales with Product Categorization

Exploring how effective categorization boosts product sales and success on Flipkart's online platform.

Sangeeta Yadav
Presenter

Ecommerce Product Categorization

Explore the complexities of Ecommerce product categorization as part of Hackathon.



Product Categorization in E-commerce

Key Insights and Observations

1 Importance of Product Categorization

Product categorization is crucial for e-commerce businesses like Flipkart and Amazon. It significantly impacts whether a product is found and purchased by customers. Incorrect categorization can lead to reduced visibility and sales.

2 Multiclass Classification

The problem of product categorization can be approached using Multiclass Classification. This is because a product can belong to multiple categories, requiring a flexible classification approach.

3 Observations from ML and DL Models

Through training various ML and DL models, insights were gathered on effectively solving the product categorization problem. These models help in accurately classifying products into appropriate categories, improving sales outcomes.

1. Dominance of Clothing and Jewelry Categories

Insight

Clothing and jewelry contribute the highest sales volumes, indicating high demand in these categories.

Action

Prioritize inventory expansion, promotional campaigns, and new product launches in these segments.

2. Seasonal Trends Impact Sales

Insight

Clothing and jewellry sales are likely seasonal—peaking during festive seasons, weddings, and holidays.

Action

Plan seasonal discounts and targeted marketing campaigns during key shopping periods (e.g., Black Friday, Christmas, Diwali).

3. Focus on High-Margin Jewelry Products

Insight

Jewelry, although lower in sales than clothing, typically has higher profit margins.

Action

Highlight jewelry in premium advertisements and offer financing options to boost sales of high-ticket items.

4. Diversify Products in Emerging Categories

Insight

Categories like electronics, home decor, and beauty products might be underrepresented but show potential for growth.

Action

Expand product offerings in emerging categories based on customer demographics and preferences.

5. Improve Product Discoverability

Insight

Ambiguous product descriptions or categories may lead to low visibility in search results, reducing sales.

Action

Optimize product categorization, include keywords, and improve SEO-friendly titles and filters for easy navigation.

6. Analyze Category-Wise Conversion Rates

Insight

Clothing might attract high traffic but could have a lower conversion rate than jewelry, which appeals to specific, committed buyers.

Action

Optimize product descriptions, sizing guides, and customer reviews to improve conversions in clothing.

Product Dataset Overview

Key Information and Characteristics

Attribute	Description
Link to the Dataset	https://docs.google.com/spreadsheets/d/1on6ApK7jNXSy20Bfw_NjKMTmeICTHyBvIK-3am9irFA/edit?usp=sharing
Number of Rows	14999
Number of Columns	15
Notes	Includes non-ASCII characters

Procedure & Observations

Steps in Data Processing and Product Categorization

- **Exploratory Data Analysis and Data Pre Processing**

  This step involves analyzing the dataset to understand its structure, identify patterns, and clean the data. It is crucial for ensuring data quality and preparing it for further analysis.

- **Machine Learning Models for Product Categorization**

  In this phase, various machine learning algorithms are employed to classify products into categories based on the preprocessed data. This helps in automating the categorization process.

- **Deep Learning Models for Product Categorization**

  Advanced deep learning techniques are applied to enhance the accuracy of product categorization. These models can capture complex patterns and improve classification performance.

Step 1: Exploratory Data Analysis & Data Preprocessing

Comprehensive guide to initial steps in data analysis and preparation for machine learning models.

Handling NaN and Duplicate Values

Addressed NaN values primarily in 'brands', 'retail_price', and 'discounted_price' columns without dropping rows due to dataset size. Focused on 'description' for product categorization.

Visualisation of Product Distribution

Used Seaborn library to plot product distribution across brands and common words in product descriptions, identifying irrelevant terms for removal.

Identifying Primary Categories

Categorized products into 8 primary categories by examining unique categories, balancing dataset entries, and removing noise.

Analysis of Text Length

Examined text length across categories, deciding against thresholds due to dataset size, ensuring no loss of useful information.

Word Clouds for Frequent Words

Generated word clouds to identify common words, aiding in creating a custom stopwords list for data cleaning.

Data Cleaning and Preprocessing

Performed steps like contraction mapping, lowercasing, stopwords removal, tokenization, and lemmatization to clean data.

Balancing the Dataset

Applied random oversampling and undersampling to balance the dataset, ensuring improved model accuracy.

Validation Accuracies of ML Algorithms

Detailed comparison of ML algorithm accuracies across different dataset types

NAME OF THE ML ALGORITHM	IMBALANCED DATASET	BALANCED DATASET (OVERSAMPLED)	BALANCED DATASET (UNDERSAMPLED)
Logistic Regression (Multiclass Classification Variant)	0.9735	0.9893	0.9654
Linear Support Vector Machine	0.9799	0.9958	0.9749
Random Forest Classifier	0.9209	0.9367	0.9235
K Nearest Neighbours	0.9564	0.9800	0.9453

Step 3: Deep Learning Models for Product Categorization

Exploring Advanced Deep Learning Techniques for Effective Product Categorization

1

Recurrent Neural Network - LSTM

LSTM, a type of Recurrent Neural Network, is used for sequence prediction tasks. It is trained using the same dataset and evaluated through similar metrics as Transformer models.

2

Dataset Preparation

The dataset used is cleaned, preprocessed, and balanced through undersampling. This preparation ensures consistent training and evaluation of the models.

3

Evaluation Metrics

Models are compared using accuracy scores, and classification reports. These metrics provide insights into model performance and reliability.

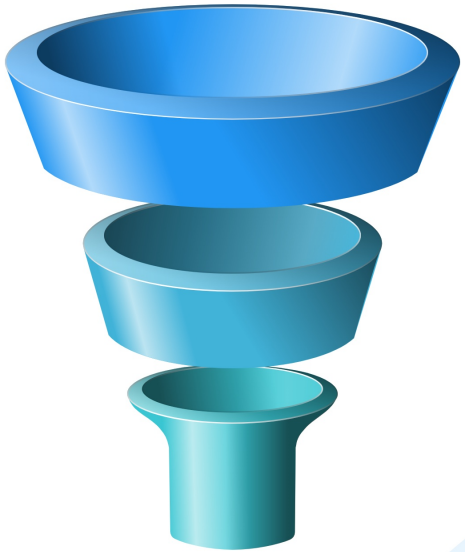
Implementation Steps for RNN Based LSTM Network

In-depth Look at the Steps for Implementing an RNN Based LSTM Network

Text Preprocessing

Text preprocessing is crucial for LSTM implementation. It involves using Field and LabelField objects to prepare data, splitting it into training, validation, and test sets, and building a vocabulary. Each word is assigned an index and converted into integer sequences with pretrained embeddings. Data is processed in batches using BucketIterator with a BATCH_SIZE of 16.

1



2

Model Training and Evaluation

The LSTM model architecture is set up for multiclass classification using init() and forward() functions. The model captures long-term dependencies with parameters like input_size, hidden_size, num_layers, and bi-direction set to true. The Adam optimizer and CrossEntropyLoss measure performance. Training includes activating dropout layers, while evaluation deactivates them. The model runs for 4 epochs tracking minimum Validation Loss.

Computational Limitations

Due to Google Colab's computational restrictions, documenting the model's accuracy was challenging. Despite this, the model was constructed following a structured process as described.

3

Thank You