Project Exposé

AI-Driven Text-Based Classification of Digital Content in E-Commerce

Course: Spring 2025 CSC 8980 Topics in CS: Advanced Text Processing and LLMs

Instructor: Dr. Mario Kubek

Team Members:	Nagamedha Sakhamuri	002828574	nsakhamuri1@student.gsu.edu
	Nookaraju Guttula	002822604	nguttula1@student.gsu.edu
	LaxmiPrasanna Ravikanti	002847070	lravikanti1@student.gsu.edu

1. Problem Statement

Manual categorization of products in e-commerce platforms is inefficient, time-consuming, and prone to errors. As product inventories grow, the need for an automated text-based classification system becomes crucial to ensure accurate, scalable, and efficient categorization of consumer goods. Misclassified products result in poor searchability and a suboptimal shopping experience.

This project aims to develop an NLP-powered classification system that categorizes products based on their textual descriptions. Our solution leverages machine learning (ML) models to automate the classification process, improving consistency and reducing the dependency on manual categorization.

2. Key Use Cases

Our approach focuses on solving **three critical challenges** in e-commerce:

1. **Seller Assistance:** Automating product categorization to reduce human errors and enhance efficiency for sellers listing products.

2. **Enhanced Search & Recommendations:** Ensuring products are placed in the correct categories to improve customer searchability and recommendations.

3. **Inventory Management:** Organizing large-scale product listings into structured categories, helping e-commerce platforms streamline their inventory.

3. Proposed Solution & Methodology

This project follows a **top-down approach**, ensuring that the applied NLP and ML techniques are justified and directly linked to addressing the above use cases.

Data Collection & Preprocessing:

- **Dataset and Justification**

  - We use the **Flipkart Products Dataset** from Kaggle (1050 entries), which includes product titles, descriptions, image URLs, and category labels. It is well-suited for our text and image-based classification pipeline.

  - **Link:** https://www.kaggle.com/datasets/PromptCloudHQ/flipkart-products

  - The dataset is **real-world, well-structured, and lightweight**, making it suitable for rapid experimentation and aligned with our course timeline.

- **Preprocessing Techniques:** Text cleaning (removal of stop words, punctuation, and case normalization), tokenization, and lemmatization.

- **Feature Engineering:** Representation of text using **TF-IDF, Word2Vec, and BERT embeddings** to extract meaningful patterns for classification.

Machine Learning Models (Use-Case Driven Approach):

- **Naïve Bayes & Logistic Regression:** Establishing a baseline for text classification.

- **Support Vector Machines (SVM):** Handling high-dimensional text data to support structured categorization.

- **Deep Learning Models (BERT & LSTMs):** Enhancing contextual understanding for **search optimization and recommendations**.

- **Ensemble Learning:** Combining multiple models to increase classification accuracy and reliability.

Evaluation & Optimization:

- Metrics: **Accuracy, Precision, Recall, F1-Score, Confusion Matrix.**

- Hyperparameter tuning to prevent overfitting and improve generalization.

- Testing with real-world datasets to validate efficiency and scalability.

4. Project Plan & Timeline

Phase	Task	Timeline
Phase 1	Data Collection & Preprocessing	March 15 - March 25
Phase 2	Feature Engineering & Embeddings	March 26 - April 5
Phase 3	Model Development & Training	April 6 - April 15
Phase 4	Model Evaluation & Optimization	April 16 - April 20
Phase 5	Final Documentation & Submission	April 21 - April 25

5. Team Roles & Responsibilities

- **Nagamedha Sakhamuri:** Data collection, preprocessing, and feature extraction.

- **Nookaraju:** Model training, evaluation, and hyperparameter tuning.

- **Lakshmi Prasanna:** System integration, deployment, and final documentation.

6. Conclusion

This project leverages **advanced NLP techniques** to create an **automated, scalable, and efficient classification system** for e-commerce platforms. By replacing manual product categorization with an **AI-driven model**, we aim to improve **searchability, accuracy, and inventory management**. Our structured **top-down approach ensures practical usability** while maintaining robust **ML-backed automation**.