

MID TERM INTERNSHIP REPORT

PRODUCT CLASSIFICATION SYSTEM FOR



LATENTVIEW ANALYTICS

Submitted by

RUJUL DWIVEDI - 2103319

Bachelor of Technology

School of Mathematics & Computer Science

Indian Institute of Technology Goa



Contents

| | | |
|----------|--|-----------|
| 1 | Abstract | 1 |
| 2 | Introduction | 2 |
| 2.1 | Organization Overview | 2 |
| 2.2 | Project Scope | 2 |
| 2.3 | Objectives | 3 |
| 3 | Internship Experience | 4 |
| 3.1 | Internship Description | 4 |
| 3.1.1 | Phase 01: Data Collection and Preparation . . . | 4 |
| 3.1.2 | Phase 02: EDA and Baseline Model Development | 4 |
| 3.2 | Project Description | 5 |
| 3.3 | Learnings & Challenges | 5 |
| 3.3.1 | Business Perspective | 6 |
| 3.3.2 | Technical Perspective | 6 |
| 4 | Results and Analysis | 7 |
| 4.1 | Classes Histogram | 7 |
| 4.2 | Correlation Between N/A in Titles and Empty Descriptions . | 8 |
| 4.3 | Peek at the Data | 9 |
| 4.4 | Accuracy Score | 10 |
| 4.5 | Testing with a Description | 10 |
| 4.6 | Confusion Matrix | 11 |
| 4.7 | Classification Report | 12 |
| 4.8 | Word Cloud | 12 |
| 4.9 | Skewness in Misclassifications | 13 |
| 4.10 | Error Distribution | 14 |
| 5 | Conclusion | 15 |
| | Acknowledgements | 16 |

Chapter 1

Abstract

This report provides a comprehensive overview of the progress made during the initial phase of my internship at **LatentView Analytics**, under the *BET Entity*. It highlights key tasks, methodologies, challenges encountered, and significant learnings.

The primary objective of this work is to develop an **AI-driven product classification system** for the retail industry. This involves multiple stages, including extensive *data collection*, *preprocessing*, *exploratory data analysis (EDA)*, and **model optimization**. Various machine learning techniques have been explored, ranging from traditional classification models to advanced approaches incorporating *word embeddings*, *feature engineering*, and **hyperparameter tuning**.

This report presents a structured breakdown of the weekly progress, detailing the methodologies employed, insights gained, and results achieved. **Key challenges**, such as handling data imbalance, improving feature representations, and mitigating model overfitting, have been addressed through rigorous experimentation.

Although substantial progress has been made, the internship is ongoing. The upcoming phase will focus on **refining classification models**, integrating *transformer-based architectures* like **BERT** and **GPT**, and implementing *Generative AI-based classification*. This shift aims to enhance model accuracy, scalability, and automation in retail product categorization.

Chapter 2

Introduction

This section provides an in-depth overview of the organization, the scope of the project, and the key objectives set for the internship.

2.1 Organization Overview

LatentView Analytics is a leading provider of data analytics and artificial intelligence-driven solutions. The company specializes in implementing end-to-end analytics strategies to help businesses derive meaningful insights and drive growth. By leveraging advanced data science methodologies, LatentView empowers organizations to make informed, data-driven decisions.

With a strong focus on **AI, machine learning, and analytics**, LatentView is committed to delivering scalable, sustainable, and impactful solutions that optimize operations and enhance customer experiences.

2.2 Project Scope

The retail industry is witnessing rapid digital transformation, resulting in an exponential increase in the number of products listed across online and offline platforms. However, ***manual product categorization*** remains a significant bottleneck, as it is:

- Time-consuming and labor-intensive.
- Prone to human errors, leading to inconsistencies in product classification.
- Challenged by the dynamic nature of product catalogs, requiring frequent updates.

This project aims to develop an **AI-driven product classification system** that automates the categorization of new products with high accuracy. The key functionalities of the system include:

- Automatically assigning products to the correct categories based on their descriptions and attributes.
- Enhancing efficiency and accuracy by reducing dependency on manual classification.
- Analyzing classification patterns to extract **valuable market insights**, such as emerging product trends and category distributions.

To achieve this, multiple machine learning techniques are (and will be) explored, ranging from traditional classifiers like **Random Forest and Long Short Term Memory (LSTM)** to advanced **Transformer-based models (BERT, GPT)** for improved classification performance.

2.3 Objectives

The primary objectives of this internship are:

- **Develop a robust AI-powered classification system** capable of accurately categorizing retail products.
- **Implement and refine NLP-based feature extraction** techniques, such as *TF-IDF*, *word embeddings*, and *named entity recognition (NER)*.
- **Improve classification accuracy** by (in future) leveraging *Generative AI* models for automated product categorization.
- **Address challenges** such as *data imbalance*, *misclassification trends*, and *overfitting* through data augmentation and sampling techniques.

The implementation details, experiments, and results of the project can be accessed via the following links:

- **GitHub Repository:** Project Code and Documentation
- **Dataset Source:** Retail Product Classification Dataset

The subsequent sections of this report will provide a detailed breakdown of the methodologies, weekly progress, challenges encountered, and future directions for improving the system's capabilities.

Chapter 3

Internship Experience

This chapter provides a detailed account of the structured phases of the internship, outlining key activities, challenges encountered, and learnings from both a business and technical perspective.

3.1 Internship Description

The internship was structured into multiple phases, ensuring a strong foundation before transitioning to model development and evaluation.

3.1.1 Phase 01: Data Collection and Preparation

The first phase was dedicated to gathering and organizing datasets for further analysis. This involved:

- **Data Acquisition:** Explored product-related datasets from various sources, including *Amazon Reviews* and *Flipkart Products*.
- **Data Quality Checks:** Identified and handled missing values, inconsistencies, and redundant data points.
- **Dataset Organization:** Structured the collected data into meaningful formats, ensuring compatibility with NLP preprocessing techniques.

3.1.2 Phase 02: EDA and Baseline Model Development

After data collection, the next phase focused on *understanding data characteristics and building initial classification models*:

- **Exploratory Data Analysis (EDA):**
 - Utilized visualization techniques such as *word clouds*, *bigram/trigram frequency plots*, and *category distribution graphs*.
 - Examined key patterns in product descriptions to enhance feature selection.
- **Text Preprocessing:**
 - Implemented NLP-based text preprocessing techniques, including *tokenization*, *stopword removal*, and *stemming/lemmatization*.
 - Explored feature extraction methods such as **TF-IDF**, **word embeddings**, and *named entity recognition (NER)*.
- **Baseline Model Development:**
 - Built initial classification models within business constraints.
 - Focused on performance evaluation using **precision**, **recall**, and **F1-score**.

3.2 Project Description

Key highlights of the project:

- The initial phases focused on dataset preparation, ensuring high-quality input data.
- Various feature extraction techniques and classification algorithms were explored.
- Due to ***business constraints***, traditional models such as *Random Forest* were not prioritized. Instead, the focus shifted to LSTMs (and later **Generative AI-based classification techniques**) to improve accuracy and adaptability.

In the next phases, by leveraging ***transformer-based models*** like **BERT** and **GPT**, the project aims to enhance product classification capabilities beyond conventional machine learning approaches.

3.3 Learnings & Challenges

Throughout the internship, several challenges were encountered, providing valuable insights from both ***business*** and ***technical*** perspectives.

3.3.1 Business Perspective

Understanding the real-world implications of AI-driven classification was crucial. Key business takeaways include:

- **Operational Efficiency:** AI-based automation significantly reduces the time and effort required for manual product categorization.
- **Business Constraints:** Model selection and deployment were influenced by infrastructure limitations, computational costs, and industry-specific requirements.
- **Scalability Considerations:** Ensuring that AI models remain efficient and adaptable when scaled for production environments.

3.3.2 Technical Perspective

From a technical standpoint, the project involved addressing multiple machine learning and data processing challenges:

- **Class Imbalance:** Dealing with underrepresented product categories required the use of *oversampling, undersampling, and data augmentation* strategies.
- **Overfitting Mitigation:** Experimented with different *regularization techniques, dropout layers, and dataset fractions* to improve generalization.
- **Feature Extraction Techniques:**
 - Compared different text representation methods such as *TF-IDF, Word2Vec, and contextual embeddings*.
 - Evaluated the effectiveness of **named entity recognition (NER)** in extracting meaningful product attributes.
- **Model Robustness:** Ensured the model performs effectively under real-world constraints, such as *limited computational resources and large-scale datasets*.

This section highlights the structured approach taken during the internship, the key challenges faced, and the solutions explored to enhance AI-powered classification. The following chapters will further elaborate on the methodology, results, and future scope of the project.

Chapter 4

Results and Analysis

This section presents the key findings from the exploratory data analysis (EDA) and model evaluation. Various visualizations and performance metrics have been included to assess the effectiveness of the product classification system.

4.1 Classes Histogram

The histogram below shows the distribution of product categories in the dataset, highlighting class imbalances that need to be addressed.

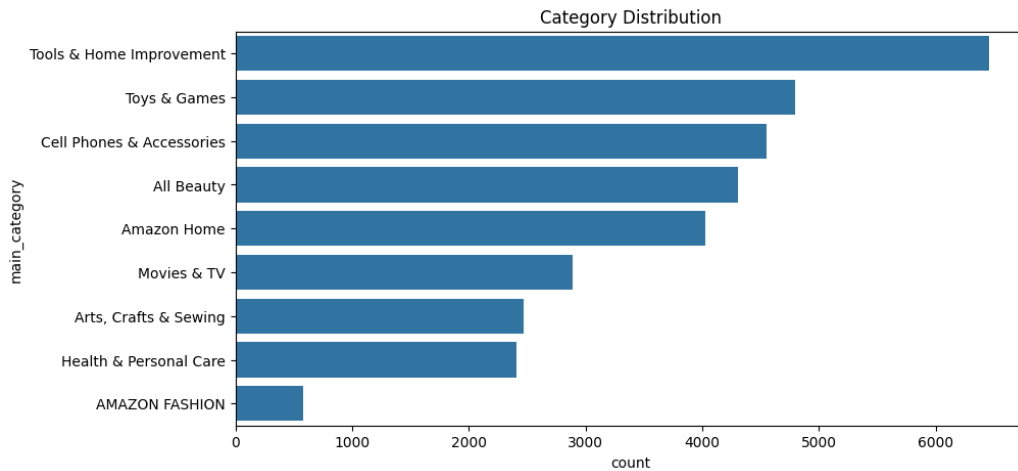


Figure 4.1: Histogram of Product Categories

4.2 Correlation Between N/A in Titles and Empty Descriptions

A correlation matrix was generated to analyze the relationship between missing values in the 'title' column and empty strings in the 'description' column. This helps in understanding whether missing data patterns affect classification accuracy.

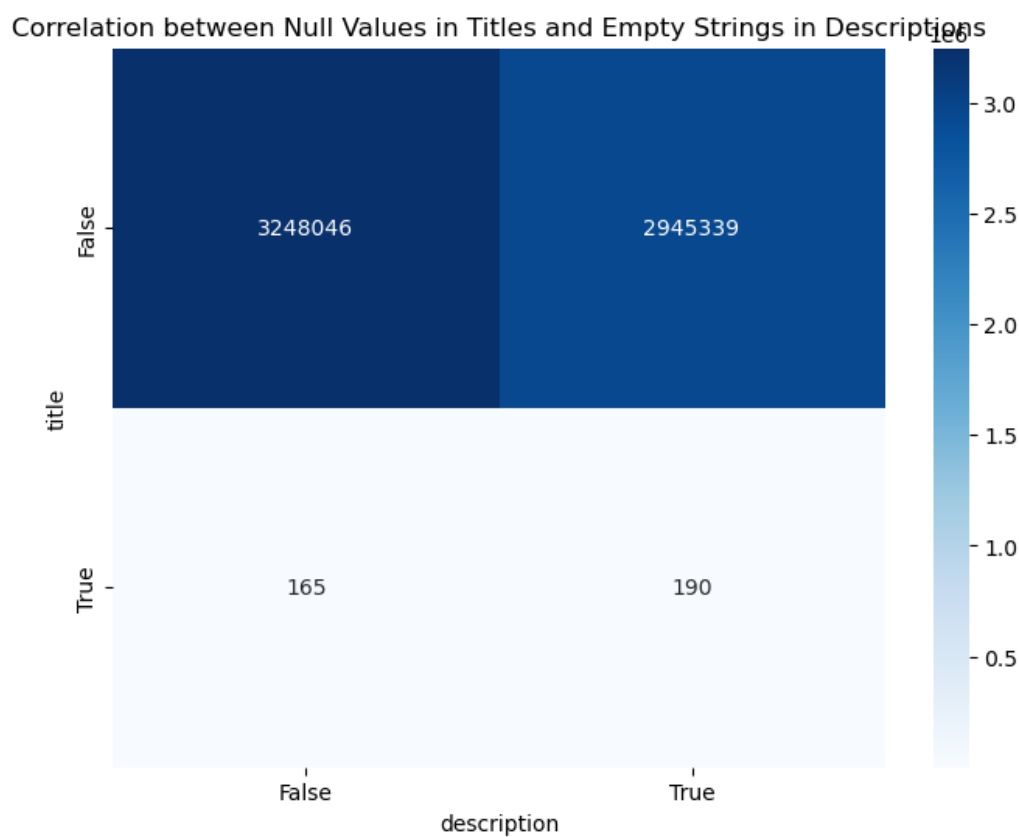


Figure 4.2: Correlation Between Null Values in Title and Empty Descriptions

4.3 Peek at the Data

A sample of the dataset is shown below, providing an overview of the text-based attributes and their structure.

```
<class 'pandas.core.frame.DataFrame'>
Index: 32480 entries, 37538 to 160665
Data columns (total 5 columns):
#   Column          Non-Null Count  Dtype
---  -
0   main_category    32480 non-null  object
1   title            32480 non-null  object
2   description       32480 non-null  object
3   images           32480 non-null  object
4   details          32480 non-null  object
dtypes: object(5)
memory usage: 1.5+ MB

   main_category \
37538  Cell Phones & Accessories
124693  Toys & Games
249067  Amazon Home
103904  Tools & Home Improvement
126448  Cell Phones & Accessories

   title \
37538  Galaxy S8 Active Case - R2D2 Droid Robot Patte...
124693  Magic: the Gathering - Arcbound Lancer - Darks...
249067  Fall Harvest Sitting Scarecrow with Dangling L...
103904  Leviton 80530-BLK EB WP 2G POWER OUTLET 2,465 ...
126448  Alcatel Idol 4s Screen Protector [6-Pack], Kle...

   description \
37538  ['Protection:', 'This Dual Layer protector hyb...
124693  ['Magic: the Gathering is a collectible card g...
249067  ['27" Fall Harvest Scarecrow with Poseable Arm...
103904  ['Product Description', 'THERMOSET/PLASTIC MAT...
126448  ['Specially designed to significantly reduce g...

   images \
37538  [{'thumb': 'https://m.media-amazon.com/images/...
124693  [{'thumb': 'https://m.media-amazon.com/images/...
249067  [{'thumb': 'https://m.media-amazon.com/images/...
103904  [{'thumb': 'https://m.media-amazon.com/images/...
126448  [{'thumb': 'https://m.media-amazon.com/images/...

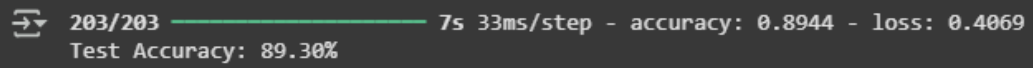
   details
37538  {'Package Dimensions': '6.5 x 4 x 0.5 inches',...
124693  {'Package Dimensions': '3.46 x 2.48 x 0.01 inc...
249067  {'Package Dimensions': '13.5 x 7 x 4.8 inches'...
103904  {'Switch Type': 'Decorator', 'Material': 'Ther...
126448  {'Item Weight': '2 ounces', 'Other display fea...
```

Figure 4.3: Sample Data Preview

4.4 Accuracy Score

The accuracy score of the classification model provides an initial evaluation of its performance on the test dataset.

```
[ ] # Evaluate model
    loss, accuracy = model.evaluate(X_test, y_test)
    print(f'Test Accuracy: {accuracy * 100:.2f}%')
```



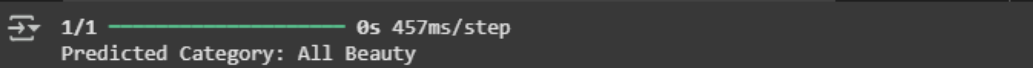
203/203 ————— 7s 33ms/step - accuracy: 0.8944 - loss: 0.4069
Test Accuracy: 89.30%

Figure 4.4: Model Accuracy Score

4.5 Testing with a Description

To validate the model's effectiveness, a sample product description was tested, and its predicted category was recorded.

```
[ ] # Test the model with a new description
    new_description = 'Unveil a radiant, youthful complexion with our luxurious Elixir
    new_description = clean_text(new_description)
    new_sequence = tokenizer.texts_to_sequences([new_description])
    new_padded_sequence = pad_sequences(new_sequence, maxlen=100)
    prediction = model.predict(new_padded_sequence)
    predicted_category = data['main_category'].unique()[np.argmax(prediction)]
    print(f'Predicted Category: {predicted_category}')
```



1/1 ————— 0s 457ms/step
Predicted Category: All Beauty

Figure 4.5: Model Prediction on a Sample Product Description

4.6 Confusion Matrix

The confusion matrix below illustrates the model’s performance in classifying various product categories, showing misclassifications and correct predictions.

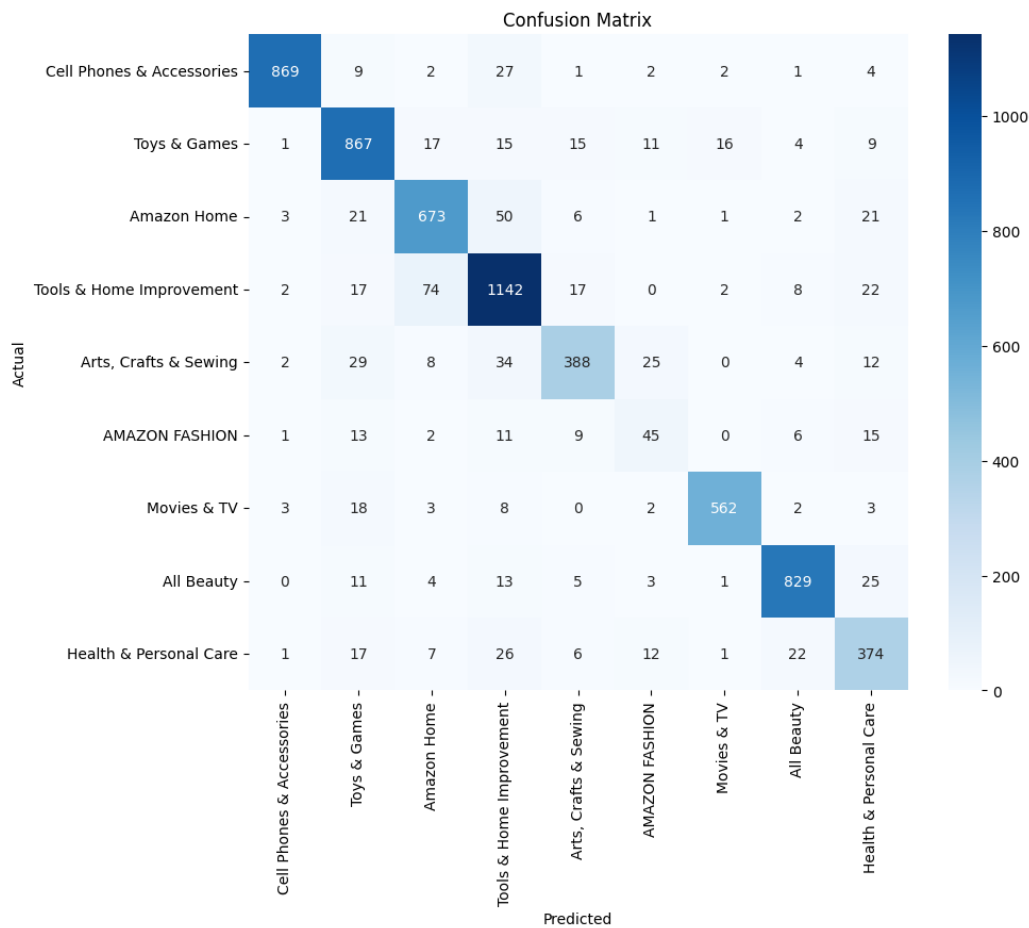


Figure 4.6: Confusion Matrix of Classification Model

4.7 Classification Report

The classification report provides a detailed breakdown of the model's performance metrics, including precision, recall, and F1-score for each category.

| | precision | recall | f1-score | support |
|---------------------------|-----------|--------|----------|---------|
| Cell Phones & Accessories | 0.99 | 0.95 | 0.97 | 917 |
| Toys & Games | 0.87 | 0.91 | 0.89 | 955 |
| Amazon Home | 0.85 | 0.87 | 0.86 | 778 |
| Tools & Home Improvement | 0.86 | 0.89 | 0.88 | 1284 |
| Arts, Crafts & Sewing | 0.87 | 0.77 | 0.82 | 502 |
| AMAZON FASHION | 0.45 | 0.44 | 0.44 | 102 |
| Movies & TV | 0.96 | 0.94 | 0.95 | 601 |
| All Beauty | 0.94 | 0.93 | 0.94 | 891 |
| Health & Personal Care | 0.77 | 0.80 | 0.79 | 466 |
| | | | | |
| accuracy | | | 0.89 | 6496 |
| macro avg | 0.84 | 0.83 | 0.84 | 6496 |
| weighted avg | 0.89 | 0.89 | 0.89 | 6496 |

Figure 4.7: Classification Report of Model Performance

4.8 Word Cloud

A word cloud was generated to visualize the most frequent words in the dataset, helping to understand important keywords that contribute to classification.



Figure 4.8: Word Cloud of Product Descriptions

4.9 Skewness in Misclassifications

To identify biases in the classification model, an analysis was conducted to determine whether certain categories were being misclassified into specific groups at a higher rate.

```
Top misclassified categories for 'Cell Phones & Accessories':  
- Toys & Games: 44.44%  
- Health & Personal Care: 27.78%  
- Tools & Home Improvement: 27.78%  
  
Top misclassified categories for 'Toys & Games':  
- Amazon Home: 23.17%  
- Movies & TV: 23.17%  
- Arts, Crafts & Sewing: 20.73%  
  
Top misclassified categories for 'Amazon Home':  
- Tools & Home Improvement: 57.73%  
- Health & Personal Care: 21.65%  
- Toys & Games: 13.40%  
  
Top misclassified categories for 'Tools & Home Improvement':  
- Amazon Home: 43.81%  
- Toys & Games: 19.05%  
- Health & Personal Care: 15.24%  
  
Top misclassified categories for 'Arts, Crafts & Sewing':  
- Tools & Home Improvement: 35.79%  
- Toys & Games: 28.42%  
- Amazon Home: 10.53%  
  
Top misclassified categories for 'AMAZON FASHION':  
- Arts, Crafts & Sewing: 40.28%  
- Tools & Home Improvement: 16.67%  
- Toys & Games: 16.67%  
  
Top misclassified categories for 'Movies & TV':  
- Toys & Games: 71.43%  
- Tools & Home Improvement: 28.57%  
  
Top misclassified categories for 'All Beauty':  
- Health & Personal Care: 47.62%  
- Tools & Home Improvement: 33.33%  
- Toys & Games: 19.05%  
  
Top misclassified categories for 'Health & Personal Care':  
- All Beauty: 30.61%  
- Tools & Home Improvement: 23.47%  
- Arts, Crafts & Sewing: 13.27%
```

Figure 4.9: Skewness in Misclassification Errors

4.10 Error Distribution

An error distribution plot was created to examine the spread of misclassified predictions across different categories.

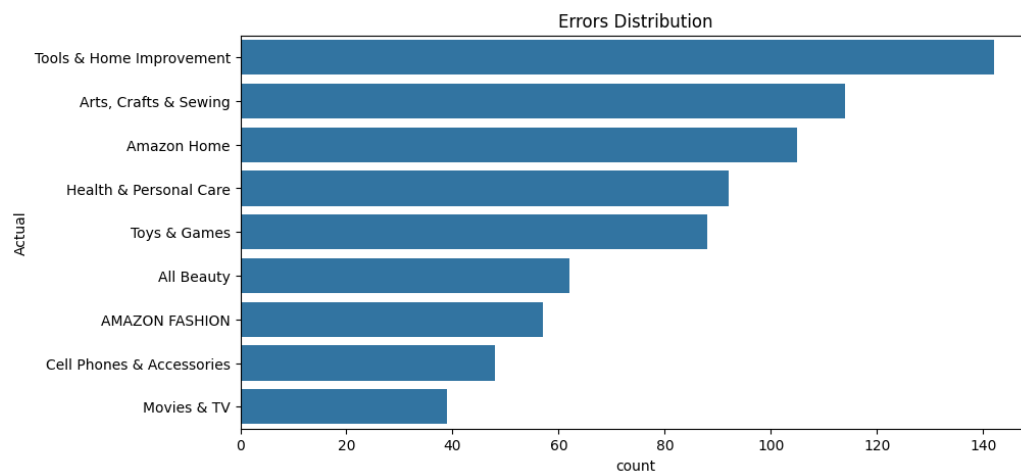


Figure 4.10: Error Distribution Across Categories

Chapter 5

Conclusion

The first half of the internship has been a transformative experience, providing valuable exposure to key concepts in **data preprocessing**, **Natural Language Processing (NLP)**, and **AI-driven classification models**. The structured approach—from dataset acquisition to feature engineering, exploratory data analysis (EDA), and baseline model implementation—has laid a strong foundation for the next phase of the project.

A key takeaway from this phase has been the *importance of data quality and feature representation in classification accuracy*. Experiments with different *feature extraction techniques (TF-IDF, embeddings, named entity recognition)* have provided insights into how textual attributes influence model performance.

Moving forward, the focus will shift towards *integrating more advanced AI techniques*, particularly transformer-based models such as **BERT** and **GPT**. These models will be leveraged to improve classification accuracy and adaptability.

This report marks the completion of a *mid-semester milestone*, but there is still much more to accomplish in the second half of the internship. The next phase will involve *refining the model architecture, implementing Generative AI techniques, and ensuring that the classification system meets industry standards for scalability and efficiency*.

Acknowledgments

I would like to express my heartfelt gratitude to my industry advisor, **Mr. Sanyam Jain**, and my internal advisor, **Ms. Saumya Bajpai**, for their invaluable guidance, mentorship, and continuous support throughout this internship. Their insightful feedback and expert advice have played a crucial role in shaping my understanding of AI-driven classification and its real-world applications.

I would also like to extend my sincere appreciation to my colleagues at **LatentView Analytics** for their collaborative spirit and willingness to share knowledge. Their expertise and constructive discussions have been instrumental in overcoming various technical and analytical challenges encountered during the project.

Finally, I am grateful for the opportunity to work on this exciting project, which has provided me with a platform to enhance my skills and contribute meaningfully to the world of AI and Data Science. Looking forward to an even more productive and enriching second half of the internship!

Thank You