

Analyzing Unstructured Data

Group 1

Client Name: Craigslist
06 – Dec – 2023

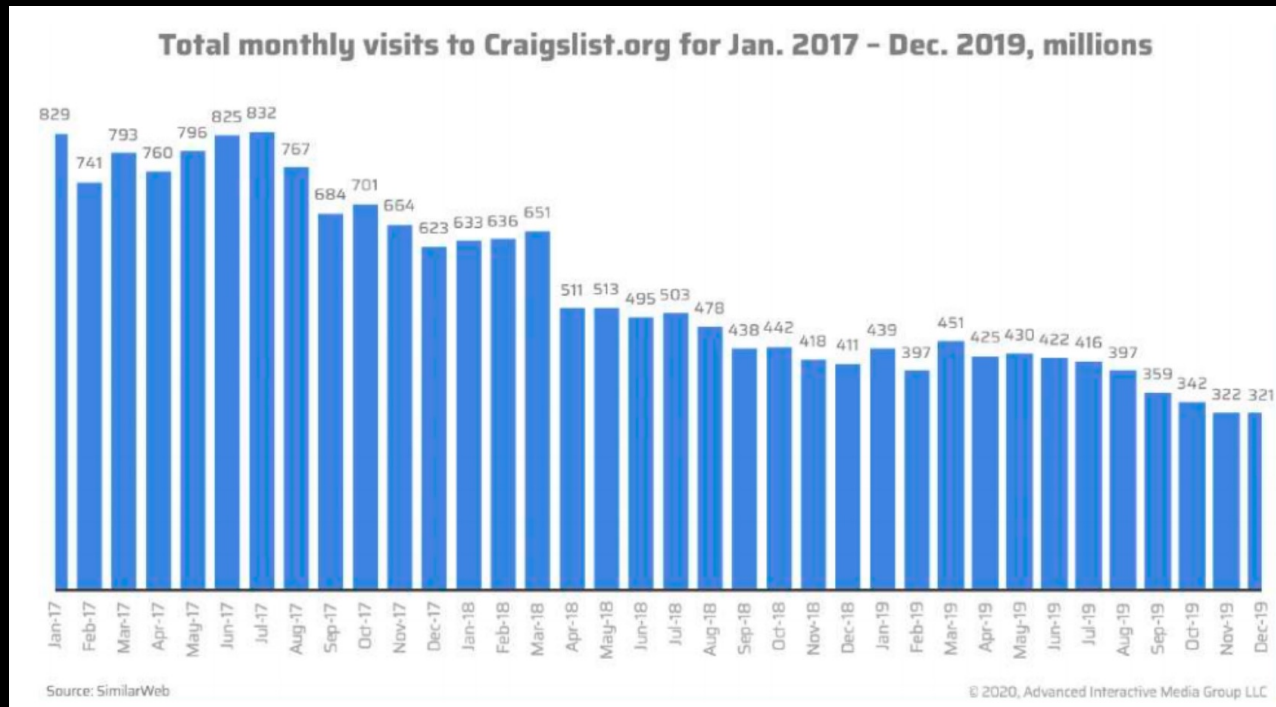


AGENDA

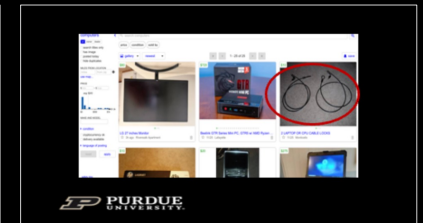
- Problem Understanding
- Solution Approach
 - Framework Overview
 - Framework Components
 - Text Classification
 - Image Classification
 - Final Heuristic (Image + Text Classification) model
- Conclusion
- Recommendations
- Future Scope

PROBLEM UNDERSTANDING

Optimizing Craigslist's categorization to accurately differentiate computer products and accessories for a more efficient and relevant user search experience

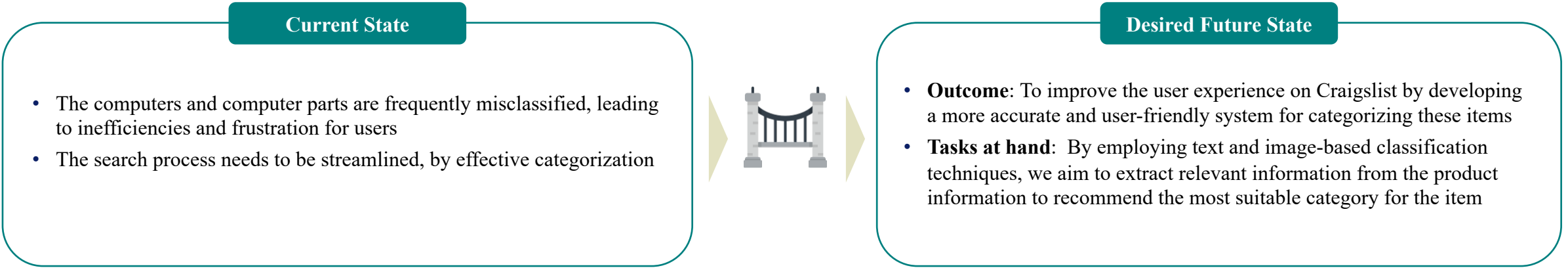


- More than **30%** of Computers are misclassified
- More than **15%** Computer Parts are misclassified



Current State & Desired Future State

Optimizing Craigslist's categorization to accurately differentiate computer products and accessories for a more efficient and relevant user search experience



What are we solving?



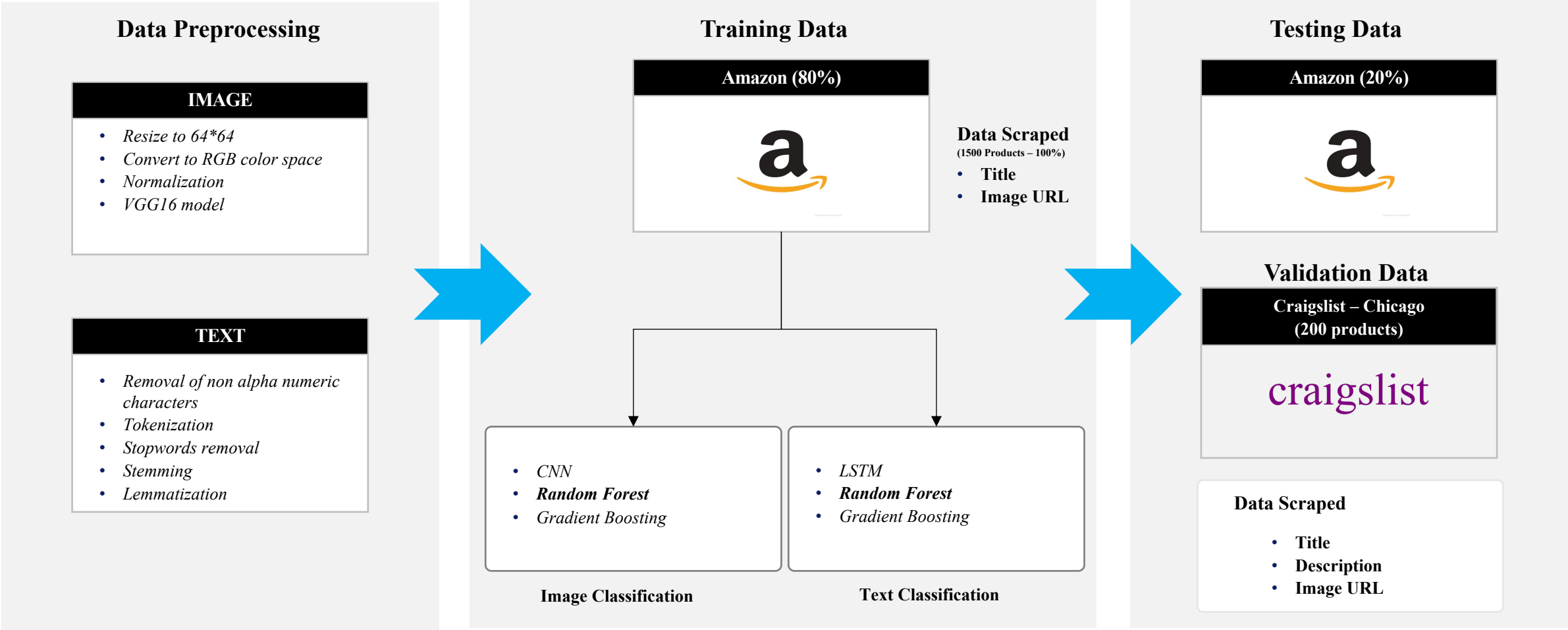
FRAMEWORK OVERVIEW

We propose to develop a heuristic model combining Text and Image Classification Techniques to effectively categorize computers and computer parts on Craigslist



Proposed Model Framework

Framework Skeleton



FRAMEWORK COMPONENTS

Image Classification

The 3-step process would be the base of the framework – training and testing on Amazon data and validating on Craigslist data

Image Preprocessing

1.Reshaping to 64x64 Pixels: Reshaping images to a fixed size (like 64x64) ensures uniformity in input dimensions, which is crucial for most machine learning models, including convolutional neural networks (CNNs).

2.Normalizing Pixel Values from 0 to 1: Normalization scales the pixel values to a range of 0 to 1. This process aids in the convergence of the model during training by ensuring that pixel values do not disproportionately influence the model’s learning.


3.Converting to RGB Color: Converting images to RGB ensures that all images are in the same color space. RGB (Red, Green, Blue) is a widely used color space in image processing.

CONVOLUTIONAL NEURAL NETWORK

Leverages layered processes and pattern recognition to effectively classify images through feature detection & spatial hierarchies

- 256 neurons and ReLU (Rectified Linear Unit) activation
- Dropout – 0.5
- Optimizer='adam'

- Validation Accuracy: 0.74**
- Misclassification Rate: 0.26**
- Precision: 0.74**
- Recall: 0.74**
- F1 Score: 0.71**

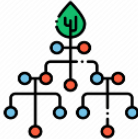


RANDOM FOREST

Classifies images by aggregating decisions from multiple decision trees, enhancing accuracy and robustness against overfitting

- max_depth: [10, 15, 20]
- n_estimators: [100, 150, 200]
- min_samples_split: [2,4,6]
- min_samples_leaf: [1,2,3]
- random_state=42: Ensures reproducibility

- Mean CV Score: 0.84**
- Validation Accuracy: 0.82**
- Precision: 0.83**
- Recall: 0.83**
- F1 Score: 0.831**




XG BOOST CLASSIFIER

It incrementally improves image classification by sequentially correcting errors from previous models, optimizing accuracy and performance

- n_estimators: [100, 150, 200]
- max_depth: [3, 6, 9]
- learning_rate: [0.01, 0.1, 0.2]
- subsample: [0.5, 0.7, 1.0]

- Validation Accuracy: 0.83**
- Precision: 0.85**
- Recall: 0.83**
- F1 Score: 0.84**



Text Classification

The 3-step process would be cleaning

Text Pre-Processing

Cleaning and processing the Text columns to remove unnecessary words



Amazon Dataset

Data Cleaning & Merging

- Removing Non alpha numeric characters and emojis
- Tokenization
- Removal of stopwords
- Stemming
- Lemmatization

Training & Validation

We tried different classification models



Craigslist Data : Out of Sample Validation

Model Building

- Combined Text and Description Columns
- Ran multiple classifier models
- TF-IDF Vectorization

Text Classification Models : Random Forests

Classification Models

Predicting product class using different models

Random Forests

n_estimators = 100
Random_Seed = 42
Min_samples_leaf = 1
Min_samples_Split = 2

Test Accuracy: 88%

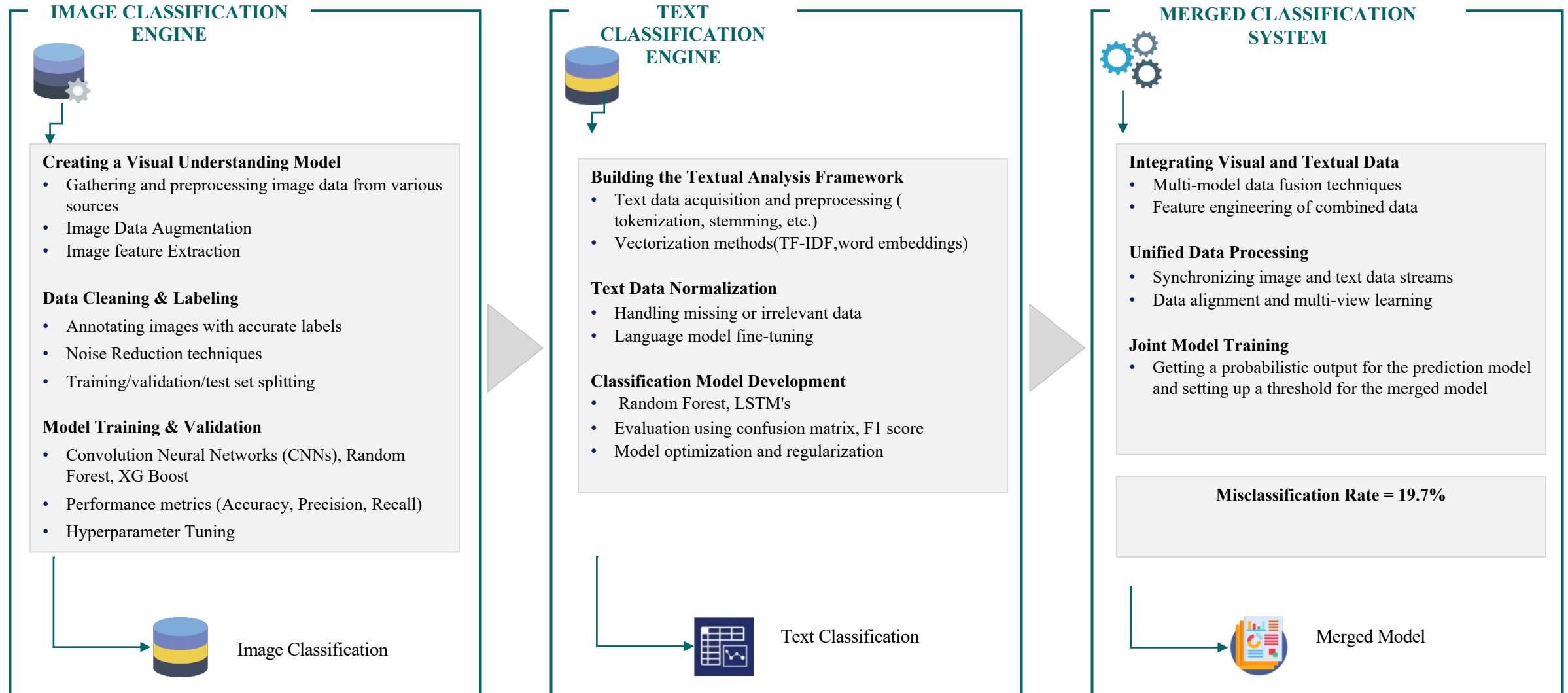
LSTM

Activation = soft_max
Input_dim = 5000
Neurons: 100

Test Accuracy: 90%

Final (Image + Text Classification)

The 3-step process would be the base of the framework - ingesting data, analyzing reports and recommend suitable actions



Value Generated



Refining Categorization

By deploying sophisticated algorithms, our aim transcends mere correct placement of listings; we seek to ensure these listings are easily discoverable by potential buyers, thereby enhancing the overall marketplace efficiency.



Optimizing the User Journey

A user's journey, spanning from the moment of listing an item to finding the desired product, should be a streamlined and intuitive experience. Presently, this journey is often hindered by issues like the misplacement of computer accessories or the appearance of unrelated products in specific search queries. Our project directly addresses these challenges, seeking to provide a seamless and effective process for both buyers and sellers.



Technical Implementation: Advanced Categorization and Tagging

The dual-pronged approach is designed to intelligently suggest the most fitting categories for listings while generating relevant tags to improve searchability and visibility.



Data Analysis: The Foundation of Intelligent Categorization

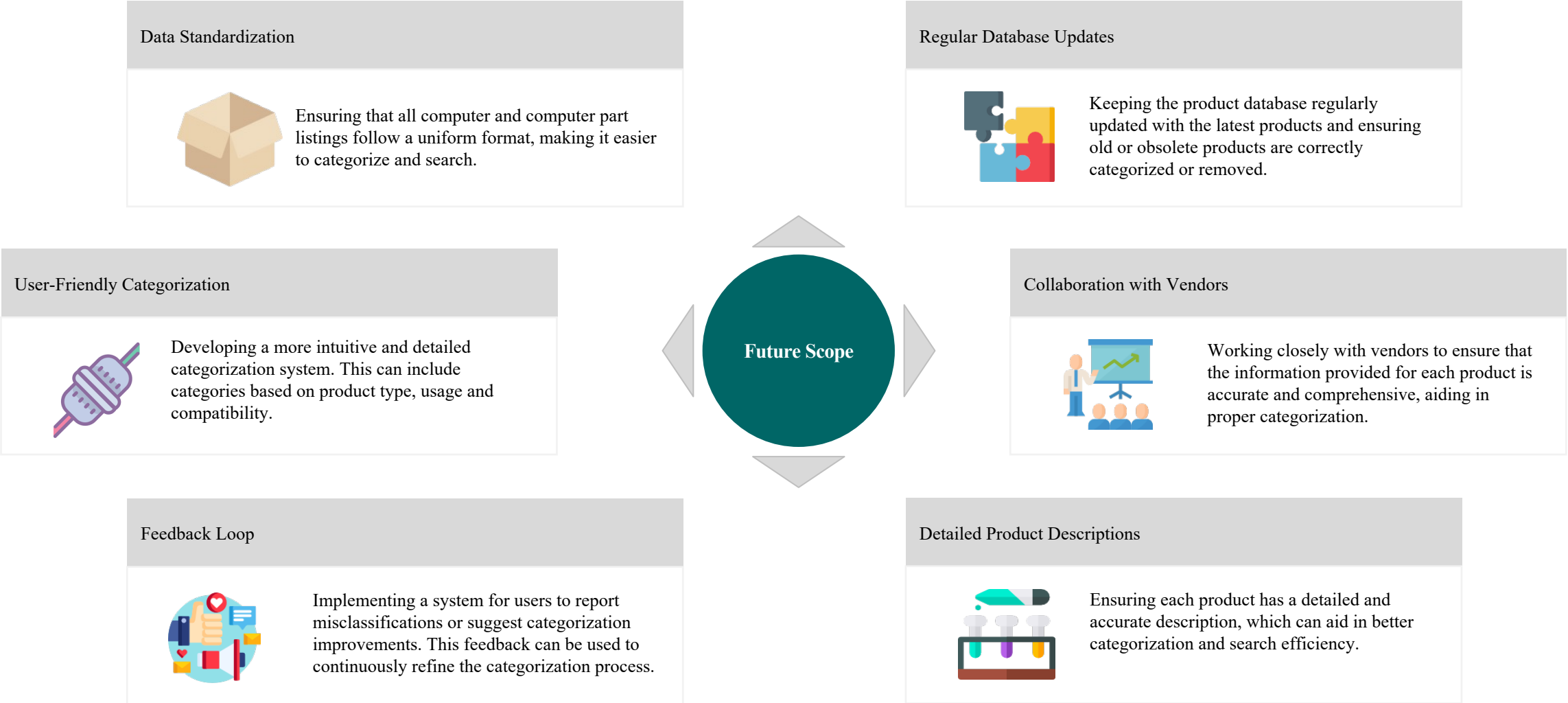
This data forms the backbone of our categorization algorithms, ensuring the system is not only theoretically robust but also practically attuned to the real content of the listings.



Addressing Classified Ad Challenges

By improving the precision of our categorization, we aim to ensure each ad reaches its intended audience more effectively, reducing inefficiencies and enhancing the likelihood of successful transactions.

Recommendations | Future Scope



Thank you!

All the best for your final exams!



Q/A



Appendix

Framework Components

The framework consists of Text and Image Classification techniques

Data Sources identified by SABIC



We also recommend using

Socialgist

Socialgist provides a search API that can provide data from various types of sources like News articles, Blogs, Review boards, and forums and has access to *over 2000+ sources*

Some of their key data partners include **tumblr, Disqus, Tencent, Weibo, Vertical Scope** etc.

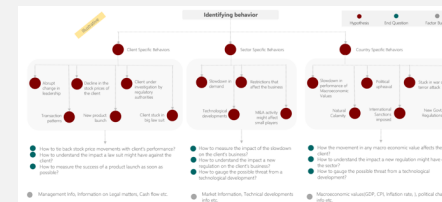
BUSINESS INTELLIGENCE



AoPST™

Using Mu Sigma's AoPST™ framework allows us to identify the key factors that affect market demand and price changes to **design the hypotheses**

Behavioral Patterns



MARKET INTELLIGENCE MODULE

- Determine key competitors across regions
- Determine product offerings and market share
 - Range of products
 - Applications of the product range, grades etc. (based on rigidity and flexibility index)
- Determine sales and pricing strategies
- Analyze customs regulation changes
 - Review Boards and Forums information from SocialGist
 - Chatter on Twitter and SocialGist to identify latest sentiment around policy changes
- Analyze product market trends and results based on region
- Perform SWOT Analysis to learn competitor strengths, weaknesses, opportunities, and threats

MA and CS Scoring

The MA and CS scoring will be done by performing **hypotheses-based testing** post **business interviews** and identifying behavioral patterns using the **AoPST™** framework

CL madison for sale computers post account

computers

all owner dealer

☐ search titles only
☐ has image
☐ posted today
☐ hide duplicates

MILES FROM LOCATION
miles from zip
use map...

PRICE
\$ min - \$ max
avg: \$214

MAKE AND MODEL

condition

☐ delivery available
☐ cryptocurrency ok

language of posting

reset apply

price condition sold by

gallery

1 - 44 of 44

save

\$500

Samsonite Aluminum Computer Laptop Briefcase
☆ 1h ago DeForest

\$1,500

MacBook Air 15" Laptop - M2 chip - 8GB Memory - ...
☆ 6h ago

\$450

HP Pavilion ze5700 Laptop
☆ 11/26 Madison

\$500

\$336

\$316

computers

all owner dealer

search titles only

has image

posted today

hide duplicates

MILES FROM LOCATION

miles

from zip

use map...

PRICE

\$min

–

\$max

avg: \$243

MAKE AND MODEL

condition

cryptocurrency ok

delivery available

language of posting

reset

apply

search computers

price

condition

sold by

gallery

newest

1 - 25 of 25

save

\$80

LG 27 inches Monitor

☆ 3h ago Riverwalk Apartment

\$720

Beelink GTR Series Mini PC, GTR5 w/ AMD Ryzen ...

☆ 11/25 Lafayette

\$10

2 LAPTOP OR CPU CABLE LOCKS

☆ 11/25 Monticello

\$10

\$20

\$275

PURDUE
UNIVERSITY®