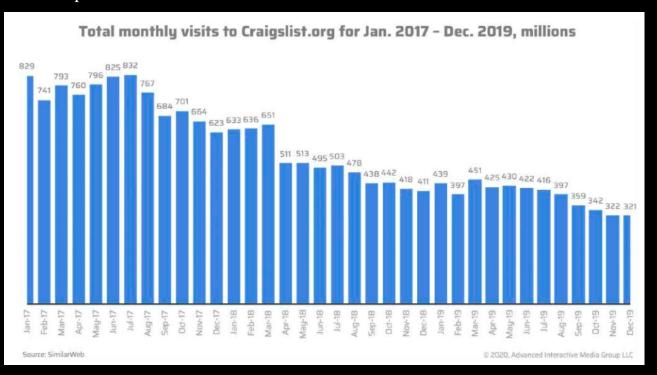
# **AGENDA**

- 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

#### **Current State**

- The computers and computer parts are frequently misclassified, leading to inefficiencies and frustration for users
- The search process needs to be streamlined, by effective categorization



#### **Desired Future State**

- Outcome: To improve the user experience on Craigslist by developing a more accurate and user-friendly system for categorizing these items
- Tasks at hand: By employing text and image-based classification techniques, we aim to extract relevant information from the product information to recommend the most suitable category for the item

What are we solving?



**Increased Search Time** 



**Purchase of Incorrect Items** 



**Increased Customer Support Queries** 



**Frustration & User Experience Deterioration** 





**Lost Sales for Retailers** 



**Inefficiency in Comparison Shopping** 

# 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**

#### **Testing Data Data Preprocessing Training Data Amazon (80%) Amazon (20%) IMAGE** • Resize to 64\*64 **Data Scraped** (1500 Products – 100%) • Convert to RGB color space • Title Normalization Image URL VGG16 model **Validation Data** Craigslist - Chicago (200 products) TEXT • Removal of non alpha numeric craigslist characters Tokenization Stopwords removal • CNN • LSTM • Stemming • Random Forest • Random Forest Lemmatization **Data Scraped** • Gradient Boosting • Gradient Boosting Title Description Image URL **Image Classification Text Classification**

# 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

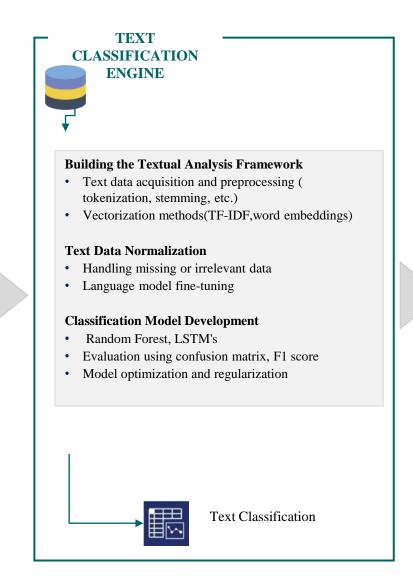
Test Accuracy: 90%

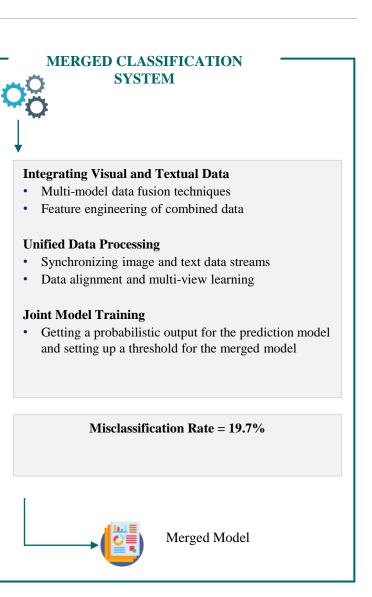
## **Final (Image + Text Classification)**



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

# IMAGE CLASSIFICATION **ENGINE Creating a Visual Understanding Model** • Gathering and preprocessing image data from various sources Image Data Augmentation • Image feature Extraction **Data Cleaning & Labeling** • Annotating images with accurate labels · Noise Reduction techniques • Training/validation/test set splitting **Model Training & Validation** • Convolution Neural Networks (CNNs), Random Forest, XG Boost • Performance metrics (Accuracy, Precision, Recall) • Hyperparameter Tuning **Image Classification**







### 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**



#### Data Standardization



Ensuring that all computer and computer part listings follow a uniform format, making it easier to categorize and search.

#### User-Friendly Categorization



Developing a more intuitive and detailed categorization system. This can include categories based on product type, usage and compatibility.

# Future Scope

#### Regular Database Updates



Keeping the product database regularly updated with the latest products and ensuring old or obsolete products are correctly categorized or removed.

#### Collaboration with Vendors



Working closely with vendors to ensure that the information provided for each product is accurate and comprehensive, aiding in proper categorization.

#### Feedback Loop



Implementing a system for users to report misclassifications or suggest categorization improvements. This feedback can be used to continuously refine the categorization process.

#### **Detailed Product Descriptions**



Ensuring each product has a detailed and accurate description, which can aid in better categorization and search efficiency.

# Appendix



