



# Analyzing Unstructured Data

## Project Report



### Group Members

Ashvin Raj, Hanbing Yang, Monika Madugula, Pratik Borkar, Samridhi Vads, Zeeshan Husein

## About Craigslist

Craigslist, an American online platform for classified advertisements, encompasses sections dedicated to jobs, housing, sales, wanted items, services, community engagement, gigs, résumés, and discussion forums. It stands as one of the largest user-generated advertisement websites, spanning 570 cities across 70 countries.

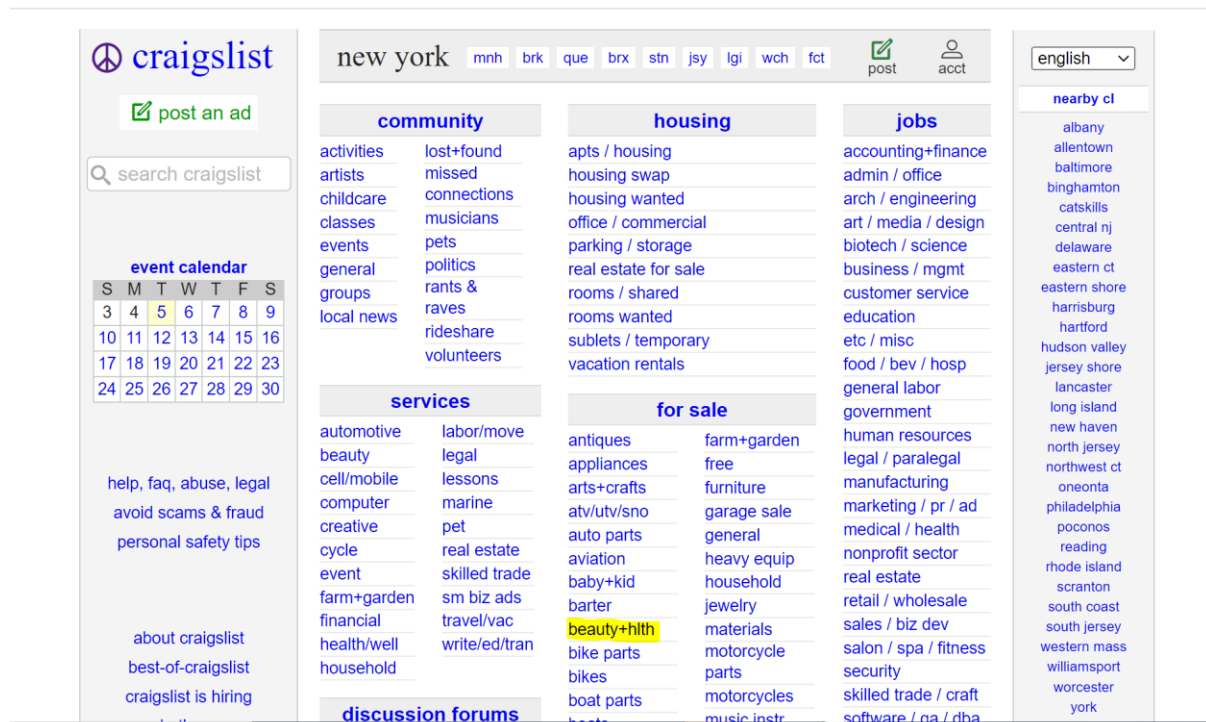
The site primarily caters to advertisers, providing them with a user-friendly interface to publish flexible and loosely formatted ads, commonly displayed as text descriptions and/or uploaded images.

**Website:** <https://newyork.craigslist.org/>

The Craigslist website, in contrast to eBay, is less visually attractive and more challenging to navigate due to its text-dense and text-heavy layout. This can make it more difficult for users to spot individual listings. On the other hand, eBay's platform is more oriented towards images and has listings that are sorted into categories, enhancing visibility.

## Business Problem

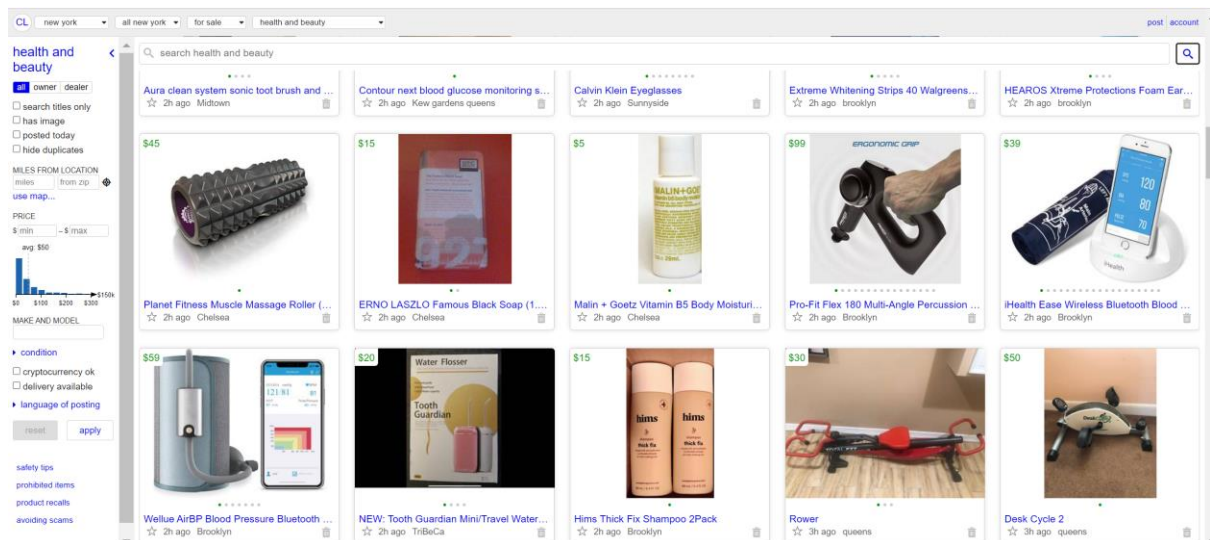
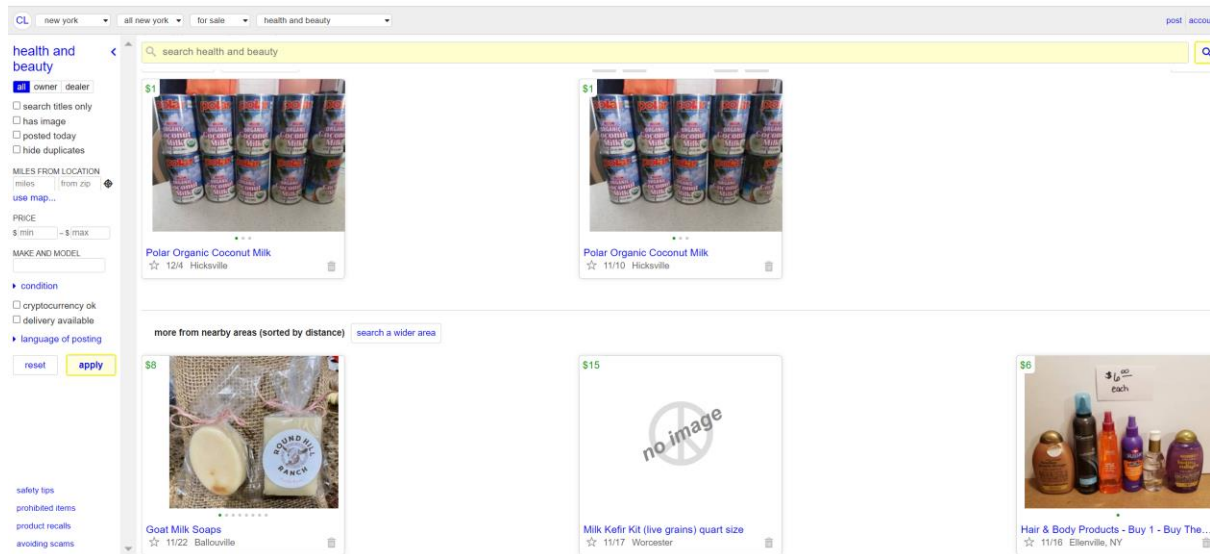
Our primary goal is to delve into the beauty and health section featured on Craigslist, specifically within the New York segment. This choice was made deliberately, leveraging New York's extensive array of data points. By concentrating on this section, we aim to thoroughly explore and analyze the different offerings within the beauty and health domain on Craigslist in one of its most active and diverse regions.



The screenshot displays the Craigslist New York homepage. The header includes the Craigslist logo, a 'post an ad' button, and a search bar. Below the header, there's a navigation bar with city links (new york, mnh, brk, que, brx, stn, jsy, lgi, wch, fct) and user options (post, acct). The main content area is divided into several columns of category links. The 'community' column lists activities, artists, childcare, classes, events, general, groups, and local news. The 'housing' column lists apts / housing, housing swap, housing wanted, office / commercial, parking / storage, real estate for sale, rooms / shared, rooms wanted, sublets / temporary, and vacation rentals. The 'jobs' column lists accounting+finance, admin / office, arch / engineering, art / media / design, biotech / science, business / mgmt, customer service, education, etc / misc, food / bev / hosp, general labor, government, human resources, legal / paralegal, manufacturing, marketing / pr / ad, medical / health, nonprofit sector, real estate, retail / wholesale, sales / biz dev, salon / spa / fitness, security, skilled trade / craft, and software / qa / dba. The 'services' column lists automotive, beauty, cell/mobile, computer, creative, cycle, event, farm+garden, financial, health/well, and household. The 'for sale' column lists antiques, appliances, arts+crafts, atv/utv/sno, auto parts, aviation, baby+kid, barter, beauty+hlth (highlighted in yellow), bike parts, bikes, boat parts, farm+garden, free, furniture, garage sale, general, heavy equip, household, jewelry, materials, motorcycle, parts, motorcycles, and music instr. The 'discussion forums' column lists activities, artists, childcare, classes, events, general, groups, and local news. On the right side, there's a 'nearby cl' dropdown menu showing a list of nearby cities including albany, allentown, baltimore, binghamton, catskills, central nj, delaware, eastern ct, eastern shore, harrisburg, hartford, hudson valley, jersey shore, lancaster, long island, new haven, north jersey, northwest ct, oneonta, philadelphia, poconos, reading, rhode island, scranton, south coast, south jersey, western mass, williamSPORT, worcester, and york.

The beauty and health section presents a significant level of disarray, showcasing a diverse range of products that span from weighing machines to shaving trimmers. Furthermore, amidst this assortment, unrelated items like coconut milk are interspersed on the same page. This lack of categorization and coherence creates a challenging navigational experience for users perusing the website's offerings in this section.

Additionally, the number of filters on this section is also very generic and not very helpful. A customer looking to shop for make-up or skin care for example will also see unrelated things such as blood pressure monitor or muscle massage roller

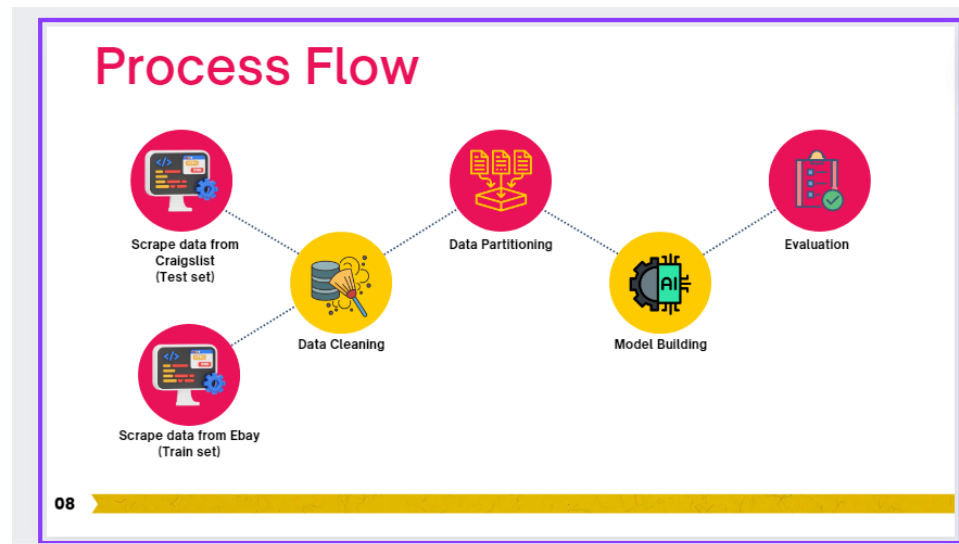


## Project objective

Our project aims to introduce descriptive tags to categorize products based on their titles and descriptions. These tags will serve as effective filters on the website, simplifying the search for

particular items. Additionally, this system will enable users to conduct comparisons among items within the same category. By doing so, it will provide users with essential details to help them make informed decisions when selecting the most suitable product.

## Process Flow Diagram



In this project, our initial stage focused on developing a robust training dataset. We aimed to achieve this by extracting data from eBay listings. Due to issues with scraping data from eBay we manually labeled our training data which we scraped from craigslist website (more mentioned in data collection part)

Following the compilation of our training set, we shifted our attention to gathering a test dataset from Craigslist. This dataset is critical for objectively assessing our model's performance later in the process, ensuring that our evaluations reflect real-world application and effectiveness.

Data integrity is paramount; hence, we dedicated a significant effort to refining our datasets. The cleaning process involved standardizing the data format, rectifying inconsistencies, and treating any missing or anomalous data points. This step was crucial to ensure the quality and reliability of our datasets before further processing.

Subsequently, we approached the challenge of converting our textual data into a structured format that our machine learning algorithms could interpret. This text representation and processing phase included transforming raw text into numerical values that encapsulate the essence of the information contained within the text, ensuring our model can efficiently learn from the data.

With our data prepared, we embarked on the model building phase. Here, we selected appropriate algorithms and employed the training set to construct a predictive model. Our methodology was guided by the principles of machine learning, seeking to develop a model that not only learns from the data but also generalizes well to new, unseen information.

The culmination of our process was the evaluation stage, where the model was subjected to the test set sourced from Craigslist. The evaluation criteria involved AUC, ROC and validation accuracy

## Data Collection

The first step in our project was to scrape data for the beauty and health section in the craigslist website.

### Gathering details about the products:

Utilizing Selenium and BeautifulSoup, we gathered the URLs of product pages specifically within the website's health and beauty section. This extensive task involved extracting approximately 1680 product URLs spread across 19 distinct pages on the site. Our process involved techniques to systematically collect and compile these links, enabling comprehensive coverage of the available products within this specific section of the website.

Subsequently, we extracted more detailed information from the product pages, including product descriptions and distinctive Posting IDs. Furthermore, we concentrated on key factors that might impact the analysis, such as Post\_date, multi\_ads\_user, latitude, longitude.

The entire dataset, encompassing both the influencing variables and collected data, was structured and compiled into CSV files for improved organization and ease of analysis.

```
def product_details(url_main):
    title = soup.find('span', {'id': 'titletextonly'})
    if title != None:
        title_temp=(title.get_text(strip=True)).replace(',','')
        product_title.append(title_temp)
    else:
        product_title.append(0)
    prod_price=soup.find('span',{'class':'price'})
    if prod_price!= None:
        price.append(int(prod_price.get_text(strip=True).split('$')[1].replace(',','')))
    else:
        price.append(0)
    Post_info=soup.find('div',{'class':"postinginfos"})
    if Post_info!=None:
        Post_id=Post_info.find('p',{'class': "postinginfo"})
        if Post_id != None:
            post_id.append(int(Post_id.get_text(strip=True).split('post id: ')[1]))
        else:
            post_id.append(0)
        datetime_str = Post_info.find('time')['datetime']
        if datetime_str!= None:
            parsed_datetime = datetime.strptime(datetime_str, "%Y-%m-%dT%H:%M:%S%z")
            post_date.append(parsed_datetime)
        else:
            post_date.append(0)
    else:
        post_id.append(0)
        post_date.append(0)
    post_body=soup.find('section',{'id':"postingbody"})
    if post_body!= None:
        desc_temp=(post_body.get_text(strip=True).split('Post')[1]).replace(',','')
        description.append(desc_temp)
    else:
        description.append(0)
```

product_title	Label	Post_ID	Post_date	description	multi_ads	latitude	longitude
PHILIPS Respironics Breathing Gadget	Health Equipments	7.69E+09	2023-11-13	Like new. Gently used. No problem	1	40.57962	-74.0037
Foam Roller 36" white	Health Equipments	7.69E+09	2023-11-25	The white longer roller left. Great c	0	40.6521	-74.0018
Kolua Wax (large package for hair removal)	Skin Care & Makeup	7.68E+09	2023-11-02	Full. Tried it once. I'd rather go to a	0	40.6521	-74.0018
The Yoga Deck: 50 Poses & Meditations for	Other Health Care	7.69E+09	2023-11-08	The Yoga Deck: 50 Poses & Meditat	0	40.6521	-74.0018
Medical Rolling Portable Folding Adult Mobi	Health Equipments	7.69E+09	2023-11-25	Good conditionComes from a pet fi	0	40.62573	-73.9564
Calvin Klein Eyeglasses	Vision Care	7.69E+09	2023-11-15	Used in very good good condition f	0	40.7416	-73.9238
Sit N cycle Exercise Bike	Health Equipments	7.69E+09	2023-11-18	Sit N cycle by Smooth Fitness exerci	1	40.6588	-73.8438
Door Doorway Frame Mount Pull Up Exercis	Health Equipments	7.69E+09	2023-11-25	Iron Gym Proxifit fitness bars for a c	0	40.62573	-73.9564
New in Box - L'occitane Verbena EDT - 3.4 oz	Fragrances	7.69E+09	2023-11-20	New in box and unused L'occitane	0	40.6424	-73.9758
2 New Adult Girls Womans Blond White Dar	Hair Care	7.69E+09	2023-11-25	\$15 each. 2 left - golden and pinkLo	0	40.62575	-73.9564
Curling Iron XTAVA \$18	Hair Care	7.69E+09	2023-11-25	Xtava curling wandBox was lost bet	0	40.6816	-73.9798
Root Branch and Blossom	Skin Care & Makeup	7.68E+09	2023-10-29	If purchased individually Body Refr	0	40.7807	-73.7812
Makes Enuresis Alarm	Other Health Care	7.69E+09	2023-11-18	Made in England.	0	40.7807	-73.7812
Conair Double Ceramic 1.5" Flat Iron Electric	Hair Care	7.69E+09	2023-11-25	Used but not abused. Works wellSe	0	40.62573	-73.9564
Covidien Kangaroo Gastrostomy Feeding Tu	Health Equipments	7.69E+09	2023-11-07	REFshow contact infoEFeaturesWit	0	40.7229	-73.8473
Acupuncture	Other Health Care	7.69E+09	2023-11-25	Acupuncture is using a needle to ac	1	40.7443	-73.9781
Cloud Massage Shiatsu Foot Massager	Health Equipments	7.69E+09	2023-11-25	This is a fantastic foot massager. A	0	40.7975	-73.9683
Facial Steamer	Skin Care & Makeup	7.68E+09	2023-11-06	Facial steamer used in spas and it v	1	40.7443	-73.9781
Estee Lauder Skincare Makeup Lot "BRAND	Skin Care & Makeup	7.68E+09	2023-10-29	BRAND NEWSEALED	0	40.6011	-73.9475
Caudalie Resveratrol Lift Anti Wrinkle Firmi	Skin Care & Makeup	7.69E+09	2023-11-08	Brand newNever openedHave a rec	0	40.6011	-73.9475
NEW Ray-Ban aviator RB-3625 58mm blue le	Vision Care	7.68E+09	2023-10-29	NEWNEVER USED	0	40.6011	-73.9475
Ray-Ban RB3664CH Chromance Polarized M	Vision Care	7.69E+09	2023-11-21	perfect like new conditionno scratc	0	40.6011	-73.9475
Dior Sauvage Men's Parfum Spray LARGE 20	Fragrances	7.68E+09	2023-10-29	brand newfull bottlewithout box	0	40.6011	-73.9475
Ray-Ban aviator RB-3025 Authentic RARE BL	Vision Care	7.69E+09	2023-11-25	Lenses are in perfect conditionChee	0	40.6011	-73.9475
Ray-Ban aviator RB-3666 Gold 56mm Polariz	Vision Care	7.69E+09	2023-11-21	perfect like new conditionno scratc	0	40.6011	-73.9475
Professional Permanent Makeup Machine K	Skin Care & Makeup	7.69E+09	2023-11-14	brand newnever used	0	40.6011	-73.9475
Sonic Electric Toothbrush BRAND NEW SEAL	Health Equipments	7.69E+09	2023-11-10	BRAND NEW SEALED	0	40.6011	-73.9475
Prada PARADOXE Parfum 90ml. Brand New	Fragrances	7.69E+09	2023-11-08	Brand new sealedHave a receipt fr	0	40.6011	-73.9475
Valentino Donna Born in Roma Eau de Parfu	Fragrances	7.69E+09	2023-11-11	Brand new sealedRegular or ENTEE	0	40.6011	-73.9475
Digital Upper Arm Blood Pressure Monitor B	Health Equipments	7.69E+09	2023-11-25	Brand newSealed	0	40.6011	-73.9475
BIOSWISS VENTED QUICK DRY BRUSH #9175	Hair Care	7.69E+09	2023-11-23	NEW OLD STOCKPROBABLY PURCHA	1	40.6548	-73.6097

## Scraped Dataset

### Label Identification

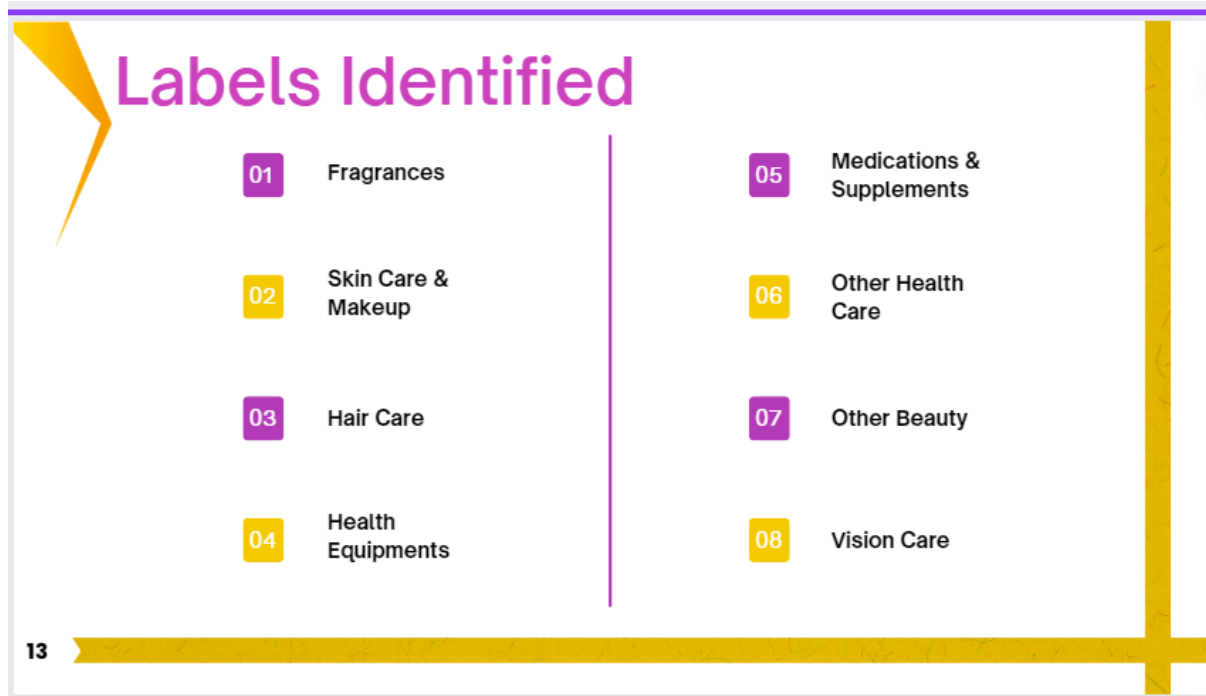
After the data collection phase, our subsequent task involved categorizing the acquired products by assigning specific labels. To achieve this, we initially outlined various label categories corresponding to different product types, which included:

1. Fragrances
2. Skin Care & Makeup
3. Hair Care
4. Health Equipments
5. Medications & Supplements
6. Other Health Care
7. Other Beauty
8. Vision Care

Initially, our plan was to extract data from eBay to train our model. However, we encountered challenges during the scraping process (insert details about the specific scraping issues faced here).



Subsequently, we opted to manually categorize the data based on the product descriptions into these eight defined categories. This approach was chosen to ensure accurate labeling despite the initial setbacks encountered during the scraping process.



## Model Identification

For our modeling part we tried SVM, Logistic Regression as well as LSTM model.

## LSTM Model

Text, by its very nature, is a form of sequential data, making LSTM an appropriate choice for many text-related applications, including classification.

We started off by doing some data cleaning, preprocessing the text by converting it to lowercase, removing punctuation and symbols, and eliminating stop words (common words that do not carry much meaning)

The next step was tokenizing and padding the document. The labels are converted into a one-hot encoded format, suitable for multi-class classification. The dataset is then split into a training set (70%) and a testing set (30%).

The model is then built including an embedding layer, a spatial dropout layer (to reduce overfitting), an LSTM layer, and a dense output layer with a softmax activation function for multi-class classification.

```

model = Sequential()

model.add(Embedding(50000, 100, input_length=X.shape[1]))

model.add(SpatialDropout1D(0.2))

model.add(LSTM(100, dropout=0.2, recurrent_dropout=0.2))

model.add(Dense(len(Y[0]), activation='softmax'))

model.compile(loss='categorical_crossentropy', optimizer='adam', metrics=['accuracy'])

```

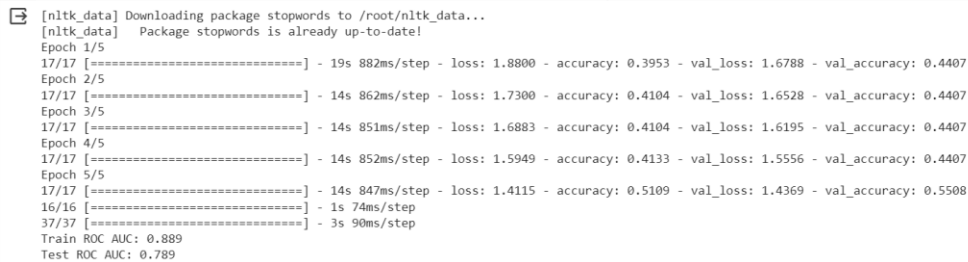
The model is trained on the training data for a specified number of epochs and batch size. Early stopping is used to prevent overfitting.

Finally, the model's performance is evaluated on the test set, and the loss and accuracy are printed.

```

print(f'Train ROC AUC: {roc_auc_train:.3f}')
print(f'Test ROC AUC: {roc_auc_test:.3f}')

```



A terminal window showing the execution of a Python script. It starts with a message from the 'nltk\_data' package indicating it is downloading stopwords. The output then shows the training progress for 5 epochs, with metrics like loss, accuracy, and val\_loss printed for each epoch. The final output shows the Train ROC AUC as 0.889 and the Test ROC AUC as 0.789.

```

[nltk_data] Downloading package stopwords to /root/nltk_data...
[nltk_data] Package stopwords is already up-to-date!
Epoch 1/5
17/17 [=====] - 19s 882ms/step - loss: 1.8800 - accuracy: 0.3953 - val_loss: 1.6788 - val_accuracy: 0.4407
Epoch 2/5
17/17 [=====] - 14s 862ms/step - loss: 1.7300 - accuracy: 0.4104 - val_loss: 1.6528 - val_accuracy: 0.4407
Epoch 3/5
17/17 [=====] - 14s 851ms/step - loss: 1.6883 - accuracy: 0.4104 - val_loss: 1.6195 - val_accuracy: 0.4407
Epoch 4/5
17/17 [=====] - 14s 852ms/step - loss: 1.5949 - accuracy: 0.4133 - val_loss: 1.5556 - val_accuracy: 0.4407
Epoch 5/5
17/17 [=====] - 14s 847ms/step - loss: 1.4115 - accuracy: 0.5109 - val_loss: 1.4369 - val_accuracy: 0.5508
16/16 [=====] - 1s 74ms/step
37/37 [=====] - 3s 90ms/step
Train ROC AUC: 0.889
Test ROC AUC: 0.789

```

We also tried to tweak a bit with the embedding method and model settings to try to see how it improves the model score. One major change was using pre-trained glove embeddings (specifically, the 50, 100, 150 and 200-dimensional version).

Embedding layer is configured with `weights=[embedding_matrix]` and `trainable=False`. This means the GloVe embeddings are used as-is and are not further trained during the model training process.

We also tried The Bidirectional wrapper. A bidirectional LSTM processes the text in both forward and backward directions (as opposed to the standard LSTM which processes text in one direction).



## Support Vector Machine

Accuracy: 0.6666666666666666

Classification Report:

	precision	recall	f1-score	support
0	1.00	0.67	0.80	18
1	0.81	0.69	0.74	51
2	0.72	0.90	0.80	229
3	0.50	0.38	0.43	37
4	0.48	0.24	0.32	41
5	0.43	0.29	0.35	69
6	0.56	0.70	0.62	50
7	0.71	0.56	0.63	9
accuracy			0.67	504
macro avg	0.65	0.55	0.59	504
weighted avg	0.65	0.67	0.65	504

Number: 0, Label: Fragrances  
Number: 1, Label: Hair Care  
Number: 2, Label: Health Equipments  
Number: 3, Label: Medications & Supplements  
Number: 5, Label: Other Health Care  
Number: 6, Label: Skin Care & Makeup  
Number: 7, Label: Vision Care

```
[26] # Retrieve the best parameters from the grid search
best_params = grid_search.best_params_
print("Best Parameters:", best_params)

Best Parameters: {'C': 100, 'gamma': 0.01}

# Create and train the SVM model with the best parameters
tuned_svm_model = SVC(kernel='rbf', C=best_params['C'], gamma=best_params['gamma'])
tuned_svm_model.fit(X_train, y_train)

# Predict and evaluate the model
y_pred = tuned_svm_model.predict(X_test)
print("Accuracy:", accuracy_score(y_test, y_pred))
print("\nClassification Report:\n", classification_report(y_test, y_pred))
```

Text Preprocessing: Our initial step involved standardizing the text data by converting all characters to lowercase, tokenizing the text into individual terms, removing common words that add little

informational value (stopwords), and lemmatizing words to their base or dictionary form. This process ensures a consistent and analyzable format for our textual data.

**Label Encoding:** We then applied Label Encoding to transform categorical labels within our 'Label' column into a numeric format. This conversion is essential because machine learning models inherently require numerical input for processing and pattern recognition.

**Word Weighting by TF-IDF:** Following the encoding, we advanced to the word weighting stage, employing the TF-IDF (Term Frequency-Inverse Document Frequency) technique. This method evaluates how important a word is to a document within a collection of documents, thereby converting the preprocessed text data into a set of meaningful numerical features.

**Model Training:** With the features prepared, we proceeded to the construction and training of the SVM model. We opted for a linear kernel due to its effectiveness in high-dimensional spaces, which is often the case with text data. The command `svm_model = SVC(kernel='linear')` was executed to instantiate and fit the model.

**Model Evaluation:** Finally, we assessed the performance of our SVM model using the test data, achieving an accuracy of 0.66. This metric reflects the proportion of correctly predicted instances and is a fundamental indicator of our model's effectiveness.

## Logistic Regression

**Text Preprocessing:** First, we made sure our text data was in a neat and understandable format. We changed all the letters to lowercase, split sentences into individual words, removed common words/stop words and simplified words to their basic form (lemmatization)

**Word Weighting TF-IDF:** We then converted our data into numerical format, using the method called TF-IDF, which basically gives each word a score based on how important it is in our text.

**Building the Model:** With our data ready, we built our Logistic Regression model. We fine-tuned our Logistic Regression model by carefully setting hyperparameters to ensure optimal performance. To address any potential imbalances between classes, we implemented a balanced class weight strategy. This approach gives more importance to minority classes, preventing them from being overshadowed by dominant ones. Additionally, to seamlessly handle multi-label classification scenarios, we encapsulated our Logistic Regression model within a `OneVsRestClassifier`. This ensemble technique enables us to manage multiple classification tasks effectively, making our model versatile and flexible

**Evaluation:** Finally, we tested our model with some test data. It performed with an accuracy of 0.58, which was not as high as the SVN model.

C	solver	max_iter	penalty	Accuracy	random_state
1	liblinear	10000	l2	0.5565	0
0.05	newton-cg	10000	l2	0.5595	0
1	liblinear	10000	l1	0.3988	0
1	sag	5000	l2	0.5625	0
0.1	lbfgs	5000	l2	0.5595	0
0.05	sag	5000	None	0.5416	0

```

names_list = []
descriptions_list = []
categories_list = []
column_names = ['Name', 'Description', 'Category']
df2 = pd.DataFrame(columns = column_names)
for n,d,c in zip(df['product_title'], df['description'], df['Label']):
    names_list.append(n)
    descriptions_list.append(d)
    categories_list.append(c)
df2['Name'] = names_list
df2['Description'] = descriptions_list
df2['Category'] = categories_list
df2

```

	Name	Description	Category
0	PHILIPS Recprionics Breathing Gadget	Like new. Gently used. No problems at all. Cle...	Health Equipments
1	Foam Roller 36" white	The white longer roller left. Great condition...	Health Equipments
2	Kolua Wax (large package for hair removal)	Full. Tried it once. I'd rather go to a spa bu...	Skin Care & Makeup
3	The Yoga Deck: 50 Poses & Meditations for Body...	The Yoga Deck: 50 Poses & Meditations for Body...	Other Health Care

Name	Description	Health Equipments	Skin Care & Makeup	Other Health Care	Vision Care	Fragrances	Hair Care	Medications & Supplements	Other Beauty	Information
philip recprionics breathe gadget	like new gently use problem clean neat ready use	1	0	0	0	0	0	0	0	philip recprionics breathe gadgetlike new gent...
foam roller white	white long roller leave great condition sunset.	1	0	0	0	0	0	0	0	foam roller whitewhite long roller leave great...
kolua wax large package hair removal	full try id rather go spa work great	0	1	0	0	0	0	0	0	kolua wax large package hair removalfull try i...
yoga deck pose meditation body mind spirit card	yoga deck pose meditation body mind spirit car...	0	0	1	0	0	0	0	0	yoga deck pose meditation body mind spirit car...
medical roll portable fold adult mobility walk...	good conditioncomes pet free smoke free home l...	1	0	0	0	0	0	0	0	medical roll portable fold adult mobility walk...

Accuracy: 0.5833333333333334					
	precision	recall	f1-score	support	
0	0.73	0.84	0.78	135	
1	0.80	0.63	0.71	52	
2	0.53	0.49	0.51	47	
3	0.50	0.17	0.25	6	
4	0.77	0.71	0.74	14	
5	0.62	0.57	0.59	28	
6	0.47	0.33	0.39	27	
7	0.56	0.19	0.28	27	
micro avg	0.68	0.63	0.65	336	
macro avg	0.62	0.49	0.53	336	
weighted avg	0.67	0.63	0.63	336	
samples avg	0.61	0.63	0.61	336	

We also tried working on a pre-trained model based on logistic regression. This model was computationally very extensive and we were not able to get results for it. Hence, we decided to train a model ourselves on the data we had scraped.

```

tokenizer = BertTokenizer.from_pretrained('bert-base-uncased')

def encode_data(tokenizer, texts, labels, max_length=512):
    input_ids = []
    attention_masks = []
    label_list = []

    for text, label in zip(texts, labels):
        encoded_data = tokenizer.encode_plus(
            text,
            add_special_tokens=True,
            max_length=max_length,
            padding='max_length',
            truncation=True,
            return_attention_mask=True,
            return_tensors='pt'
        )
        input_ids.append(encoded_data['input_ids'])
        attention_masks.append(encoded_data['attention_mask'])
        label_list.append(label)

    return torch.cat(input_ids, dim=0), torch.cat(attention_masks, dim=0), torch.tensor(label_list)

X_train_ids, X_train_masks, y_train_tensor = encode_data(tokenizer, X_train.tolist(), y_train_array)
X_test_ids, X_test_masks, y_test_tensor = encode_data(tokenizer, X_test.tolist(), y_test_array)

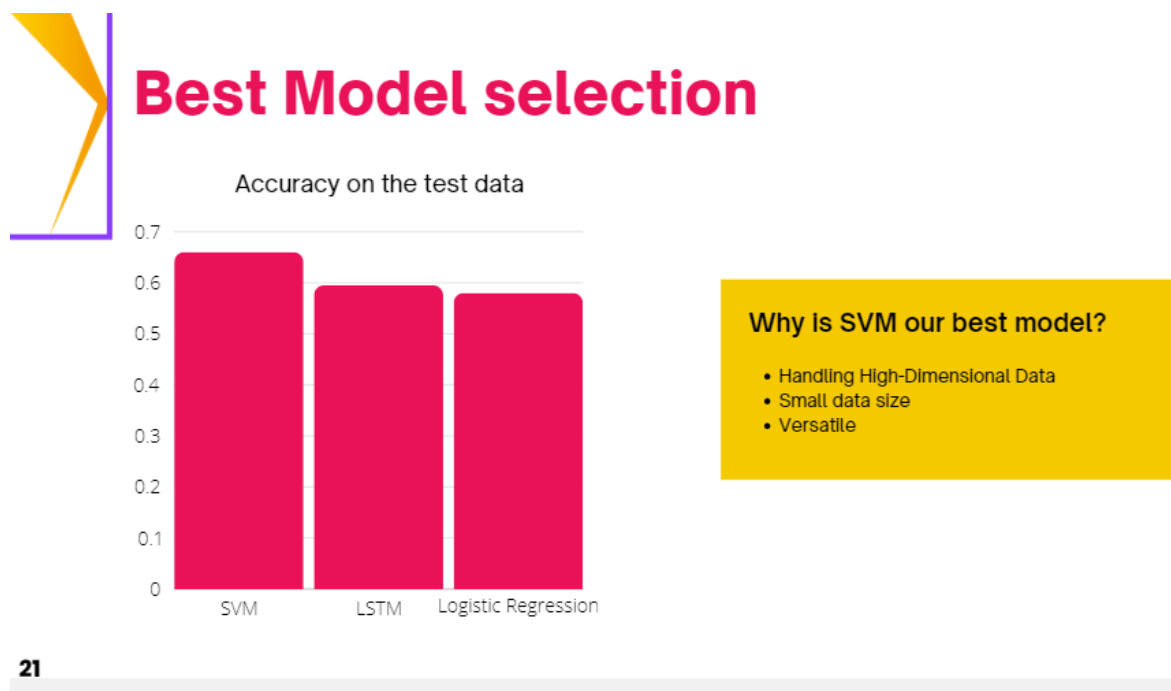
train_dataset = TensorDataset(X_train_ids, X_train_masks, y_train_tensor)
test_dataset = TensorDataset(X_test_ids, X_test_masks, y_test_tensor)

batch_size = 16

x = dataset,
est_data,
n.from_p
batch_size=batch_size)
taset), batch_size=batch_size)
els=len(df)
Error occurred when installing package 'torch'. Details...

```

## Best Model Selection



The best model we got was SVM, having accuracy of 0.66. LSTM was 0.595, and logistic regression was 0.58.

In addition to this, the added benefits of SVM are :

- Handling High-Dimensional Data: SVM is effective for data with a large number of features.
- Small data size: SVM can perform well with a limited amount of data.

- Versatile: SVM is adaptable and can be used for various types of classification problems.

### Demo Implementation:

```

Product Index: 1075
Original Text: DRIVE POWER CHAIR ..CIRRUS E C DRIVE POWER CHAIRFOLDS300 LB CAPACITYMINT CONDITIONNEW BATTERIES..COMES WITY CHARGERANTI TIPPING WHEELS
True Label: Health Equipments
Predicted Label: Health Equipments

Product Index: 1526
Original Text: first lady perfume First lady fragrance
True Label: Fragrances
Predicted Label: Fragrances

Product Index: 1377
Original Text: HoMedics BB-2K Bubble Bliss Deluxe Luxury Foot Bubbler with Heat HoMedics BB-2K Bubble Bliss Deluxe Luxury Foot Bubbler with Heathttps
True Label: Health Equipments
Predicted Label: Health Equipments

Product Index: 1632
Original Text: API POND SIMPLY CLEAR Pond Water Clarifier 16-Ounce Bottle (248B) • Contains one (1) API POND SIMPLY CLEAR Pond Water Clarifier 1
True Label: Other Health Care
Predicted Label: Other Health Care

Product Index: 1204
Original Text: VINTAGE WOOD ROLLER FOOT MASSAGER VINTAGE APRIL BATH AND SHOWER (A B AND S) ROLLER MASSAGER5.5 X 4.5 X 1 3/4"
True Label: Health Equipments
Predicted Label: Health Equipments

```

We randomly chose 5 product descriptions and based on that we compared the actual label with the true label.

## Conclusion

The benefits we have of correctly classifying the beauty and health section of Craigslist are as follows:

**Ease of Finding Products:** The introduced category filter streamlines the search process, allowing users to effortlessly locate desired products via an organized structure instead of combing through cluttered listings.

**Improved Search Results:** The category filters refine the search process, leading to a more streamlined and effective shopping experience.

**Better Visibility for Listings:** Implementing product filters improves the prominence of listings and offers a convenient way for shoppers to uncover new products, simplifying the hunt for great deals.