



MGMT 59000: Analyzing Unstructured Data

Final Report

Date of Submission: 08/12/2021

Submitted by: Data Poets

Arpan Datta
Gokul Harindranath
Neha Gupta
Padmapriya Pannerselvam
Sagar Baronia

Contents

The client: Craigslist	2
Project Topic	2
Specifics:.....	2
Describe the business scenario and identify a problem	2
Propose a specific improvement for the problem	3
Describe the data requirement and attempt to collect the data.....	3
Sample text descriptions from Craigslist for identifying the condition of the appliance:.....	3
Sample images from Craigslist Appliances section to be used for sub-categorization:.....	4
Web Scraping	5
Classification of appliances – Using Image	5
CNN Model.....	5
Condition of Appliances – Using text reviews.....	7
Conclusion.....	10

The client: Craigslist

Craigslist is an American based online platform that showcases classified advertisements. The sections covered by the website ranges from jobs, housing, services to resumes. Due to the high volume of unstructured formatting within the website it is very hard for platform managers to maintain and organize data. Moreover, consumers find it hard to go through all the advertisements in a messy/unorganized fashion.

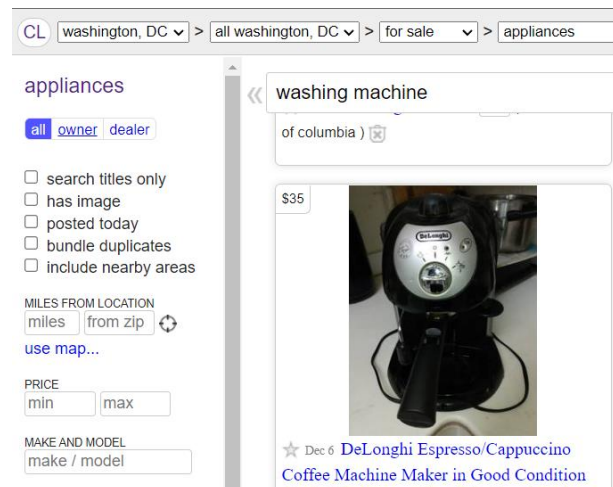
Project Topic

Optimizing search results for 'Appliances' subsection in Craigslist for sale items by classifying appliances into sub-categories and further segregating them based on their condition.

Specifics:

Describe the business scenario and identify a problem

At present, Craigslist for sale has a subsection for Appliance listings. Each listing has a description where the seller provides some description of the appliance like how old it is, what is the condition, whether there are any defects, etc. Along with the description, the seller provides few images of the product. However, there are few problems with the present interface which restricts a smooth search experience for the users.



- **Sub-category classification:** There isn't a segregation of listings based on sub-categories like Washing & Drying machines, Refrigerators, stove, ovens, fans, etc. So, if a buyer is looking for a particular appliance, he/she goes through all the appliance listings irrespective of which category it belongs to.
- **Condition identification:** Even though there is a condition tag for products, more than 50% of the listings do not have a condition specified. Sometimes, a user doesn't mind paying a little extra to get a machine in good condition and it would also lead to better negotiations based on the product's condition. It also helps buyers make quicker decisions which will improve the sell through rate and lead to a win-win situation for both buyers and sellers. Therefore, this is one feature which would be of prime importance to most of the users.

Propose a specific improvement for the problem

We propose a 2-step improvement in the present search process which involves first categorizing the appliance into sub-category and then identifying its condition. We will use images provided by the user to classify the appliance into a particular sub-category and then identify its condition using the text description provided by the seller. Therefore, to summarize, the scope of our project involves two broad areas.

- Classification of appliances into sub-category using Image classification
- Identification of condition of the appliance based on listing description using NLP

Business value for different stakeholders:

Seller: The seller gets extra insight into the pricing for the product by benchmarking their product with the average price for the items that are in the same condition. This lets them give competitive prices for their product and aids them to sell faster and optimize their profit.

Buyer: The buyer would be able to negotiate better since they could look at the average price for the items in a particular condition and can bargain to lower the price of a used product.



Craigslist: Since the feature is beneficial to both buyers and sellers, it will attract more traffic from both sides and set off a virtuous cycle which will increase activity and in turn increase revenue for Craigslist.

Describe the data requirement and attempt to collect the data

For the sub-category classification, we will pick around 100 images for each sub-category from Craigslist as well as Google images to train the last layer of the model trained on the ImageNet data, as the ImageNet database already has all the sub-categories which are needed to classify all the appliances. We further intend to use Image augmentation techniques, and hence expect to have a good accuracy with just 100 images for each sub-category.

We used python libraries such as Scrapy, BeautifulSoup and Selenium to procure data, i.e., text description and images of each listing via web scraping from Craigslist. We will attempt to sub-categorize first 'n' pages/listings within the appliances category.

Sample text descriptions from Craigslist for identifying the condition of the appliance:

Text description	Link
 4.3 cu ft front load washer and dryer Washer leaks and dryer is in good working condition With pedestals <ul style="list-style-type: none">• do NOT contact me with unsolicited services or offers	https://tippecanoe.craigslist.org/app/d/lafayette-lg-front-load-washer-and-dryer/7395571352.html
 Whirlpool Roper Washer and Electric Dryer. About 3 years old. Both work perfectly and are extremely clean inside and out. Super capacity models. Cords and hoses included. \$325. Call or text show contact info	https://tippecanoe.craigslist.org/app/d/fishers-roper-washer-and-dryer/7409378538.html



Nice High Efficiency Washer and Electric Dryer. Both work A1. Extremely clean and well maintained. Super capacity models. Cords and hoses included. ([show contact info](#))

<https://tippecanoe.craigslist.org/app/d/fishers-nice-he-washer-and-dryer/7409026014.html>

Sample images from Craigslist Appliances section to be used for sub-categorization:





Web Scraping

We scraped 1200 listings from Craigslist from top 10 locations in the US. We specifically got fields like Heading, Description, Price, Condition associated with each listing under Appliance section as we required it for the text analysis part.

For image classification, we manually downloaded 100 images for each appliance category. Alongside the training images, we downloaded some additional images for each category to be used as our validation set for evaluating our model.

The code that we used for scraping data from Craigslist has been included in the submission files with the name “Craigslist_crawler.ipynb”.

Classification of appliances – Using Image

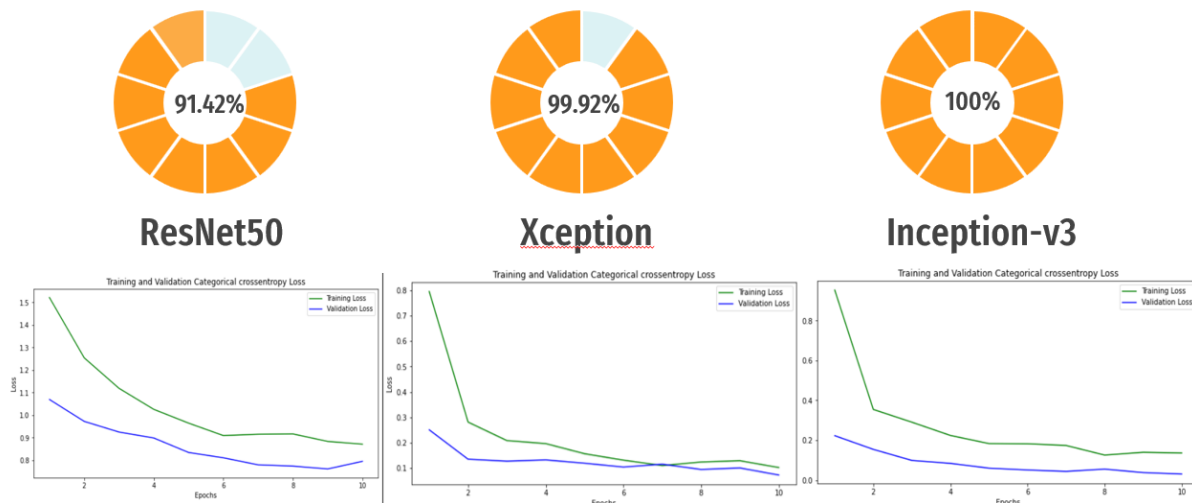
CNN Model

We used a pre-trained model using ImageNet Database which is trained over 14 million images and classifies images into 1000 objects/categories. The appliances which we wanted to classify were also present in the 1000 categories. However, since our project restricts to 5 specific categories, we dropped the last layer of the base model and trained the last layer using ‘softmax’ function to classify objects into one of these 5 categories.

We tried different base models like Inception V3, Xception, ResNet50 taking weights from imagenet and input image shape as 224*224. To the base model, we added a global average pooling layer, a dropout layer of 0.5 and a final dense layer with ‘softmax’ as activation function to classify into 5 categories which is Refrigerator, Stove oven, Washing Machine, Fan and Others category. Finally, we compiled the model using ‘categorical_crossentropy’ as loss function and Adam optimizer with accuracy as evaluation metric.

Image Augmentation: We have used Image Data Generator to consider different versions of an image using flipping, rotation, etc. This would help the model learn better and also reduce overfitting. The preprocessing function processes and returns image in a format that the model expects.

We performed the same analysis using IncpetionV3, Xception and ResNet50 and got an accuracy of 100%, 99.92% and 91.42% respectively. Inception v3 reached an accuracy of 100% in the third epoch itself.



When someone posts an image on Craigslist, it is not necessary that the image would fall in either of the 4 classes. Therefore, such cases had to be handled and we had to classify them as “Others”. Since most of our predictions for the 4 classes were with very high accuracy, which is more than 98%, we took a threshold of 0.8. So, if our model predicted a particular appliance with a probability less than 80%, it gets classified in the “Others” category. Below table shows how our model performed on the test set.

	fname	predicted_appliance_class	class_prob	modified_prediction
0	fan_1.jpeg	fan	0.999757	fan
1	fan_2.jpeg	fan	0.999775	fan
2	fan_20_2.jpg	fan	0.999926	fan
3	fan_5.jpg	fan	0.997992	fan
4	fan_6.jpeg	fan	0.996407	fan
5	other_1.jpg	stove_oven	0.996646	stove_oven
6	other_2.jpg	fan	0.646375	other
7	other_3.jpg	washer_dryer	0.300179	other
8	other_4.jpg	fan	0.505819	other
9	other_5.jpg	fan	0.681658	other
10	refrigerator1.jpg	refrigerator	0.999875	refrigerator
11	refrigerator82.jpg	refrigerator	0.959925	refrigerator
12	refrigerator_10 (1).png	refrigerator	0.992078	refrigerator
13	refrigerator_10.png	refrigerator	0.997584	refrigerator
14	stove_oven.png	stove_oven	0.994631	stove_oven
15	stove_oven3.png	stove_oven	0.984352	stove_oven
16	stove_oven4.png	stove_oven	0.999873	stove_oven
17	washer_dryer2.jpg	washer_dryer	0.999791	washer_dryer
18	washer_dryer3.jpg	washer_dryer	0.999192	washer_dryer
19	washer_dryer_1.jpg	washer_dryer	0.997612	washer_dryer

The predictions that the model threw for images belonging to the “Others” category were with a probability of 0.64, 0.3, 0.5, 0.68 and hence classified as “Others”. However, there is one image from the others category which is being classified as a stove oven with a probability of 0.99. This image was actually very similar to stove oven and it could have been possible with human judgement also. Therefore, we decided to go ahead with this way of handling images which does not belong to the 4 classes.

Using web scrapping technique, we were able to get 1200 records of description. After analyzing the distribution of condition tags on these distribution, we merged two of these tags, i.e. Excellent and New into excellent and Like new, good and Fair into one tag good. Below is the screen shot of the word cloud from both of these categories.

With our initial approach we vectorized all our text data using count vectorizer and TF-IDF both imported functions from SKLearn. Eventually for all the description records that did not have any condition label in it, we segregated it from our training data set, instead used it for our testing purpose after our final model was ready. The description from the dataset with condition label was further portioned in Train and cross validation, with ratio of 4:1. We started with all the conventional models like Ridge classifier, Naive bayes multinomial, Gaussian Naïve bayes, Decision Tree, Random Forest, SVM, and logistic regression. A few of them performed at accuracy of more than 55% of cross validation dataset, we have listed these below in the screenshot. Additionally, in order of leverage Neural Net, we used word embedding to represent our text and used GRU as our model to train and predict.

Basic Preprocessing	Accuracy	
	Naïve Bayes (Multinomial)	59.37%
	Gaussian Naïve Bayes	62.25%
	Word embedding	67.5%
	Ridge Classifier	68%
	Random Forest	70%

As can be seen in the above table, the best accuracy we received is 70%. In order to take this ahead we included a few more variable withdrawn from the description itself. First new parameter we introduced was length of each description, once we got this value we normalized it, in order to keep this value within 0 to 1. Secondly, we introduced another parameter as flag if there is a presence of digit in the description. Third, we bring in another parameter as number of characters present in the description as this was an integer value varying from 1 to a large value, we decided to normalize it before taking ahead for the modelling. Fourth parameter that we introduced is sentiment score of each description started with placing all 3 negative, compound and positive scores as 3 different parameters. However, on trying various combinations from these 3, score on positive sentiment alone helped us generate a better model. Lastly, we introduced a parameter to give extra weightage to words which are highly likely to be present in category – 1 or category -2. Finally with these 5 new variables and vector generated with countVectorizer (enabling n-grams) for cleaned descriptions, we started training our model with best accuracy on basic pre-processing.

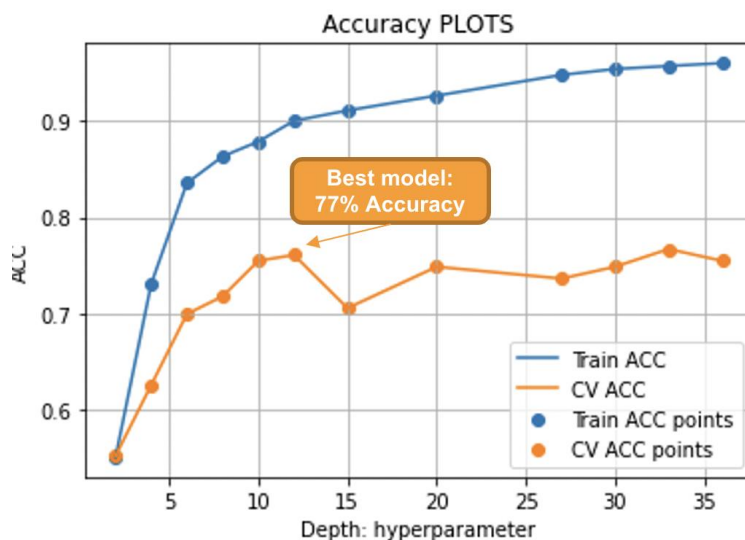
Below is the code-snippet of training and. Hyperparameter tuning of random forest which resulted in final accuracy of 77%.

```

1 #applying the Random Forest Classifier on the prepared data set
2 import math
3 from tqdm import tqdm
4 import matplotlib.pyplot as plt
5 from sklearn.naive_bayes import MultinomialNB
6 from sklearn.metrics import roc_auc_score
7 from sklearn.metrics import accuracy_score
8 from sklearn.ensemble import RandomForestClassifier
9 from sklearn.datasets import make_classification
10
11 train_acc = []
12 cv_acc = []
13 depth = []
14 K = [2,4,6,8,10,12,15,20,27,30,33,36]
15
16 for i in tqdm(K):
17     Rf_clf = RandomForestClassifier(max_depth=i, random_state=0)
18     Rf_clf.fit(X_tr_bow, y_train)
19     y_train_pred = Rf_clf.predict(X_tr_bow)
20     y_cv_pred = Rf_clf.predict(X_cr_bow)
21     train_acc.append(accuracy_score(y_train, y_train_pred))
22     cv_acc.append(accuracy_score(y_cv, y_cv_pred))
23
24 for i in K:
25     depth.append(i)
26
27 plt.plot(depth, train_acc, label='Train ACC')
28 plt.plot(depth, cv_acc, label='CV ACC')
29 plt.scatter(depth, train_acc, label='Train ACC points')
30 plt.scatter(depth, cv_acc, label='CV ACC points')
31 plt.legend()
32 plt.xlabel("Depth: hyperparameter")
33 plt.ylabel("ACC")
34 plt.title("Accuracy PLOTS")
35 plt.grid()
36 plt.show()

```

Below is the accuracy over various value of depth.



After our model was ready, we tried making prediction on the dataset which has no condition tag attached. Below are 2 such descriptions that were predicted as category 1 i.e., excellent and category 2 i.e good. After reading these predicted descriptions, we noticed that the tagging was well aligned with what we had

imagined it to be if inputted manually.

Outcome

Used once for a party. It's in brand new condition showing no wear at all. Pickup near Harlem & Devon, 60631 zip

Category 1

It may have been used. Looks new. It's presently \$420 plus tax everywhere.

Category 2

Conclusion

With a model in place for classifying the condition of the appliance and subcategory segregation, we expect the following improvements in the experiences of the different stakeholders:

- i) Buyers:
 - a. While buyers were previously only able to navigate through the results in the appliances section at an overall level, with the classification of subcategories, there could be a new subcategory filter which will help them navigate to the specific products that they are searching for such as washing machine, fridge and fan
 - b. It would also help them to identify the condition of the product which would help them filter according to their budget and negotiate better with the sellers.
- ii) Sellers:
 - a. Sellers are currently losing sales since potential buyers might not always reach to their results in the first page. But with the granular subcategorization in place, they would be able to get more and meaningful traffic for their listings.
 - b. The condition tag would also help them to benchmark their price against the products in their category.
- iii) Craigslist:
 - a. Craigslist is currently not utilizing 100% of the footprint on the website since users dropout due to frustration of not finding the specific products. With the inclusion of these features more buyers will start buying from craigslist. Since there are more buyers it will attract more suppliers and create a virtuous cycle. This would lead to a positive network effect, which will increase traffic and will eventually increase their revenue.

Some of the future proposals we have are to scale this to other categories in craigslist such as mobile phones, vehicles, kitchen appliances etc. Using more sophisticated models which does character embedding such as Facebook's Fast text would improve accuracy significantly. After this initial implementation we would propose doing the condition tagging with image analysis.