**M590 - Project Report**

Correctly Classifying Ads on Craigslist

Group 1

Mohinder Pal Goyal  
Robin Jindal  
Rahul Madhu  
Tirthankar Mukhopadhyay  
Abhinav Chanda

**Introduction**

Craigslist is an American classified advertisement website- with different sections for jobs, housing, sale, services and many others. It was started by Craig Newmark in 1995 as an email distribution list to friends and has now expanded to more than 70 countries, over 700 cities, and in more than ten languages.

“Place to buy, sell, rent, hire, share, meet, swap, discuss, find, serve, connect, give away, announce, work, collect, care, perform, learn, marvel, mentor, befriend, fall in love, and/or save the world” - Craigslist Official website.

Craigslist serves more than 50 billion page views every month. It is ranked 72nd among the most visited websites worldwide and 11th place among websites in the US. It has more than 60 million unique monthly visitors in the United States alone. With more than 80 million new classified advertisements each month, Craigslist is the leading classifieds service in any medium. The site receives more than 2 million new job listings each month, making it one of the top job boards in the world. Craigslist also dominates the US rental housing market, with millions of new listings every month. The 23 largest US cities listed on the Craigslist home page collectively receive more than 300,000 postings per day just in the "for sale" and "housing" sections. The classified advertisements range from traditional buy/sell ads and community announcements to [personal ads](https://en.wikipedia.org/wiki/Personals). Craigslist's estimated value is at least $3 billion. Moreover, a surprising fact is that only 50 employees manage the website.

Different sections of a craigslist home page are-

* For Sale
* Jobs
* Gigs
* Resume
* Community
* Services
* Discussion Forums
* Housing

In the “For sale” 40+ sections are belonging to different categories-

* Antiques
* Appliances
* Arts+Crafts
* Atv/Utv/sno
* Auto Parts
* Aviation
* Baby+kid
* Barter
* Beauty+hlth
* Bike parts
* Boats
* Books
* Business
* Furniture
* Garage Sale
* General
* Heavy Equip
* Household
* Jewellery
* Materials
* Motorcycle parts
* Motorcycles
* Music instr
* Sporting
* Tickets
* Many more

Because of so many sections on the craigslist home page, a user is often unable to decide on segments of interest. Additionally, all the sections are not mutually exclusive, and a single advertisement may belong to different categories. Hence, there is scope for misclassified ads.

From a seller’s point of view- If a seller needs to sell, for example, a vacuum cleaner, then the advertisement can be posted either in Appliances or in General or Household. Having so many options creates confusion, which leads to a misclassified post.

For a buyer, a person looking for a product needs to check all the different categories where a product may belong. It is time-consuming, and often not possible to check all the sections.

Because of the misclassified ads, there is a loss of business to Craigslist as a trade is not feasible. In addition to that, it gives a tiresome experience both to the seller and the consumer. Misclassified ads lead to loss of revenue to the company, and sometimes even a loss of customer as a customer may switch to another website to sell/ buy the product. Furthermore, the issue of misclassified ads can be seen on most of the websites where a user needs to post an advertisement like ‘Classified Ads,’ ‘Backpage.com,’ and many others.

Such a massive inflow of advertisements every day means it is impossible to classify each ad manually. One of the solutions to this can be the generation of tags, but even then, the problem persists.

A solution that is scalable and can speedily solve the issue of misclassified ads with accuracy is to generate a machine learning algorithm. One can then run the model on all the postings to identify misclassified ads.

**Business Analysis**

For Chicago’s craigslist, through our exploratory data analysis, we found that around 5% of the advertisements in the ‘Car+Truck’ section do not belong to the ‘Car’ category and would fit better in the ‘General’ category. Similarly, around 2% of advertisements from the ‘General’ category can be switched over placed in the ‘Car+Truck’ category. Hence there is a total of approximately 7% misclassification in both sections.

There are 80 million ads posted[1] on craigslist every month. If we assume a 7% misclassification rate across all the cities and sections Craigslist operates in, then there are around 560,000 misclassified ads on craigslist. Having such a vast number of misclassifications will not only degrade the buyers' experience as they will spend higher amounts of time searching for suitable advertisements but will also lead to loss of opportunity for sellers to sell those products if they were correctly tagged. This could decrease the worth of the company in consumer’s eyes, and they can switch to other similar platforms like Letgo and Swappa.

Craigslist identified this issue and devised a procedure to handle misclassification. The procedure currently deployed has the following steps:

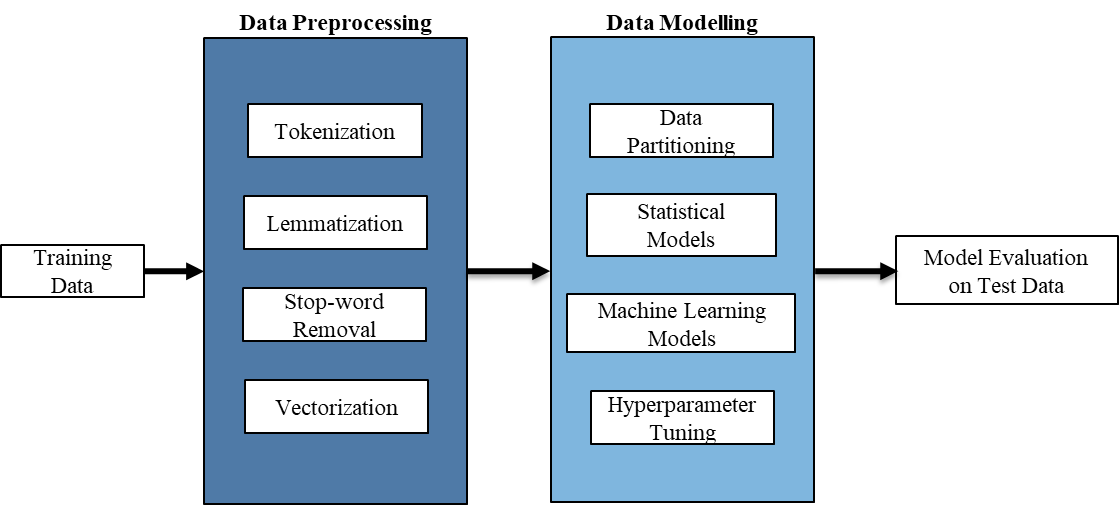
*Step1: Flagging of misclassified ads by customers:*Customers, when encountered with an advertisement, has an option to flag the advertisement if they feel it does not belong in that category.

*Step2: Development team fixes the issue:*Development team checks all the advertisements with flags and manually move them to the correct location. After manually checking all the flags, the updated advertisement is posted in the correct category.

Though the procedure currently deployed theoretically solves the issue at hand but has the following limitations:

* Limited amount of flagging: Users, when encountered with a misclassified advertisement, might ignore the advertisement rather than flagging the advertisement. Hence, craigslist does not know about the wrongly categorized advertisement. A high number of such misclassified advertisements display can result in customer dissatisfaction, and they can switch to other platforms.
* Limited team size:As per 2017, craigslist only had 50 employees [2]. Manually checking all the flagged advertisements from all over the world becomes a very time-consuming task. This could delay the correct classification of the advertisement. Since the classification is done at the manual judgment, the process is prone to human errors.
* Identification of problem after it reaches the end-user:In today’s world, most of the firms try to identify the potential problems before it even occurs. Craigslist being able to detect such issues after it has already reached the end consumer, would decrease its brand image and decrease customer retention.

Taking all these factors into account, our team developed a data-driven approach to not only identify misclassified ads before it reaches our customers but also find the correct categories they need to be moved to. We believe our model will make craigslist platform much smarter and would also enhance ad-viewers’ experience. Our model would still allow users to publish advertisements in a flexible and unstructured format easily. Sellers can still publish advertisements by only completed filling out the text description of the advertisement and maybe uploading some images. Buyers' cost to extract relevant information would also decrease by this approach. Our model would also be helpful for platform managers to manage and organize the platform easily. Hence, deploying this model would not only improve user experience but would also decrease the need for manual work.



**Methodology Overview**

**Data**

Our team chose advertisements from Chicago as the training set. Chicago had just over 3000 advertisements in the “Car+Truck” and “General” categories each. We scraped all of these ads directly from the webpage using a “Scrapy” based spider in Python.

The data was pre-processed to a limited extent during the extraction phase itself, utilizing string manipulations to remove leading and trailing spaces and unwanted characters.

Both sets were downloaded separately and were manually tagged with the following labels – ‘General’ and ‘Car.’ We included all sorts of automobiles under the ‘Car’ label, including snow removers, lawnmowers, recreational vehicles, and trucks. The label tags acted as our response variable. Post labeling, we combined the separate data sets into a single one of just over 6000 records.

The Chicago data was split into training and validation sets of equal sizes. We chose a 50% split because natural text can have many variations based on a user’s vocabulary and writing style. Hence, it would help in obtaining a generalizable model.

For our test data, we chose Tippecanoe county. Tippecanoe being a low population area, had around 900 records.

**Data Preprocessing**

Due to the use of multiple data analysis models in our analysis, we implemented two different pre-processing jobs,

1. *Statistical Models and Multi-Layer Perceptron* *Classifier:* a word tokenizer, followed by lemmatization, stop word removal and a Term Frequency-Inverse Document Frequency (TF-IDF) vectorizer
2. *Long Short Term Memory:*text cleaning custom functions based on regular expressions to remove extra spaces and wrong symbols (emojis, other symbols). Followed by a word tokenizer, stop word removal, and a TF-IDF vectorizer. We did not lemmatize words to preserve context for sequence generation, which is essential for an LSTM model.

**Vectorizer**

Text representation is a critical process in converting the text to be analyzed into a mathematical form and ultimately structure the unstructured data we have at hand. There are various vectorizers available to be used, which all use different techniques serving different forms of representation. We have used the TFIDF vectorizer.

**Theory**

TF-IDF(t,d,D) = TF(t,d) X IDF(t,D)

Where,

t 🡪 each term

d 🡪 each document

D 🡪 collection of documents

TF(t,d) 🡪 Term frequency

IDF(t,D) 🡪 Inverse Document Frequency = log

This vectorizer scores the importance of tokens in the collection of documents. This method takes into consideration both the frequency of the word in a document and penalizes based on the number of documents it appears in for a collection of documents.

The n-gram of tokens created, and the minimum document frequency required for a token to be considered as a feature are the two tuning parameters that have been used.

**Baseline Models**

The strategy we have adopted was to run basic models of various algorithms to set a benchmark and then use hyper-parameter optimization to tune the best basic model we got along with the tuning parameters of the vectorizer as discussed earlier. In line with the business model that we have framed to solve, ‘accuracy’ was the statistic we used for model comparison and calibration.

The following are the details of the base models we had run,

|  |  |
| --- | --- |
| Model | Accuracy |
| Naïve Bayes Classifier | 88.66% |
| Random Forest Classifier | 90.70% |
| *Support Vector Classifier* | *92.53%* |
| *LSTM* | *94.30%* |
| *MLPClassifier* | *93.50%* |

We chose the *MLPClassifier* as the best model on the training set. It was further tuned to improve the results as it yielded the best results out of the five algorithms we used.

**Hyperparameter Tuning**

We have integrated the vectorizer with the MLP classifier in our optimization process to get the best n-grams range and the best minimum document frequency along with the MLPClassifier. The following are the parameters and the values that have been used to optimize the model.

*TFIDF vectorizer*

Python Library used: sklearn.feature\_extraction.text

Function used: TfidfVectorizer

Parameters tuned:

* ngram\_range = { (1,2) , (1,3) , (1,4) }
* min\_df = {3, 4, 5}

*MLPCLassifier*

Python Library used: sklearn.neural\_network

Function used: MLPClassifier

Parameters tuned:

* nodes = { first\_layer = [5,10,15,20,25,30,35], second\_layer = [2,3,4,5,6]}
* solver = {‘adam’, ‘nadam’, ‘sgd’}

We used a brute force technique to try out all the possible combinations and get the best set of parameters which fetched the best accuracy.

The following are the optimized parameters:

1. ngram\_range = (1,4)
2. min\_df = 3
3. nodes = (15,3)
4. solver = ‘adam’

This model gave an accuracy of *96.90%* on the validation set and *96.05%* on the test set.

**Model Evaluation**

The following are the results of the model predicting on the test data (Tippecanoe dataset):

|  |  |
| --- | --- |
| Model | Accuracy |
| *MLPClassifier* | *95.8%* |
| LSTM | 76% |
| Support Vector Classifier | 96.1% |

We saw that the SVC had performed the best out of the three. However, there is a big difference between the validation set error and test set error for SVC, which points out that the model is not generalizable. Additionally, we cannot rule out LSTM as with hyperparameter optimization; it may give better results on the test data.

Hence, the MLP classification model, which has shown consistent performance in both the validation set and the test set is chosen as the model. Furthermore, once our solution is implemented at a national or global level, the increase in the size of datasets will mean that the deep learning models will learn better.

**Conclusion**

Craigslist would greatly benefit from implementing our model. Firstly, the misclassified ads in the cars section and the wrongly categorized car ads in the general section would no longer occur. All car ads would be correctly classified and placed in the desired section. Since our model would recognize the misclassified ads, there is no longer any necessity for Airbnb employees to detect and correct these ads manually. This would greatly reduce the manual efforts expended.

Furthermore, proper classification of ads would improve the user experience. Users are no longer bombarded with spam and irrelevant ads when they are browsing and also, they would be able to find desired ads more efficiently. Lastly, even the users who post ads would be benefitted. Their ads, when classified correctly, would reach desired audiences better and would sell faster and more efficiently.

**Limitations**

All our training data was scraped from a single location, Chicago. This would lead to the model being trained better for ad types in this region and might not perform as effectively in other regions, where the ads are of slightly different types.

Our data of 6000 records were manually categorized as either a car ad or a general ad. With such a large data set, there is prone to be some human error and some data might have been misclassified.

Our model would ignore incomplete ads, those that have no description. Our model cannot account for images and tag information, even if correct. It only looks at the description.

**Future Scope**

We could implement our model in the future in the following ways:

* Implement external ads from other sites in our training model to better account for all sorts of ads
* Expand the model to perform categorization of more Craigslist categories other than just cars
* Improve the model to detect spam ads as well
* Complement the model by using image recognition to help find categories by processing the attached images