VTLT Consulting

**Craigslist Sales Posts**

**Auto Categorization**

**Project Report**

12/5/2019

Group 3

Ziyun Huang | huang747@purdue.edu

Cheng Cheng | cheng436@purdue.edu

Luqi Chen | chen1297@purdue.edu

Wei-Cheng Chen | chen1614@purdue.edu

Wenying Huang | huang814@purdue.edu

# **abstract**

# Auto-categorization is becoming a prevalent technology that e-commerce companies used to regulate the deluge of unstructured and unorganized digital content on the web pages. These unstructured data are challenging to store and manage, resulting in high storage and management costs as well as other potential security threats for the companies. Meanwhile, it brings many inconveniences to individual users of websites like inefficient browsing, scams, and fraudulent information. Due to the loose business operating model without strict management, Craigslist has been cumulating a large amount of unstructured data. With the pursuit of enhancing website usability and increasing customer satisfaction, Craigslist needs to implement the auto-categorization functionality in its ad post service. To deploy the auto-categorization feature for Craigslist, VTLT Consulting developed text and image classification models for unstructured ad posting data. This proposal discusses both business insights and technical procedures of Craigslist auto-categorization for sale post.

# **Background**

## **Client Overview**

## Launched in 1995, Craigslist now has become a world-prominent advertisement website that offers online posting service for housings, jobs, personals, and community discussion forum. According to Business Insider, it has consistently ranked the top 20 in the US and around 118th world-wide, and it reached annual revenue of one billion in 2018(Johnson, 2019).

## Every month, Craigslist has more than 25 million new classified ads, over 1.5 million job postings, 75 million user postings in the discussion forum, and more than 10 million photos updated (Strickland, 2007). In addition to the abundant and miscellaneous collection of ad postings, Craigslist also has incredible high online traffic compared to other online marketplace websites. In 2018, Craigslist acquired an estimated 50 billion page views per month (Smith, 2019). While pageviews of e-commerce giants Amazon and eBay are 2.6 billion and 1.2 billion, respectively. Pageviews of less well-known companies Wish, and OfferUp is 92.9 million and 19.4 million (Godin, 2018). Among other online marketplace competitors, craigslist take the advantages of diverse categories, free services, fast transaction without intermediaries, and available local trading to survive in the industry.

## Despite the massive website traffic, Craigslist has long been criticized for its outdated UI design and business model. This controversy exists due to a large amount of unstructured data, which raised problems of miscategorized postings, spams, repeated postings, and prohibited contents. Reviews and complaints from Craigslist users are mainly focusing on its low usability in navigation, content organization, and layout design. In the recent decade, numbers of active users and visits have been steadily decreasing while other online marketplaces were emerging and thriving. Regardless of trading quality, massive user dissatisfaction toward Craigslist has a high proportion in its miscategorized content.

## **Business problem**

# Based on consumer feedback analysis, the “general” category under the for-sale section contains highly disorganized ad postings, leading to low click-through rate and user dissatisfaction. Sellers with limited understanding of their merchandise usually label their ads as “general”. Correspondingly, buyers are less likely to search for "general" products. Therefore, ad postings under “general” categories contribute least sales and traffic to the total Craigslist trading and visits. With a lower likelihood of a successful sale, which would result in lower revenue, the seller’s satisfaction would also be compromised. With dissatisfactions from both sellers and buyers, Craigslist’s reputation and market share as an online trading community would be threatened.

# From the operating perspective of Craigslist, uncategorized “general” ads are also useless and costly. Because of lacking defined labels and descriptions, the ads become unstructured data accumulating in the storage space of Craigslist. Ads under the general section are challenging to manage and sensor. For this reason, the ads are very likely to contain the information of prostitution, spams, and other prohibited content which violate laws and terms of Craigslist policy.

# **Business Analysis**

As other e-commerce businesses are continually updating their technology and business model to embrace the era of big data, Craigslist needs to improve data management and processing capabilities through upgrading the categorization features in their system design. Specifically**, Craigslist needs an auto-categorizing function for their posts that autonomously categorizes user’s posts based on advertisement descriptions and/or images.**

The auto-categorizing feature allows Craigslist sellers to post advertisements more efficiently as well as to facilitate buyers’ experience. Implementation of auto-categorization can replace the “general” categories under the for sale section; this not only reduces data storage and management but also improves visibility of website layout for Craigslist.

**CAPABILITY FOR AUTO-CATEGORIZATION**

With the auto-categorization, the seller would upload the photo and enter the title or description of the merchandise, and click post. The auto-categorization system at the backend would make their advertisement publish at the corresponding category. This process could also include filtering potential spam posts and blocking them from being published.

This function would not only lower the possibility of misclassifications and solve the issue of unclassified posts but also minimize user intervention. With this function, Craiglist sellers will no longer worry about how to categorize their posting themselves. For the buyers, a more accurately categorized shopping platform could significantly improve their browsing experience.

**DEVELOPING THE FUNCTION**

To help the software engineering team to construct this auto-categorization function, VTLT would build a classification model based on both texts and images.

The text classification model would analyze a user’s entry in title and description and label them with one of the 45 categories. The image classification model would analyze the photo uploaded by the seller and label them with the correct categories. Since the model accuracy varies depending on the completeness of text entries and image quality, the final classification result would be assessed based on the feature weights and model accuracy comparison. With the combination of text and image analysis, the team would not just deliver a higher classification accuracy model for the function, but also endue a tremendous future potential for the model to be utilized for other purposes.

# **Data Analysis**

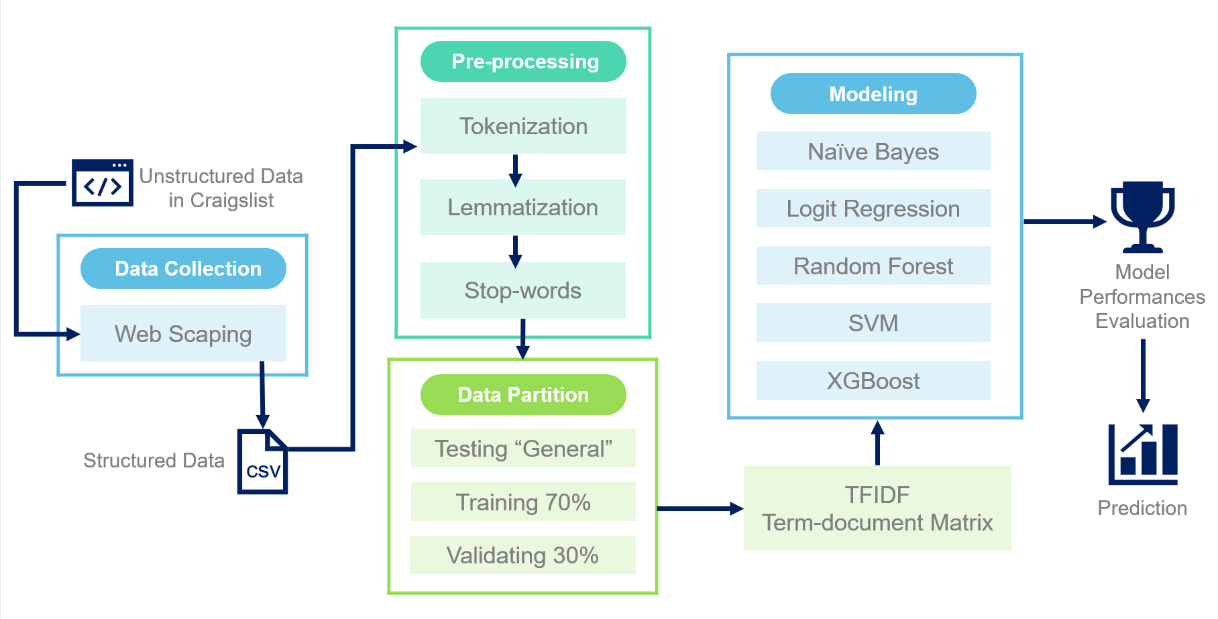


Image1. Model Design Flow Chart

**Text Classification Models**

First, we utilized Scrapy to collect existing posts. We built a spider with three nested layers to crawl data. There are 45 categories under the “for sale” section, and each category has hundreds of records. We collected titles, descriptions, and categories labels for all the records (more than 84,000 posts till 12/03/2019) and saved them into a CSV file.

Due to the computational limitation, we decided to shrink the data size. We extracted 500 records from each category and ended up with around 22,000 records. We believed this sample is large enough to get an accurate result. In data pre-processing, we applied the nltk package to tokenize the titles with lower case on word level, used WordNet Lemmatizer to lookup lemmas, removed less important words, which are those high-frequency words and low-frequency words. Next, we took out the records under the general category as the test set, did data partition on rest records, and split them with 70% on the training set and 30% on the validation set. In the following step, we converted training, validation, and test set to a term-document matrix with Tfidf Vectorizer and specified two minimum document frequency and 6-grams. After the conversion, we have finished the data processing.

We then build five models, which are Naïve Bayes, Logit Regression, Random Forest, Support Vector Machine, and XGBoost. We also trained these models with various tuning parameters. For Support Vector Machine model, we tried linear and polynomial kernel to the degree of 2 and 3. For Random Forest model, we tested different combinations of numbers of trees and maximum depth.

**Image Classification Model**

Our team decided to only scrape images from the Bike and Motorcycle categories due to limited GPU capacity and more accessible demonstration purposes. We also would like to find out if the model could distinguish between the two products, which are similar in terms of appearance. We used an application called ParseHub to scrape a total of 1,600 images from Craiglist, with 800 from each category.

The method we chose to perform image classification is by using the pre-trained model. These models utilized some of the most common convolutional neural network (CNN) architectures, including VGG, Resnet, InceptionNet. The significant advantage of using a pre-trained model is that they have been developed and tested using 14M hand-labeled images.

The main objective of CNN is to output the probability of an image being a particular class. CNN works by first converting an input image into arrays of numbers, next the network will process the image to reduce the form and capture image features. The process is known as convolution, where a filter or kernel, having the same depth as the input image and has an assigned weight, is applied to portions of an image. Each time when the kernel is applied to the image, it will perform a matrix multiplication between the kernel and the portion. The convolution process will continue until the filter covers all parts of the image. The resulting array of numbers is referred to as an activation map. Convolution can have multiple layers; generally, the first layer will capture low-level features such as edges and colors. Deeper layers will capture more features, such as more complicated shapes.

With the pre-trained model from the Keras application library, we can perform transfer learning. We first prepare the images downloaded from Craigslist and prepare a training set with labels and a test set without the labels. Then we adopted the VGG19 model to extract image features of the training set, allowing the model to learn the features of different labels. The final step of prediction is utilizing MLP classifier with two hidden layers and pipeline from sklearn to generate final predictions in a simple and accurate method.

# **Model performance**

# **Text Classification Models**

Among all the models we built, the Support Vector Machine performed the best, which the accuracy rate is 61.80%. The rest models performed relatively well, and the accuracy rates are all over 50%.



Image2. Prediction result using text

Then we applied the Support Vector Machine on our holdout set, which is the data extracted from the general category, to generate predicted category labels for each record based on its title. Overall, the predicted results look good.

一張含有 自行車, 停車, 貨車, 騎馬 的圖片

描述是以非常高的可信度產生**Image Classification Model**

Image3. Prediction result using images

After the model was developed, we randomly selected 20 images from both categories to test if the model could accurately categorize the image. VGG19 model showed a high level of accuracy, and the result showed that the model could correctly classify if the image is a bike or motorcycle every trial based on the features we extracted from the training set.

# **conclusion**

**AUTO-CATEGORIZATION FUNCTION**

VTLT Consulting has identified the complex “general” category under the for-sale section as an inconvenience to Craigslist’s user experience. With the issue identified, VTLT proposed Craigslist to develop an auto-categorization function to allocate this issue.

VTLT Consulting expects the analysis conducted and delivered to Craigslist to be a solid capstone for the development of the auto-categorization function. With the auto-categorization, Craigslist would no longer have the problems brought by the “general” category of the sale posts. This function could also minimize human intervention in classifying the post as the function would categorize all the future posts autonomously.

**FUTURE APPLICATION OF MODEL**

With the classification model developed, there is a potential application to enhance user experience.

A batch upload and post function could be another utilization of the image part of the classification model. A typical scenario nowadays is that a seller has a batch of items that belonged to different categories for sale. Currently, sellers need to post multiple advertisement posts individually, and this is a tedious process for them. It also increases the risk of misclassification and lacking details. The batch upload and post function would allow the seller to upload all photos at once, and the system would categorize them individually and post them under corresponding categories with default description. The average history price of a similar item with the same condition would be used as a temporary list price for the advertisements.

The model could also be utilized for spam detection. With both text analysis and image analysis results available, an apparent discrepancy in result from text classification and image classification could raise a red flag of the possibility of spam.

VTLT Consulting profoundly believes that the thorough analysis and valuable model could help Craigslist to improve the user experience dramatically and become a more friendly online community.

Reference

Johnson, D. (2019, May 17). 11 mind-blowing facts about Craigslist, which makes more than $1 billion a year and employs just 50 people. Retrieved from <https://www.businessinsider.com/craigslist-facts-2019-5>.

Strickland, J. (2007, November 1). How Craigslist Works. Retrieved from <https://money.howstuffworks.com/craigslist.htm>.

Smith, C. (2019, May 11). 20 Amazing Craigslist Statistics. Retrieved from <https://expandedramblings.com/index.php/craigslist-statistics/>.

Godin, M. (2018, September 9). 21 Top Online Marketplaces You Can Actually Make Money on Today. Retrieved from <https://crazylister.com/blog/online-marketplaces-ecommerce/>.

Appendix

Text Scraping Terminal Screenshot:

A screenshot of a computer

Description automatically generated

Text Model Python Screenshots:

A screenshot of a cell phone

Description automatically generated

A screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generatedA close up of a computer

Description automatically generated

A screenshot of a cell phone

Description automatically generated

Image Model Python ScreenShots:



