Scam Identification: Automobile Classifieds

Table of Contents

craigslist

Background:

Today the online classified ads form an integral part for the market of advertisements. Craigslist is one of the online advertising market that is hugely popular. When opposed to more conventional media like newspapers and printed media, the World Wide Web offers people a handy and quickly available means for listing and browsing advertisements. The ease of access to the internet has the unintended consequence of luring online con artists who use false advertisements to pass themselves off as legitimate sellers and deceive consumers. Scammers can take millions of dollars from naive consumers, endangering the utility and credibility of internet marketing businesses. For online classified advertisements, there is no systematic reporting of market or fraud information. Companies that run classified ads typically don't release information about their revenue or fraud rates to the public. One of the primary challenges preventing the future development of online advertising is the fact that scammers target the online classified advertisement domain because of its huge economic potential. As a result, scam identification in internet advertising is a serious issue.

According to Alexa, a web data administration, Craigslist is currently ranked seventh in the United States. According to a fact sheet from Craigslist, the site receives over 250 million visits each month and well over 80 million listings . Even if a tiny percentage of the advertisements posted each year, which include exchanges worth billions of dollars, are false, it can cheat customers of a sizable sum of money. These platforms must monitor and moderate possibly fraudulent activity when dealing with sales in the millions.

The success of Craigslist is constrained by devious techniques. Exchanges might not be completed due to the possibility of being deceived. In terms of finance, it is possible to think of the probable tax on future transactions as being the realistic calculation of misfortune brought on by a trick. Given Craigslist's positive effects on the economy, local governments should make plans to reduce this common disaster. Therefore, local governments should either reduce the chance of exploitation or reduce the severity of problems.

Craigslist started up in 1995 as a simple email appropriation system for advertising local events. It was released the following year as a web application. Since then, the administrations provided have expanded and adhere to a non-exclusive viewpoint that emphasizes willing acquisition of goods and ventures. County and city divisions have been made on the Craigslist website. This geographic separation encourages local commerce, increasing protection from scam. Everybody has access to these classifications and can browse and post new listings. Posting is valid for 30 days after which it is removed from the website. A post includes information on the good or service being sold or needed, as well as a way for people to get in

touch with the banner. Since the start of the online administration, it has been plagued by a series of security problems, such as email gathering and mass phishing schemes. The extent to which tips are written on tactics throughout the website demonstrates glaring security flaws in the framework. However, there has hardly been any alteration to the system in the past very few years.

In this project, we have focused on whether advertisements of cars and trucks are scam or not based on pre-defined business rules.Previous methods for Web scam detection mainly used link-based features and content-based features like n-grams to distinguish between scam and non-scam pages. However, link-based capabilities are useless in this particular sector because online marketing posts rarely link to one another. The fact that scam posts frequently contain misleading information sets them apart from non-scam advertisement posts in terms of substance. For instance, a scam advertisement post can draw customers by setting an inflated asking price. The content-based features are unable to capture this trait. Consequently, typical methods for detecting Web scam are ineffective in this field.

Business Analysis

Online classified ads are an essential component of every business' digital marketing strategy. Due to the platform's enormous potential, fraudsters have been drawn to it, which has hampered the growth of online advertising. Because of this, spotting fake advertisements on such platforms is a critical problem that has to be efficiently filtered out. We'll be putting into practice a model that might assist us in spotting auto sales scams. We provide a creative approach to developing a scam detection program that will mark an advertisement as potentially fraudulent based on particular criteria. We have discovered characteristics that are most frequently found in fake advertising based on a few study publications[1], for example, phone and email not provided in the listings and price-based heuristics.

Our project scope is limited to cars and pickup trucks which are one the most sold items on craigslist and we are only focusing on these items being sold within 60-mile radius of Chicago. The goal of this project is multi-fold.

- First, we want to gain the trust of the audience and increase the credibility of the platform.
- Second, we want to be able to sell more quickly as genuine ads gather more traction from the buyers.
- Finally, we would be able to save people from getting scammed and increase awareness among the customers.

We will categorize each advertisement as "Scam" or "Not Scam" based on the established rules, and subsequently train a supervised model using the labels. The description (embedded text) of the advertising will be the model's first input.

After this we move on to performing topic-modeling for the listings which have been identified as genuine. This will give us insights about the listings and find the features that can be attributed to monitor scams on craigslist.

Web Scraping

To effectively harvest that data, we used Python libraries requests and BeautifulSoup. Here are the steps used to scrape data:

Part 1: URL Extraction of ad listings

- Inspected the **HTML structure** of Craiglist/Cars+Trucks with Chrome's browser's **developer tools**
- Deciphered data encoded in **URLs**
- Used requests and BeautifulSoup for **scraping and parsing URL information of each ad listing** from the Web
- Stepped through a **web scraping pipeline** from start to finish
- **Built a script** that fetches URLs from the website and saved the URLs in Excel
- **Saved the URLs** scraped from anchor tags of each listing

To create our target URL, we split it into two parts:

- **The base URL** represents the path to the search functionality of the website. The base URL is https://chicago.craigslist.org/search/chicago-il/cta? This URL directly goes to our desired category of cars and automobiles.
- **The specific site location** that limits our results for only 60 mile radius of Chicago: "lat=41.7434&lon=-87.7104&search_distance=60"

As one page had only 120 records, we noticed that when we clicked "Next" at the bottom of the page, our URL changed which further appended "s=240&" in the URL. We then tweaked our code to include s=240 then 360 and so on.

Part 2: Data Extraction of ad listings

Once we have extracted links from each of the job listings, the next step is traversing through the URLs saved in the previous step and then extracting features like description, price, title, color of the automobile, etc. from each of the links.

Similar to the steps taken in Part 1, we identified the HTML elements we want to scrape the data. This data is then saved in an Excel file which is then pre-processed and used for modeling.

Libraries used:

1. **BeautifulSoup**
2. **Requests**
3. **Pandas**

Data Analysis

Phase 1: Predicting Scam Listings

Text Pre-Processing:

Before feeding the textual columns into the models we had to perform the following pre-processing steps on the "Clean Description" columns:

- Removal of HTML Tags such as '/a',</br>,
- Removal of Punctuations
- Removal of most common words occurring across all the documents
- Tokenization of sentences
- Lemmatization of tokens (*this step was not followed for topic modelling*)
- Stopwords removal
- TF-IDF vectorization to finally convert the textual data into numerical vectors

The above pre-processing steps were performed for feeding the data into classical ML Models. For preparing the data for advanced modelling techniques such as Recurrent LSTM NN, we had to perform further pre-processing on the textual data such as:

- Converting text to sequence (Instead of TF-IDF vectorization implemented for Classical models)
- Pre-Padding the sequences

Classification Models

Since our data was highly imbalanced (4% Probable Scam labelled as True), we had to stratify the data set and split it into training and testing. We split the vectorized data in a ratio of 80-20 of training and validation. The distribution of label in training and testing set is as follows:

	Probable Scam=True	Probable Scam=False	Total
Training Set	64	1892	1,956
Testing Set	13	476	489
Total	77	2368	2,445

We implemented seven machine learning algorithms out of which 6 are classical ones such as tree-based/kernel based or an ensemble of these. We implemented a state of the art technique as well - Recurrent LSTM Neural Network[2], which generally works well on textual data.

Model Name	Parameters Used
Random Forest	○ Estimators=300 ○ Max Depth=8 ○ Criterion =Entropy
Extra TreeClassifier	○ Estimators=100 ○ Criterion=Entropy ○ Max Depth=4
Support VectorClassifier	○ Kernel=RBF
XG Boost RF Classifier	○ Learning Rate=0.5 ○ Estimators=200 ○ Max Depth=6
Voting Classifier	○ Voting=Hard
LSTM Neural Network	○ Number of LSTM units=128 ○ Loss=Binary Cross-entropy ○ Optimizer=Adam ○ Metrics=AUC

Phase 1: Validation

We used the 20% test data set kept aside to validate the performance of different models we trained on the training data set. The evaluation metric we chose for selection of best model out of all the models we implemented was AUC ROC score or "The Area Under ROC Curve". We wanted to maintain a balance between both False Positive Rate and True Positive Rate, as the genuine ad listings being predicted as scam (False Positive) would prove to be an impairment to the craigslist's business. Based on all the models we trained, LSTM gives us the best validation ROC AUC Score of **0.7398** on test data set. The score of other models implemented are as follows:

Models	Test ROC AUC Score
Random Forest Classifier	0.57
Extra Tree Classifier	0.65
Support Vector Machine	0.57
XG Boost RF Classifier	0.68
Voting Ensemble Classifier	0.61
Recurrent LSTM NN	0.74

Based on our best model, we created a Confusion Matrix based on the predictions of validation dataset:

	Predicted Labels: False	Predicted Labels: True
True Labels: False	466	7
True Labels: True	11	5

If we calculate the True Positive Rate (TPR) or Recall based on the above best model comes out to be 5/11=0.31. We can further fine tune the model parameters using transfer learning techniques to achieve higher TPR. False Positive Rate (FPR) comes out to be 7/473=0.01 which is extremely low and makes our model robust against False Positives.

Phase 2: Identifying Topics in Legitimate Listings

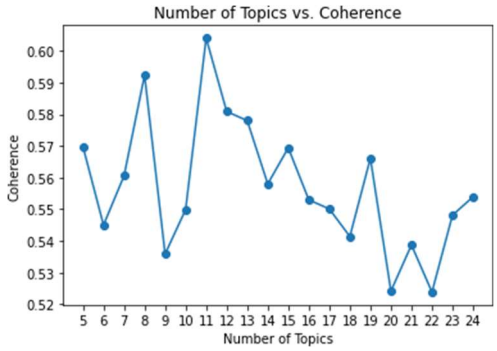
To further process the legitimate listings, we conducted topic modelling to identify the features that were more closely related with authentic ads.

Preprocessing:

We first filtered the test dataset to only analyze the non-scam ads and then processed the 'Clean Description' of the legitimate ads. Then we conducted Bigram and Trigram noun phrasal analysis to make the topics more human interpretable.

Topic Modelling:

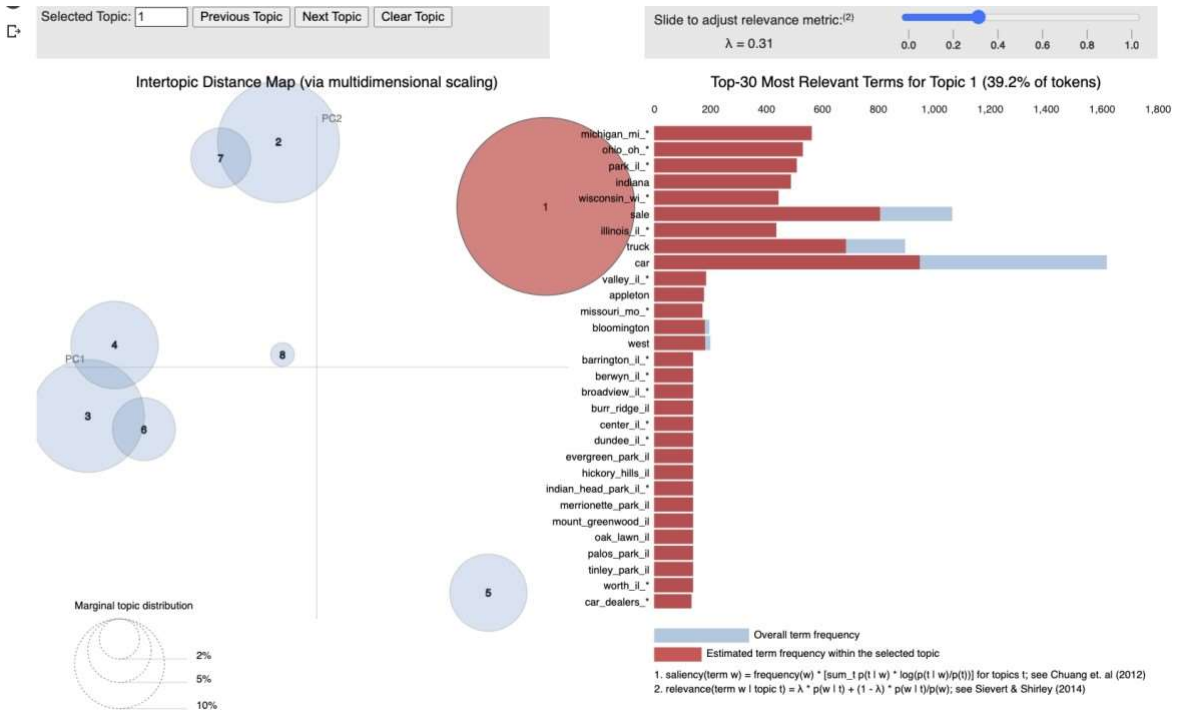
Using the trigram phrases, we then conducted a LDA analysis and tried to assess the right number of topics using the coherence as a measurement of topic identification. Here is the coherence with the different number of topics in the LDA model.



We see that after 10 topics the coherence improvement stops, we base our model from these results to work out the ideal number of topics for the LDA model. We further tuned the data using hyperparameters such as passes, chunksize and iterations.

Phase 2: Validation

We performed the topical analysis and found the various topics relevant to predicted genuine classified ad on craigslist. The results of the same are presented below with the 8 topics brought in from the chosen LDA model as they made more sense and were more human interpretable:



Another metric to fine tune the result for the model is to tune the relevancy score to prioritize the terms more exclusive to a topic. Words across a corpus representing a given topic may be ranked high because they are globally frequent. Relevancy score helps prioritize terms that belong more exclusively to a given topic, making the topic more obvious.
We used a **priority score of 0.31** for our analysis.

Recommendations and Conclusion

Online classified advertisements have become an essential part of the advertisement market. In this project we proposed an approach for scam detection in online advertisement posts. To the best of our knowledge, our work is an initial attempt for scam detection in this important domain. First, we identify special characteristics of a fraudulent advertisement as compared to other legitimate advertisements and analyze which business rules could be applied to segregate such advertisements. Second, we propose a novel set of domain-specific features that discriminate scam from non-scam advertisement posts.

craigslist

There is a need for Craiglist to implement effective scam filters in their website so that they don't lose their credibility and gain the trust of audience. They could employ other state-of-the-art techniques like text mining on the description and recognize trends in image/text posts to identify scams. This a vast domain of research and one could apply different business rules for scam identification. Although Craiglist offers guidelines for potential buyers to save themselves from getting scammed but clearly it is not enough. FTC reported that ~2.8 million people were scammed in the year 2021 alone. Out of these, Auto related scams also amounted to ~137k reports being filed. This brings in a need for Craigslist to not only host credible sellers but also regularly scrutinize the ads posted on the platform

For future work, we plan to extend the experimental dataset. Identifying good instances for judgment is itself an interesting problem. Since this is a highly imbalanced classification problem, if we randomly pick a sample of instances for judgement, there are very few scam instances in the sample. To overcome this, we plan to use other advanced models to pick instances that are likely to be scam for judgment. We can also include image classification as a potential method to identify scams. Like for instance, do the images contain any text, if they do is it legitimate or not, are the pictures available on Google like stock images or not (if yes, they are likely scams). There is definitely a lot of scope, and this area is a vast area of research and our project was an initial step towards this.

Appendix

1. Alsaleh, Hamad, and Lina Zhou. "A Heuristic Method for Identifying Scam Ads on Craigslist." *2018 European Intelligence and Security Informatics Conference (EISIC)*, 2018, https://doi.org/10.1109/eisic.2018.00019.

2. **LSTM Architecture:**

```
Model: "sequential_21"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 embedding_19 (Embedding)    (None, 500, 16)           8000

 lstm_17 (LSTM)              (None, 128)               74240

 dense_15 (Dense)            (None, 1)                 129

=================================================================
Total params: 82,369
Trainable params: 82,369
Non-trainable params: 0
```