MGMT 59000

ANALYZING UNSTRUCTURED DATA

FINAL PROJECT REPORT

CRAIGSLIST - ELECTRONICS

CLASSIFICATION AND SPAM DETECTION

AUD ONES OUT

Amrit Singh, Deepa Narayanan, Mrinmoy Dalal,

Sanchit Agarwal, Vedanti Gulalkari, Anusha Reddy

Introduction

Online advertising has developed into one of the most significant kinds of advertising because of the widespread use of the internet and the popularity of online platforms. The market for internet advertising worldwide increased by 21% in 2017 to reach $87 billion, and it is anticipated to reach $117 billion by 2021. A handy & well-liked way to offer goods or services is through online classified websites like Craigslist (www.craigslist.org), Backpage (www.backpage.com), Oodle (www.oodle.com), and eBay Classifieds (www.ebayclassifieds.com). These competitor websites' popularity is only growing. When opposed to more conventional media like newspapers and printed booklets, the World Wide Web offers people a handy and quickly available means for listing and browsing advertisements.

The increased accessibility of the internet has the unintended consequence of luring online con artists, who publish bogus adverts while posing as legitimate sellers to con unsuspecting consumers. Scammers can take millions of dollars from unwary consumers, endangering the utility and credibility of internet marketing businesses.

About Craigslist:

Craigslist is essentially like the classifieds section in a newspaper, with the exception that it is free, allows for the use of images, and instantly reaches millions of people. The website is divided into several major categories, and each category includes several subcategories, making it simple to find the desired area quickly. 450 cities are currently served by the website. A calendar of nearby events and discussion boards are also included.

One can use Craigslist to purchase and sell goods, hunt for a job, find local events, find an apartment, start a discussion, and more. In 1993, a software programmer by the name of Craig Newmark created Craigslist to inform people about activities taking place in and around the San Francisco Bay Area. The site's subscription base surged because of word-of-mouth marketing, and in 1999 it expanded to Boston. Later that year, Craigslist became incorporated and concurrently reached nine significant American cities. At over nine billion page views each month, Craigslist is currently the ninth most frequented website in the US.

Why Electronics?

There are several benefits to purchasing used electronics from Craigslist, whether you're a college student or a retiree, and in recent years, the demand for electronic devices in the US rental market has increased. About 150 billion USD worth of electronic goods were purchased in America in 2022.

According to your needs and tastes, you can find a variety of classified ads in the popular category of "Electronics" on Craigslist. The USA's leading source of information on the electronic sales market is now Craigslist. Each month, billions of new listings are added, and this industry is expanding annually.

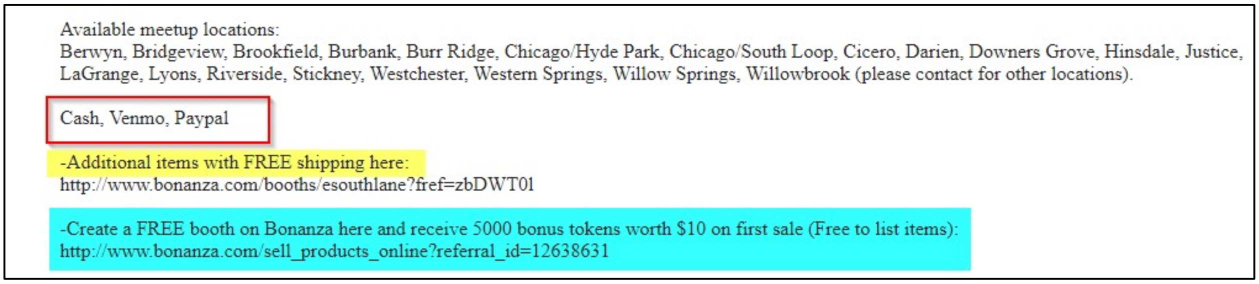
Problem Statement

Every day, millions of ads are anonymously submitted on Craigslist across the globe. In a single month, millions of electronic products are listed. It is challenging for those seeking for a new home to search the postings. 11% of product advertisements, on average, are spam. They are unable to travel the globe arresting and charging people, though. The goal of this project is to eliminate spam by filtering it based on the descriptions of electronic goods listings.

Only a small portion of the posted ads that are processed by classified ad websites on a regular basis are false. Online con artists frequently put a lot of work into making their listings appear authentic. Examples include utilizing stolen account information to purchase additional services or replicating existing adverts from other services. The listing that is given below more closely resembles a piece of art. By resolving some of the major problems, this project will attempt to add value for both its client (Craigslist) and its users. High numbers of electronic product frauds harm Craigslist's reputation and raise user attrition. Users must spend a lot of time and resources searching through thousands of postings to uncover authentic ads, which might take hours. For this project, online classified fraud for electronics ad postings will be detected by using data analysis and text analysis concepts and methodologies.

To extract pertinent attributes from a database of online classified ads, traditional data mining techniques are used. Machine learning algorithms are then used to identify patterns and connections between fraudulent activity. With the help of our suggested strategy, we will show how well data mining techniques can be used to spot fraud in online classified ads for electrical product ads in the Chicago area.

Ways to identify spam messages



Available meetup locations:
Berwyn, Bridgeview, Brookfield, Burbank, Burr Ridge, Chicago/Hyde Park, Chicago/South Loop, Cicero, Darien, Downers Grove, Hinsdale, Justice, LaGrange, Lyons, Riverside, Stickney, Westchester, Western Springs, Willow Springs, Willowbrook (please contact for other locations).

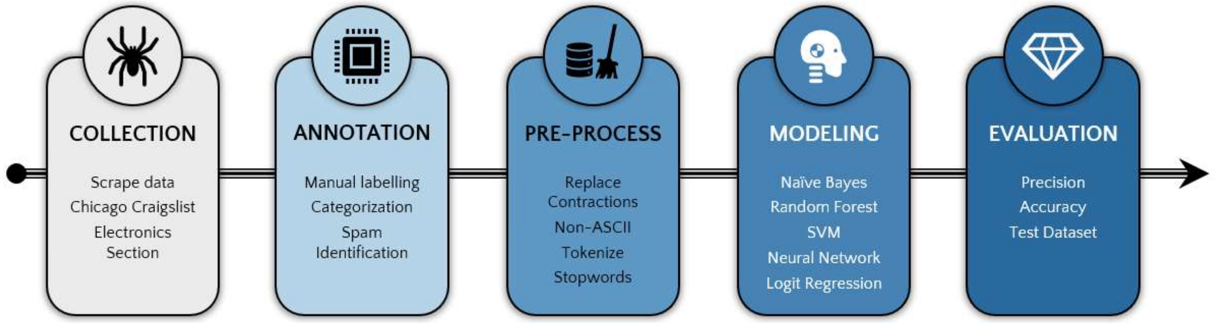
Cash, Venmo, Paypal

-Additional items with FREE shipping here:
http://www.bonanza.com/booths/esouthlane?fref=zbDWT01

-Create a FREE booth on Bonanza here and receive 5000 bonus tokens worth $10 on first sale (Free to list items):
http://www.bonanza.com/sell_products_online?referral_id=12638631

The above image is an example of a spam ad. We see at least three indications for the same. First one being the presence of stop words like paypal and free. The second indication showcasing the free shipping option made available. The third indicator is presence of a link that redirects user to another page. These three are classic indications of spam; along with a few more indications we carried out spam identification in our project.

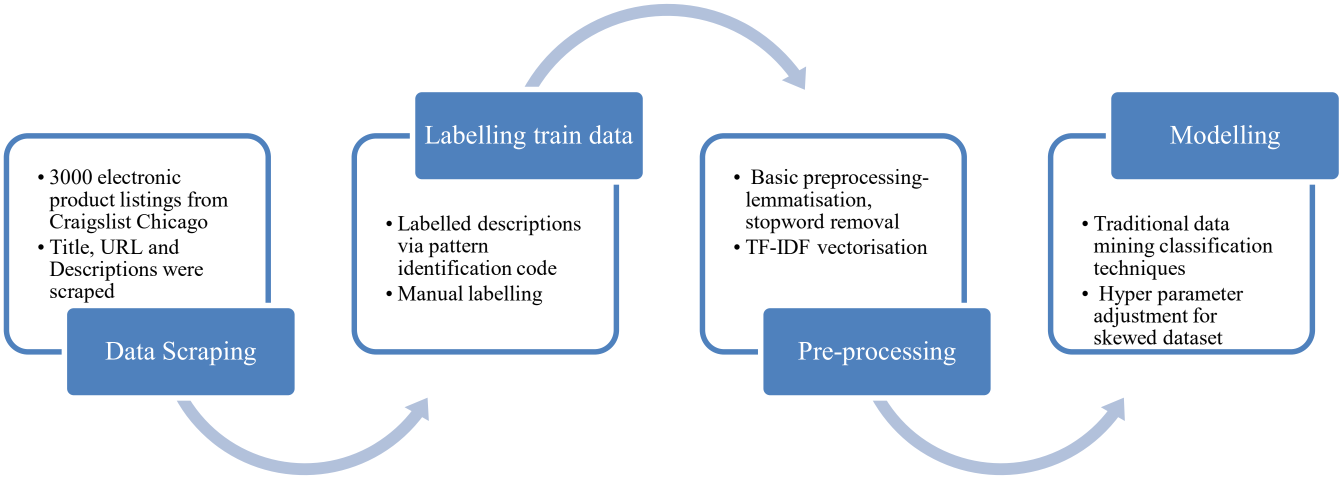
Project Methodology



Data Pre-processing:

1. Performed manual and code assistant annotation to label categories
2. Removed embedded URLs and footer texts from the description
3. Converted text to lower case
4. Tokenized each word
5. Removed non-ASCII characters
6. Removed stop words and punctuations
7. Removed unwanted digits
8. Converted variables to categorial variables

Project Methodology:


• 3000 electronic product listings from Craigslist Chicago
• Title, URL and Descriptions were scraped
Data Scraping
Labelling train data
• Labelled descriptions via pattern identification code
• Manual labelling
• Basic preprocessing- lemmatisation, stopword removal
• TF-IDF vectorisation
Pre-processing
Modelling
• Traditional data mining classification techniques
• Hyper parameter adjustment for skewed dataset

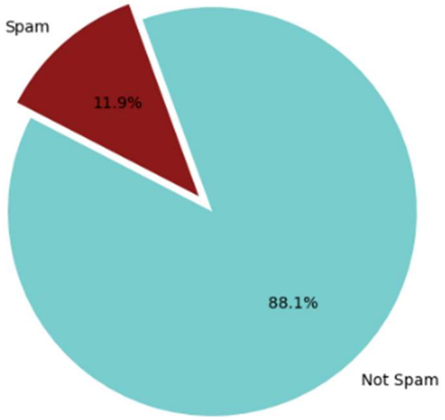
For the project, we scraped the Chicago Electronics category product listings from the Craigslist website (https://chicago.craigslist.org/search/ela).

The scraped file consisted of product title, URL link and description. For training our model, we required a labelled sample of sizeable amount (3000 in this case). The labelling was done using two methods- a) Coded labelling through pattern identification b) Manual labelling.

For code assisted labelling, a generic pattern of spam recognized through the data set was taken into consideration:

- Descriptions not having an authentic QR Code Scan
- Use of words such as free, voice mailing, credit card details, checks, western union
- No price information
- Grammar and spelling mistakes
- Offered free shipping
- Checked the correlation of post description with spam email examples (cosine similarity)

The rest were done manually. Post labelling and data exploration, we found our dataset to be heavily imbalanced:



This was taken into consideration as we went ahead with our model

Step by Step Process of Modelling the Solution:

In our modelling solution, we have 3 steps:

a. Data Pre-processing: Standard process: Lemmatization and Stop word removal

b. Vectorization: We chose TF-IDF vectorization over CountVectorizer since we wanted to highlight the importance of terms within each document (in this case description).

c. Traditional data mining methods: We used classification models like Naïve Bayes, Random Forest, Linear and Polynomial SVM and XG Boost for model training. Due to high class imbalance, we tuned our hyperparameters to classify minority class better. In the XGBoost model, we used weighted **scale_pos_weight**

hyperparameter adjustment. We also tried SMOTE oversampling to improve results, and a stratified Train-test split ration of 80:20. Our final model, which gave the best recall output was Random Forest with class weighting.

- **Spam Detection Model Comparison:**

Model Comparison					
	Precision		Recall		Accuracy
	0	1	0	1	
Naïve Bayes	91%	31%	89%	37%	83%
Random Forest- Class weighting (0:1, 1:8)	98%	43%	84%	87%	85%
Linear SVM	89%	100%	100%	7%	89%
Polynomial SVM (Degree 3)	0%	12%	0%	12%	12%
XG Boost (weighted)	91%	73%	99%	23%	90%

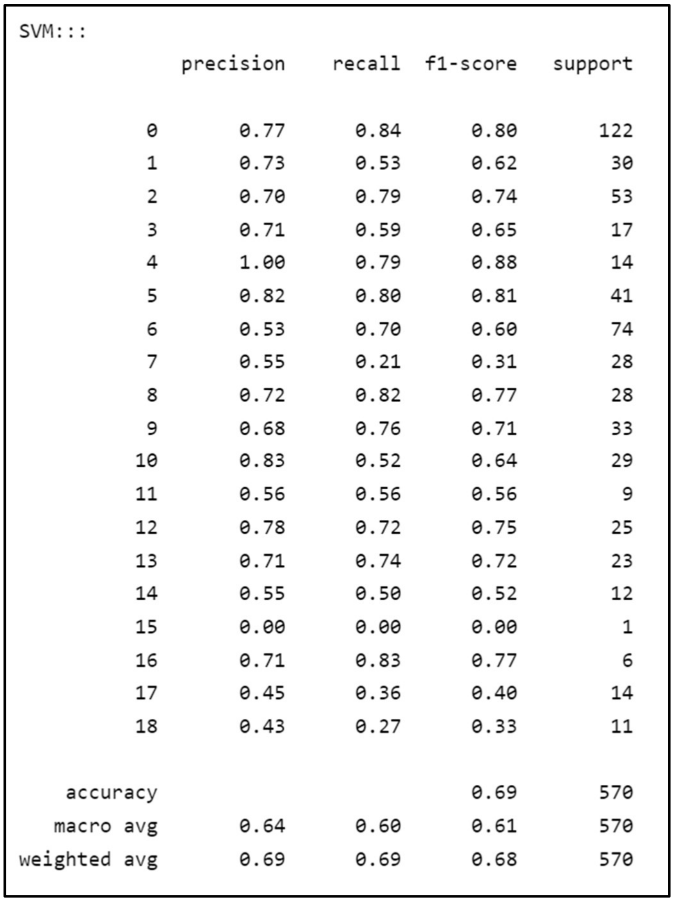
- **Classification Model Comparison:**

Models	Precision	Accuracy
Naïve Bayes	45%	39.12%
Random Forest	57%	51.05%
Linear SVM	69%	69.12%
XG Boost	67%	65.09%
KNN	58%	50.18%
Neural Network	50%	50%
Logistic Regression	67%	57.19%

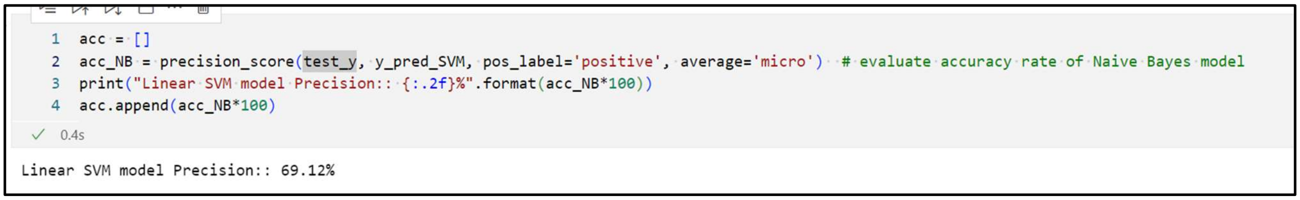
- **Results and validation:**

Classification: We tried multiple classification models such as Naive Bayes, Random Forest, Linear SVC and XG Boost. To evaluate the model performance, we decided on the accuracy metric to facilitate a better user experience since it would not be the end of the world if a few listings were falsely classified.

Based on our analysis, Naive Bayes had the least accuracy of about 39% and our best performing model was the support vector classifier which is used for both classification and regression. We used a linear kernel since electronics could be easily bucketed into disparate categories. Since we had about Linear SVC works best on small datasets. By running this model, we were able to accurately predict categories such as printer, phone, and audio equipment and this is since the postings in the category were very specific and correctly structured.



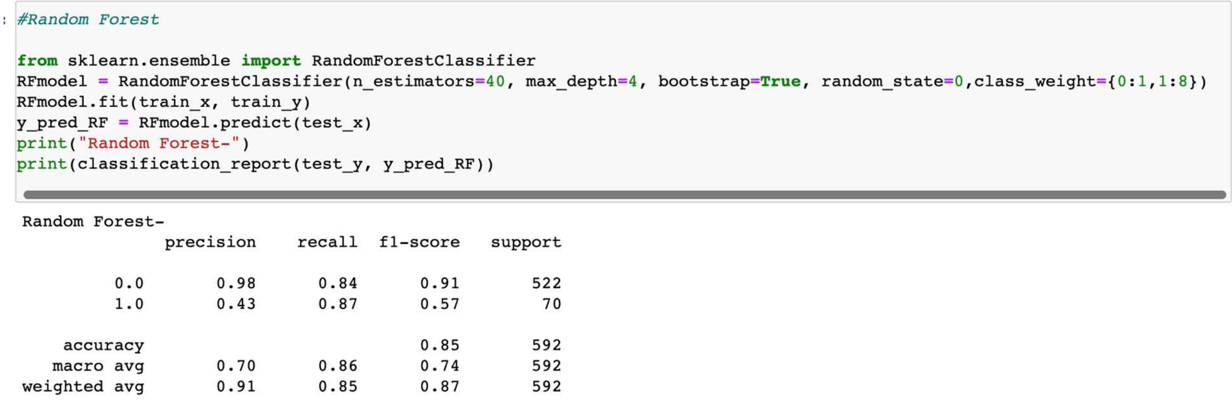
```
SVM:::
              precision    recall  f1-score   support

           0       0.77      0.84      0.80       122
           1       0.73      0.53      0.62        30
           2       0.70      0.79      0.74        53
           3       0.71      0.59      0.65        17
           4       1.00      0.79      0.88        14
           5       0.82      0.80      0.81        41
           6       0.53      0.70      0.60        74
           7       0.55      0.21      0.31        28
           8       0.72      0.82      0.77        28
           9       0.68      0.76      0.71        33
          10       0.83      0.52      0.64        29
          11       0.56      0.56      0.56         9
          12       0.78      0.72      0.75        25
          13       0.71      0.74      0.72        23
          14       0.55      0.50      0.52        12
          15       0.00      0.00      0.00         1
          16       0.71      0.83      0.77         6
          17       0.45      0.36      0.40        14
          18       0.43      0.27      0.33        11

    accuracy                           0.69       570
   macro avg       0.64      0.60      0.61       570
weighted avg       0.69      0.69      0.68       570
```



```
acc = []
acc_NB = precision_score(test_y, y_pred_SVM, pos_label='positive', average='micro')  # evaluate accuracy rate of Naive Bayes model
print("Linear SVM model Precision:: {:.2f}%".format(acc_NB*100))
acc.append(acc_NB*100)
```

```
Linear SVM model Precision:: 69.12%
```

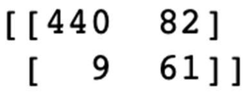
Spam Detection: The class weighted Random Forest model was what gave us the best results in terms of classification.



```
#Random Forest

from sklearn.ensemble import RandomForestClassifier
RFmodel = RandomForestClassifier(n_estimators=40, max_depth=4, bootstrap=True, random_state=0,class_weight={0:1,1:8})
RFmodel.fit(train_x, train_y)
y_pred_RF = RFmodel.predict(test_x)
print("Random Forest-")
print(classification_report(test_y, y_pred_RF))
```

```
Random Forest-
              precision    recall  f1-score   support

         0.0       0.98      0.84      0.91       522
         1.0       0.43      0.87      0.57        70

    accuracy                           0.85       592
   macro avg       0.70      0.86      0.74       592
weighted avg       0.91      0.85      0.87       592
```

We chose Recall as our business consideration, since our ultimate objective was to not miss out unidentified Spam in our test data set. The confusion matrix for our results looked like this:



```
[[440  82]
 [  9  61]]
```

You can see here that maximum 1s have been truly classified, however the model does suffer from low Precision. This is a trade-off we took.

This has been our biggest learning out of the modelling, which we'll consider as we proceed with further classification-based learning models in the future.

Recommendations and Future scope:

- Categories clustered for electronic devices should be provided as filters
- There should be a structured way for the sellers to sell their products with some mandatory information like price, category
- Have validation checks to avoid duplicate posts by tweaking the 30-day retention post for users
- Currently classification has been done via manual annotation. A sophisticated way to classify the images would be to use image captioning, which requires higher computational cost. (GPU: $3.96/hr)

Appendix

1. https://www.dummies.com/article/technology/internet-basics/spot-scam-craigslist-242902/
2. https://www.craigslist.org/about/scams
3. https://par.nsf.gov/servlets/purl/10095443
4. https://diamondvalleyfcu.org/blog/8-ways-avoid-getting-scammed-craigslist
5. https://towardsdatascience.com/spam-detection-in-emails-de0398ea3b48