AUTHENTICATING CATEGORIZATION OF SALES ADVERTISEMENTS

A Project Report

Submitted by

DATA.AI

Agrawal, Sakshee

Gawande, Surbhi

Harlalka, Shipra

Lathi, Saumya

Patel, Prerak

Kalra, Harsh Vardhan Singh

ACKNOWLEDGEMENT

During this work, the constant association with Dr. Jinyang Zheng and Mr. Weilong Wang has been most pleasurable. We thank them for this help and counsel and wish to convey our gratitude for the same.

We would also like to extend our thanks to our classmates for their unconditional support unstintingly given during the completion of this work without which it would have been immeasurably difficult for us.

TABLE OF CONTENTS

ABBREVIATIONS USED

ABBREVIATION	FULLFORM
SVM	Support Vector Machine
RF	Random Forest
NB	Naïve Bayes
SVMP	Support Vector Machine Polynomial
CNN	Convolutional Neural Network
TF-IDF	Term frequency–inverse document frequency
SMOTE	Synthetic Minority Over Sampling Technique

1. BACKGROUND

Craigslist is an American classified advertisements website with sections devoted to jobs, housing, for sale, items wanted, services, community service, gigs, résumés, and discussion forums. 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. It assists users in buying and selling products and services in over 450 cities worldwide.

Craigslist's vision is to "Restore the human voice to the Internet, in a humane, non-commercial environment.", this makes it clear that the company is more focused on user satisfaction rather than just monetary gain.

Every month, it has more than 25 million new classified ads, 75 million user postings and over 10 million photos updated. Craigslist has a wide variety of diverse categories, free services, and local trading options due to which it has an extremely high online traffic in comparison to its other online marketplace competitors.

However, despite the massive web traffic, Craigslist has been criticized for its outdated UI design and business model. This could be due to the large amount of unstructured data they have to deal with daily, which raised problems of spam ads, wrongly categorized postings, and prohibited content. The user dissatisfaction towards Craigslist has largely been due to spam and miscategorized data, and we aim to solve this issue for them. Since the company values the user experience on their website, we believe, by helping them remove spam ad postings, our service will add great value to their company.

2. INTRODUCTION

2.1 Business Problem

Craigslist aims to provide a very simple, honest and non- commercial environment to the user. They understand what their users want and tend to keep things simple, real and common-sense based. They provide a free non-commercial marketplace that connecting buyers and sellers across geographies. Millions of ads are posted every single day on Craigslist worldwide to a large extent anonymously. It is difficult to check the listings for the people looking for them. This project intends to solve the problem of spam advertisements and give the users a good experience. This project consists of applying data analysis and text analysis concepts & techniques towards the detection of spam advertisements.

2.2 Project Objectives

- To create a model to filter spam advertisements
- Applying Data analysis concepts to solve the problem
- Enrich the user experience

2.3 Process Flow

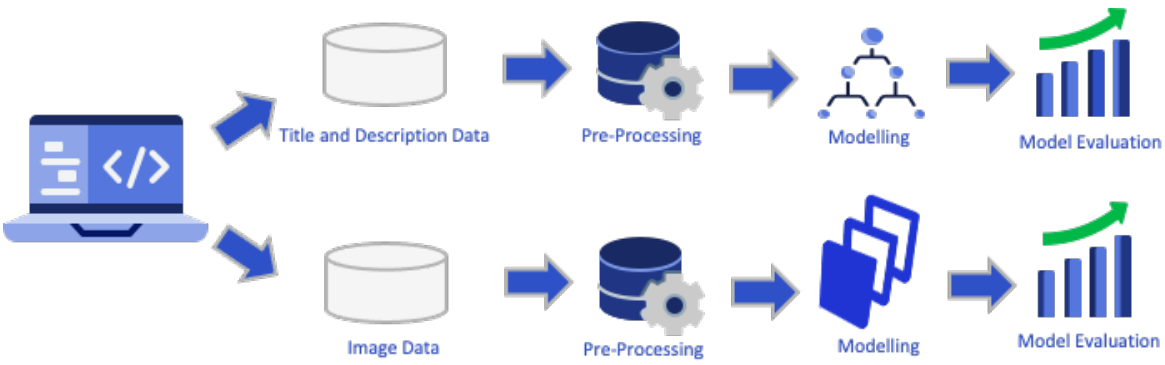

Figure 1 Process Flow Diagram

3. DATA ANALYSIS

3.1 Text Classification Models

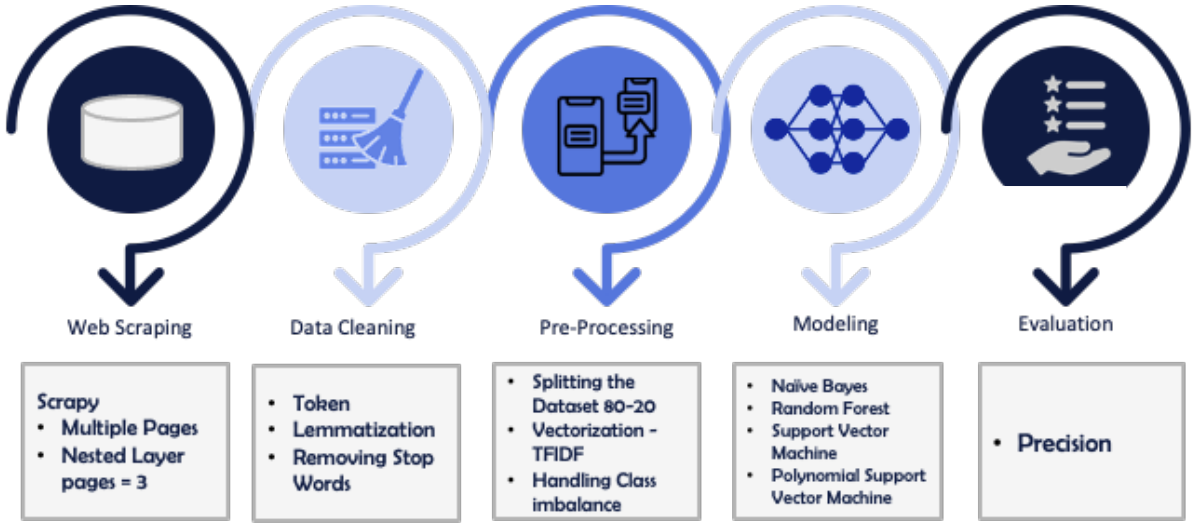

Figure 2 Text Classification Model

We first started the text analysis by scraping the title and description from several existing ad posts in the boat section. Using the scrapy library we crawled through over 45 categories. Among the 45 categories we shortlisted data for boats section for 5 different counties. Each category had over 100 results and we collected the data for title, description, price and hyperlinks. After collecting the data we labelled the data and prepared a csv file that could be used to train our model.

Next, for data pre-processing, we applied the nltk package to tokenize the titles, convert them to lower case on a word level, and then used WordNet Lemmatizer to lookup lemmas. We proceeded to further reduce the dimionsionality by removing the stop words, which comprise of the less important, high-frequency words or low-frequency words.

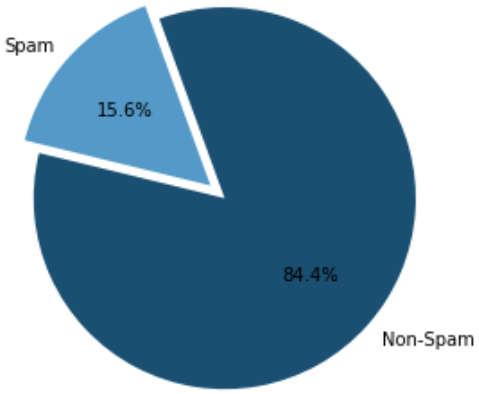

Figure 3 Distribution of Spam and Non-Spam Advertisements

We then vectorized our data using the TF-IDF vectorizer. We observed low precision scores but high accuracy scores initially, which made us realize that our data had a severe class imbalance in terms of the target column, the proportion of no spam to spam was around 85%-15%. To overcome this imbalance, we applied SMOTE which stands for Synthetic Minority Oversampling Technique, this improved our ratio upto 25%.

We then used this preprocessed and cleaned data to train several models on it to find the best fit. The four models we trained were Naïve Bayes, Random forest, Support Vector Machine, and the Polynomial Support Vector Machine. Since the size of our dataset was quite small with only 500 rows, more complex models like Neural networks and deep learning did not give a satisfactory score, instead the more traditional models provided us with better results.

For the performance evaluation, we used precision as our judging metric instead of accuracy, as it would give a more realistic result with the size and skewness of our data.
Precision tells us that out of those predicted positive, how many of them are actual positive, so it's a good measure especially when the costs of False Positive is high, like it would be for Craigslist.
Hence, by following these steps we were able to analyze the text data from the description as well as title of a craigslist ad posting and classify it as spam or not spam.

3.2 Image Classification Model

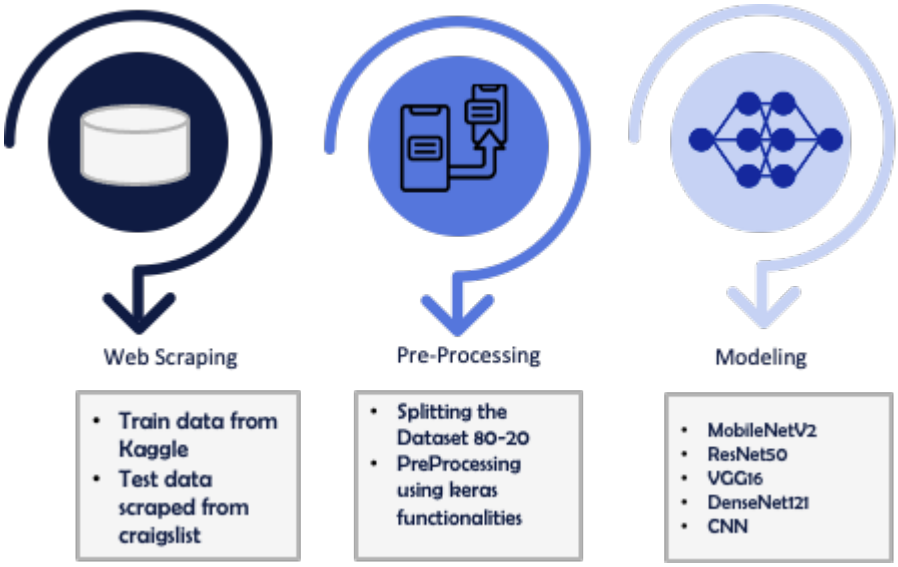

Figure 4 Image classification Model

We started the image classification model by getting our training dataset downloaded from the Kaggle which had 1500 labeled images of 8 categories of boat. For the test dataset, we scraped 195 images containing unclassified boats as well as some images of boat parts from craigslist. We then created a new in the training dataset naming "Not Boat" with photos collected from the craigslist which generally contains the image of boat parts.

For the pre-processing, the data was split into 80:20 ratio where 80 was training dataset and 20 was kept for data validation. The data was further augmented to reduce the bias towards dataset. ImageDataGenerator was used for this purpose.

We then train this model on four transfer learning models –

- MobileNetV2
- ResNet50
- VGG16
- DenseNet121
- one CNN-based model with below architecture created. by us:

```
Model: "sequential_5"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
conv2d_15 (Conv2D)           (None, 220, 220, 128)     9728
_________________________________________________________________
max_pooling2d_14 (MaxPooling (None, 110, 110, 128)     0
_________________________________________________________________
conv2d_16 (Conv2D)           (None, 106, 106, 64)      204864
_________________________________________________________________
max_pooling2d_15 (MaxPooling (None, 53, 53, 64)        0
_________________________________________________________________
conv2d_17 (Conv2D)           (None, 49, 49, 32)        51232
_________________________________________________________________
max_pooling2d_16 (MaxPooling (None, 24, 24, 32)        0
_________________________________________________________________
flatten_5 (Flatten)          (None, 18432)             0
_________________________________________________________________
dense_66 (Dense)             (None, 512)               9437696
_________________________________________________________________
dense_67 (Dense)             (None, 128)               65664
_________________________________________________________________
dense_68 (Dense)             (None, 9)                 1161
=================================================================
Total params: 9,770,345
Trainable params: 9,770,345
Non-trainable params: 0
_________________________________________________________________
```

Figure 5 Architecture of the CNN based model created

For the transfer Learning model, we have added two more layers in the default architecture, first layer contains Dense network with 128 neurones with “relu” as activation function and the second layer contains dense network of 9 neurones with SoftMax as activation function. The last layer of neurones is 9 because we have 9different categories that need to be predicted. After training the dataset on the five models, we have created different plots for checking model performance and to choose the best model for our case.

4. MODEL PERFORMANCE

4.1. Text Classification Model

To analyze the performance of our models, we validated the 4 trained models against our test dataset and got the precision scores of 30%, 67%, 67% and 60% for Naïve Bayes, Random forest, Support Vector Machine and the Polynomial Support Vector Machine, respectively. Naïve Bayes was the least performing at 30% while SVM and Random Forest were giving similar scores of 67% precision.

Naïve Bayes probably didn't work well because it works by assuming all the predictors are independent of each other, which is not the case with our data. The polynomial SVM is more flexible than the linear SVM, and hence might be overfitting the data resulting in a lower score.

Our final performance evaluation led us to select the Support Vector Machine as our best model since it gave better validation results and higher precision. SVM is also faster than the Random Forest and is more memory efficient.

The scores of the model performance can also be observed in the graph below.

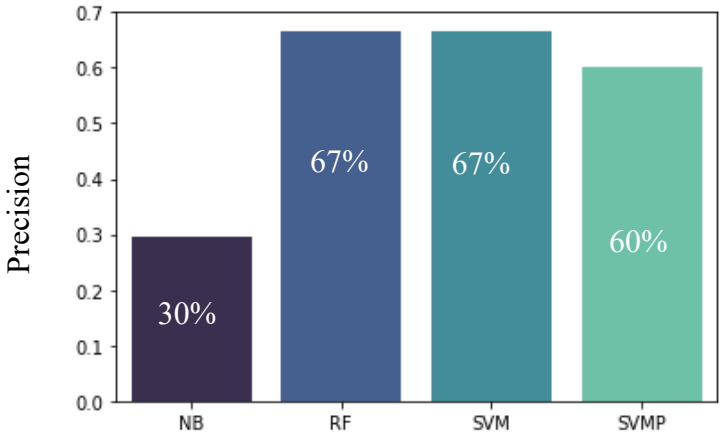

Figure 6 Precision of Text classification Model

4.2. Image Classification Model

From the classification report of all the model we can see that the MobileNetV2 performed best on both validation and train data. The average precision for predicting the class was 0.89 while the recall was also 0.86 with accuracy of 86%.

The other model which had comparable accuracy was DenseNet121 but the accuracy for the validation data is little bit low as compared to the MobilenetV2. All other model performed significantly worse than MobileNetV2 therefore, we have chosen MobileNetV2 as our final model for image classification.

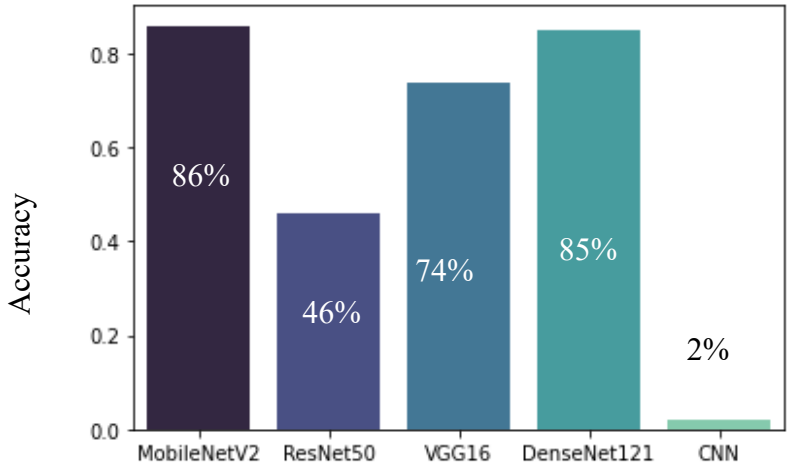

Figure 7 Acccuracy of image classification Model

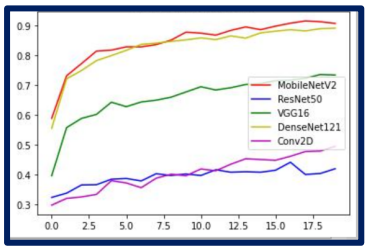

Figure 9 Accuracy Comparison

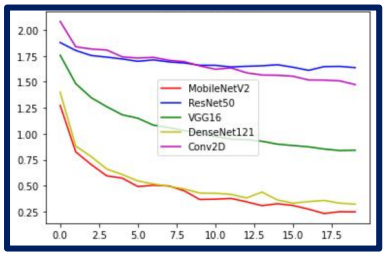

Figure 8 Loss Comparison

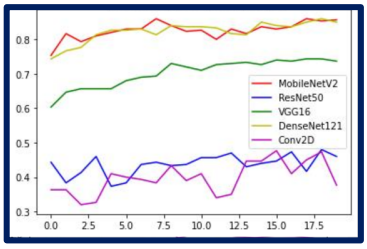

Figure 10 Validation Accuracy Comparison

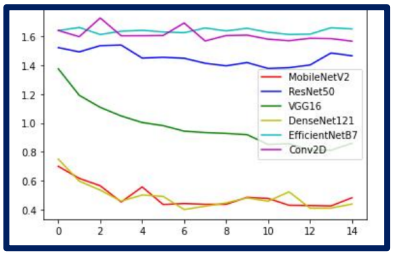

Figure 11Validation Loss Comparison

To test the model, we have scraped the images from the craigslist and run the model on few images that were either boat or a boat part. The model was successful in predicting the boat type correctly or if it is a boat part than the model predicted it to Not boat.

MobileNet v2	0	1	2	3	4	5	6	7	8
Precision	0.74	0.87	0.74	0.75	0.93	0.67	0.93	0.73	0.87
Recall	0.88	0.89	0.56	0.50	0.93	0.80	0.76	1.00	0.92
F1-Score	0.80	0.88	0.60	0.60	0.93	0.73	0.84	0.84	0.89

ResNet50	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Precision	0	0.33	1	0	0.5	0	0	0	0.46
Recall	0	0.24	0.12	0	0.65	0	0	0	0.91
F1-Score	0	0.12	0.22	0	0.57	0	0	0	0.61

VGG16	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Precision	0.7	0.77	0.75	0	0.67	0	0.71	1.00	0.75
Recall	0.44	0.91	0.19	0	0.88	0	0.64	0.62	0.87
F1-Score	0.54	0.84	0.30	0	0.76	0	0.67	0.77	0.81

DenseNet21	0	1	2	3	4	5	6	7	8
Precision	0.71	0.92	0.48	1	0.88	1.00	0.82	0.89	0.92
Recall	0.62	0.80	0.81	0.33	0.95	0.80	0.84	1.00	0.90
F1-Score	0.67	0.86	0.60	00.5	0.92	0.89	0.83	0.94	0.91

CNN	0	1	2	3	4	5	6	7	8
Precision	0	0	0	0.02	0	0	0	0	0
Recall	0	0	0	1.00	0	0	0	0	0
F1-Score	0	0	0	0.04	0	0	0	0	0

5. CONCLUSION

Aligning with the vision of Craigslist, we have created an authentication categorisation model which could help Craigslist in giving the users an enriched experience on the website keeping everything simple. As the False negatives are more harmful for the company, we decided on evaluating the model on precision.

We have created a model on a very small dataset therefore the traditional models worked better than a more advanced model like Neural network, Gradient boosting which could be used for the further studies. Apart from this, our model is scalable and can be applied to other categories as well. 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.

Data.AI profoundly believes that the thorough analysis and valuable model could help Craigslist not only to improve the user experience dramatically but also helping the advertisement posters to know if the images uploaded by them are proper not hence making Craigslist a more friendly online community.

6. REFERENCES

[1] 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.

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

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

[4] 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/

[5] Kaggle Image Classification Training Dataset- Boat Types Recognition from https://www.kaggle.com/clorichel/boat-types-recognition