Analyzing Unstructured Data

Final Report

Logo

Description automatically generated

BAIM AUDitors

Team 9

Chethan Manjunath

Rohan Walyat

Yamini Ambati

Yash Kulshreshtha

Yi Chen Chu

# **Background**

Online advertising, a marketing technique that utilizes the internet to promote services and products, is still growing rapidly, expected to reach US$ 349.6 Billion by 2027 by IMARC Group. Some of the most popular advertising websites in USA are Craigslist, Free Ads Time, FinderMaster, AdvertiseEra and so on. To assist our client Craigslist to have a bigger share of the market, we strive to improve the users’ experience on Craigslist. Job postings, as the very first addition to Craigslist’s website, catch our attention for advancement. We noticed that there are lots of misclassified job postings on the website, which would have caused lower customer engagement and customer retention. If people keep seeing misclassified job posts on the website, they might think it’s an untrustworthy place to look for jobs. Thus, it is very important for Craigslist to help employers put their job postings in the right section on the website.

**Who is Craigslist?**

Craigslist is an American classified advertisements website with different sections such as jobs, housing, sales, services, and a whole lot more. It works like an online version of a classified section of a newspaper and can be filtered by your local region. Since it is free, people will be willing to put things that people might bother to put if it’s a paid ad. As a result, it is very convenient for people to connect with a huge range of people around the world to find, buy, or sell just about anything.

**Our subsection of the project:**

Craigslist can be a great source to find a job. People can find listings for a wide variety of fair opportunities, from part-time jobs to proper full-time ones. However, you can find a plethora of misclassification on Craigslist, which might cause the users’ inconvenience and inefficiency while searching for jobs in the correct criteria they want. Although Craigslist is not a website especially for job-hunting like LinkedIn, we’d like to improve the job-hunting experiences for people so that we can retain both employers and job seekers and make them keep coming back to the website. It is very important since job listings is one of the sections that Craigslist can ask for fees.

**Why educational jobs?**

With a looming recession and economic turbulence, comes the risk of job uncertainty and losses. Multiple sectors are affected in these times, especially the non-essential ones such as automobile, tourism, and real estate. Education/Private Tutoring domain is no different when in a budget crunch. Even schools cut down on staff.

Craigslist is an established platform that supports a big chunk of US gig economy, including the education sub-sector. A considerable chunk of the population relies on Craigslist to connect jobseekers and employers, and so do people from the fields of education/tutoring. Given the increasing need for this platform in current economic scenario, it is imperative that the right listings are shown to people looking for job/employees. Improperly classified records become a source of frustration and demotivation for people.

We’ve noticed around 20% of misclassification of jobs in education on the website. It would be a time-consuming job if Craigslist hire people to detect those posts and put them to the section that they belong to. The project objective is to catch posts that should not be put in the section of education of job listings, which we believe will add a big value to both our client (Craigslist) and its users. Users will have to spend less time finding the right job they want to apply for, which will lead to better reputation for Craigslist and increase its revenue.

# **Business Analysis**

# **Project objective** Using ML, we’ve tried to lower the misclassification rate of education-related job postings on Craigslist for the benefit of people trying to get their next employment through the website. We scraped the relevant data from Craigslist for 82 different US cities. To ensure that our analytics were robust, we took a random sample including cities of all sizes and demographics. Our objective is to enhance users’ experience in finding an education-related job on Craigslist by showing them correct job listings. We will remove jobs that are not related to education to reduce the users’ job investigation time and efforts. To obtain this, we need to train a robust classification model to ensure we reach a decent correctness in classifying the jobs. We have used various machine learning algorithms such as Naive Bayes, Support Vector Machines, Random Forest, and Logistic Regression to classify the jobs into two categories: education-related and non-education related. We have also used various techniques such as hyperparameter tuning, and cross-validation to improve the accuracy of our model. We have also used various evaluation metrics such as ROC Curve and accuracy to evaluate our model's performance. Finally, we have compared our model's performance with other existing models and found that our model outperforms them in terms of accuracy and F1 score.

# **Data Analysis**

Given the business problem and objective decided, a framework was constructed to develop the solution. Identifying this as a classic example of classification problem under supervised learning, the below approach has been used.

A picture containing icon

Description automatically generated

**Scraping:** The Craigslist website has been used to scrape exclusively the ‘education’ sub section under the ‘job’ category. A combination of Beautiful Soup in addition with Selenium HTML parsers are used for the same. List of cities were selected at the first place and then links for all the posts were obtained. Consequently, all the relevant information about the Job postings such as employment type, compensation, title, description etc. were collected.

**Data Pre-Processing:** This step in the approach involved multiple sub steps given the data has lot of text and thereby not clean. The stages included are as follows:

1. Removal of Punctuation: As lot of posts had whole lot of descriptions, the first stage excluded the punctuations (a set of 32 different symbols) before proceeding
2. Tokenization: Now the data is clean from symbols and noise related to it, the sentences are split into tokens to be able to convert them into vectors
3. Lemmatize: Word net lemmatizer was used to take care of inflected variations of single words. Thereby helping in reducing the dimensionality of the text dataset
4. Stop words: The final stage within the data pre-processing was the removal of stop words in English. This ensured the elimination of commonly used words.

Performing these steps addressed the dimensionality reduction problem in the earlier steps to be more precise and optimized.

**Manual Labelling:** For the machine to correctly classify, a proper grouping of correctly and incorrectly classified postings need to be identified. A manual review of all the data was done to correctly tag the misclassified posts. Given the number of negatives or mis classified are comparatively in good number (~20%), any additional sampling technique wasn’t used.

**EDA:** Before moving on to the modelling, a preliminary understanding of what kind of postings were being posted was analyzed. Based on the collected 1000 postings, topic modelling, an unsupervised learning technique was performed across the descriptions and titles text data. This helped us uncover abstract topics and cluster of words within the dataset which thereby can even use to tag the posts as an additional step in future. The analysis from the topic modelling are as follows:

The top 8 words for each of the topics modelled for job description gives us as idea that most of the posts consist of requirements seeking experienced schoolteachers. The other topics cover a range of requirements consisting of tutors, contractors, beginner teachers, sports coaches etc. While the probability of each post being mapped is higher for Topic 0 (as shown in the graph below), assumptions can be made for misclassified jobs posted in the education section may also consist of relevant words, but the function of the jobs may not be directly related to the education category.

Table

Description automatically generated with medium confidenceChart

Description automatically generated

Further 3 topics consisting of top 8 words were analyzed for job titles. This analysis indicated that most job titles are similar or standard for jobs that truly belong to the education category or those that don’t belong to the education category. As shown in the graph below, the distribution of most probable job titles in each topic is almost the same.

Graphical user interface, text

Description automatically generated with medium confidenceChart, bar chart

Description automatically generated

**Test Train Split:** Moving towards the end objective, it is required to split the labelled dataset into test and train dataset and avoid overfitting the training dataset. The 1000 postings were split in the 70:30 ratio with test data set having 30% of misclassified labels and rest 70% with training data.

**Word2Vec:** Any supervised or unsupervised modelling technique requires a set of numbers to build the model. As a result, all the text currently present in the data needs to be moved into numbers. Different set of techniques, algorithms and pre-trained models are present for the conversion. However, the results from topic modelling showed a higher similarity to the education posts we were analyzing, thereby to capture that information, we proceeded with utilizing the TF\_IDF technique for text vectorization. Bigrams were used to capture more information with constraint on minimum document frequency to eliminate unnecessary terms.

For the description text documents, min\_df of 6 was given. Whereas a min\_df of 3 was used for the titles of these postings which are usually much shorter than the descriptions.

**Training Classification Model:** The most crucial stage in the entire methodology is the building of a classification model that can accurately identify misclassified posts within the education subcategory. Different classification machine learning models were built on the training dataset.

All the vectorized text columns from the training were combined to form the final independent set of X variables. Both the manually labeled column, Y and vectorized text, X are fit into supervised classification algorithms.

Machine learning techniques used to train the training dataset are as follows:

1. Logistic Regression: Since the target variable is only 0 or 1, miss classified vs correctly classified, logistic regression is used with multiple penalty functions available. Estimated based on log odds of an event
2. Support Vector Machine: Since SVM tries to find a plane that splits the class, SVM has been used as well.
3. Gaussian Naïve-Bayes: Naïve Bayes also being a probabilistic machine learning model used for classification, it’s been used to train the data
4. Multinomial Naïve-Bayes: When compared to Gaussian, since these are distinct words, multinomial was used as they tend to do better on such data
5. Random Forest: Since it’s an ensemble of random decision trees, this also has been used
6. XG Boost Classifier: It being the ultimate ensemble model that outperforms most of the supervised learning techniques, XG Boost classifier is used as well
7. Multilayer Perceptron – Fully Connected Neural Network: This neural network was used to capture the non-linearity between the dependent and independent variables.

But there needs to be a way to evaluate which model is performing the best before finalizing it with the client.

# **Validation**

This final stage in the process flow helped us analyze which model performed good when compared to the rest. The validation is performed using the rest 30% data which was split earlier.

Before predicting the labels, validation set which are words are also converted to vectors by transforming on to the TF\_IDF words vocabulary. As a preliminary analysis, the scores of the different models were analyzed to understand which models performed overall better. The initial accuracy scores for various models are highlighted below:

Initial Models:

|  |  |
| --- | --- |
| **Model** | **Accuracy** |
| Logistic Regression | 86.3% |
| Support Vector Machine | 84.9% |
| Multinomial Naïve Bayes | 81% |
| Gaussian Naïve Bayes | 85.6% |
| XG Boost Classifier | 87% |
| Random Forest | 83.15% |
| MLP Neural Network | 86.7% |

Given the initial iterations gave good scores for most models, we proceeded with doing hyper parameter tuning to identify the best parameter for the best models.

Since the main business context is to identify the misclassified posts and then help client to flag them in the future, looking at only accuracy score can often be misleading in case of classification problems, given the bias in positive and negative samples. So, we considered the area under ROC curve which isn’t affected by the sample bias.

|  |  |  |
| --- | --- | --- |
| **Model** | **Grid Search Parameters** | **Best Parameter- ROC Curve** |
| Logistic Regression | param=  {'penalty': ['l1',**'l2'**],  'solver’: ['newton-cg','lbfgs',  **'liblinear'**]} | Chart, line chart  Description automatically generated |
| XGBoost Classifier | param= {'learning\_rate': [0.001, **0.1**, 1, 10]} | Chart, line chart  Description automatically generated |
| Multi-Layer Perceptron | param=  {'hidden\_layer\_sizes’: [(**5,3**), (6,5),(10,3)], 'activation':['tanh',**'relu'**], 'solver':['sgd', **'adam'**], 'alpha':[0.001,0.05,**0.1**]} | Chart, line chart  Description automatically generated |

Along with ROC Curves, looking at confusion matrix; help us chose the **XGBoost Classifier** with a learning rate of 0.1. Keeping in mind the reduction of miss classified posts, reduction of false negatives was of the critical importance in this case.

# **Conclusion**

Using the model built is beneficial to not just the client, Craigslist in this case, but also to the users, which is kind of a win-win situation. It helps users find the relevant posts by escaping the myriad of irrelevant content and thereby improving user experience with the website, which in turn benefits the Craigslist. This additional filtering or flagging thereby provides Craigslist in creating a competitive advantage over others.

Value addition that can be brought to the client Craigslist by implementing this automatic misclassification are as follows:

* Overall improvement in Craigslist reputation for better handling of postings
* Higher customer engagement given the presence of comparatively more relevant content
* Improves the customer lifetime value, in terms of relying on craigslist for longer durations
* Enhances the customer trust due to higher percentage of correctly tagged posts in the results
* Become an authorized or huge trusted resource for any jobs starting from lower wage to moderately waged contracts

Coming to the value addition for user by flagging the misclassified posts are as follows:

* Finding the right content in relatively lower times
* Enhancement in reliable number of postings to go through and ease to apply
* Elimination of irrelevant or misclassified content thereby narrowing down on the number of posts to look for.

This way, adopting the model keeps Craigslist and users in a win-win strategy. Also, this increases in trust and reputation, could potentially turn into higher revenues for the craigslist, thereby more reason to implement it.

# **Future Scope**

Given the constraints on the computation power to scrape the entire Craigslist, relatively lower volumes of data were used than any traditional machine learning model. So, gathering more and more data could be a potential enhancement for the machine to better classify the data.

Also utilizing the spam repositories or spam archives from relevant local bodies could help classify the postings much accurately owing to the richer dataset present.

Finally, gathering additional data should allow in building much more densely connected deep neural network models which studies the relations beyond regular linearities.

In addition, the scope of the problem statement could be extended and thereby more categories can be included form the entire Craigslist.

**Appendix**

Tuning the parameters to improve the performance:

1. Word2vec Representation: Changing Minimum Document Frequency from 6 to 3 incase to identify the loosing of any information: But the confusion matrix wasn’t better than the 6 document frequency :

[[ 34 21]

[ 18 206]]

**ROC Curves:**

Logistic Regression: Using liblinear solver and l2 penalty

Chart, line chart

Description automatically generated

XGBoost Classifier: Learning Rate: 0.1

Chart, line chart

Description automatically generated

Multilayer Perceptron: With relu as the activation function, adam solver and alpha equal to 0.1. It had 5 layers of 3 nodes each.

Chart, line chart

Description automatically generated

**Confusion matrices:**

Confusion matrix showing the false positives and true negatives for the top 3 models:

Logistic Regression: Using liblinear solver and l2 penalty

|  |  |  |
| --- | --- | --- |
|  | Actual 0 | Actual 1 |
| Predicted 0 | 20 | 35 |
| Predicted 1 | 3 | 221 |

XGBoost Classifier: Learning Rate: 0.1

|  |  |  |
| --- | --- | --- |
|  | Actual 0 | Actual 1 |
| Predicted 0 | 35 | 20 |
| Predicted 1 | 17 | 207 |

Multilayer Perceptron: With relu as the activation function, adam solver and alpha equal to 0.1. It had 5 layers of 3 nodes each.

|  |  |  |
| --- | --- | --- |
|  | Actual 0 | Actual 1 |
| Predicted 0 | 35 | 20 |
| Predicted 1 | 26 | 198 |