

MGMT 590 – Analyzing Unstructured Data

Final Project Report (Milestone-2)

**Consulting Ninjas**

Niyathi Aiella

Sohini Chakraborty

Jason Crawford

Shreyas Joshi

Mahima Kriplani

Amal Tom

**Table of Contents**

Background………………………………………………………………………………………..2

Project Objective…………………………………………………………...……………………...2

Data analysis/Methodology……………………………………………………………………….3

Data scraping……………………………………………………………………………...........3

Data cleaning…………………………………………………………………………………...4

Modelling and training………………………………………………………………….…………5

Random forest…………………………………………………………………………………..5

Linear SVC...…………………………………………………………………………………...5

Multinomial NB………………………………………………………………………………...5

Logistic Regression……………………………………………………………………………..5

Decision Tree……………………………………………………………………………….......5

Bert: ……………………………………………………………………………….....................5

Classification Models……………………………………………………………………………...6

Model Validation……………………………………………………………………………….....7

Recommendation to recruiters……………………….……………………………………………9

Conclusion………………………………………………………………………………...............9

Future Scope………………………………………………………………………………..........10

Exhibits………………………………………………………………………………..................11

**BACKGROUND**

Craigslist is an open platform used for posting classified advertisements. These advertisements can be related to products, properties, skills, jobs or even discussions. One such scope of work covered on Craigslist is **Job advertisement**. The job listings on Craigslist do not follow a particular structure. Upon visiting the website, the users can only see the broad category of the job, date, title, and location. There is no specific information about the niche category of job posting, employment type, compensation, or organisation available. This means that the users need to visit each posting individually to find a job posting that suits him. This lack of structure of job postings makes it less popular than other widely used websites like indeed.com. The recruiters often end up posting the same ad twice with the exact same description which does not lead to any better response rate from candidates; this ultimately leads to driving the traffic away from the website.

**PROJECT OBJECTIVE**

We aim to use the methods for handling unstructured data that is scraped from Craigslist, extract relevant contents from it and classify it into each of its niche sub-categories. Further, we will provide recommendations to the recruiter to add features (in case they have missed it) which could possibly enhance the conversion rate of the ad posting.

We are planning to address the following categories:

1. **Education:** There are a lot of jobs available in craigslist in education sector which are not properly arranged or filtered according to the user’s needs. We are going to subcategorise them for the ease of search for the user.

Example: Text

Description automatically generated

As we can see in the above figure, there are many jobs which are tagged with titles which are difficult to comprehend. These will be tagged accordingly and categorised according to groups once we finish our project.

1. **Food/beverage/Hospitality:** This is another category which has a wide range of job listings but does not have subcategory tagging for ease of job search.

Example: Text

Description automatically generated

Here, there are jobs related to food like “Cooks wanted” or “Pizza maker”, but a person searching for a specific job he has experience in like sous-chef at a pizzeria, or a specific cuisine specialist cook may find it hard to search for jobs, requiring them to click into each posting to get more related information. Thus, we can apply filtering according to the data present inside and categorise them to make it easier for the user to find a desirable posting.

**DATA ANALYSIS / METHODOLOGY:**

**Data scraping:**

Data has been scraped from Craigslist using Selenium and BeautifulSoup. The program iterated through all the links in the Restaurants and Education categories for many cities and scraped the job descriptions from each link. One of the major constraints faced was scraping large amount of data at once from the website, as repeated efforts caused temporarily blockage from Craigslist. We were able to scrape 683 data rows from Education and 1464 data rows from Restaurants.

Graphical user interface, text, application

Description automatically generated

Fig: Scraped data for ‘education’ category

**Data cleaning:**

The data scraped in both categories had to be cleaned and pre-processed to conduct our analyses. The first step in that was to drop duplicate rows from the dataset. Following this, we tokenized job descriptions and removed punctuation associated with words after tokenization. For example, “job” and “job.” would be treated as different tokens, skewing our analysis. Hence, removing punctuation to equalize the tokens in our data was necessary. Next, we removed all html tags and stop-words. To improve accuracy of the model, we also created a list of high frequency words irrelevant to our categorization and removed them from the dataset. The cleaned data was then lemmatized for further analyses and added as a column to our dataset.

Graphical user interface, text, application, email

Description automatically generated

Fig: Cleaned and lemmatized data for ‘education’ category

Graphical user interface, text, application

Description automatically generated

Fig: Cleaned and lemmatized data for ‘Food/Beverage/Hospitality’ category

**MODELING AND TRAINING**

**TOPIC MODELING ALGORITHMS USED:**

***K-Means Clustering:*** K-means clustering divides observations into k clusters by minimizing distances between the data points within a cluster and maximizing distances between data points between different clusters. A similarity measure such as Euclidean-based distance or correlation-based distance is used to find homogeneous subgroups within the data. Different applications require different similarity measures.

***Latent Dirichlet Allocation (LDA):*** The LDA algorithm is one of the most famous topic modeling algorithms. It is a three-level hierarchical Bayesian model in which each item in a collection is represented as a finite mixture of topics. Topic probabilities are modeled as an infinite mixture over each topic. An explicit representation of a document is provided by topic probabilities in text modeling.

***Non-negative Matrix Factorization (NMF):*** A non-negative matrix, X, is factorized into a product of two lower rank matrices, A and B, so that AB approximates an optimal solution to X. NMF is an effective feature extraction technique for ambiguous and weakly predictable features. A meaningful pattern, topic, or theme can be derived from it.

***Hierarchical Dirichlet Process (HDP):*** In Hierarchical Dirichlet process (HDP) is a nonparametric Bayesian approach to clustering grouped data using statistics and machine learning concepts. For each group of data, a Dirichlet Process is used, with the Dirichlet processes for each group sharing a base distribution that itself is derived from a Dirichlet Process. Groups can share statistical strength by sharing clusters across groups. The HDP is an extension of LDA, designed to deal with cases where the number of mixture components (the number of "topics" in document-modeling terms) is unknown. As a result, both methods produce different results.

**Topic modeling in our Project:**

Two different models were built simultaneously to tackle the data in the two different categories. We start the modelling by performing a k-means clustering to identify the number of topics that can be utilized for an effective topic modelling. For this, we first convert the cleaned data into a tf-idf vector and compute the ideal number of clusters for the k-means algorithm.

NMF algorithm is usually used for shorter texts and tweets to get an accurate topic modeling. In order to check if NMF would be useful for modeling the topics in the job descriptions we calculated the length of the job description texts in our dataset. The texts are found to be long in length as depicted in the below histogram and hence, we did not go ahead with the NMF model.

Chart, histogram

Description automatically generated

Fig: Distribution for job description length

LDA and HDP algorithms were used to recognize topics in our project for both categories. To get an accurate result, we first convert the cleaned data into an input favorable for the LDA and HDP algorithms. Once the LDA model extracts topics, we compute the coherence score associated with the number of topics extracted. We use the number having the highest coherence score to decide the number of topics for modeling the data. However, for both job categories LDA outperforms HDP and leading us to reject HDP topics. Following identification of topics, we find out the dominant topic in each job description and map it to that data row. The dominant topic for each job description is calculated using the topic number with the highest percentage contribution in that document. However, we proceeded with K-means clustering for topic modelling as the clusters obtained with this method were intuitive and made more sense.

Chart, bar chart

Description automatically generatedChart, bar chart

Description automatically generated

Fig 2: Coherence score for HDP and LDA models for food/beverage category

Fig 1: Coherence score for HDP and LDA models

for education category

**CLASSFICATION MODELS**

The data in each category is split into a training dataset, which utilizes 75% of the data, and the remaining 25% data is used as test data. The former is used to train different models, while the latter helps evaluate the models built. For each category, we choose the model, providing us with the highest accuracy in correctly recommending words potentially missed in the job description but are like the dominant topic of that data.

We used the following algorithms to build the different models in our project:

* ***Random Forest:*** A random forest is a meta estimator that fits several decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting.
* ***Linear SVC:*** SVM or Support Vector Machine is a linear model for classification and regression problems. The algorithm creates a line or a hyperplane which separates the data into classes.
* ***Multinomial NB:*** Multinomial Naive Bayes algorithm is a probabilistic learning method that is mostly used in Natural Language Processing (NLP). The algorithm is based on the Bayes theorem and predicts the tag of a text. It calculates the probability of each tag for a given sample and then gives the tag with the highest probability as output.
* ***Logistic Regression:*** Logistic regression estimates the probability of an event occurring based on a given dataset of independent variables. Since the outcome is a probability, the dependent variable is bounded between 0 and 1.
* ***Decision Tree:*** Decision trees are a non-parametric supervised learning method used for classification and regression. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. A tree can be seen as a piecewise constant approximation.
* ***BERT:*** BERT (Bidirectional Encoder Representations from Transformers) is an encoder developed by researchers at Google AI Language. As opposed to directional models, which read the text input sequentially (left-to-right or right-to-left), the Transformer encoder reads the entire sequence of words at once, which makes it bidirectional. This characteristic allows the model to learn the context of a word based on all its surroundings (left and right of the word). Thus, it helps to capture semantic meaning of a word unlike other naïve encoders. There are various pre-trained models available, for our requirement we used ‘bert-base-uncased’, which uses masked language modelling and works well when the data is case insensitive. Job descriptions being case insensitive we opted for this model.

**MODEL VALIDATION**

Chart, line chart

Description automatically generatedGraphical user interface, application

Description automatically generated

Fig 1 Elbow curve or education category Fig 2 Elbow curve for food/Bev category

As observed in Fig. 1, the number of clusters from the elbow curve for Education category is found to be 10 or 18. We proceed with 10 clusters as the first elbow is observed at that number. We perform K-means clustering with the number of clusters as 10. Similarly, for the Food and Hospitality category we see that the first dip is observed at 10 and hence we perform the K-means clustering of the data with cluster size as 10. on each cleaned job description, we assigned each data row to a cluster. The results have been attached to Exhibit 1 and 2 respectively.

The LDA and HDP models were performed for both categories following the K-Means clustering. For the LDA, we arrived at the number of topics based on the coherence score calculated for each number of topics. Topic Coherence measures score a single topic by measuring the degree of semantic similarity between high scoring words in the topic. The coherence scores tables for both categories are attached below.

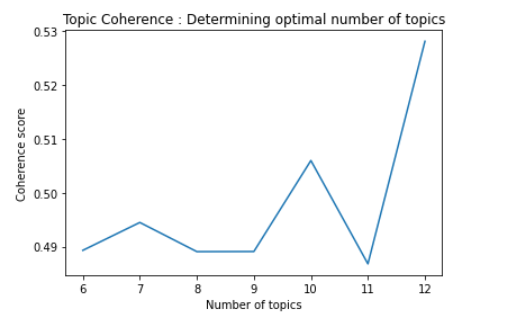
Table

Description automatically generated

Table

Description automatically generated

We can observe that for education category the highest coherence score is observed at 10 topics and for Food and Hospitality category the highest score is observed at 12 topics. The LDA models are developed using these numbers. We visually represented the coherence scores for different topics as a topic-coherence graph and these graphs are attached below for reference.



Chart, line chart

Description automatically generated

Fig 1: Topic coherence for ‘education’ category Fig 2: Topic coherence for ‘ food/Bev category

The data of both categories was used to train the different models. Post-training, the test data was used to evaluate the models and choose the best model for the particular category. As observed in the figures below, Linear Support Vector Machine performed best for Education Category and the GridSearchCV outperformed all the models in Food and Hospitality category.

Graphical user interface, application, table

Description automatically generatedGraphical user interface, application

Description automatically generated

Fig 2: Finding the best fit model for ‘education’ category

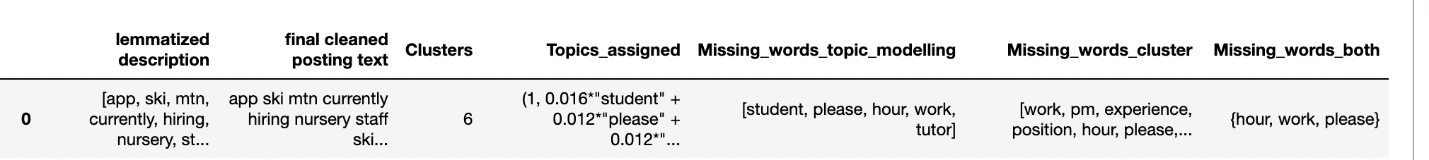
Fig 1: Finding the best fit model for ‘food and beverage’ category

BERT model gave an accuracy of ~59% for the education section and ~70% for the food/beverage and hospitality section.

**RECOMMENDATIONS TO RECRUITERS**

Apart from classifying the job description into various topics, its contents can be cross referenced against the top keywords occurring in both the cluster to which it belongs to, as well as the topic it got assigned to in the top modelling exercise. If the keywords do not appear anywhere in the job description, the job poster can be nudged to include the same so that it improves the success of the job posting.

This simple technique can be powerful as the key words are based on the historically successful job postings. The key words which are common in both the topic modelling output as well as the cluster analysis output but not in the job description can further help the job poster be specific and on point with regards to what is expected out of the candidate and the role.



In the above case for the above topic the words [hour, work, please] are common in both the cluster as well as the topic model, and could be utilized as a recommendation to make the job description better.

**CONCLUSION**

We believe that implementing this model will be helpful to Craigslist, impacting both employers as well as jobseekers. Since jobseekers will have an easier time navigating the platform and extracting relevant information without even clicking on a link, they will be able to browse through a larger number of jobs. If the job matches their criteria, they can then seek further details by clicking into the full post. In the current scenario, they need to click on all the links to find details and then decide, leading to many missed opportunities. Increased interest from jobseekers will encourage more employers to post jobs, thus perpetuating a cycle of growth.

Currently, recruiters who post available jobs on Craigslist are not sure how to structure their job descriptions, nor are they clear on how to word it to catch a potential employee’s attention. Keywork recommendations will help companies optimise their job postings by improving each posting’s efficiency in falling into a certain professional subcategory while also aiding jobseekers in finding that job.

**FUTURE SCOPE**

The model can be improved even further if we are able to provide tags based upon the generated labels. We can also add tags for **Employment Type, Salary etc**. With more granulated data, the need to navigate through multiple pages will reduce to just glancing at the search results and filtering out the ones that don’t fit the needs of the user. The advanced model can use our current model for its underlying logic. These steps will help Craigslist be a competitive and modern job board able to compete with more formal sites like Indeed and LinkedIn.

**EXHIBITS**

Exhibit 1: K-Means clustering results for Education Category

Text, letter

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

Exhibit 2: K-Means clustering results for Food/Bev/Hospitality Category

Text, letter

Description automatically generated

Exhibit 3: Topic Themes of LDA for Education Category

Graphical user interface, text, application

Description automatically generated

Exhibit 4: Topic Themes of LDA for Food and Hospitality Category

*Topic 0: Part-time for Specific Days​*

*Topic 1: Chef/Kitchen Staff​*

*Topic 2: Server​*

*Topic 3: Manager(event/restaurant)​*

*Topic 4: Outside city/Highway restaurant job​*

*Topic 5: Food delivery​*

*Topic 6: Cook/bartender​*

*Topic 7: Part-time server​*

*Topic 8: Mexican restaurant​*

*Topic 9: Employee*