MGMT 590

FINAL PROJECT REPORT- STRATIFY

Chayadeepsai Cherukupalli, Hsiao-Chien Wei, Mu-Hua Hsu,
San Martin Galindo, Jose, Srivatava, Swati, Xue Han

Background

Founded by Craig Newmark in 1995, Craigslist was originally a San Francisco community electronic newsletter. Today, Craigslist is one of the biggest free classified advertisements websites in the world. Craigslist focuses on community classifieds. Its sites cover more than 570 cities in 70 countries, and it is available in 7 languages. The contents include housing, jobs, services, for sale, item wanted, services, and general discussion forums.

Craigslist serves over 20 billion page views per month in 2017, while the company just has around 50 employees (2017). Craigslist's large amounts of information, page views and its extremely crude website design bring about high profit margin, as well as difficulties in better management of posts and improving customers' experience using the website.

The feedback forum of Craigslist is a "suggestion box" which allows users to discuss their experiences of using the website, reporting problems, or giving suggestions. The users can compose a new thread or reply to, rate or flag existing ones. Effectively utilizing those feedbacks could help our client detect imperfections, operate the forums more efficiently, and be enlightened with new business improvements. Whereas, due to the large size of the unorganized feedback comments/posts, it is so challenging for Craigslist to efficiently identify and classify major user problems.

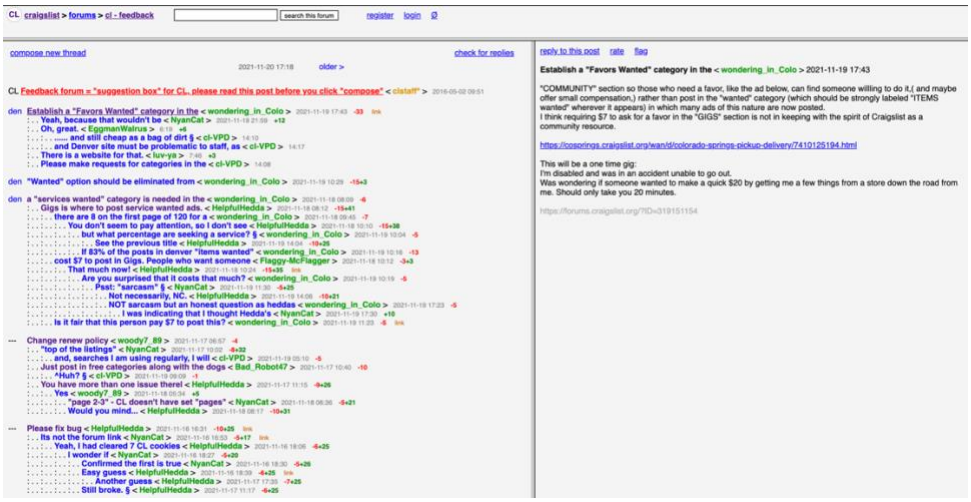


Figure 1. the feedback forum

Business Analysis

However, the unstructured data on the website brings the high cost for platform managers to maintain and organize the website. They even need to manually check misclassified advertisements and the properness of the content.

Therefore, *Stratify* decided to start from feedback. Feedback, whether it is positive and negative, works as a piece of valuable information that will be used to make important decisions. Companies should consistently search for ways to make progress. It's a true focus based on feedback from across the entire organization – internal business units as well as external stakeholders.

Currently, there are 16,000+ posts in our client's feedback forum. However, the issue of unstructured data also exists in the feedback forum. The overall user interface is designed difficult to read.

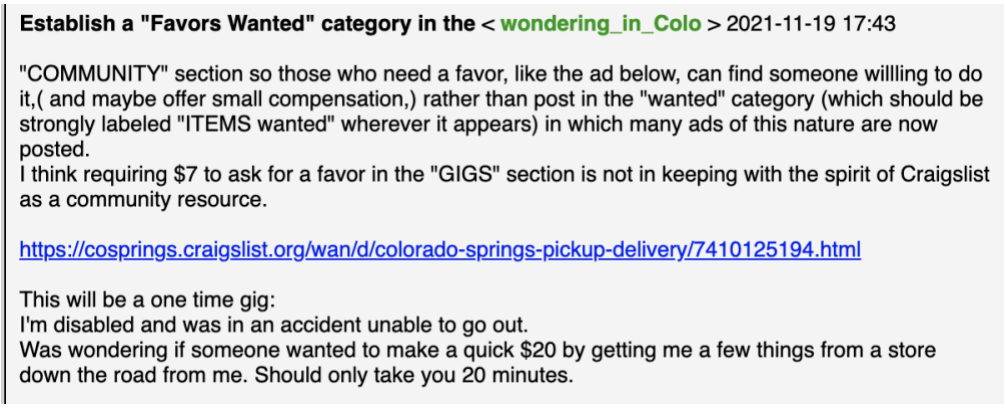
As Craigslist's consulting partner, *Stratify* proposes to extract and structure the data from feedback forum. After scraping data, we utilize the existing categories of the forum to create labels for supervised learning of the model. Eventually, we will design and train a classification model to

predict the classes of new feedback posts on the forum. We are trying to achieve the following objectives:

From the company's perspective

1. **Prioritize by identifying specific areas of improvement**
   Product development and customer service teams can utilize the labeled posts to prioritize the tasks they need to deal, and further take the number of thumbs up as well as thumbs down with the weighted value.

2. **Explore new business ideas based on user suggestions**
   Business Development teams can find potential business ideas from the feedback forum. For example, they could add a new city, section or feature on the website.


Establish a "Favors Wanted" category in the < wondering_in_Colo > 2021-11-19 17:43

"COMMUNITY" section so those who need a favor, like the ad below, can find someone willling to do it,(and maybe offer small compensation,) rather than post in the "wanted" category (which should be strongly labeled "ITEMS wanted" wherever it appears) in which many ads of this nature are now posted.
I think requiring $7 to ask for a favor in the "GIGS" section is not in keeping with the spirit of Craigslist as a community resource.

https://cosprings.craigslist.org/wan/d/colorado-springs-pickup-delivery/7410125194.html

This will be a one time gig:
I'm disabled and was in an accident unable to go out.
Was wondering if someone wanted to make a quick $20 by getting me a few things from a store down the road from me. Should only take you 20 minutes.

Figure 2. Users feedback for potential business ideas

From the users' perspective

1. The users wish to extract relevant information from the website with ease and less frequently turn to the feedback forum for help.

2. Our solutions will help enhance user experience. Once the users' feedback is taken by the company, it will then increase stickiness and web traffic to the website.

Data Analysis

In our project, we extracted data from feedback forum. Then, we utilized the existing categories of the forum to create labels for supervised learning of the model. After creating those labels, we started to design and train the classification model to predict the class of new feedback posts on the forum.

Data collection and preprocessing

We collected the data from the feedback forum, performing web scraping with Python to get a sample size as representative as possible. For this project, we got a total collection of 2355 records.

With the scrapped data from the feedback forum, we chose the top 4 classes as training datasets based on the suggestion box on the forum. These labels include suggestions of: 1. Adding a new category; 2. Adding a new city; 3. Problems or bugs reporting; 4. Flagging problems.

To avoid lacking information in the training data, we checked the class distribution of our data. As the figure below shows, there are at least 400 feedbacks for each class, and the count differences among the classes aren't large. It is a well-balanced multi-classification problem. Therefore, we don't need to do any upsampling or downsampling.

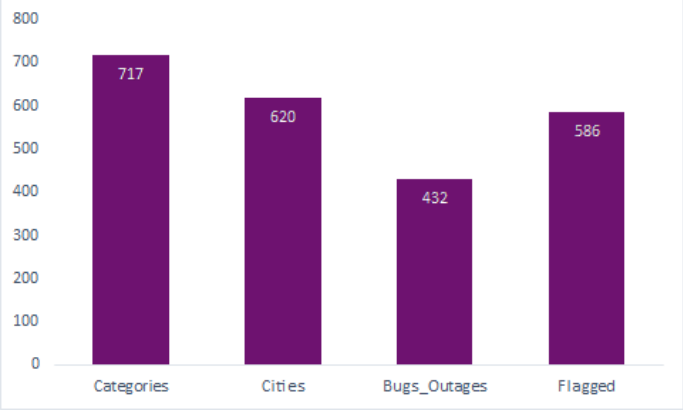


Figure 3. topic distribution histogram

We then began the basic steps of text representation, inluding tokenization, lemmatization, stopwords removal, and vectorization. We also removed the punctuations as well as the emojis. For the vectorization, we used TF-IDF vectorization, and set the minimum document frequency equals to 2, so that we could remove most of the potential typos. Besides, we also included 2-gram

to get more order information. After the preprocessing, the dimension of the document is 2355 * 17722.

Next, we used the stratify method to split our data, 80% for training and 20% for testing. By doing so, we can make sure that the proportion of each label is the same in both parts.

Model Training and Validation

From perspectives of easier implementation, better application, and higher performance, we firstly tried to run seven models individually with cross validation. The results are as follows:

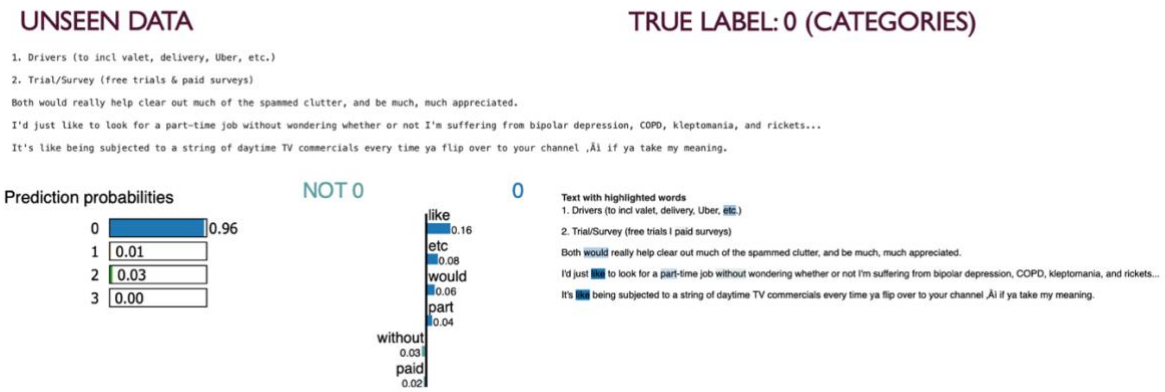
Model	Mean Accuracy	Standard Deviation
Linear SVC	**88.24%**	**0.0282**
Logistic	**88.15%**	**0.0325**
SGD	**87.56%**	**0.0266**
Complement NB	**87.47%**	**0.0176**
MLP	**86.88%**	**0.0241**
Light GBM	**85.18%**	**0.0324**
Decision Tree	**76.77%**	**0.0392**

Since Decision Tree performed far behind others, in the second step, we applied the ensemble method on the other six models. Our final ensemble model includes **Stochastic Gradient Descent**, **Linear Support Vector Classifier**, and **Light Gradient Boosting Machine**, applied with hard voting method to reach our final outputs. The mean accuracy of this combined method reaches 90.23%, and the model performs at satisfying levels and well-balanced in all the four classes.

	precision	recall	f1-score	support
class 0	0.84	0.93	0.88	135
class 1	0.94	0.90	0.92	125
class 2	0.84	0.81	0.83	91
class 3	0.99	0.93	0.96	120
accuracy			0.90	471
macro avg	0.90	0.90	0.90	471
weighted avg	0.91	0.90	0.90	471

Interpreting Predictions using LIME

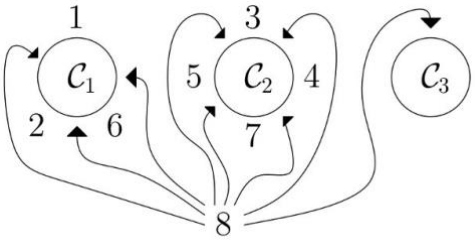
We used LIME library to interpret the results predicted from our selected model (ensemble). As demonstrated below, we predicted the class label on an unseen data. LIME explainer class allows us to understand the words that in the document that affected the prediction



Our ensemble model correctly predicted the class (categories) and the highlighted words played an important role in classification. This helps us to understand the working of our multilayered model and provide better insights into the classification.

Topic Modeling

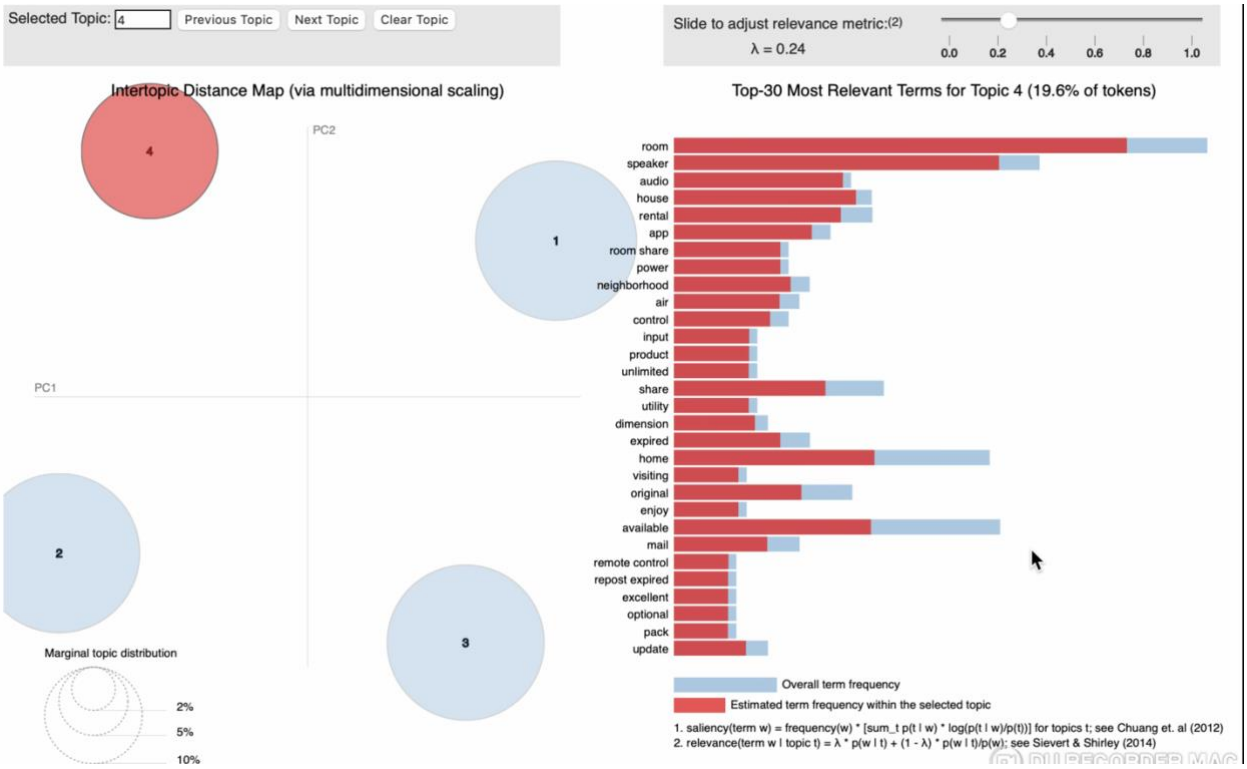
Further our team wanted to explore the theme of topics under each class. We selected *'bugs and Outages'* category to explore the broad topics associated within the corpus. We used LDA to for topic modeling. However, there were a lot of overlaps when we set the number of topics to 6. Thus, we used HDP to identify the optimal number of topics in the corpus.



HDP (Hierarchical Dirichlet Process) is a non-parametric Bayesian model that allows us to learn the topics from the data. The main advantage of HDPs is that they allow distinct corpora (groups) to share statistical strength when modelling — in this case share a common set of potentially infinite topics. So it is an extension of Dirichlet Process mixtures. In our case the optimal number of topics is 4.

Visualizing LDA Topic models

To elaborate the effectiveness of using LDA, we used *pyLDAvis* to create the visualization. This format is intuitive to understand the topic separation and helpful to identify the most frequent words within each topic for understanding the major theme of each cluster.



This visual interpretation shows how unique/differentiated each topic is (the chart on the left), and you can hover over each circle/cluster representing each topic to see which terms/ngrams were most frequent in that topic on the right. From topic 4 we can see that the major theme is feedback related to real-estate (house, room, rental) – top words. Further applying domain knowledge can supplement to take strategic initiatives in the problem areas.

Value addition to the company

1. **Cost saving - unclassified feedbacks**

As evident from its accuracy and performance, our model will efficiently classify the existing unclassified or new incoming feedbacks. This will save a lot of time and resources in terms of costs which it currently takes to classify those reviews.

2. **Elimination of manual efforts - misclassified feedbacks**

While performing data analysis, we observed that approximately 35% of the feedbacks were misclassified or belonged to wrong categories. The company has to manually check the wrongly classified feedbacks and the properness of the content. Our model will eliminate those manual efforts for Craigslist.

3. **Incremental revenue stream - better insights from feedbacks**

With improved feedback classification quality, Craigslist can rely on the feedback to generate valuable insights, develop new features in the long term and initiatives that would generate additional revenue for our client. Currently this can only be done in a limited magnitude as feedback classification is not proper.

Future Analysis

1. **Sentiment analysis**

Craigslist can stay on the top of whether people like or dislike their products and address their concerns offering a feeling that their voice is heard and valued. It can also be used to monitor the evolution of their brand reputation and how the brand image is perceived by the customers. Sentiment analysis could also be used to understand the shift in the customers' opinions and mood of as well as the transition happening around their opinions over time.

2. **Net ratings**

Net ratings could be utilized to judge the popularity of feedbacks. For example, if one feedback has received 2 thumbs up but 20 thumbs down, then the feedback is not popular among the users. So, we can calculate the net ratings to understand the popularity.

3. More classification labels

At present craigslist classifies the feedbacks into – website-related problems, flagged ads, site bugs/outages related problems and suggesting a new category or city related feedback. But as we went through user feedbacks, we realized there could be more classification labels. There were a lot of feedbacks for grievances related to spams or kudos for the good work. This will help to separate the general sentiments from particulars in the overall feedback.