Crime Prediction Using Hotel Customer Reviews

Somkiadcharoen Robroo
Faculty of Information Technology and Electrical Engineering
University of Oulu
Oulu, Finland
robroo.pc@gmail.com

Bofan Lin
Faculty of Information Technology and Electrical Engineering
University of Oulu
Oulu, Finland
bofan.lin@student.oulu.fi

***Abstract*—Can hotel customer reviews be used as a proxy for predicting crime hotspots? It becomes a hot issue recently. Tourists are prime targets for criminals. Therefore, tourists and visitors should remain alert, attentive, and vigilant to suspicious activities and can be used as reliable "human crime sensors". We proposed a novelty analysis method to enhance the features from hotel reviews dataset. And compared the London crime heat map with the hotel data map with spatial clustering and sentiment feedback. The result is likely that the features that we get from both datasets are from a different perspective and might not applicable to translate the knowledge into this domain.**

***Index Terms*—Crime Prediction, Spatial Clustering, Sentiment Analysis, Heat Map, Hotel, Review**

I. Introduction

Crimes are a potential problem in the world. The wellbeing and the trust of the society are affected by crimes, and there is no significant drop in the crime rate until now [1]. One way to reduce the crime rate can be done by putting more effort into creating more advanced analytics platform to sense and predict the crime. The problem is challenging since there is no absolute way to infer it directly, even though there are many effort spending on existing dataset such as Twitter Posts [2], Demographics, and Mobile Data [3].

In this paper, we proposed to use an NLP toolkit from Stanford to extract features and visualize the hotels on the map compared with the crime hot map. So, we can find the relationship between hotel customer reviews and crimes.

The main contribution of this work is that we give the result of "Can hotel customer reviews be used as a proxy for predicting crime hotspots?"

II. Background

A. Problems

Crimes are usually recorded according to a specific group of location, which can determine a specific group of location as a whole whether the crime has occurred or not. With the help of the hotel reviews data, heuristically, there might be some correlation with how the users feel with the crimes that are happened in the same area since the users might feel uncomfortable living around the area that is unsafe in a way.

The purpose of this work is to try to synthesize two different sets of data which are hotel review data and crime record data. This study will help us get more understanding about the crime rate and how the customers perceive hotels. Because hotel reviews are known to be a perception of how users feel about the hotel, where the feel is purely subjective and might correlate with an insecure feeling they have during the stay. The combination of these 2 data would give us more information towards predicting crimes.

B. Datasets

Here we employed the use of 2 different datasets for different purposes. The first one is a hotel review dataset which is likely to infer to the quality of the hotel service. Another one is a crime dataset from Police.

1) TripAdvisor hotel reviews dataset: We used the customer reviews dataset which contains about 140,000 customer reviews from hotels in London, and includes 15 features as shown in Table I. In this datasets, there are many missing features. For example, hotel address, we fill missing values by using the hotel name and the hotel zipcode.

TABLE I: The attributes of two datasets.

Column	Hotel Review Dataset	Metropolitan Crime Dataset
1	Hotel Name	Crime ID
2	Hotel Review Stars	Month
3	Hotel Address	Reported by
4	City	Falls within
5	ZIP	Longitude
6	Review Title	Latitude
7	Review Date	Location
8	Review Content	LSOA code
9	Review Stars	LSOA name
10	Reviewer Name	Crime type
11	Reviewer Location	Last outcome category
12	Reviewer Profile	Context
13	Reviewer Total Reviews	
14	Reviewer Hotel Reviews	
15	Helpful Votes	

2) Crime Heatmaps Visualization: Visualize crimes happened in London on a heat map.

3) England Police Metropolitan Crime Dataset: This dataset contains the England Crimes information happened in Metropolitan and downloaded from https://data.police.uk/. The data is selected from 3/2017 - 8/2017. The metropolitan-street.csv dataset includes 12 features as shown in Table I.

III. Methods

We hypothesized that hotels with more negative ratings would be the ones that are in high crime rate, and the

positive rating hotels would yield opposite result, as compared with each group. To test this hypothesis, we extracted a textual feature, use them to enhance the existing features, and combined those 2 data and spatially analysis the correlation between the data.

Geo-coding Conversion: the hotel reviews datasets only contains the hotel address, in order the harmonize those two datasets, we have to convert the address to latitude and longitude. It is also helpful for the further analysis.

Features Extraction: in order to analysis the hidden information of the hotel reviews, we have to use sentiment analysis method to extract at least three new features. For example, how was the service the hotel and how was the environment around the hotel.

Hotel Features Visualization: in order to explore the spatial dimensions of the data, we can visualize the structured data on the map. With the help of some machine learning methods, to define our observations are independent or identically distributed.

A. Geo-coding Conversion

In order to harmonize the existing 2 datasets, the preprocessing on the address of these must be done. The hotel reviews data only have the address as a text, while the crime data provide both text and spatial location. It is done by using the Google's Geo-coding API https://developers.google.com/maps/documentation/geocoding to decode the hotel address to the Latitude and Longitude. In this step, we registered to the Google Maps Platform and using the API key to request the conversion. In the original data, there are some missing values of Hotel Address. For those hotels, we converted the Hotel Name and City to Latitude and Longitude.

B. Features Extraction

The deep learning based sentiment analysis in this project is based on hotel review titles. The library in use is the StanfordCoreNLP [4]. It is an Natural Language Processing toolkit from Stanford which provides many aspects of the texts to be used. However, we are only interested in the sentimental analysis of the reviewer. The sentiment is done only to the review titles since we tried them on both review titles and review texts, but the results are not significantly better and took a longer time to inference. Also, heuristically we can infer the sentiment of the review by just reading the heading of the review. The pre-processing task is done by using a Python wrapper of StanfordCoreNLP https://github.com/Lynten/stanford-corenlp to detect the level of sentiment where the possibilities are Verynegative, Negative, Neutral, Postive, Verypositive. With the help of luck, these results can be directly mapped to the review star rating column with the value of 1-5 where the sentiment will be used to find the Effective Rating Score and perform a Fraud detection. There are 3 features that are extracted from the hotel reviews dataset as the followings

1) EFF (Effective Rating Score): The intuition behind the EFF based on the biases of the reviewers. This can be illustrated in the following case. Given that there are two reviewers gave the same review score of 3, but one of them said "Quite A Bad Hotel" and another one said "OK". We can see in this example that the meaning of 3 varies on perspectives of the people. Therefore, we would like to use the sentiment and the review ratings to compute whether they go together in the same way. The processing in this step can be seen in 2 cases. The first case is in the same direction. It means that if the review score is good (4-5) and the sentiment is also good (Positive, Verypositive), the rating score is valid. The same goes for the negative direction. If the score comes out quite bad (1-2) and the sentiment goes to the same direction of (Negative, Verynegative), the rating score is valid and should be kept, also done the same with Neutral rating and score of 3. The second one is the contradict. If the rating score is good, but the sentiment goes bad, the rating of this one is not invalid, and it won't be used to calculate anything further than this. The same goes with the rating score of 3 and sentiment other than Neutral. The intuition of this is summarized in the Table II.

TABLE II: Effective Rating Score Calculation

Review Score	Predicted Sentiment	Enhanced Result
4-5	Positive, Verypositive	same score
1-2	Verynegative, Negative	same score
3	Neutral	same score
3	Not Neutral	neglected
1-2	Positive, Verypositive	neglected
4-5	Verynegative, Negative	neglected

2) DIST (Distance from a crime spot): The intuition behind this is that the hotel that are close to crime spots are supposed to have a worse review. In the column of in the hotel data and column in the crime data, we foresee that there is a column which has a street name which can be used to use the direct text map after tokenisation, where the result is that more than 90% of them can be mapped directly. After the mapping, we use a Python package called GeoPy a to compute the geodesic distance between two spatial location [5].

3) AVG_CRIME_N (Average occurrence of crime per month): The intuition of this is just to inference another feature on the average crime that is occurred between 6 months timeframe. It can be easily seen as $\#crime/\#month$, a number of crime over a number of month. This feature will be used as a dependent value.

4) RATIO_NEGREVIEWS: (Ratio of negative reviews) The intuition of this is that to extract negative reviews features from the data. We would like to normalize the number of negative reviews, so the total number of negative reviews are used as $\#negative\_review/\#total\_review$.

C. Visualization

1) Hotel Features Visualization: We used the Statistical GIS Boundary Files for London from https://data.london.gov.uk/dataset/statistical-gis-boundary-files-london as a shape file to plot the spatial data into each city in London. The clustering

features (effective rating score, distance from a crime spot, ratio of negative reviews and average occurrence of crime per month) are grouped by each region and average all the features over the region only. Then, we cluster the hotels into four classes and plot into the map as shown in Figure 1.

We then analyze each cluster using bar charts on each features in Figure 2. The 4th cluster(labeled as 3) is the one with highest crime rate, even with the nearest distance from crime scene. The hotel ratings that they have are 66%, which is quite high while the ratio of negative reviews is quite low. The 3rd cluster which is labeled as 2 has the highest hotel star ratings compared to other clusters. It is the farest from crime scene among these clusters, resulting in lowest crime rate. The 1st and 2nd cluster (labeled as 0 and 1) are quite similar. They are quite close to the crime scene, but not significantly high crime rate. Moreover, the 1st cluster has higher hotel rating scores and lower negative reviews.

2) Crime Heatmaps Visualization: We used the England Police Metropolitan Crime Dataset which contains the data from 3/2017 to 8/2017. The result shows that the crime is almost everywhere in London. As shows in the Figure 3. However, the small holes in the heatmap are mostly parks or airports. In other words, it is safer to stay in those area in London.

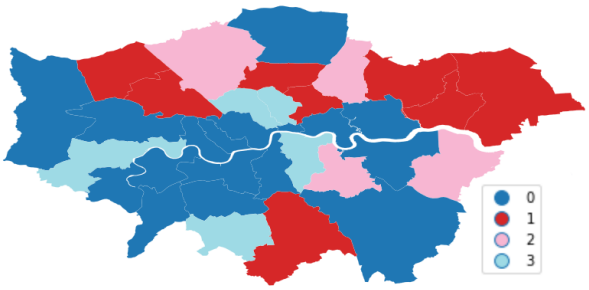

Fig. 1: The Cluster Map of Features

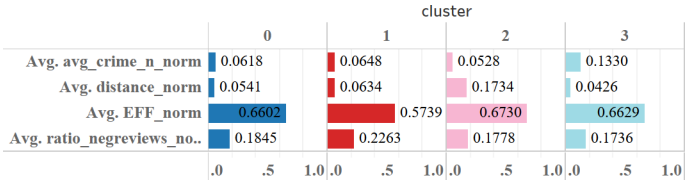

Fig. 2: The Summary of Cluster Characteristics

IV. Data Analysis

In this part, we tried to exclude the fraud users out from the platform then, we would like to measure the correlation and causality of 3 features whether they are possible to infer the crime or not.

A. Fraud Detection

We hypothesized fraud users as those who gave only negative reviews, which we tried to exclude them before the analysis. The intuition that we have is that users who only gave negative reviews on all reviews should be the fraud ones. From the analysis, we only see that the users who gave negative reviews for all their reviews are quite low in the number of reviews. Reportedly, only 2,10 users gave only 4,3 negative reviews respectively are not high enough to exclude them from the analysis.

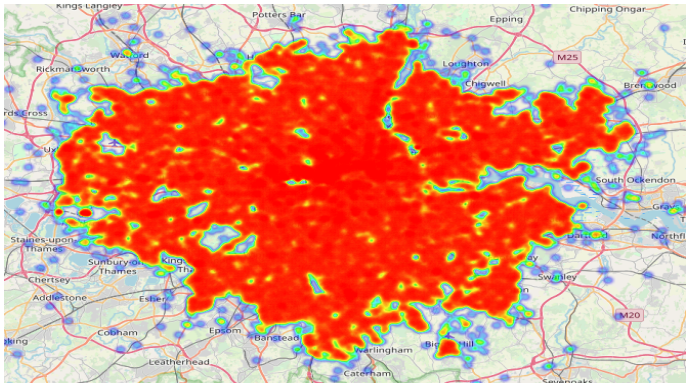

Fig. 3: Crime Heatmap of London in 6 months timeframe

B. Correlation Analysis

Using correlation measurement as defined as the following equation:

$$\rho_{xy} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 \sum_{i=1}^{n}(y_i - \overline{y})^2}} \quad (1)$$

TABLE III: Features Correlation on Crime Rate

Variable	$\hat{\rho}_{xy}$
EFF	0.01633
DIST	-0.0433
RATIO_NEGREVIEWS	-0.0089

This table shows a measure of correlation between explanable variables and dependent variables.

The estimated correlation then can be found in the Table III, and the results of correlation calculated from the Equation (1) are then interpreted that effective Star rating is positively correlated with average number of crime at 0.01633, which is contradict with our assumption that hotels with higher effective star Rating should have lower crime rate. In addition, distance and ratio of negative reviews are negatively correlated with average number of crime at -0.0433 and -0.0089 respectively. Distance is correlated with our assumption that hotels which are close to crime scene distance should have more crimes. Moreover, the ratio of negative reviews is against our assumption which we assumed that hotels with poor reviews should have more crimes.

However, the correlations are very weak, so we need to continue searching for unobservable variables that might strongly affect crime rates.

C. Multiple Regression Analysis

We ran OLS Regression on the features and dependent variables to understand more about how each feature impact crime rate. The preliminary results can be found in the Table IV

TABLE IV: Regression Analysis on Average Number of Crime

	AVG_CRIME_N
constant	2.6279*
	(0.001)
EFF	-0.0611
	(0.732)
DIST	-1.4480
	(0.151)
RATIO_NEGREVIEWS	-2.261e-05
	(0.745)
R^2	0.002

Notes: OLS minimizes squared error between predicted and actual values. The number of observations is 1196.
* Significant at the level of 0.05

As seen on Table IV, the coefficients of the effective star rating on average number of crime rate is -0.0611 which suggests that an increase in 1 unit of effective star rating would decrease the average crime rate by 0.0611. The distance showed that the average number of crime is strongly in contrast with distance, in a sense that it would decrease by 1.4480 every increasing in distance. It is the largest magnitude of all features. In addition, the ratio negative reviews would also decrease the average crime number only by a small amount of -2.261e-05. R^2 mean that all the features can explain variations of average crime only 0.2%.

However, there are no statistically significant between each features and crime rate at the level of 0.05 except for the constant. This could mean that there should be other features that can better describe the average number of crime.

V. Results & Discussion

We initially think that the hotel reviews should directly related to the fact that how the service of specific hotels are perceived by customers, and thus would not have any direct correlation or causalities on the crime dataset. After the ad-hoc analysis before the visualization phase, the data seems to us that it might not work well in this case, and we do not have enough backup plan to put any additional dataset into solving this problem. In addition, the experiences of us related to text mining in this case is quite basic, given that we only did the sentiment analysis on the hotel review and use those results to enhance the existing knowledges that we have. We feel like we could have done it better if we have more experience in this.

VI. Conclusion

We presented the methodology to detect crime using hotel reviews. The results are based on those 2 datasets only which are TripAdvisor hotel review dataset and England's Metropolitan police crime dataset. The preliminary results that we have is somehow could not select good features for the task that can infer the dependent variable easily or the OLS regression is too basic to work on this task.

VII. Future Work

We learned that this project is a pure combination between art and science. As both of us do not have prior knowledge in dealing with text data and spatial data before, we were a bit lost in the first place. However, with the rise of popularity in the NLP field, it helps us to be able to find the references to guide us on what we should work on, and that was only the guideline. This kind of work is likely to make use of both creativity and technical skills to solve.

We would love to know more on how to heuristically solve this project. Moreover, it would be great to ask for opinions from people in the industry who are working on this to roughly know how they tackle this kind of project. We suggested to develop this work further by combining more modalities of the data in addition to dig deeper on textual features which might not help getting a significantly better result due to the nature of the dataset.

Appendix
Source Code & Datasets

The Python source codes are publicly available on the GitHub link https://github.com/robroooh/txt-mng. The TripAdvisor hotel reviews dataset is available at https://drive.google.com/file/d/1T-6tAgL1_M4pRatHTGMhIwX732XpvWGq with University of Oulu's G Suite student account. Also, the crime data can be downloaded on data.police.uk at https://data.police.uk/data, where the user can choose only Metropolitan Police Service.

References

[1] Liang Ge, Junling Liu, Aoli Zhou, Hang Li, Crime Rate Inference Using Tensor Decomposition, 2018 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation, Guangzhou, China.

[2] Wang X., Gerber M.S., Brown D.E. (2012) Automatic Crime Prediction Using Events Extracted from Twitter Posts. In: Yang S.J., Greenberg A.M., Endsley M. (eds) Social Computing, Behavioral - Cultural Modeling and Prediction. SBP 2012. Lecture Notes in Computer Science, vol 7227. Springer, Berlin, Heidelberg.

[3] Andrey Bogomolov, Bruno Lepri, Jacopo Staiano, Nuria Oliver, Fabio Pianesi, Alex Pentland, Once Upon a Crime: Towards Crime Prediction from Demographics and Mobile Data, Proceedings of the 16th International Conference on Multimodal Interaction, November 12-16, 2014, Istanbul, Turkey.

[4] Manning, Christopher D., Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP Natural Language Processing Toolkit In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pp. 55-60.

[5] https://geopy.readthedocs.io