2020 2nd International Conference on Economic Management and Model Engineering (ICEMME) | 978-1-7281-9144-7/20/$31.00 ©2020 IEEE | DOI: 10.1109/ICEMME51517.2020.00215

Leveraging Multi-Modality Data to Airbnb Price Prediction

Ningxin Peng
School of Software
South China University of Technology
Guangzhou, China

Yiyuan Qin
Duke Kunshan University
Suzhou, China

Kangcheng Li*
School of Computer and Information Technology
Beijing Jiaotong University
Beijing, China
* Corresponding author: likangcheng@bjtu.edu.cn

***Abstract*—Airbnb is a short-term housing rental platform, which is gaining tremendous popularity among tourists. A reasonable prediction of Airbnb rentals could help customers make the best choices and provide homeowners with an indicator of prices to reach maximum benefits. In this paper, a variety of data are combined as inputs into machine learning algorithms and natural language processing framework to construct a reliable price prediction model for Airbnb rentals. This research shows: 1) customer reviews, house features, and geographical data can be used as predictive factors for Airbnb rentals; 2) using multi-modality data in price forecast performs higher accuracy compared with single-type data.**

Keywords-Airbnb; NLP; machine learning; price prediction

I. INTRODUCTION

With the development of the internet and the improvement of wealth, more people emphasize accommodation conditions as factors when they are traveling. They not only require the house to be safe, clean and affordable, but also add some personalized requirements. At the same time, the emergence of homestays meets the demand of those travelers. As the leading platform of homestay, Airbnb provides diversified and personalized rooms for tourists with different needs. This research's primary objective is to predict the reasonable rental price of the homestay.

Previous studies have demonstrated the feasibility of predicting Airbnb prices through various factors. Wang and Nicolau[1] confirmed that factors related to sites, services as well as customer reviews could influence the prices of hotel rentals. They identified host attributes as important price determinants. Cai et al.[2] found that additional text inputs such as comments have contributed to the performance of Airbnb price prediction. The correlation of prices among neighboring houses was proved by Dubin[3]. Based on this discovery, Ma et al.[4] estimated the rental prices of shared warehouses using geographic data, such as distance from the city center and distance from the closest house.

Scholars have adopted a variety of methods on Airbnb price prediction. Gu et al.[5] presented a hybrid of genetic algorithm and support vector machines (G-SVM) in housing price forecasting. Li et al.[6] employed Linear Regression with Normal Noise (LRNN) to predict legitimate Airbnb prices. Zhang et al.[7] used Geographically Weighted Regression (GWR) and General Linear Model (GLM) model to recognize the key factors that determine the Airbnb housing prices. Additionally, Neural Networks (NN) have also been applied to house and Airbnb price prediction[8] [9]. Luo et al.[10] showed that a price prediction model that is trained on aggregated datasets from multiple cities outperforms models that use an individual dataset.

However, previous research built at most two factors for price prediction, while this work combines comments, house attributes and geographical position of houses to create a new model. By improving the robustness and the performance of the model, the multi-modality dataset will make results more accurate.

In this research, three types of data are incorporated, including numeric data, text data, and map data, into a framework for predictive analytics in house prices. This work leverages the method of natural language processing (NLP) to extract the textual comments given by previous customers. After comparing multiple popular machine learning methods to analyze different types of data, a best-performed algorithm with the expectant output is came out. The finding is that measures can be used to determine future trends of house prices with a degree of high accuracy. This findings provide a framework that combines hybrid data to implement future price prediction, which can be useful for both house owners and customers.

Therefore, the new method presented in the study is adapted to forecast house rentals. The cases in Airbnb can be applied to validate the rental forecasting ability of the multi-modality dataset method.

II. DATA

Approximately nine million data points from Inside Airbnb[11] for 10 different cities between October 2015 and December 2019 are utilized in this research. Inside Airbnb is an investigatory website without any business purpose, and it reports and visualizes several different kinds of data on the worldwide rental marketplaces. Inside Airbnb also provides open homestay information and guest's evaluation data, which gives visitors perspectives on housing conditions. The summary of the number of various types of data in the dataset is shown in Table 1.

978-1-7281-9144-7/20/$31.00 ©2020 IEEE
DOI 10.1109/ICEMME51517.2020.00215

TABLE I. SUMMARY OF DATA QUANTITY STATISTICS

	Listing	Review	Neighborhood	Total
Total	287579	9345265	287579	9920423

Three modalities of data are used in the research: listing, review, and neighborhood. Figure 1 shows the quantity of Listing and Neighborhood data and Figure 2 explains the data information of Review. All the features of the dataset used in this research are summarized in Table 2.

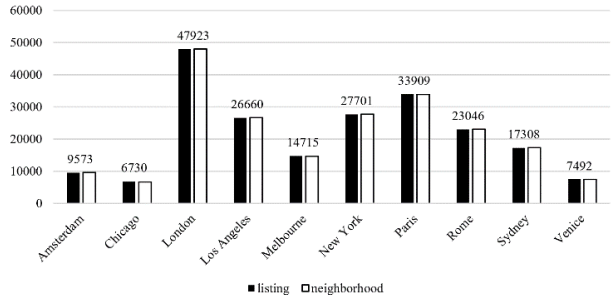

Figure 1. Quantity of Listing and Neighborhood

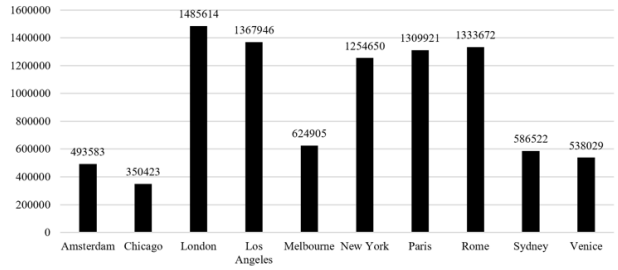

Figure 2. Quantity of Customer Reviews

TABLE II. SUMMARY OF DATA QUANTITY STATISTICS

Features	Description	Type
host_since	the date that the Airbnb house start to operate	date
host_has_profile_pic	whether the Airbnb host has profile pictures	boolean
host_listings_count	number of Airbnb houses the host operates	number
is_location_exact	whether the Airbnb house is location exactly	boolean
room_type	type of the Airbnb house	categories
accommodates	number of accommodates	number
bathrooms	number of bathrooms	number
bedrooms	number of bedrooms	number
gest_included	number of included guests	number
minimum_nights	number of minimum nights for renting	number
availability_365	number of days available for renting in a year	number
number_of_reviews	number of reviews	number
cancellation_policy	type of cancellation policy	categories
reviews_per_month	number of reviews per month	number
reviews	customer reviews text	text
neighborhood	geographical coordinates of Airbnb houses	geojson

III. METHODOLOGY

A. Data Pre-processing

Before model training, it is very necessary to ensure that the data is complete and valid. Raw data often contains some missing values, incomplete data, inconsistent data, noisy data, and outlier data, which is significant for data to be processed before being analyzed[12].

For data preprocessing, wrong data and impute missing values were deleted using the attribute mean. We eliminate data of houses with prices greater than $500 and less than or equal to $0, in order to screen out the outliers. Plus, we apply logarithmic transformation for label transformation of Airbnb prices. After preprocessing, the distribution of house prices is as Figure 3.

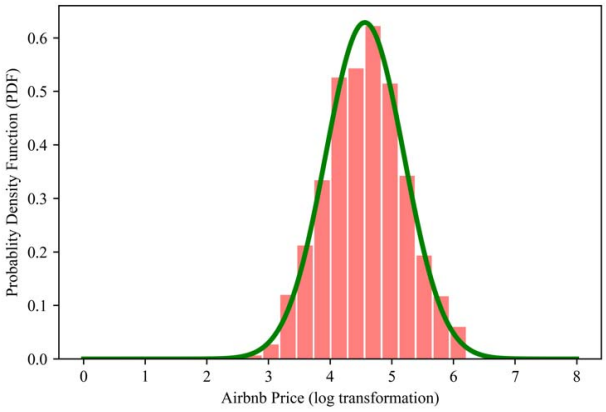

Figure 3. Distribution of Airbnb price data. After eliminating the outliers and label transformation, it is clearly to see that the data satisfies the normal distribution.

B. Text Data Processing

Some of the features that will be used in the model are textual data, so we use natural language processing (NLP) techniques to extract sentiment from user comments. Python provides a large standard library to implement data processing [13]. The analysis is based on the following Python libraries:

- Natural Language Toolkit (NLTK): an efficient pythonbuilt platform for processing human natural languages. It provides a series of simple-to-use tool interfaces that allow access to more than 50 corpora and lexical resources, as well as text classification, parsing, stem tagging, and semantic reasoning models.
- Sklearn: a data analysis package for Python. Sklearn implements a large number of built-in libraries and commonly used data models, providing us the capabilities and methods that allow us to process data quickly and easily.
- TextBlob: a Python library for textual data processing[14]. TextBlob implements numerous

frequently-used interfaces for natural language processing (NLP) tasks, including sentiment analysis, tagging and simple translation. Purposed by Loria (2018).

We firstly preprocess the dataset by deleting invalid comments (such as comments with garbled characters) and filtering some modal particles and stopwords. We analysis the comments and calculate the mean score of comments for each house. We then transform the textual data into numerical data so that it can be used as a variable into our model.

C. Geographical Data Processing

In the multi-modality data model, we also utilize the geographical data, which are the geographical coordinates of the houses. According to the research by Li et al.[6], the landmarks and houses have similar location distributions. They also found that the distance to landmarks takes a strong effect on house renting prices. Generally, the shorter the distance, the higher the price. To this extent, we add the distance to the nearest landmarks as one of the prediction factors.

Spectral Clustering (SC), which was popularized by Shi & Malik [15] and Ng et al.[16], is a novel and efficient clustering algorithm suitable for graph clustering. Compared to some rudimentary clustering algorithms such as k-means, SC algorithm has advantages of: 1) higher robustness and accuracy k; 2) effective for sparse data clustering because only the similarity matrix is needed; 3) can cluster on sample space with arbitrary shapes; 4) easy to converge.

Figure 4 roughly shows the effect of clustering on Airbnb houses in Amsterdam city. In order to make the graph look more concise, we randomly select 800 from 17229 points to plot the graph. After clustering, the exemplars could be regarded as landmarks, and we could use (1) to calculate the Euclidean Distance d from each point to the exemplars. Through normalization of (2), the normalized distances $\hat{d}$ act as part of the house features for predicting the house prices.

$$d(p, p_e) = \sqrt{(x_p - x_{p_e})^2 + (y_p - y_{p_e})^2} \tag{1}$$

where p is the house point and p e is the exemplar point;

$$\hat{d} = \frac{d - min\,(d)}{max(d) - min\,(d)} \tag{2}$$

where $min(d)$ and $max(d)$ are the minimum and maximum distances in a single cluster.

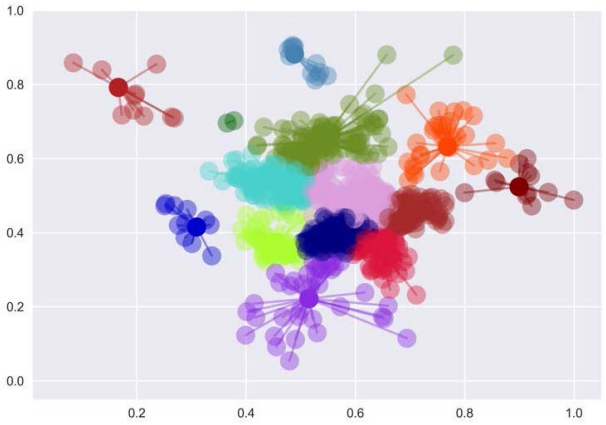

Figure 4. A schematic graph of the clustering results of the geographical coordinates of the Airbnb houses. Houses belonging to different clusters, which are marked with dots of different colors, are connected with the exemplars using straight lines.

D. Principle Components Analysis

Principle component analysis (PCA)[17] is an effective approach for Exploratory data analysis (EDA). PCA is often used to analyze the correlation between high-dimension data, and to reduce the dimension by extracting the dominant patterns of the data matrix, which are called principle components (PC). PCA can be interpreted as projecting the data into a different coordinate system, such that, on the first coordinate lies the projection of data with the greatest variance (called the first PC), and on the second coordinate lies the projection with the second greatest variance (second PC), and so on.[18].

In this research, PCA is utilized on the Airbnb listing dataset to reduce the dimension of the data, so as to remove unnecessary data redundancies and reduce computational complexity. The quality of each PC can be measured by its explained variance ratio (EVR). The EVR of a PC indicates the information that the component retains under its projection. The formula for EVR is as follow:

$$EVR(x'_m) = \frac{\|x'_m - x_m\|^2}{\sum_{i=1}^{p} \|x_i\|^2} \tag{3}$$

$$\text{s.t. } x'_m \in X', x_m \in X, 1 \le m \le k$$

The summary for EVR of each PC in the dataset is shown in Figure 5. we take the first six PCs, for their EVR sum exceeds 90% (94.12%), that is to say, these six PCs are enough to represent more than 90% of the information of the entire dataset. Therefore, we select the first 6 out of 16 principle components in the Airbnb dataset.

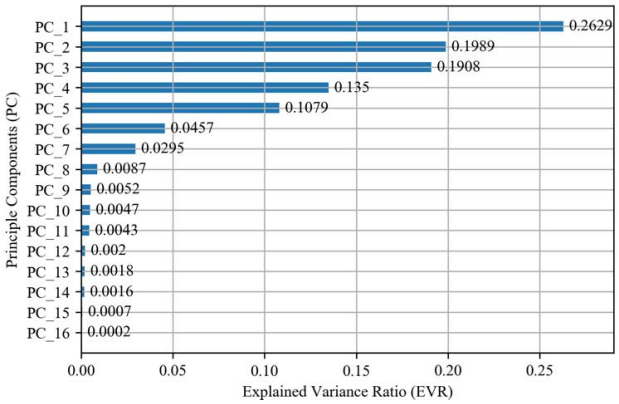

Figure 5. Summary of EVR. EVR reflects the ability that a PC could represent the entire dataset.

E. Forecasting Models

We use the simplified variables calculated by PCA as features of machine learning models trained to predict the Airbnb prices of each house. In this paper, we leverage 4 different types of machine learning models as discriminative models for Airbnb price prediction. Brief descriptions of the models are as below:

- Linear Regression: a statistical analysis approach to determine the interdependent quantitative relationship between multiple variables. We can find a best-fitting line using the method of least-squares so that the vertical distance of each data point to the straight line is minimized[19].
- Support Vector Regression (SVR): the application of support vectors in the field of functional regression. Lagrange multiplier and relaxation factor are used to perform regression analysis on data to minimize the total deviation of all sample points from the hyperplane, which performs excellent generalization ability.
- XGBoost: XGBoost[20], which is the abbreviation of eXtreme Gradient Boosting, is an improvement of GBDT (Gradient Boosting Decision Tree). XGBoost is one of the boosting algorithms, whose objective is to integrate many weak models to form a strong model.

  In this research, we use XGBRegressor to price prediction and apply GridSearchCV to adjust the parameters of the regressor. The key parameters are a number of estimators and the max depth of the regression tree.
- Deep Neural Network (DNN): purposed by Schmidhuber[21], is one of the most fundamental and widely-used frameworks of deep learning. Generally, DNN is an Artificial Neural Network with multiple hidden layers between input and output layers. Each layer contains multiple neuron nodes and every two nodes of adjacent layers are fully connected in a linear relationship:

$$z = \sum w_i x_i + b \tag{4}$$

  In addition, the neurons are applied to a non-linear activation function σ(z). Hence, DNN architectures can act on both linear and non-linear models.

  Specifically, the DNN framework accepts 6 features from Airbnb houses in the input layer and generates one output for Airbnb house price in the output layer (Fig. 6). The model possesses two hidden layers, with 9 and 10 neurons respectively. We select ReLU (Rectified Linear Unit) transformation as the activation function in the model.

Additionally, we employ 5-folds cross-validation on our data in order to avoid overfitting. The basic idea of k-folds cross-validation is to divide the original dataset into k batches. We take the first batch for the validation set and the rest for the training set. The model is trained using the training set and evaluated using the validation set. Subsequently, we select the second batch for validation and repeat the operation, and so on until all the k batches have been selected.

F. Output Evaluation

To evaluate the forecasting accuracy of the model, an output evaluation is performed. We use R^2 scores and Root Mean Square Error (RMSE) to calculate and compare the output results. A relatively credible model has a higher R^2 scores and lower RMSE values. Formulas are as below:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(P_i - \hat{P}_i)^2}{\sum_{i=1}^{n}(P_i - \bar{P}_i)^2} \tag{5}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(P_i - \hat{P}_i)^2} \tag{6}$$

where P_i is actual house price, $\hat{P}_i$ is estimated house price, $\bar{P}_i$ is the average house price and n is the number of observations.

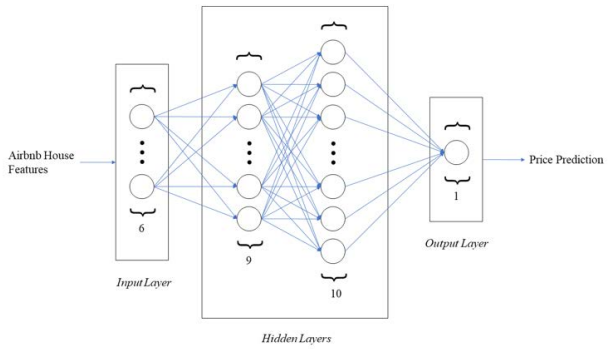

Figure 6. The DNN architecture

IV. Results

A. Comparison of Forecasting Models

We randomly selected 50 samples for each model and compared them with true values. The fit curves of True vs.

Predicted values are shown in Fig. 7. When the two curves overlap, the model is more accurate.

Table 3 shows the RMSE and R^2 scores of each fold of cross-validation as well as the average scores of all the models. More accurate results have lower RMSE and higher R^2 scores. Due to the linear models that Linear Regression and Support Vector Regression are based on, they perform poorly for relatively complex data relationships of Airbnb prices. In contrast, the Neural Network model and XGBoost model perform better on the dataset. Neural Network is a stable and plug-and-play machine learning model, and it performs better than of XGBoost (using default parameters without any adjustment). In this research, XGBoost, with its parameters adjusted by GridSearchCV, performs the best among the four models. The RMSE of XGBoost yielded the lowest and the highest R^2score.

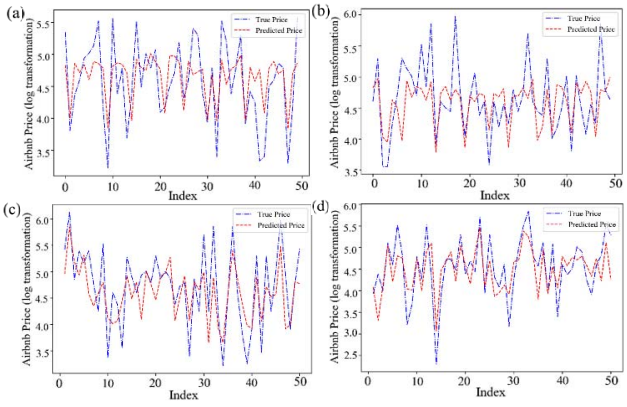

Figure 7. Fitting effects of four Multi-modality models. The figures below illustrate a direct expression of the error between the predicted value and true value. Indexes are fifty random points in the dataset. The subplots correspond to: (a) Linear Regression, (b) Support Vector Regression, (c) Deep Neural Network, (d) XGBoost Regressor.

B. Comparison of Different Modality of Data

Figure 8 illustrates how well varying forms of modality performed with respect to one another. All modalities of data were trained using XGBoost.

Compared with other modalities, multi-modality yielded the highest R^2 score and lowest RMSE, 0.4768 and 0.4647, respectively. When used in conjunction with one another, XGBoost and multi-modality yielded accurate predictions for house prices on Airbnb.

TABLE III. MODEL COMPARISON. THIS TABLE SHOWS A COMPARISON BETWEEN THE PREDICTED VALUE AND TRUE VALUE.

Models	**SVM**		**Linear Regression**		**XGBoost**		**DNN**	
Folds / Scores	**RMSE**	**R2**	**RMSE**	**R2**	**RMSE**	**R2**	**RMSE**	**R2**
1	0.517236	0.349339	0.518393	0.350079	0.464357	0.481330	0.499073	0.400879
2	0.520084	0.343782	0.516363	0.348781	0.463836	0.481881	0.486134	0.430868
3	0.518937	0.351808	0.520333	0.344740	0.464583	0.471651	0.486502	0.420617
4	0.521322	0.341390	0.518566	0.348502	0.465275	0.473915	0.486688	0.424378
5	0.518529	0.348084	0.520234	0.347939	0.465623	0.474997	0.486205	0.427558
Average	0.519222	0.346881	0.518778	0.348008	0.464735	0.476779	0.485784	0.428310

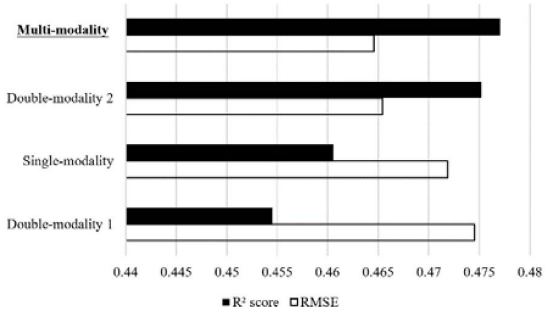

Figure 8. The comparison of various modalities of the dataset. Singlemodality stands for using only listing data for price prediction; Double-modality 1 and Double-modality 2 represent using listing data, with text data and geographical data respectively; Multimodality represents a mixture of the three types of data.

V. CONCLUSION

Our research developed a model that can predict house prices on Airbnb. These results suggest that homeowners can price homes more accurately and enable users to determine fair prices. Thus, creating a fair marketplace.

The more exotic algorithms, DNN and XGBoost, performed better than linear-based-models, Linear Regression and Support Vector Regression. Using multi-modality data enhanced the performance of the Airbnb price prediction model.

The limitation of this research is that we only use data from Inside Airbnb which restricts the market range. As a result, future research will be focused on expanding data scope so that

it takes convenience to more people, as well as the improvement of our model in the meantime.

REFERENCES

[1] D. Wang and J. L. Nicolau, "Price determinants of sharing economy based accommodation rental: A study of listings from 33 cities on airbnb. com," *International Journal of Hospitality Management*, vol. 62, pp. 120–131, 2017.

[2] T. Cai, K. Han, and H. Wu, "Melbourne airbnb price prediction," 2019.

[3] R. A. Dubin, "Predicting house prices using multiple listings data," *The Journal of Real Estate Finance and Economics*, vol. 17, no. 1, pp. 35–59, 1998.

[4] Y. Ma, Z. Zhang, A. Ihler, and B. Pan, "Estimating warehouse rental price using machine learning techniques." *International Journal of Computers, Communications & Control*, vol. 13, no. 2, 2018.

[5] J. Gu, M. Zhu, and L. Jiang, "Housing price forecasting based on genetic algorithm and support vector machine," *Expert Systems with Applications*, vol. 38, no. 4, pp. 3383–3386, 2011.

[6] Y. Li, Q. Pan, T. Yang, and L. Guo, "Reasonable price recommendation on airbnb using multi-scale clustering," in *2016 35th Chinese Control Conference (CCC)*. IEEE, 2016, pp. 7038–7041.

[7] Z. Zhang, R. J. Chen, L. D. Han, and L. Yang, "Key factors affecting the price of airbnb listings: A geographically weighted approach," *Sustainability*, vol. 9, no. 9, p. 1635, 2017.

[8] A. Varma, A. Sarma, S. Doshi, and R. Nair, "House price prediction using machine learning and neural networks," in *2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT)*. IEEE, 2018, pp. 1936–1939.

[9] P. R. Kalehbasti, L. Nikolenko, and H. Rezaei, "Airbnb price prediction using machine learning and sentiment analysis," *arXiv preprint arXiv:1907.12665*, 2019.

[10] Y. Luo, X. Zhou, and Y. Zhou, "Predicting airbnb listing price across different cities," 2019.

[11] I. Airbnb, "Get the Data - Inside Airbnb. Adding data to the debate." 2015, retrieved January 15, 2020, from http://insideairbnb.com/get-thedata.html.

[12] V. Limsombunchai, "House price prediction: hedonic price model vs.arti fi cial neural network," in *New Zealand agricultural and resource economics society conference*, 2004, pp. 25–26.

[13] B. Gupta, M. Negi, K. Vishwakarma, G. Rawat, P. Badhani, and B. Tech, "Study of twitter sentiment analysis using machine learning algorithms on python," *International Journal of Computer Applications*, vol. 165, no. 9, pp. 29–34, 2017.

[14] S. Loria, "textblob documentation," Release 0.15, vol. 2, 2018.

[15] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 22, no. 8, pp. 888–905, 2000.

[16] A. Y. Ng, M. I. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," in *Advances in neural information processing systems*, 2002, pp. 849–856.

[17] K. Pearson, "Liii. on lines and planes of closest fit to systems of points in space," *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, vol. 2, no. 11, pp. 559–572, 1901.

[18] I. T. Jolliffe, "Principal components in regression analysis," in *Principal component analysis*. Springer, 1986, pp. 129–155.

[19] S. H. Brown, "Multiple linear regression analysis: a matrix approach with matlab," *Alabama Journal of Mathematics*, vol. 34, pp. 1–3, 2009.

[20] T. Chen, T. He, M. Benesty, V. Khotilovich, and Y. Tang, "Xgboost:extreme gradient boosting," *R package version 0.4-2*, pp. 1–4, 2015.

[21] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural networks*, vol. 61, pp. 85–117, 2015.