Comparison of Machine Learning Models for Sentimental Analysis of Hotel Reviews

Xiaoyu Li
Department of Automotive Engineering
Tongji University
Shanghai, 200092, China
1352116@tongji.edu.cn

Cai Liu
Lvmama Co., Ltd
Shanghai, 200033, China
liucai@lvmama.com

***Abstract* — It is widely acknowledged that the enormous boost of Internet technology has greatly changed tourism industry, especially for hotel booking. A plenty of research have addressed that most tourists do decision-making based on online hotel reviews. With the goal of predicting satisfaction of hotel customers precisely and efficiently, in this study ten thousand hotel reviews across US with one to five stars are collected from TripAdvisor, and the performance of a variety of machine learning models, namely Logistic Regression, Naive Bayesian, Decision Tree, Random Forest, Support Vector Machine (SVM) and Neural Network, are compared through two 10-folder-cross-validation experiments. The result indicates that models directly trained on review titles which have only 25.7 byte lengths on average could achieve promising classification accuracies ranging from 84% to 87%, and with review contents and irrelevant word filtering, the SVM model is able to reach maximal predictive accuracy nearly 92%. Furthermore, this paper demonstrates that classical machine learning models, such as SVM and Naive Bayesian, are computationally efficient and perform well in terms of accuracy, recall and precision, while neural network requires careful structure design as well as parameter tuning.**

Keywords-Machine Learning; SVM; Neural Network; Sentimental Analysis; Hotel;

I. INTRODUCTION

With rapid development of Internet technology, tourism industry, specifically hotel booking, has been transformed greatly [1]. According to statistics, approximately 150 million tourist bookings have been made online at year 2016 [2], and this tendency has grown more and more heated so far. Moreover, researchers have found out that review rating from social media can explain hotel performance metric better than traditional customer satisfaction evaluations [3]. Furthermore, through survey method, a study [4] has revealed that 70 percent of potential customers worldwide trust hotel reviews from website for decision-making, especially for the cases where tourists have never been to the travelling spots [5]. Therefore, extracting textual information and analysing sentiments from hotel reviews is beneficial for the whole tourism industry chain.

There have been excellent research works on textual analytics of hotel guest experience according to online reviews [6][7]. However, the existing tourist experience evaluation system have the following problems: (1) customer reviews can be informal and fuzzy, and traditional factor analysis and linear models are not able to capture the majority of high-dimensional data patterns. (2) since deep neural network and the domain of Natural Language Processing gain unprecedented popularity during this decade, neural network models with millions, even billions of parameters are able to achieve brilliant performance [8]. However, the training process could take up to several months, and the models fail to be guaranteed that they do not overfitting.

In this study, ten thousand hotel reviews were collected from TripAdvisor, and 1-3 star feedbacks were labelled as 'negative' while 4-5 star reviews were regarded as 'positive'. To investigate the performance of classical and up-to-date machine learning models, a variety of machine learning models, namely Logistic Regression, Naive Bayesian, Decision Tree, Random Forest, Support Vector Machine (SVM) and Neural Network, are compared through two 10-folder-cross-validation experiments which can reduce the risk of overfitting. This study aims to compare the predictive power and training efficiency of widely-used machine learning models, and discuss how text processing impact the model performances.

II. DATA PREPARATION

A. Data Collection

The data was collected from the database: data.world [9] and this database acquired as well as formatted 10,000 hotel comments from TripAdvisor into csv files. As **Fig.1** illustrates, approximately 7800 reviews are 4 to 5 stars and were labelled as 'positive (1)' while around 2200 reviews are 1-3 stars and were treated as 'negative (0)'.

B. Experimental Preparation

To overcome the problem of overfitting, the data is randomly shuffled and divided into ten trunks. Each time 9 chunks (9,000 data) was utilized for training and the labels of the rest chunk (1000 data) was applied for testing. This design

Figure 1. Customer rating data

is called 10-folder-cross-validation and it helps to reduce the risk of excessively fitting and local optimal solution.

As Table 1 shows, because in the first experiment only titles were used, each single data vector merely consists of a few words and the size of bytes (1 English letter or numerical digit requires 1 byte for storage) has mean 25.7 and standard deviation 14.6. For the second experiment, the contents of review were firstly filtered by punctuations (comma, period, semi-colon, colon, hyphen, etc) and neutral emotional words with high frequency ('hotel','inn','motel', etc). After adding filtered review contents for training, the average bytes of each sample increase from 25.7 to 442.4 and the sample size varies more intensively with the standard deviation being 394.0.

TABLE I. EXPERIMENTS PREPARATION

Designs	Details		
	Training/Testing dataset size	*Mean of bytes*	*Standard deviation of bytes*
Experiment 1	9000/1000	25.7	14.6
Experiment 2	9000/1000	442.4	394.0

III. EXPERIMENTS

A. *Analysis on Experiment 1*

As shown in **Table 2**, the six machine learning models are imported from Python module 'sklearn' [10] and the parameters for models are listed in detail for readers' experimental reproduction. For instance, for Neural Network a 3-layer structure with 5,4,2 neurons on each layer was applied, relu function was chosen as activation function and the learning rate was 10^{-5}.

The predictive performance of six models, namely Decision Tree, Naive Bayesian, Logistic Regression, SVM, Neural Network and Random Forest, are depicted in **Fig.2** as boxplot format which contains five horizontal lines to indicate minimum, first quartile, median, third quartile, and maximum of data. First of all, neural network performs worst among all five models, and this phenomena attributes to small training data which is just 25.7 bytes on average. Furthermore, due to the fact that Random Forest consists of average results of 100 decision trees given specific randomness, Random Forest outperforms Decision Tree in terms of accuracy. Thirdly, as mentioned before, the training data is tiny and monotonous, linear model such as Logistic Regression performs above non-linear models such as neural network. Apart from that, SVM has the advantage of dealing with unstructured data and scales relatively well to high dimensional data, therefore for textual analysis SVM has the best predictive performance. Last but as significant, each review titles merely consists of a few words and feeding them into Machine Learning models only takes time in the magnitude of seconds, but five of six models are able to achieve prediction accuracy ranging from 84% to 87%. This result indicates that without dataset in the scale of gigabytes or deep neural network trained for months with billions of parameters, classical models with relatively tiny structures have the power to achieve satisfying predictive performance.

TABLE II. 'SKLEARN' MODULE PARAMETERS

Models	Parameters
Decision Tree	criterion='gini', min_samples_leaf=1, min_samples_split=2
Naive Bayesian	alpha=1.0, fit_prior=True
Logistic Regression	max_iter=100, multi_class='warn', penalty='l2', solver='lbfgs'
SVM	degree=3, gamma='scale', kernel='rbf'
Neural Network	activation='relu', alpha=10-5, hidden_layer_sizes=(5, 4, 2)
Random Forest	criterion='gini', n_estimators=100

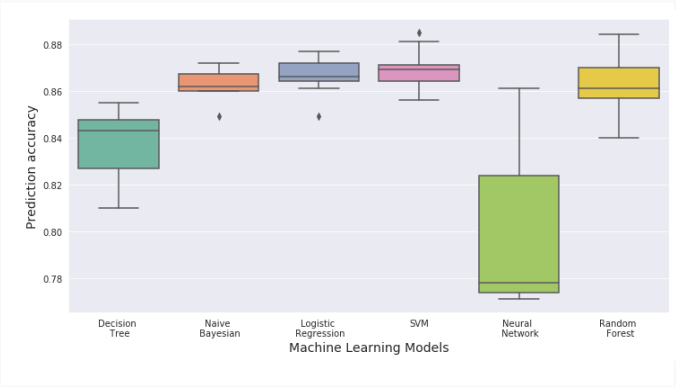

Figure 2. Accuracy performance of ML models on title texts

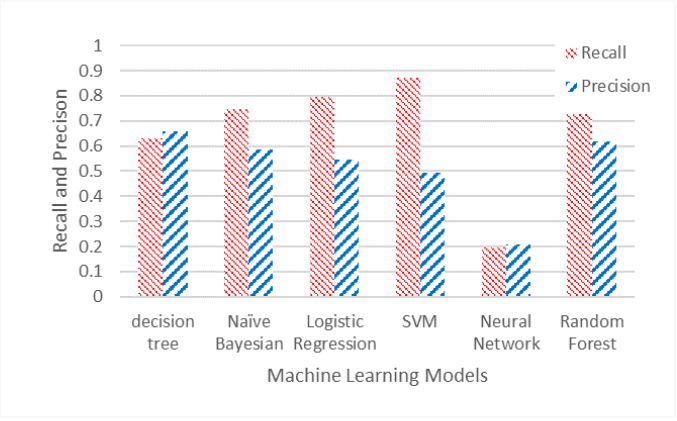

Figure 3. Recall and Precision of ML models on title texts

When it comes to **Fig.3**, even though SVM achieves optimal predictive accuracy, this methodology behaves apparent imbalance between precision and recall. Also, in terms of Logistic Regression which reaches secondary accuracy performance, the approach still has non-negligible distinction between precision and recall. Therefore, it can be concluded that although SVM as well as Logistic Regression models obtain satisfying accuracy based on small training sample bytes, there is a disturbing tendency for these two models to predict positive customer reviews as negative. In real practices, the two models would result in unwilling bias.

B. Analysis on Experiment 2

As described in Section 2.2, a lot more textural information was supplemented for training and disruptive bytes such as punctuations and neutrally common words have been filtered before the experiment.

Judging from the discrepancies between **Fig.4** and **Fig.5**, first and foremost, SVM and Naive Bayesian achieves nearly 92% maximal prediction accuracy and around 89% median predictive accuracy. Hence, by feeding more relevant data into machine learning models there will be distinguishable improvement on performance. Secondly, neural network reached 86% accuracy which raised by 8% from 78% during which merely titles were utilized for training. This finding corresponds to the property of Neural Network that with larger related datasets, the approach will behave steadily better predictive power.

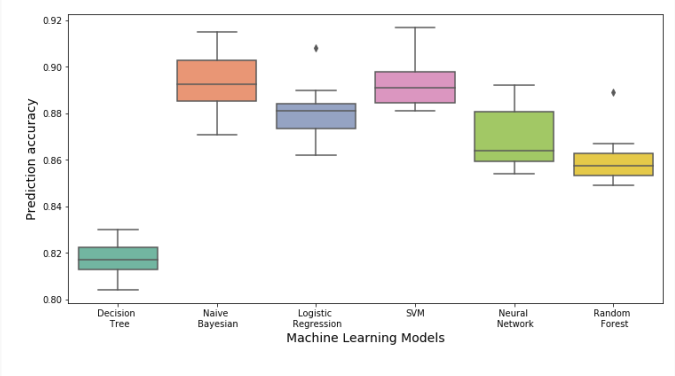

Figure 4. Accuracy performance of ML models on filtered review contents

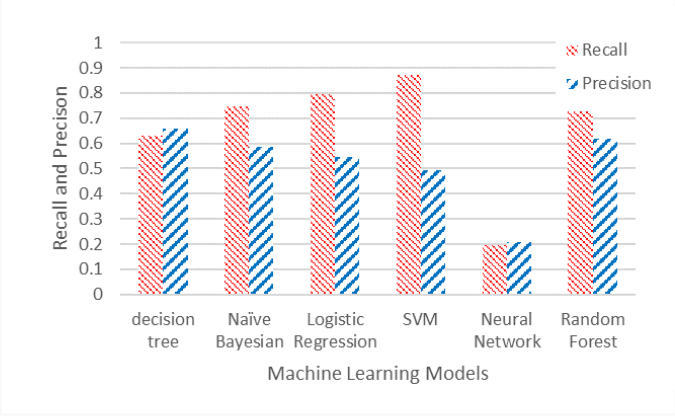

Figure 5. Recall and Precision of ML models on filtered review contents

In contrast to **Fig.3**, according to **Fig.5**, both Logistic Regression and SVM display harmonious differences between recall and precision. Hence, the two models will behave less bias towards True positive and have few errors to mistakenly treat actual positive reviews as negative.

IV. CONCLUSION

In this paper, ten thousand hotel reviews in English were collected and labelled as negative or positive, and then review titles as well as filtered review contents were trained by Logistic Regression, Naive Bayesian, Decision Tree, Random Forest, SVM and Neural Network models through 10-folder-cross-validation scheme respectively. The conclusions are summarized below:

- When training samples have small byte sizes, the 3-layer Neural Network model distinctly performs worse than other five classical models. With more textual information, predictive accuracy of Neural Network model improves massively. However, Neural Network still cannot surpass traditional models such as SVM and Logistic Regression in the two experiments, which, on the contrary, indicates that neutral network requires careful structure design as well as parameter tuning for excellent performance.
- Among all six models, SVM outperforms others in terms of accuracy, precision and recall, due to its model strengths in dealing with unstructured data and high-dimensional features.
- When models were directly trained on review titles which have only 25.7 byte lengths on average, five out of six models are able to achieve predictive accuracy ranging from 84% to 87%. Considering the fact that training merely takes a few seconds, this finding casts light on business application of classical machine learning models for individual hotels to identify customer satisfaction.
- With more textual data and irrelevant punctuations as well as neutral words filtered, SVM model could reach maximal predictive accuracy nearly 92%.

REFERENCES

[1] A. Emir, H. Halim, A. Hedre, D. Abdullah, A. Azmi, and S. Kamal. Factors influencing online hotel booking intention: A conceptual framework from stimulus-organism-response perspective. International Academic Research Journal of Business and Technology, 2016, 2(2), 129-134.

[2] Statistic Brain, 2016. Internet Travel & Hotel Booking Statistics. https://www.statisticbrain.com/ internet-travel-hotel-booking-statistics/

[3] W. G. Kim, and S. A. Park. Social media review rating versus traditional customer satisfaction: which one has more incremental predictive power in explaining hotel performance?. International Journal of Contemporary Hospitality Management, 2017, 29(2), 784-802.

[4] A. J. Flanagin, and M. J. Metzger. Trusting expert-versus user-generated ratings online: The role of information volume, valence, and consumer characteristics. Computers in Human Behavior, 2013, 29(4), 1626-1634.

[5] R. K. Nielsen, and K. C. Schrøder. The relative importance of social media for accessing, finding, and engaging with news: An eight-country cross-media comparison. Digital journalism, 2014, 2(4), 472-489.

[6] X. Tian, W. He, R. Tao, and V. Akula. Mining online hotel reviews: a case study from hotels in China, 2016.

[7] Z. Xiang, Z. Schwartz, Jr, J. H. Gerdes, and M. Uysal. What can big data and text analytics tell us about hotel guest experience and satisfaction?. International Journal of Hospitality Management, 2015, 44, 120-130.

[8] Z. Jianqiang, G. Xiaolin, and Z. Xuejun. Deep convolution neural networks for twitter sentiment analysis. IEEE Access, 2018, 6, 23253-23260.

[9] data.world, 2019. Hotel Reviews – dataset in DATAFINITY. https://data.world/datafiniti/hotel-reviews.

[10] J. Hao, and T. K. Ho. Machine Learning Made Easy: A Review of Scikit-learn Package in Python Programming Language. Journal of Educational and Behavioral Statistics, 2019, 44(3), 348-361.