Sentiment Analysis of E-commerce Customer Reviews Based on Natural Language Processing

Xiaoxin Lin
Department of Computing and Decision Sciences
Hong Kong Lingnan University
Shenzhen, China
dingdinglin23@yahoo.com

ABSTRACT

E-commerce can largely boost the economic development and customer behavior analysis is necessary for e-commerce marketing strategy. We used the dataset of Women's E-Commerce Clothing Reviews to study the sentiment analysis of customer recommendation. Five popular machine learning algorithms were applied to solve the problem, including Logistic Regression, Support Vector Machine (SVM), Random Forest, XGBoost and LightGBM. These algorithms aimed at figuring out the insight correlation between review features and product recommendation based on natural language processing (NLP). The best result was achieved by LightGBM algorithm with highest AUC value and accuracy. The precision, recall and F1 score were all 0.97. Ridge Regression, Linear Kernel SVM and XGboost algorithms which had close performances with the accuracy of 0.94. This research can help generate a deeper comprehension of customer sentiment and grasp customer psychology in e-commerce transaction industry

CCS CONCEPTS

• Computing methodologies • Artificial intelligence • Natural language processing • Lexical semantics

KEYWORDS

E-commerce, Reviews, Machine learning, Natural language processing

ACM Reference format:

Xiaoxin Lin. 2020. Sentiment Analysis of E-commerce Customer Reviews Based on Natural Language Processing. In *Proceedings of ACM ISBDAI conference (ISBDAI 2020). ACM, Johannesburg, South Africa, 5 pages.* https://doi.org/10.1145/3436286.3436293

1 Introduction

The market share of the traditional apparel selling industry is being eroded by the e-commerce industry, facing the problem of high costs including the labor cost, store rental cost, operation cost, marketing promotion cost, etc [1-2]. And the fierce competition in the market forces the sellers to find out another economic way to further their business. Many physical apparel stores are closed and tend to sell their products on the e-commerce platform to save the cost and obtain customers more reliably and efficiently. The e-commerce category capacity is enormous, but sales of apparel always account for a large proportion of customers' online consumption considerably because clothes are regarded as necessities of life commonly. Especially, women are the major customers of apparel consumption rather than men no matter online or offline as women are more likely to be influenced by the fashion trend and brand temptations than men [3]. Thus, it is necessary to study women customers' purchasing behavior in the apparel selling industry.

For gaining a better understanding of online customer behavior, many types of research have done to analyze online reviews by using various machine learning algorithms. Agarap et al. adopted a bidirectional recurrent neural network (RNN) with long-short term memory unit (LSTM) for recommendation and sentiment classification based on Women's E-commerce Clothing Reviews dataset and got a result that an F1 score of 0.88 [4]. Zubrinic et al. analyzed the identical dataset and they achieved the best results by using the SVM algorithm and maximum entropy classifier with the accuracy of 0.84 [5]. Noor et al. applied several algorithms to evaluate the dataset as well. As a result, Logistic Regression provided the best result with an accuracy of 0.88 and KNN delivered the highest recall value of 0.99 [6]. To detect the unfair views and evaluate the performance of sentiment classification, Elmurngi et al. analyzed the datasets of reviews from Amazon. The results showed that the Logistic Regression algorithm is the best classifier with the highest accuracy [7]. Ye et al. finished sentiment analysis of online reviews of seven popular travel destinations by three supervised machine learning algorithms including Naive Bayes, SVM, and the character-based N-gram model. SVM and N-gram performed better than Naive Bayes [8].

Regarding this research, it concentrates on applying machine learning algorithms including Logistic Regression, SVM, Random Forest, XGboost, and LightGBM to analyze Women's E-Commerce Clothing Reviews dataset. The main objective is to figure out the correlation between review features and product recommendation based on natural language processing (NLP). NLP is a branch of computer science and artificial intelligence that is closely related to the interaction between computers and humans using natural language [9]. Extracting and analyzing the keywords from the reviews by customers based on NLP can benefit a deeper comprehension of customer emotions and sentiments.

ISBDAI '20, ISBDAI '20, April 28–30, 2020, Johannesburg, South Africa
 ACM ISBN 978-1-4503-7645-7/20/…$15.00
https://doi.org/10.1145/3436286.3436293

The layout of this paper is designed to be clear and logical for a better understanding. Its structure is as follows. After the introduction, the second part illustrates the data source, basic features of the dataset and a tool we used to process the raw texts. The models used in this research is detailedly introduced in the third part with formulas. The achieved results with figures and some discussions are presented and demonstrated in the fourth part. Finally, conclusions and a possible improvement in the future are described in the last part.

2 Data Research

This research is based on a dataset related to Women's E-Commerce Clothing Reviews from Kaggle [10]. This dataset has 19,675 samples in total with 10 columns. Several important variables in this dataset are as follows. Title represents the title of the review text. Rating means the score granted by the customer from 1 worst to 5 best. Recommended IND is the binary variable stating whether the customer recommends the product where 1 is recommended, 0 is not recommended. Positive Feedback Count stands for the number of other customers who found this review positive.

Table 1 illustrated several details of this dataset. The average age of this group of women customers is 43 where the minimum age is 18 and the maximum age is 99. The mean of rating and recommended IND are 4.18 and 0.82 respectively which means that most of the customers tend to show the satisfaction on the clothes they buy. With regard to the positive feedback count, the minimum number is 0 while the maximum number is 122 and it shows a huge gap.

Table 1: Statistical results of the dataset

	Age	Rating	Recommended IND	Positive Feedback Count	Division Name	Department Name	Class Name
mean	43.26	4.18	0.82	2.65	0.47	2.35	6.93
std	12.26	1.11	0.39	5.83	0.61	1.63	5.23
min	18	1	0	0	0	0	0
25%	34	4	1	0	0	1	3
50%	41	5	1	1	0	2	7
75%	52	5	1	3	1	4	10
max	99	5	1	122	3	6	20

To analyze the raw text, the text must be encoded as some specific values or vectors for use as input to a machine learning algorithm, this is called feature extraction or vectorization. TfidfVectorizer is a tool from sklearn library for transforming the text into vectors. 'TF' in the word is the term frequency and it stands for the times a specific word showing in the dataset. 'IDF' is the inverse document frequency and it means that the inverse proportion of that word over the entire document corpus. TF-IDF stands for the result that TF multiplies IDF, and the higher the result is, the word is more likely to be the keyword [11]. TfidfVectorizer uses this algorithm to vectorize the raw text and it can be illustrated in a matrix form. It is an efficient and practical function for pre-processing the text because the TF-IDF algorithm is reliable and it can always return documents that are highly relevant to a particular query.

3 Model

3.1 Logistic Regression

Logistic Regression is one of the supervised machine learning models for dealing with classification tasks. It is based on a probability which its range has to be from 0 to 1 and the range of liner regression function is from negative infinity to positive infinity [12]. To be specific, if the positive probability is higher than the negative probability, then the prediction tends to be positive, vice versa. In other word, if the probability is larger than 0.5, then this prediction should be the correct one [13].

Logistic regression function is:

$$P\left(Y = Yes \mid X; w\right) = S\left(w^t x + b\right) = \frac{1}{1 + e^{-(w^t x + b)}} \quad (1)$$

$$1 - P\left(Y = Yes \mid X; w\right) = P\left(Y = No \mid X; w\right) \quad (2)$$

where P is the probability, w is a weight vector, b is a bias, e is a constant.

However, the bias and variance are usually easy to cause overfitting problems. But regularization can reduce the complexity and instability of the model and avoid the risk of overfitting. L1-norm and L2-norm are two methods of regularization [14]. L1-norm is the sum of the absolute values of the vectors and it is also called Lasso Regression. It helps generate a sparse weight matrix that can be used for feature selection. L2-norm is also called Ridge Regression which is the sum of the squares of the vectors and it can help to reduce the overfitting. Lambda (λ) is a parameter for adjusting regularization. By increasing λ starting from 0, the classification accuracy will decrease at first and increase later and this is for optimizing the minimum classification error rate.

L1-norm function is:

$$j(w) = \sum_{i=1}^{N} (y_i - w_0 - \sum_{j=1}^{p} w_j x_{ij})^2 + \lambda \sum_{j=1}^{p} |w_j| \quad (3)$$

L2-norm function is:

$$j(w) = \sum_{i=1}^{N} (y_i - w_0 - \sum_{j=1}^{p} w_j x_{ij})^2 + \lambda \sum_{j=1}^{p} w_j^2 \quad (4)$$

3.2 Random Forest

Decision Tree is a supervised machine learning method used for regression as well as classification and helps to predict the value of a target variable by abstracting data features from training data. A decision tree is a flowchart splitting from a root node and the branches always point downwards to the next decision node. The process will end up when the features cannot be split any more or the model has been optimal. One of the Decision Tree algorithms is called CART and Gini coefficient is a metric for this algorithm. It is calculated by subtracting the sum of the squared probabilities of each class from one. We can formulate it as illustrated below:

$$Gini(x) = 1 - \sum_{i=1}^{k} p_i^2 \tag{5}$$

where k is the number of classes in the feature and i is the sequence of the feature.

Parent node corresponding sample collection for D, CART choose features A to split into two child nodes, the corresponding collection for D_L and D_R. The Gini index after splitting is defined as follows:

$$Gini(D, A) = \frac{|D_L|}{|D|} Gini(D_L) + \frac{|D_R|}{|D|} Gini(D_R) \tag{6}$$

Random forest is an ensemble learning algorithm consisting of abundant Decision Trees by randomly choosing the specimens and features and output the class which is the majority class from each tree. Bagging and Boosting are two machine learning ensemble meta-algorithms designed to improve stability, reduce variance and bias, and avoid overfitting problems [16-17]. Random Forest can avoid overfitting problems. Pruning can reduce the size of Decision Trees by two techniques which are pre-pruning and post-pruning to remove parts of the tree that do not provide power to build the tree.

3.3 Support Vector Machine

SVM can be utilized to handle the classification problems. It aims at finding out a hyperplane which is one dimension smaller than the data dimension and this hyperplane has a maximum margin between two classes of the classification training data [18]. The training data at the decision boundary of hyperplane called support vector. The function of hyperplane is:

$$w^T x + b = 0 \tag{7}$$

where w is a weight vector, b is a bias. The function of the decision boundary is:

$$f(x) = sign(w \cdot x + b) = sign(\sum_{i=1}^{N} \alpha_i y_i x_i \cdot x + b) \tag{8}$$

where α_i is a Lagrange multiplier, $y_i \in \{+1, -1\}$, $i = 1, 2 \ldots. x_i$ is the training data, x is the testing data. N stands for a set of input vectors with corresponding class labels.

Not only can SVM be applied for the linear separable training data, but also it can solve the non-linear separable problems by using a technique called kernel trick which can transform the data into a higher-dimensional feature space so that it can be classified [19]. Linear function, polynomial function, radial basis function and sigmoid function are four kernel functions which can be used in the SVM algorithm.

3.4 XGboost

XGboost is an open-source package based on a tree gradient boosting framework. It is not uncommon that the predicted value has an error compared to the real value. To optimize and build a more sophisticated model, the main principle of gradient boosting is to adjust the target of the fitting process to the residual error of the current sample circularly. XGboost is an upgraded algorithm of Gradient Boosting Decision Tree (GBDT) which is one of the implementations of gradient boosting as well. Compared to GBDT, XGboost uses both the first and second derivatives and performs a second Taylor expansion of the loss function, while GBDT only applies the first derivative [20].

The object function of XGboost is:

$$L(\phi) = \sum_i l(y_i, \hat{y_i}) + \sum_k \Omega(f_k) \tag{9}$$

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda ||w||^2 \tag{10}$$

where i is the sequence of the sample, $l(\hat{y_i} - y_i)$ is the prediction error of this sample which means that it is the training loss. k is the number of trees. $\sum_k \Omega(f_k)$ represents the complexity of the trees. T stands for the number of leaves and $\frac{1}{2}\lambda ||w||^2$ is L2 norm of leaf scores.

XGboost can process the sparse data and missing values flexibly and it allows users to define custom optimization goals and metrics. In addition, the L2 norm applied by XGboost can help to control the complexity of the model and reduce the risk of overfitting. As a result, in most situations, XGboost usually can generate an outstanding result.

3.5 LightGBM

LightGBM was launched by Microsoft company in January 2017. It refers to two core techniques, which are exclusive feature bundling (EFB) and gradient-based one-side sampling (GOSS). GOSS can reduce the computation amount by distinguishing the instances with different gradients, retaining the larger gradient instances and sampling the smaller gradient instances randomly [21]. EFB improves training speed by reducing the feature dimension through feature bundling and the bundled features are mutually exclusive so that they will not lose information when they are bundled together.

LightGBM uses the histogram optimization algorithm. It means that the values are separated into different intervals which will be regarded as the bins. Then the bins will be illustrated by a histogram. Compared with the pre-sorted algorithm applied by XGboost, histogram algorithm can reduce computation amount and avoid overfitting to some extent. Furthermore, it adopts the Leaf-wise tree growth method to find a leaf with the largest split gain, which can get better accuracy and control the depth of the tree than the Level-wise algorithm. And it is capable of handling large-scale data and it has the support of parallel and GPU learning as well.

The function of the split gain is:

$$Gain = \frac{1}{2}\left[\frac{G_L^{\,2}}{H_L+\lambda}+\frac{G_R^{\,2}}{H_R+\lambda}-\frac{(G_L+G_R)^2}{H_L+H_R+\lambda}\right]-\gamma \quad (11)$$

where $\frac{G_L^{\,2}}{H_L+\lambda}$ stands for the left tree, $\frac{G_R^{\,2}}{H_R+\lambda}$ stands for the right tree and γ is the cost of complexity

4 Result And Discussion

Figure 1 is the correlation matrix which can reflect the specific quantitative correlation between each feature. When the absolute value is higher, the correlation between two features is stronger, vice versa. There are three groups of features had a significant and positive figure. It is clear that the correlation between Recommended IND and Rating is 0.79. It means that the prediction of whether the customer recommends the product is highly related to the rating given by the customer. Besides, Department Name and Clothing ID have a strong correlation with 0.87 as well. Division Name and Class Name have a slight correlation with 0.16.

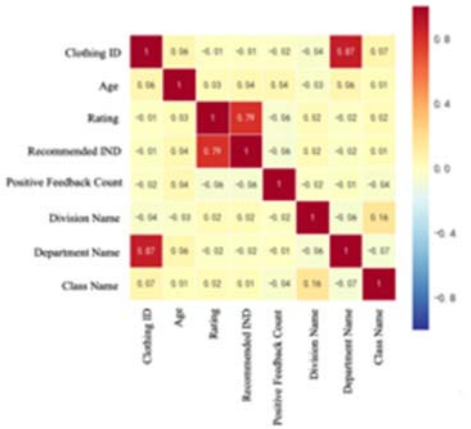

Figure 1: Correlation between each feature

For each algorithm, we need to set these important parameters to optimize the model.

Regarding the Logistic Regression algorithm, 'penalty' is used to decide the norm used in the regularization including L1-norm and L2-norm and its value is L2-norm. 'C' is the reciprocal of regularization extent and it is set to be 13.0. 'max_iter' controls the maximum frequency of iterations and its value is 100. Logistic Regression can generate a numeric and more accessible result so it can define the classification more easily and even can be ranked in order. Besides, Logistic Regression does not require a complex calculation process or large storage resources and it is more efficient and timesaving.

For the SVM algorithm, the gamma parameter defines how much the influence of a single training example has and its value is 'scale'. Kernel parameter can be adjusted to different kernel functions including linear function and Gauss function.

For Random Forest algorithm, 'n_estimators' means the number of trees in the forest and its value is 100. 'min_samples_split' is the least number of samples needed to split a node and its value is 5. 'min_weight_fraction_leaf' is the least number of samples requested for a leaf node and its value is 0.1.

For the XGboost algorithm, 'n_estimators' is the frequency of iterations and its value is 10. 'learning_rate' stands for the step size of each iteration and if it is too large then the accuracy will decrease, and its value is 0.01. 'gamma' parameter can be used to assign the least loss reduction to a split and its value is 0.

For the LightGBM algorithm, 'boost_type' can be defined as the type of algorithm we want to run, and it is set to be gbdt (traditional Gradient Boosting Decision Tree). Lambda parameter specifies regularization and its typical values range from 0 to 1 and its value is 0.6. The metric parameter can specify the loss and it is set to be 'auc'.

Table 2 and Table 3 showed the result by using different algorithms. It is clear from the first tables that the LightGBM algorithm had the best prediction performance with the highest accuracy of 0.98, and precision, recall, and F1 score were all 0.97. In addition, the Ridge Regression algorithm had a good performance as well, with the accuracy of 0.94, recall of 0.97, and both precision and F1 score were 0.96. XGboost algorithm achieved a closely similar and excellent result compared with Ridge Regression algorithm with the only one difference from the recall of 0.96. By contrast, however, SVM with the RBF kernel algorithm got the worst result. Four indexes of it were the lowest where both the accuracy and recall were 0.81 and both the precision and F1 score were 0.90. SVM with Linear kernel and Random Forest algorithms just got intermediate results.

In regard to the Area Under Curve (AUC) value, the LightGBM algorithm still had the best result of 0.96 followed by XGboost with the result of 0.92. Ridge regression and linear kernel SVM, Lasso Regression and RBF kernel SVM got the same result of 0.90 and 0.89 respectively. However, the Random Forest algorithm only got a 0.86 AUC value.

Table 2: Evaluation of different models

Model	Accuracy	Precision	Recall	F1 score
Lasso Regression	0.93	0.96	0.95	0.96
Ridge Regression	0.94	0.96	0.97	0.96
Linear Kernel SVM	0.94	0.97	0.95	0.96
RBF Kernel SVM	0.81	0.92	0.81	0.90
Random Forest	0.90	0.97	0.91	0.94
XGBoost	0.94	0.96	0.96	0.96
LightGBM	0.98	0.97	0.97	0.97

Table 3: AUC value of different models

Model	AUC value
Lasso Regression	0.89
Ridge Regression	0.90
Linear Kernel SVM	0.90
RBF Kernel SVM	0.89
Random Forest	0.86
XGBoost	0.92
LightGBM	0.96

Compared with other research based on the same dataset, Random Forest, SVM, and Logistic Regression algorithms are used in every research but this research adopts extra two algorithms called XGboost and LightGBM which the other research did not use, especially, LightGBM got the more excellent results with the accuracy of 0.98 than the other research. However, in the other

research, KNN, Naïve Bayes and Adaboost algorithms have been used but they are not used in this research and KNN delivered a similar recall value.

5 Conclusion

This research focused on figuring out the correlation between the reviews and product recommendation by applying machine learning algorithms including Logistic Regression, SVM, Random Forest, XGboost, and LightGBM to analyze the Women's E-Commerce Clothing Reviews dataset. LightGBM algorithm got the best result with the highest accuracy and AUC value of 0.98 and 0.96. Ridge Regression, Linear Kernel SVM and XGboost algorithms had close performances. While the RBF Kernel SVM algorithm got the worst result with the lowest accuracy of 0.81.

Although this research got a high accuracy with the help LightGBM algorithm, we can conduct the in-depth processing of the raw review texts to gain a more accurate result and fulfill NLP better in the research. Deleting words such as prepositions, pronouns, and other words without really great meaning but often appear and analyzing words of emotional significance in the text pre-processing stage can help increase the accuracy and comprehension of the reviews. In addition, distinguishing the real difference between rating and recommendation options can reduce the loss when the model is training and predicting.

REFERENCES

[1] Hammond J, Kohler K. 2000. E-commerce in the textile and apparel industries[J]. *Tracking a Transformation: E-commerce and the Terms of Competition in Industries*.

[2] Jinfu W, Aixiang Z. 2009. E-commerce in the textile and apparel supply chain management: Framework and case study[C]//*2009 Second International Symposium on Electronic Commerce and Security*. IEEE, 1: 374-378.

[3] Abraham L B, Mörn M P, Vollman A. 2010. Women on the web: How women are shaping the internet[M]. *ComScore, Incorporated*.

[4] Agarap A F, Grafilon P. 2018. Statistical Analysis on E-Commerce Reviews, with Sentiment Classification using Bidirectional Recurrent Neural Network (RNN)[J]. *arXiv preprint arXiv*:1805.03687.

[5] Žubrinić K, Miličević M, Sjekavica T. 2018. A Comparison of Machine Learning Algorithms in Opinion Polarity Classification of Customer Reviews[C]//*18th International Conference on Applied Computer Science (ACS'18)*.

[6] Dey U K, Noor A. 2019. Comparative Exploration Of Prediction Algorithms For Sentiment Analysis Using NLP[C]//*2019 International Conference on Computer Communication and Informatics (ICCCI)*. IEEE, 1-6.

[7] Elmurngi E I, Gherbi A. 2018. Unfair Reviews Detection on Amazon Reviews using Sentiment Analysis with Supervised Learning Techniques[J]. *JCS*, 14(5): 714-726.

[8] Ye Q, Zhang Z, Law R. 2009. Sentiment classification of online reviews to travel destinations by supervised machine learning approaches[J]. *Expert systems with applications*, 36(3): 6527-6535.

[9] Bird S, Klein E, Loper E. 2009. Natural language processing with Python: analyzing text with the natural language toolkit[M]. " O'Reilly Media, Inc.".

[10] Kaggle Dataset available at: *https://www.kaggle.com/nicapotato/womens-ecommerce-clothing*.

[11] Dichiu D, Rancea I. 2016. Using Machine Learning Algorithms for Author Profiling In Social Media[C]//*CLEF (Working Notes)*. 858-863.

[12] Menard S. 2002. Applied logistic regression analysis[M]. *Sage*.

[13] Press S J, Wilson S. 1978. Choosing between logistic regression and discriminant analysis[J]. *Journal of the American Statistical Association*, 73(364): 699-705.

[14] Ravikumar P, Wainwright M J, Lafferty J D. 2010,. High-dimensional Ising model selection using ℓ1-regularized logistic regression[J]. *The Annals of Statistics*, 38(3): 1287-1319.

[15] Du W, Zhan Z. 2002. Building decision tree classifier on private data[C]//*Proceedings of the IEEE international conference on Privacy, security and data mining-Volume 14*. Australian Computer Society, Inc. 1-8.

[16] Liaw A, Wiener M. 2002. *Classification and regression by randomForest*[J]. R news, 2(3): 18-22.

[17] Cootes T F, Ionita M C, Lindner C, et al. 2012. Robust and accurate shape model fitting using random forest regression voting[C]//*European Conference on Computer Vision. Springer*, Berlin, Heidelberg, 278-291.

[18] Scholkopf B, Smola A J. 2001. *Learning with kernels: support vector machines, regularization, optimization, and beyond*[M]. MIT press.

[19] Campbell W M, Sturim D E, Reynolds D A, et al. 2006. SVM based speaker verification using a GMM supervector kernel and NAP variability compensation[C]//*2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*. IEEE, 1: I-I.

[20] Chen T, Guestrin C. 2016, Xgboost: A scalable tree boosting system[C]//*Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. ACM, 785-794.

[21] Ke G, Meng Q, Finley T, et al. 2017. Lightgbm: A highly efficient gradient boosting decision tree[C]//*Advances in Neural Information Processing Systems*. 3146-3154