Proceedings of the 5th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/17/20230927

A series of machine learning methods for user rating prediction

Jingbo Wang

Leighton School Shanghai, Shanghai, 200000, China

tomy@djiun.org

Abstract. Users' feedback online for services or products reflects their satisfaction. For example, higher rating implies better user satisfaction and experience. It is crucial to provide users satisfying services or products to ensure user experience and thus improve the long-term value of the platform. Manually reviewing all candidate services and products to select potentially user-preferred ones it time-costuming due to large amount of candidates. Therefore, this paper studies a series of machine learning based algorithms to automatically predict the user feedback for an online shopping website, i.e., user ratings in this scenario. To be specific, three machine learning models are investigated, including logistic regression, multi-layer perceptron, and convolution neural network. Starting with the pre-processing of raw data crawled from Amazon, an international online shopping website such three models are first trained over the training set and then evaluated on the testing set. The experiments demonstrate the convolutional neural networks give the highest accuracy for the unseen data.

Keywords: user rating prediction, logistic regression, multi-layer perceptron, convolutional neural networks.

1. Introduction

Nowadays, mass user review is being generated every second. People leave a comment after they finish a service, buy products, and do any activities that need feedback. Websites, service providers, and platforms use the reviews to elevate the feeling of customers. This kind of data can be used to analyse the satisfaction of the customers to a creating product, which is related to the rating scores customers have given to a product.

Shopping websites, like Amazon and Taobao, can provide sentimental analysis services to their own platform. But they do not sell such services as APIs to other shopping websites where such analytic tools are needed. Around the world, numbers of personal e-commerce stores are open on social media or self-hosted platforms that are run by individuals. It usually takes up a lot of time to analyse the review made by the users manually, since there are no available services that do this simple take for the store runners. By applying machine learning methods to automat the data analysis process can save considerable time and allow individuals to focus more on producing better products. Thus, this work presents a series of machine learning based methods to predict user ratings for a product or a service, so that shoppers are able to provide products with higher user rating scores.

The process of developing a user rating prediction model usually has three phases. In the first phase, the program loads in datasets and conducts pre-processing on the datasets. This includes splitting

DOI: 10.54254/2755-2721/17/20230927

datasets into test and training sets, removing the empty row, dropping useless columns, and converting the string sentences into a vector representation. Such an action is crucial to ensure a model is not affected by the noises in datasets. In the second phase, the program trains models under specific hyper-parameters settings. To be specific, different models can be trained using a GPU accelerator that shouter the time of training. As in the last phase, the program will use the test sets to evaluate the models and give the accuracy scores at the end.

In this project, the three models are considered, including logistic regression, multi-layer perceptron as well as convolutional neural networks. Through experiments, logistic regression gives the lowest accuracy score, while convolutional neural networks obtains the best performance which demonstrates the effectiveness of such a deep learning based model.

2. Method

This paper follows a flow [1] to study three machine learning based models, including logistic regression, multi-layer perceptron and convolutional neural network. This process starts with retrieving data from online websites or a local database. Then the data pre-processing is conducted by removing empty rows. Then three different models are trained over training data and evaluated on the test set.

2.1. Data collection and pre-processing

This research is based on a dataset that includes thirty-five million reviews gathered from Amazon in a period of 18 years up to March 2013. The content of datasets includes product and user information, review text, as well as user ratings. For ratings, one stands for positive, and two stands for negative. Each instance includes 1.8 million and 2 million reviews individually. The dataset has previously been divided into training and test sets, with 20% as test sets and 80% as training sets. The review text is anonymous. Due to the available computing power on hand, this research includes the top ten thousand lines of data as the datasets of this project. The positive and negative review in both the test and training datasets account for 50% each, as shown in figure 1.

The histograms in figure 2 show the length distribution of review data. In Natural Language Processing (i.e., NLP) activities, like sentiment analysis, the length of input text is crucial. Longer reviews may contain more information that might be pertinent to predicting sentiment. The shorter length of data is also useful. It is easier to process and analyse shorter reviews, as they contain less noise and irrelevant information. This can be especially important in tasks that involve text classification or sentiment analysis, which is relevant to the research in this paper. In addition, shorter review data is more efficient to be processed, especially when dealing with large datasets, since shorter reviews require fewer computational resources.

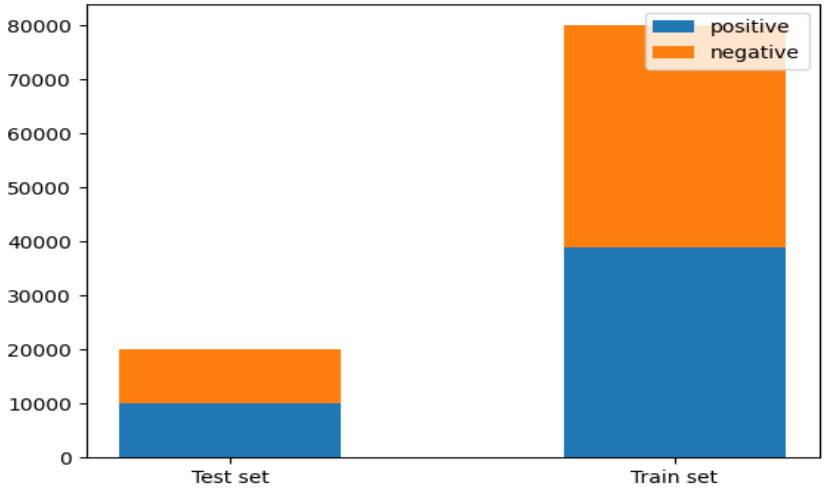

Figure 1. The histograms of positive and negative reviews in dataset.

DOI: 10.54254/2755-2721/17/20230927

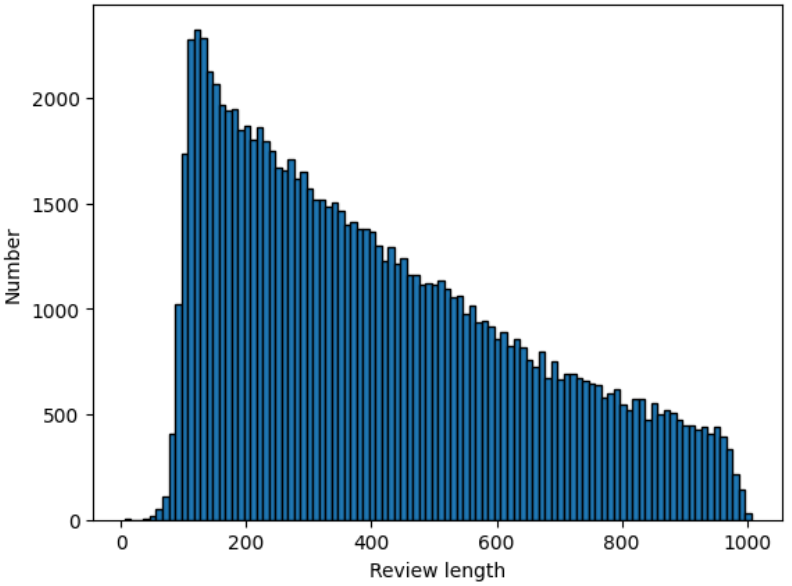

Figure 2. The histograms of review length.

2.2. Tools

There are many packages used in this research. The code is implemented using TensorFlow based on the Keras framework. TensorFlow and Keras are deep learning frameworks used for training machine learning models. TensorFlow is an open-source software library that enables dataflow and differentiable programming for various tasks. It is especially useful for building and running deep neural networks. The framework offers a wide range of tools and capabilities of building and training deep neural networks, including high-level APIs for constructing and training models fast and efficiently and low-level APIs for defining and executing computational graphs. A high-level Python neural network API, called Keras, makes it simpler to develop and train neural networks. It offers a simple, straightforward interface for creating and refining deep learning models. Other packages used in this project are NumPy and Pandas, used for data pre-processing, as well as Matplotlib for data visualization.

2.3. Models

A supervised machine-learning method, called logistic regression, could be used to group texts according to their emotion. It works by learning a function that maps a text input to a probability of belonging to a certain class, like 0 for negative and 1 for positive. Logistic regression can handle binary or multi-class classification problems and incorporate features such as word frequencies, n-grams, or word embeddings for text input representations [2].

One reason to choose logistic regression for sentiment analysis is that it is simple, fast, and interpretable. Such a method can be regarded as a baseline for comparison with more complex models, like neural network models. It can also efficiently handle large and sparse feature vectors that can be easily regularized to prevent over-fitting issues. Logistic regression can also output the probability of each class, which can be useful for measuring the confidence or uncertainty of the predictions.

Building a logistic regression model includes three steps, i.e., feeding the data into the model, parameter estimation, and predictions. The model can be defined using the equation (1) below. μ is the location parameter, which is the middle point of the curve. s is the scale parameter. Then the model can be rewritten as equation (2). Then the model is trained by fitting the data. The measure of goodness of the fitting is logistic loss. As given $p_k = p(x_k)$, where p_k is the probability that the y_k will be the same as true label, the goal is to find the values of two parameters, i.e., β_0 and β_1.

DOI: 10.54254/2755-2721/17/20230927

$$p(x) = \frac{1}{1+e^{-(x-m)/s}} \tag{1}$$

$$p(x) = \frac{1}{1+e^{-(\beta_0+\beta_1 x)}} \tag{2}$$

The following step is parameter estimation. Parameter estimation in logistic regression is the process of finding the values of the regression coefficients that best fit the observed data and the underlying probability model. The most popular method is maximum likelihood estimation (MLE), which optimizes the chance of witnessing the data provided in the model as shown in equation (3) below. Solving the equation for β_0 and β_1 can find the result.

$$0 = \frac{\delta l}{\delta \beta_0} = \sum_{k=1}^{K}(y_k - p_k) \qquad 0 = \frac{\delta l}{\delta \beta_1} = \sum_{k=1}^{K}(y_k - p_k)\, x_k \tag{3}$$

The last step is to perform predictions. The accuracy score of the trained logistic regression model can be measured.

The second model used in this research is multi-layer perceptron. The way multi-layer perceptron works is similar to human brains' work. It is made up of layers of linked neurons that process information and develop over training. Neural networks can be used for sentiment analysis by training them to recognize and categorize opinions in a piece of text. Figure 3 shows the structure of the multi-layer perceptron model.

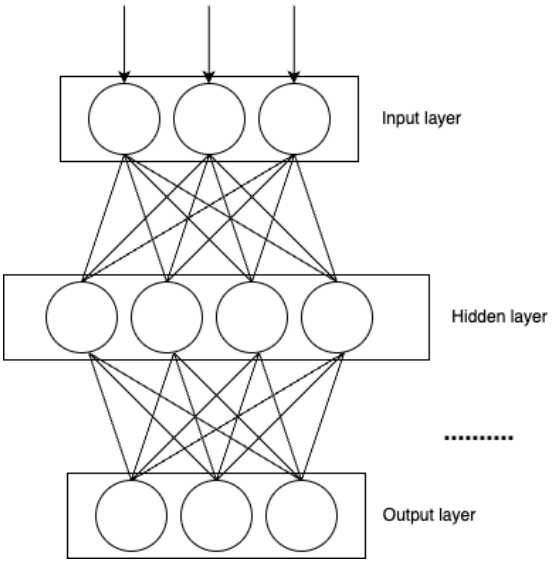

Figure 3. Structure of Multi-layer Perceptron.

Each layer consists of nodes that perform some computation on the data. A multi-layer perceptron input layer obtains input data and then feeds it through the network's hidden layers. Through a network of weighted connections, processing occurs in the hidden layers. The data from the input layer is then combined with a set of coefficients by nodes in the hidden layer, and the inputs are subsequently provided with the proper weights.

To process text, neural networks need to convert discrete sequences of words or characters into numerical representations that can be used as model inputs, which can be achieved by using techniques like one-hot encoding, word or character embeddings. Such representations capture some semantic and syntactic information about the text. Different types of operations can be performed on these representations, such as convolution, pooling, and recurrent connections, to extract features and patterns from the text, which can be then used by the output layer to perform a specific task, such as classification, generation, or translation [3].

DOI: 10.54254/2755-2721/17/20230927

One of the challenges of natural language processing is to represent words in a way that captures their meaning and usage. A common solution is to use word embeddings, which are vector representations of words learned by neural networks. In the following sections, this paper will explain what word embeddings are, how they are learned, and how they can be used for natural language processing tasks.

Word embeddings are a way of representing words as vectors of numbers in a high-dimensional space. They capture the semantic and syntactic information between words based on their contexts. Word embeddings can be learned using neural networks that take as input a large corpus of text and output a vector for each word. Some examples of neural network models that can learn word embeddings are Word2Vec and Glove. Word embeddings are useful for several applications related to NLP, including sentiment analysis, machine translation and text categorization [4].

One example of using word embeddings for text classification is to train a Word2Vec model on a collection of shopping reviews and use the resulting vectors as features for a classifier. For instance, given a review text "This product is amazing, I love it", the Word2Vec model can be utilized to convert each word into a vector and then average them to get a single vector for the whole review. This vector may then be used to predict the review's sentiment (positive or negative) by feeding it into a classifier like a logistic regression or a multi-layer perceptron models. The classifier will learn to relate vectors to related feelings based on the training data. Word embeddings have the benefit of capturing the meaning and usage of words with other words in the corpus, which can increase the generalization and accuracy of the classifier [5].

However, word embeddings alone are not enough to capture the structure and context of natural language. A more advanced technique is to use convolutional neural network (i.e., CNN), which are neural networks that apply convolutional layers to local features. This work will describe CNN in the parts that follow, along with how they work and how they can be used to address problems for natural language processing.

A convolutional neural network is capable of processing structured data, including text or pictures. The structure of the CNN is shown in figure 4 [6]. CNN use convolutional layers, which are composed of filters that slide over the input data and produce output features that highlight important patterns. Convolutional layers perform an action called convolution on the input, which entails multiplying a filter or kernel by a portion of the input and adding the output. The filter can have numerous channels to match the depth of the input and can be any size smaller than the input. A feature map, which shows the positions and strengths of each identified feature in the input, is the convolution's output. CNN can learn these filters automatically during training and extract relevant information from the data. CNs are commonly used for tasks that call for a comprehension of the context and structure of the input, such as image recognition and NLP [7].

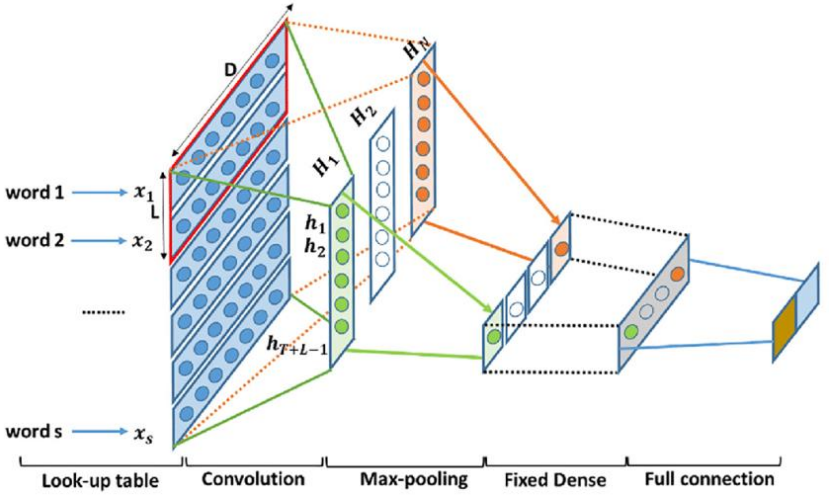

Figure 4. Structure of convolutional neural network.

DOI: 10.54254/2755-2721/17/20230927

3. Result

Based on the model structure and earlier research, this paper presents that the logistic regression model has lower accuracy than the neural network based models. A recent study [8] shows the accuracy of the neural network model could obtain better performance than the logistic regression model, in terms of accuracy, recall, and score. In another paper [9], the authors contrast a few neural network models and logistic regression, which demonstrates that the accuracy of the neural network model is greater than that of the logistic regression model.

The other factor is word embedding. This paper presents the models that work with word embeddings achieve higher accuracy scores than the one that does not. Word embeddings have improved the linking between words and increases the computer's understanding of a whole sentence. The paper [10] proves that word embeddings can have improvement when processing Turkish words from Wikipedia articles.

The results of the actual training prove the hypothesis, as listed in table 1. Logistic regression and multi-layer perceptron models with word embedding achieve accuracy score of around 80%. CNN models result in the highest accuracy of 85.98%. The graph of the result is shown in figure 5.

Table 1. Accuracy scores of models.

Model	Testing Accuracy
Logistic Regression	79.59%
Multi-layer Perceptron	80.81%
Multi-layer Perceptron with pre-training word embeddings	82.24%
Multi-layer Perceptron with word embeddings	79.51%
Convolutional Neural Network	85.98%

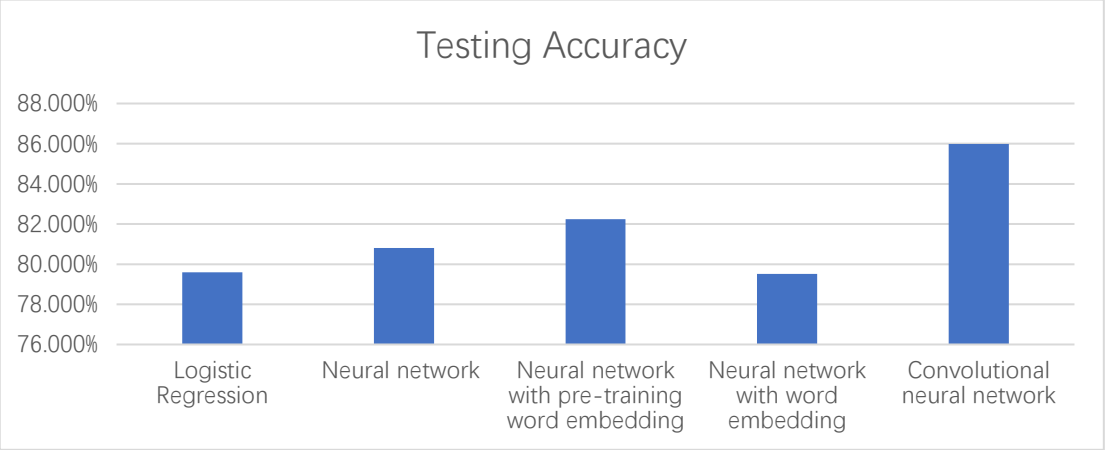

Figure 5. Accuracy of models presented in bar chart.

4. Conclusion

This paper investigates three different models for user rating prediction in e-commerce scenario, demonstrating a reasonable solution to automatically predict the user review for browsed items. Logistic regression, multi-layer perceptron network, and convolutional neural networks are the methods considered. Based on experiments, the CNN is the solution with the highest accuracy of user rating prediction, compared to the other two methods and their variants. In a conclusion, CNN can give promising performance for the task of user rating prediction.

There are several actions that can be done to further improve the performance. An expansion of the training set's size can enhance the model's performance.

The size of training data has an impact on how well the model generalizes and forecasts unseen data. The other factor that can be improved is hyper-parameters. Hyper-parameters are settings that determine

DOI: 10.54254/2755-2721/17/20230927

how the model is trained. Optimal hyper-parameters can significantly improve a model's accuracy and reduce the risk of over-fitting issues.

Long Short-Term Memory (i.e., LSTM) networks are an additional option that can significantly improve model performance. A memory cell can be added to LSTM, a form of recurrent neural network, to capture long-term relationships between words. This allows the model to maintain context and retain important information over longer sequences of data. By incorporating LSTM into the model architecture, the model's ability to understand and process sequential data can be potentially enhanced.

References

[1] Ramadhan W, Novianty S A, and Setianingsih S C 2017 Int. Conf. on Control, Electronics, Renewable Energy and Communications

[2] Goswami M and Sajwan P 2020 Trends in Wireless Communication and Information Security In Proc. of EWCIS 2020 p 165-74

[3] Jacovi A, Shalom O S and Goldberg Y 2018 Understanding convolutional neural networks for text classification Arxiv Preprint arXiv:1809.08037

[4] Nedjah N, Santos I and de Macedo Mourelle L 2019 Sentiment analysis using convolutional neural network via word embeddings Evolutionary Intelligence p 1-25

[5] Tang D, Wei F, Qin B, Yang N. Liu T and Zhou M 2015 Sentiment embeddings with applications to sentiment analysis IEEE transactions on knowledge and data Engineering vol 28.2 p 496-509

[6] Nguyen V Q, Anh T N and Yang H J 2019 Real-time event detection using recurrent neural network in social sensors International Journal of Distributed Sensor Networks vol 15.6 p 1550147719856492

[7] Severyn A and Alessandro M 2015 Twitter sentiment analysis with deep convolutional neural networks In Proceedings of ACM SIGIR p 959-62

[8] Smitha N and Bharath R 2020 Performance comparison of machine learning classifiers for fake news detection International Conference on Inventive Research in Computing Applications p 696-700

[9] Saif M A, Medvedev A N, Medvedev M A and Atanasova T 2018 Classification of online toxic comments using the logistic regression and neural networks models In AIP Conference Proceedings vol. 2048.1 p 060011

[10] Maroco J, Silva D, Rodrigues A, Guerreiro M, Santana I and de Mendonça A 2011 Data mining methods in the prediction of Dementia: A real-data comparison of the accuracy, sensitivity and specificity of linear discriminant analysis, logistic regression, neural networks, support vector machines, classification trees and random forests BMC Research Notes vol 4.1 p 1-14