Georgetown University

The Proposal of Product Review Sentiment Analyzer based on amazon

amazon

Zijing Cheng zc233

Yanfeng Zhang yz1045

Xin Xiang xx123

1 Project Introduction

1.1 Abstract

In this Internet age, online shopping has already become the main way of shopping for people. People trade products on major business platforms, and product ratings and reviews have become important factors for users to purchase products. We also found that there are merchants who write false positive reviews for products while peer merchants who post negative reviews on the platform to intentionally reduce the rating of the product. If the quality of a product is judged by 1-5 stars alone like now, the ratings containing malicious competition and false comments are not reliable and it is likely to damage the interests of users and affect the long-term competitiveness of the product. To address this problem, we turn our attention to text reviews that are informative. We hope to directly perform sentiment analysis on the text to grab information from the reviews, combined with the scoring situation, so as to filter out behavior characteristics of posting positive and negative reviews, so as to provide users with more real and effective information about whether they can purchase products and also help sellers to have a more comprehensive understanding of their own products to improve their competitiveness.

This project takes the review data of the headphones category on Amazon.com as an example, and crawls the reviews of the headphones category. The crawled text data is processed for data pre-processing (including Cleaning, Tokenizing, Normalizing and Vectorizing). Then we apply various machine learning and deep learning methods (including SVM, SGD, RF, NB, LR, RNN, LSTM, BiLSTM and CNN) to classify the text into three categories: positive, neutral, and negative. Finally, the best classifier is selected by the evaluation methods of Accuracy, Recall, Negative Recall, Precision, and F1-score and predict on testing set with the best classifier.

The selection of the best classifier can well explain the problem we want to solve, which has the following meanings:

1) For users, after filtering false information, users can get relatively real comments from several users who have purchased products, so as to judge their own purchasing needs. It can also reduce the negative impact of not seeing the real thing because of online shopping.

2) For merchants, we can find important factors that affect user purchases and give positive or negative reviews by analyzing reviews. And in a practical sense, it can truthfully report problems such as slow express delivery and poor product quality control to the merchants that are not doing well. Improving the competitiveness of products is conducive to the better development of products.

In general, sentiment analysis of product reviews can promote both users and merchants in both directions.

1.2 Background Introduction

1.2.1 Amazon Product Reviews

Amazon is one of the largest e-commerce companies in the world. Since its creation as an online platform in 1994, Amazon has grown rapidly. With this vast user base and huge product collection, Amazon has become a microcosm for user-supplied reviews. There is tremendous interest in sentiment analysis of these Amazon product reviews across various domains such as commerce, health, and social behavior study.

Amazon allows its users to rate products from 1 to 5 stars (1 being the lowest review, five being the best) and provides a written summary of their experience and opinions about the product as well as the seller. This scoring system is universal, regardless of product category. With no guidance on how Amazon Web users should use this scoring system, Amazon product reviews are very personal and subjective. A person can rate a good product 1 star, but have a bad buying experience, such as high price, or late delivery, and vice versa. This lack of guidelines makes it challenging to determine user sentiment about different aspects of a product, and different parts of the shopping experience, but it also makes Amazon product reviews a very rich source of data on how people perceive products and services.

1.2.2 Sentiment Analysis

An emotion is an attitude, thought, or judgment triggered by a feeling. Sentiment analysis, also known as opinion mining, studies how people feel about certain entities. The internet is a resourceful place when it comes to emotional information. From the user's point of view, people can post their own content or comment on a certain content through online social networking sites. From a researcher's perspective, many online sites publish their application programming interfaces (APIs), enabling researchers and developers to collect and analyze data. Thus, sentiment analysis seems to have a strong foundation, backed by massive amounts of online data.

And the most popular way to gain insight into product reviews is to perform sentiment analysis to determine whether a review is positive or negative, or neutral. Furthermore, the sheer size of the user-generated content library and its continued rapid growth make it very labor-intensive to manually monitor and extract sentiment in user-generated content. Automatic classification of textual content becomes the only practical method for effective data classification and insight.

2 Data

2.1 Data Introduction

Data used in this project is a set of product reviews collected from amazon.com using selenium. To maintain the diversity and authenticity of the dataset, we plan to collect a total of over 10,000 latest product reviews. consists of reviews of multiple different products belonging to the same electronic category of Audio and Accessories.

Each review includes the following information: 1) reviewer ID; 2) rating; 3) date; 4) verification of purchase; 5) helpfulness; 6) time of the review; 7) review text; 8) review title. Every rating is based on a 5-star scale, resulting in all the ratings to be ranged from 1-star to 5-star with no existence of a half-star or a quarter-star.

2.2 Data Collection

The data is collected from web scraping using selenium, which refers to a number of different open-source projects used for browser automation. Selenium is a powerful tool for controlling web browsers through programs and performing browser automation, as it can work like a real user starting from opening a browser, typing a keyword in a search box, and then clicking to find the result.

2.3 Data Preprocessing

The metadata will be saved in the JSON files, with 7 features in text format, and there will be steps to get the clean and format data.

2.3.1 Missing values

Since we are using sentiment as our target (the variable we will be predicting, also the rating of the products), we cannot have any null values, so these are dropped.

2.3.2 Clean the data

In this part, we will process data into clean text format by removing punctuation, removing emojis.

2.3.3 Tokenization

Tokenization is the process of converting text into tokens before transforming it into vectors. It is also easier to filter out unnecessary tokens. In this case, we are tokenizing the reviews into words.

After we get tokenizing all reviews into words, we need to drop the stop words among tokens, as stop words are the most commonly occurring words that are not relevant in the context of the data and do not contribute any deeper meaning to the phrase.

2.3.4 Normalization

Words that look different due to casing or written another way but are the same in meaning need to be processed correctly. Normalization processes ensure that these words are treated equally. So we need to convert all text into lowercase, handle negation (turn words like "can't" into "can not"), and do lemmatization that finds the base or dictionary form of the word known as the lemma.

2.3.5 Vectorization

After getting the clean and tidy format of tokens, each of the words can represent a feature for the whole review and these features are in the form of a string. So we need to convert these string features into numerical data by vectorization. And here we can use Bag of Words, TF-IDF, or Word2Vec.

3 Methodology

After vectorization, we can build a feature vector that can be used for testing the models. We will split our data into training and testing sets and use classification models that have already been trained by the training set and then we compare the models based on accuracy, f1-score, recall and precision to select the best model and predict the testing set with the best model. We will divide this part into two parts, machine learning methods and deep learning methods. In machine learning part, we choose SVM, SGD, RF, NB and LR while in deep learning part, we choose SimpleRNN, LSTM, BidirectionalLSTM and Conv1D.

3.1 Sentiment Labeling

As the dataset used for the sentiment analysis has a feature, "rating", which is an integer range from 1 to 5, and we label the product reviews with the rating of 4 or 5 as positive, 3 as neutral and 1 or 2 as negative.

3.2 Support Vector Machine (SVM)

In Support Vector Machine, every data point is represented as a point in n-dimensional space, there are n number of features, with its coordinate used for SVM algorithm. By training labeled training set for each category, usually used for two categories, SVM models can categorize new texts. The main reason we choose them is because they are faster and perform better with fewer data over modern methods. Thus, they are well suited to text classification problems, as there are just a few thousand tagged available examples in text classification problem. As the problem we want to solve is multi-class problem, we will compare SVC, nuSVC and linearSVC in this part with sklearn module.

3.3 Stochastic Gradient Descent (SGD)

Stochastic Gradient Descent is an iterative machine learning optimization method and it can help us to find the optimal model parameters that match expected and actual outcomes. We choose it as it is efficient in high-dimensional optimization problems as it can reduce high computational burden and implement faster iterations.

3.4 Random forest (RF)

Random forest is a supervised machine learning algorithm and it is used widely in classification problems. The logic behind random forest is that it builds decision trees on different samples and takes the majority vote for classification. We choose this method because of its diversity and Immunity to the curse of dimensionality. First, Diversity, as not all attributes will be considered when building a individual tree, different tree can help us to get Relatively fair results. Second, by using partial features in a individual tree can effectively reduce dimension as textual analysis always having high dimensions.

3.5 Naive Bayes (NB)

Naive Bayes is a supervised machine learning method used for classification, which is based on the conditional probability theorem. The basic assumption of Naive Bayes in textual analysis is that in BOW, the presence of a particular word in a class is unrelated to the presence of any other word. In other words, this model predicts the probability of each word in a text sentence and considers it a feature of any one of the dataset classes. We choose it since it is a simple model but it is useful in text classification. And in this case, the equation is expressed as a conditional probability for the NB algorithm as follows:

$$P(A \mid B) = \frac{P(B \mid A)(A)}{P(B)}$$

where A is the class label, positive, negative or neutral and B is the piece of text. As we have multiple classes, we will compare MultinomialNB in sklearn module and our custom NB classifier in lab03.

3.6 Logistic Regression (LR)

Logistic Regression is a supervised machine learning algorithm that can be used to model the probability of a class. It is used when the data is linearly separable and used for binary classification problems. However, with our multi-class output, we should use multinomial logistic regression, which is an extension of logistic regression. For multinomial logistic regression, a popular approach is to split this multi-class classification problem into multiple binary classification problems and fit logistic regression models on sub-problems. In this part, we will use LogisticRegression with class equal to multinomial in sklearn Module adjusted by different punishment methods.

3.7 RNN

A recurrent neural network (RNN) is a special type of artificial neural network (ANN) adapted to work for time series data or data that involves sequences. And in this case, we will choose to use simple RNN in keras module.

3.8 LSTM

Long Short-Term Memory (LSTM) model is a prominent RNN architecture that was developed to deal with the issue of long-term dependence and solve the vanishing gradient problem. Because the RNN model may not forecast the present state well when the previous state influencing the current prediction is not recent, we introduce LSTM to compare with simple RNN model. In LSTM, there are three gates in the deep levels of the neural network, that is, an input gate, an output gate and a forget gate. These gates control the flow of data needed to forecast the network's output. In this case, we choose to use LSTM in keras module.

3.9 BiLSTM

A Bidirectional LSTM is a sequence processing model that comprises two LSTMs. The one is forwarding the input while the other one is reversing it. BiLSTM can effectively improve the amount of data available to the network. In this case, we choose to use Bidirectional and LSTM in keras module.

3.10 CNN

A CNN is an effective model in detecting simple patterns in data. We choose to use 1D CNN model as we want to extract valuable features from small chunks and we recognize that the location of the feature inside the segment is not important in this case as we choose it to train our models. A CNN comprises three layers, that is, input, output, and hidden layer. The middle layers act as a feedforward neural network. The hidden layers also include convolutional layers. The dot product of the convolution kernel with the input matrix of the layer is performed here. ReLU and the Frobenius inner product act as the activation functions. A feature map is generated by the convolution operation as the convolution kernel slides along the input matrix for the layer, later contributing to the input of the following layer.

4 Evaluation

In order to find the most effective model to describe the sentiment of the reviews, here are various evaluation metrics that will be used in the evaluation part, including accuracy, recall, negative recall, precision, and F1-score.

4.1 Accuracy

The Accuracy measure gives how many data values are correctly predicted. It is computed by the following expression.

$$Accuracy = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}}$$

4.2 Recall (Sensitivity)

The Recall (or Sensitivity) computes how many test case samples are predicted correctly among all the positive classes. It is computed as follows:

$$Recall = \frac{\text{Number of True Positives}}{\text{Number of True Positives} + \text{Number of False Negatives}}$$

4.3 Negative Recall (Specificity)

The Negative Recall (or Specificity) computes how many test case samples are predicted correctly among all the negative classes. It is computed by the following expression.

$$Negative\ Recall = \frac{\text{Number of True Negatives}}{\text{Number of True Negatives} + \text{Number of False Positives}}$$

4.4 Precision

The precision measure computes the number of actually positive samples among all the predicted positive class samples as follows.

$$Precision = \frac{\text{Number of True Positives}}{\text{Number of True Positives} + \text{Number of False Positives}}$$

4.5 F1-score

The F1-score is the harmonic mean of Precision and Recall. It is also known as the Sorensen-Dice Coefficient or Dise Similarity Coefficient. The perfect value is 1. F1-score is computed as shown in the following.

$$F1 - score = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

5 Computational Considerations

The data set used in this project is the review data of a certain type of product crawled from amazon, and the amount of data is not large. Therefore, there is no special requirement for the computing power of the device, and it can run smoothly on the device.

6 Biggest Unknowns Dictate the Project

The first situation is that we cannot completely avoid merchants manually reviewing positive and negative reviews, because generally positive reviews and negative reviews are machine reviews, which can have a certain similarity in text information. However, if the merchant asks users who have purchased this product to give a good review, we cannot predict this situation. Since it also appears to be a real and valid review, this will have a certain negative impact on our classifier, and it is also an error from the real situation.

The second situation is not all the comments are in English, there may be other languages. When processing data, it is very likely that we cannot completely convert other languages into English for classification. Thus, this will also have a certain negative impact on our classifier, or an unavoidable error.

7 Final Presentation

The final prestation of this project will be presentation video with a slide deck.

References:

[1] Rodrigues, A. P., Fernandes, R., A, A., B, A., Shetty, A., K, A., Lakshmanna, K., & Shafi, R. M. (2022). Real-time Twitter spam detection and sentiment analysis using machine learning and Deep Learning Techniques. *Computational Intelligence and Neuroscience*, *2022*, 1–14. https://doi.org/10.1155/2022/5211949

[2] Nagi Alsubari, S., N. Deshmukh, S., Abdullah Alqarni, A., Alsharif, N., H. H. Aldhyani, T., Waselallah Alsaade, F., & I. Khalaf, O. (2021). Data Analytics for the identification of fake reviews using supervised learning. *Computers, Materials & Continua*, *70*(2), 3189–3204. https://doi.org/10.32604/cmc.2022.019625

[3] Fang, X., Zhan, J. Sentiment analysis using product review data. *Journal of Big Data* **2**, 5 (2015). https://doi.org/10.1186/s40537-015-0015-2

[4] Nguyen, Heidi; Veluchamy, Aravind; Diop, Mamadou; and Iqbal, Rashed (2018) "Comparative Study of Sentiment Analysis with Product Reviews Using Machine Learning and Lexicon-Based Approaches," SMU Data Science Review: Vol. 1: No. 4, Article 7. https://scholar.smu.edu/datasciencereview/vol1/iss4/7