1.Introduction

Our project's name is Artificial Intelligence Based Amazon Consumer Sentiment and Behavior Analysis. The goal is to use artificial intelligence to analyze the sentiment of Amazon Consumers by the comment they write. The project has very high practical value since the sellers can use the result to better understand their product.

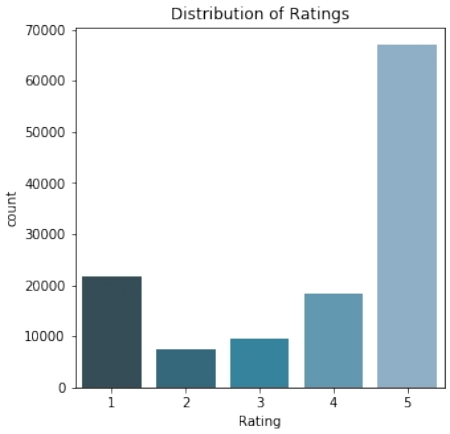
The method for this project is machine learning, and it is the most important type of Artificial Intelligence. The source of machine learning is data from the big data collected on internet. There are two types of machine: learning-supervised study and non-supervised study. The former means the data is labelled by human and the latter means the data is not labelled by human. In this project, we use supervised study because the rate system on amazon naturally provides a perfect label. Our research area is NLP, which mean natural language processing. This is the most significant problem in artificial language. The main goal is to let computer understand or interpret human language.

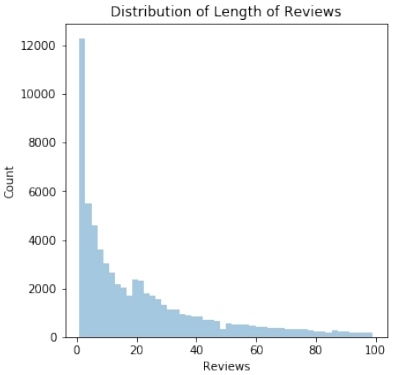
2.Dataset

We use data from reviews of unlocked mobile phones sold on Amazon.com. These data are acquired by the crawlers in December 2016. The data is stored in a csv file including 124151 different reviews. Each review is constituted by the product name, brand name, price, rating, reviews and review notes. The reviews are mainly on electronic devices.

3.Data Visualization and Exploratory Data Analysis

First, we do some exploratory data analysis to learn about the data. We use packages such as seaborn, matplotlib and pandas. Data visualization is important because we can use it to explore our data. Though the data itself is abstract, we can easily interpret it by drawing graph to show the certain information we want from the data. By doing data visualization, we can learn our data well excavate the potential of our data. Here are two examples of data visualization:





4.Problem Formulation

There are two kinds of problems in machine learning, regression and classification. Regression problem wants an answer with an exact value. In opposite, our problem is a binary classification problem, which means than the answer is either 0 or 1. Here 0 represents the customer dislikes the goods while 1 represents the customer likes the goods. Our goal is to predict the 0 or 1 from the text of customer's review as input data.

5. Feature Engineering

5.1 Bag-of-words

In computer program, we can't process the words in each sentence directly. Instead, we changed the words into vector by using bow (bag of word). Each vector has length of all unique words and each element in the vector represents one unique word. We first pick all unique words from the reviews. Then, we score each review into a vector by setting the correspondent position of unique words 0 or 1. Our scoring method is binary scoring. There are also other methods such as TF-IDF or count scoring.

5.2 Bag-of-Ngrams

The basic bag of word method has a problem, it omits the function of words order in the sentence. To solve this, we devise bag of n-grams. It is like bag of words, but the only difference is that each element in the vector is represented by n continuous words in reviews. Therefore, the elements can include the order of words. There are still problems with bag of n-grams. This method generates much more elements than bag of words, and some of these are meaningless. Therefore, it is important to choose bag of words and bag of n-grams wisely.

6. Algorithm

6.1 Logistics Regression

Logistics regression is a linear model for prediction. Function h is the whole process, it

represents the probability the result is 1, which means that the review is positive. There are various choices for Function g. We adopt sigmoid function in the experiment. it ranges from 0 to 1. When the result of function g, or function h is bigger than 0.5, we predict that the review is positive. When the value is smaller than 0.5, we predict that the review is negative. The input of sigmoid function is x times transpose of θ. X is the vector composed by the feature of each review. Θ is parameter we are going to learn in gradient descent.

$$h_\theta(x) = P(y = 1|x) = g(\theta^T x)$$

6.2 Gradient Descent

To get the value of θ, we use gradient descent method. Since we want to make the prediction as real as possible, we use a cost function J(θ), to represent the difference between the predicted value and true value. Then we keep subtracting a portion of the value of cost function from θ until the cost function is minimized. The portion is hyperparameter, which is set by human. To prevent underfitting and overfitting, we use regularization, which means that we add θ into the cost function to ensure that θ is not too big or too small.

7. Experiment Design and Results

7.1 Experimental Design

Dataset is divided into training, validation and test dataset. The three data sets have different functions Models are trained and tuned on training data set. Hyper-parameters are selected using validation dataset. Then, we use tuned models to predict labels on the test dataset. The division of dataset ensures that the model can predict from the new data instead of remembering the data it has seen before.
We used two methods of feature engineering. One is bag-of-words and another is bag-of-ngrams. Experiment is set up for comparing those different feature engineering methods
The metric we are using is F1-Score. F1-Score is the harmonic mean of recall and precision. The values of F1-Score is between 0 and 1. The higher the F1-Score is, the better the model predicts.

7.2 Result and Analysis

Feature Engineering Method	Training F1 Score	Testing F1 Score
Bag of Words	0.927	0.919
Bag of Ngrams	0.932	0.921

This table shows the result of experiment. The experiment is successful because F1 score is high. The F1 Score of Bag of Ngrams is higher that of the Bag of Words, which means that Bag of Ngrams is better than the Bag of Words is this experiment since the former considers the order of word.

8. Conclusion

In the project Artificial Intelligence Based Amazon Consumer Sentiment and Behavior Analysis, we create a machine learning model and train it by using the data from Amazon. Our result is a model that can predict the consumer sentiment based on the reviews