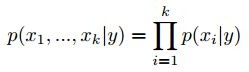
SENTIMENT ANALYSIS FOR AMAZON REVIEWS

Analysis

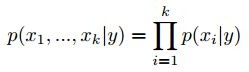
* The dataset is unbalanced. Based on the result, the model may not have a good generalization of these data. That’s why even the highest accuracy is around 70%.
* The increase of the dictionary's length did not have too much effect on the accuracy. Because the length of dictionary has only increased by 720 when we decrease the times that the word appears in one sentence, which is small compared to the original length.
* The result using glove mean is worse than the method of normal word count. The possible reason is that if we use the average, the individual word feature will be weakened, then the distance between different reviews will be inaccurate.





Models

* Naive Bayes



This algorithm assumes that xi is are conditionally independent given y.

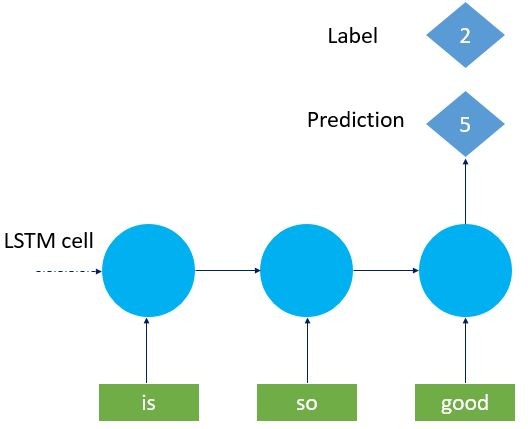
* SVM

Geometrically given two types of points, circles and xi, in a space, it tries to maximize the minimum distance from one of the points to the other. Here, we used both linear kernel and radial kernel.

* KNN

This algorithm looks for the K = n nearest neighbours of the input. Then, it will assign the majority of that n neighbours’ class. We tuned the n and compared the results.

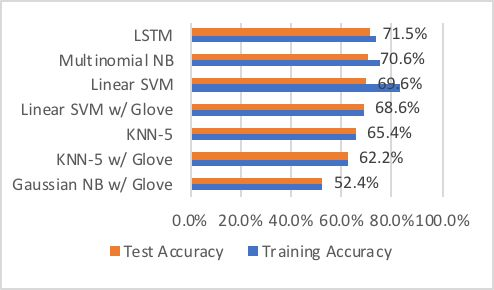
* LSTM



A common LSTM unit is composed of a cell, an input gate, an output gate and a forget gate.

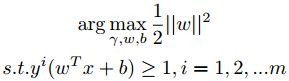
Introduction

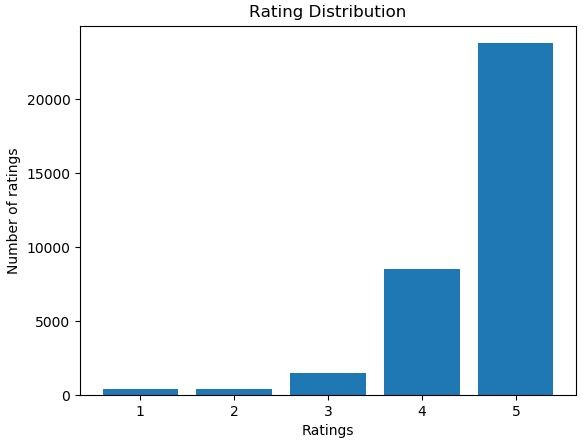
* Machine learning is sub domain of working in computer science. Machine Learning and artificial intelligence have boosted the technological aspects of computer science.
* Machine learning is Core domain for Artificial intelligence and deep learning. It is the most important field in todays advanced environment and foreseen future in technologies. It is a method to analysis data and understand it. Machine learning also includes predicting and understanding data and forecasting information and relevant problem statement. It Involves computers discovering how a data can perform and extract necessary information with the help of pervious datasets.



Results

* The entire dataset of 34,627 reviews was divided into a training set of size 21000 (70%), a validation set of size 6814 (30%) and a test set of size 6813 (30%).
* For the first way of representing review text, we implemented Multinomial Naive Bayes, SVM with Linear Kernel, SVM with RBF Kernel, KNN-4, 5,& 6 and LSTM with 4223-d features. LSTM performs best in term of test accuracy among them.
* For the second way using glove dictionary, we run Gaussian Naive Bayes, SVM with Linear Kernel and KNN-4, 5 & 6 with 50-d features. It turned out that SVM with Linear Kernel generated best predictions
  + Training accuracy and test accuracy of each model are shown in the table to the left.
  + Models are sorted by test accuracy in the chart below.





Dataset

* The dataset used is a sample of from nasdaqa and xignite illustrates all the details of prediction stock prices for large-cap technology companies according to the features of all the stock prices for large-cap technology companies.

•

The dataset has the information about 2000 mobile details and their price range. Each Price range is take under consideration as 10k, 20k, 30k, 40k respect for the numbers 0,1,2,3.

|  |  |  |
| --- | --- | --- |
|  |  |  |
|  | 75.1% | 70.6% |
|  | 83.4% | 69.6% |
|  | 69.7% | 69.2% |
|  | 61.7% | 61.7% |
|  | 65.5% | 65.4% |
|  | 64.9% | 64.6% |
|  | 73.5% | 71.5% |
|  | 52.2% | 52.4% |
|  | 68.7% | 68.6% |
|  | 58.1% | 57.6% |
|  | 62.6% | 62.2% |
|  | 61.3% | 61.6% |

Future Work

If we have more time, we want add Chatbot in our system so that if any one has any doubt they can directly chat with bot. We also want to go deeper in the LSTM neural network in which case we might get better accuracy.

Features

* The system is use to predict future stock prices.
* Its being predicting long term stock prices for large cap companies .