

RATINGS PREDICTION

Submitted by:

Lokesh Bisen

**ACKNOWLEDGMENT**

n machine learning, classification is used to classify a new observation into a specific set/category based on a training set of data containing observations whose category is known in advance. The most common example is “spam” or “non-spam” classes for emails. In E-Commerce, classifier algorithms can be used to classify sentiments of review based on words. The specific words in the language are categorized in advance for their positive or negative sentiments.

Classification is an instance of supervised learning. Training set has correctly identified observations. Classifier algorithms are used to create cluster/sets from the uncategorized unsupervised data based on similarity and/or distance from the training data set.

**INTRODUCTION**

* Business Problem Framing

Nowadays, a massive amount of reviews is available online. Besides offering a valuable source of information, these informational contents generated by users, also called User Generated Contents (UGC) strongly impact the purchase decision of customers. As a matter of fact, a recent survey (Hinckley, 2015) revealed that 67.7% of consumers are effectively influenced by online reviews when making their purchase decisions. More precisely, 54.7% recognized that these reviews were either fairly, very or absolutely important in their purchase decision making.Relying on online reviews has thus become a second nature for consumers.

In their research process, consumers want to find useful information as quickly as possible. However, searching and comparing text reviews can be frustrating for users as they feel submerged with information (Ganu, Elhada & Marian, 2009). Indeed, the massive amount of text reviews as well as its unstructured text format prevent the user from choosing a product with ease. The star-rating, i.e. stars from 1 to 5 on Amazon, rather than its text content gives a quick overview of the product quality. This numerical information is the number one factor used in an early phase by consumers to compare products before making their purchase decision.

* Conceptual Background of the Domain Problem

However, many product reviews (from other platforms than Amazon) are not accompanied by a scale rating system, consisting only of a textual evaluation. In this case, it becomes daunting and time-consuming to compare different products in order to eventually make a choice between them. Therefore, models able to predict the user rating from the text review are critically important (Baccianella, Esuli & Sebastiani, 2009). Getting an overall sense of a textual review could in turn improve consumer experience.

Nevertheless, this predictive task presents some challenges. Firstly, because reviews are human feedbacks, it may be difficult to accurately predict the rating from the text content. Indeed, users all have different standards and do not rate a product the same way. For instance a user may rate a product as good and assign a 5-star score while another user may write the same comment and give only 3 stars. In addition, reviews may contain anecdotal information, which do not provide any helpful information and complicates the predictive task. Finally the vocabulary used can also be very specific according to the product category.

**Analytical Problem Framing**

* Mathematical/ Analytical Modeling of the Problem

Our data set comes from Consumer Reviews of Amazon Products1 . This data set has 21367 data points in total. Each example includes the type, Number of ratings and Full Review of the product. To better utilize the data, first we extract the rating and review column since these two are the essential part of this project. Then, we found that there are some data points which has no ratings when we went through the data.Besides, to have a brief overview of the data set, we have plot the distribution of the ratings. In Figure 2, it shows that we have 5 classes - rating 1 to 5 as well as the distribution among them. Also, these five classes are actually imbalanced as class 1 and class 2 have small amount of data while class 5 has more reviews.

* Data Preprocessing Done

customers. Research has considered traditional machine learning algorithms along with Naive Bayes analysis, SVM, Knearest neighbor mechanism. Research has also considered deep neural networks along with Recurrent Neural Network (RNN). Research is supposed to provide better solution for sentiment analysis.

Data is collected from e-commerce websites through web scarping tools like beautiful soap and selenium. This assessment is done on the products which are provided by Amazon and Flipkart. Near about twenty thousand informational points are enclosed by such type of data . Each example includes the reviews and given rating of the product. For the utilization of data in a refine way, two, most important column related to this project are extracted by us initially. These two columns are rating and assessment. After that, when the data is checked by us, we noticed those information points which remain unrated. When we eliminate such sample, we generally left with thirty four thousand and six hundred twenty seven information points. On the other hand , for outlining the data, allocation of ratings has been plotted by us. There have been 5 classes. Rating from 1 to 5 has been distributed in the middle of these classes. In reality, such type of classes are uneven. The basic reason behind this unevenness is the availability of less number of information in class one as well as in two. At the same time, twenty thousand reviews is possessed by class five . During research review text has been converted into an input vector.

* Hardware and Software Requirements and Tools Used

In order to build a good model it is necessary to have good computational power hardware and software’s. we have used python and its respective libraries. Our data was huge hence we I have used google colab pro as it offers good computational power.

**Model/s Development and Evaluation**

* Identification of possible problem-solving approaches (methods)

After exploratory data analysis and preprocessing the main steps comes up that is model building and evaluation is the technique where we make assumptions that which model giving the best accuracy. We have performed some punctuation and stopwords removal. And we have used Lemmatization also, Lemmatization is the process of grouping together the different inflected forms of a word so they can be analyzed as a single item.

**CONCLUSION**

To conclude, In this project we have performed various techniques like Data analysis which has proven way for any organizations and enterprises to gain the information they need to make better decisions, serve their customers, and increase productivity and revenue. The benefits of data analysis are almost too numerous to count, and some of the most rewarding benefits include getting the right information for business, getting more value out of it. And it creates more effective marketing campaigns, gaining a better understanding of customers, and so on. In addition, Stemming is the process of converting inflected/derived words to their word stem or the root form. Basically, a large number of similar origin words are converted to the same word. E.g. words like “stems”, “stemmer”, “stemming”, “stemmed” are based on “stem”. This helps in achieving the training process with a better accuracy.

Lemmatising is the process of grouping together the inflected forms of a word so they can be analyzed as a single item. This is quite similar to stemming in its working but not exactly same. Lemmatising depends on correctly identifying the intended part of speech and meaning of a word in a sentence, as well as within the larger context surrounding that sentence, such as neighboring sentences or even an entire document. I used the word-net library in nltk for this purpose. Stemmer and Lemmatizer were imported from nltk.

Tfidf Vectorizer is used for converting a string of words into a matrix of words. Column headers have the words themselves and the cell values signify the frequency of occurrence of the word.Finally I reached the core part of the project, where I could start building the classifier. I had use 4 major algorithm . Where, our random forest model giving the good accuracy hence selected it as a final predictor.