Imperial Journal of Interdisciplinary Research (IJIR)
Vol-3, Issue-4, 2017
ISSN: 2454-1362, http://www.onlinejournal.in

User Review Sentiment Classification and Aspect Extraction

J.Vijayabhaskar[1], R.Sridhar[2] & P.Vijayaragavan[3]
[1,2] B.E - Student /CSE, Dhanalakshmi College of Engineering, Chennai.(India)
[3] Assistant Professor, Dept of Computer Science & engineering, Dhanalakshmi College of Engineering, Chennai.(India)

Abstract*: Online User reviews is a great platform for collecting large volume of information for sentiment analysis. Here we propose an system which classifies the user reviews into two main categories: Positive and Negative reviews. The Classification algorithm uses only the overall review scores to understand sentiment behind each review and extract the important aspects about the product. We developed an efficient classifier model to classify the provided review is either a positive review or negative review by analysing the performance of various classification algorithm on the review data corpus. Clustering techniques are then used to identify key sentiment characteristics to provide them to the users, which helps the user to understand the aspects of the products/service they wish to buy or experience. .*

1. Introduction

User reviews on ecommerce websites such as Amazon or Flipkart provide a great platform for user to provide information about the specifications and experience about the products. In many cases when a new customer tries to buy a product from an ecommerce website, they first read the description about the product and if they are interested in the product they then directly read through the user reviews of that particular product, Because more than the description, the experience and the satisfaction of the other users matters a lot. Even any single product in the ecommerce website could have large amount of reviews. For instance, the Oneplus One available on Amazon India has over 18,500 product reviews as of this writing[1].

A new user who wish to buy the product now has to read through a lot of reviews, So the user spends lots of his valuable time by reading reviews of the product itself We automate this process by analyzing all the reviews provided for the particular product we extract all the important aspects of the product that are being discussed in the review segment and a summary of the all the import aspects of the product or the service is generated and provided to the customer.

In this project we propose an project to extract the important aspects of the product, aggregate them and provide them to use customer which saves a lots of time to the customer.

In Section 2 we discuss about the background inspirations and work behind this project. In Section 3 and 4 we discuss about the dataset which we used to train the model and the various analysis process we did on that dataset to extract as much knowledge from the available resource.

2. Background

Liu and Zhang[3] describes about the five important elements of sentiment analysis: entity extraction and grouping, opinion holder and time extraction, aspect sentiment classification, aspect extraction and grouping, and opinion quintuple generation. In this paper we will focus on sentiment classification and aspect extraction to summarize the reviews. Pang et al.[4] used the bag-of-words model to predict positive and negative labels for movie reviews. They find little difference between Naive Bayes and SVM classifiers. Based on this work of theirs our experiments begin with bag-of-words model and Gaussian Naive Bayes classifier.

We experiment with several models for our classifier and choose our best model using precision and recall metrics, We also use cross validation technique to test the performance(accuracy) of the model on the testing data.

3. Dataset

In this work we experiment with the Amazon Fine foods data set which is available freely on kaggle[2] and also the raw data can be mined from the Amazon websites itself. The Amazon fine food data set have a huge variety of products and about 500,000 user reviews. The Review dataset contains the review text, product rating provided unique users along with their unique user id and the unique product Amazon Standard Identification number.

The dataset is checked for missing values, if exist they are removed to provide meaningful data. We split the data set into two data sets as training and testing data sets each of 80% and 20% of the whole dataset. The classifier model and the

clustering model are developed upon the training data set to avoid over fitting of the data the k-fold cross validation is used to measure the performance of the model on the testing dataset.

4. Data Analysis

First in order to experiment the classifier and clustering model we should understand the structure of the data we have. We have a large amount of data which contains reviews with 1,2,3,4 and 5 star reviews of individual products. In order to understand the sentiment behind the reviews we need to train the model, to do so we need to remove all the 3 star reviews which often contains mixed reviews and words which can't be distinguished as a positive or a negative words.[6]

Before removing the 3 star reviews we visualize the distribution of the stars among the reviews in the dataset.

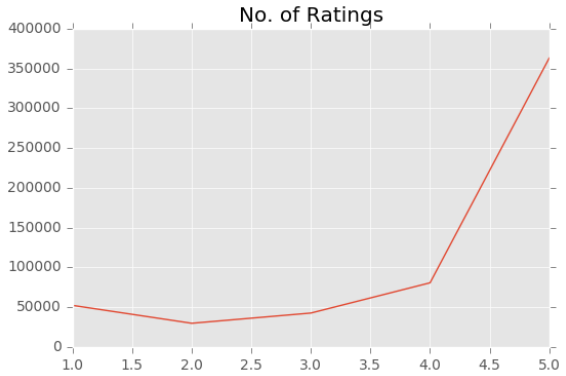


Figure 1:Distribution of ratings among the reviews

5. Methods

First our goal is to write an efficient classifier model to classify the reviews but to classify the reviews we need a labeled dataset. In a review all we can get is a Rating which is given along with each reviews. To make a labeled dataset we tell the moel to skip the 3 star reviews which contain lot of mixed reviews which in turn may confuse the model, so we ignore it and label all the 4 and 5 reviews as positive reviews and all the 1 and 2 star reviews as negative reviews. Even with some amount of mixed reviews we can expect a decent classifier as we have a large dataset.

Second, We build two clusters for good reviews and bad reviews. All the reviews are tested for the distance between centroid of their respective clusters and their position to extract the aspects.

The Tfidfvectorizer splits reviews into sparse matrix, since the machine learning algorithms cant process textual data without being converted to numbers. The Classifier then assigns a *positive*, or *negative* label to each sentence.[5]

The Clustering model groups together common words of each labeled class. Finally, Aspect Extraction is done on the clusters of data and outputs a representative sentence for each selected clusters.

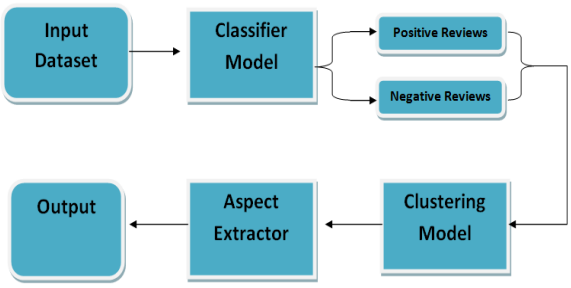


Figure 2: Block diagram of the entire process

5.1 Classification

In this work we experiment with four major classification algorithms. They are Linear SVM classifier, Stochastic Gradient descent Classifier, Naive Bayes Classifier and Logistic regression. Out of which three algorithms provided promising results on testing the performance using k-fold cross validation on the testing data. They are Linear SVM classifier, Stochastic Gradient descent Classifier and Logistic regression.

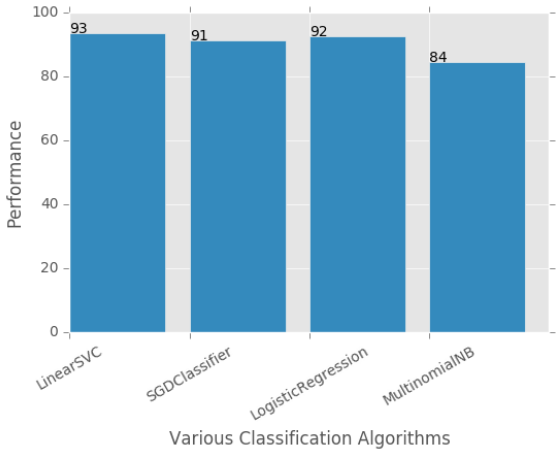


Figure 3: Comparison of Classifier models

We now analyze the top three performing algorithms by plotting the confusion matrix. Before plotting the Confusion matrix the values must be normalized to be descriptive of the data. After the normalization, the values in the matrix are expressed in the range of 0 to 1.

The confusion matrix can be used to visualize the amount of data predicted correctly. A prediction is said to be correct if the positive value is predicted as positive and the negative value is predicted as negative value.

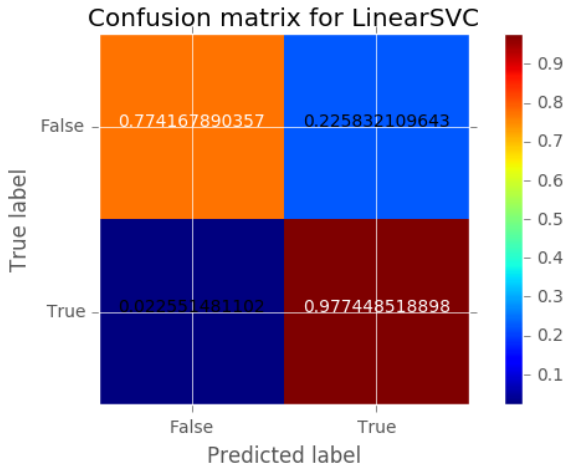
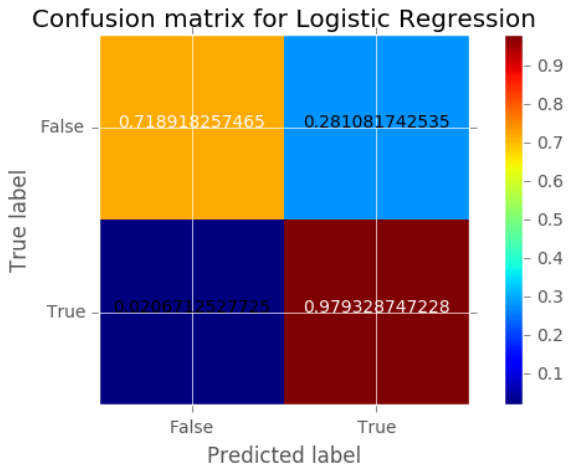


Figure4: Confusion Matrix for LinearSVC



o

Figure 5: Confusion Matrix for Logistic Regression

From the confusion matrix we can conclude that the Naive Bayes algorithm performs well on predicting True Positives but failed awfully while predicting false values.

By plotting the Precision and Recall we can summarize the performance of the classifier model we have discussed.

5.2 Clustering and Aspect Extraction

The output of the classification algorithm is a labeled set of user reviews. The labeled positive reviews and the negative reviews are taken separately and two clustering model is trained upon the them individually.

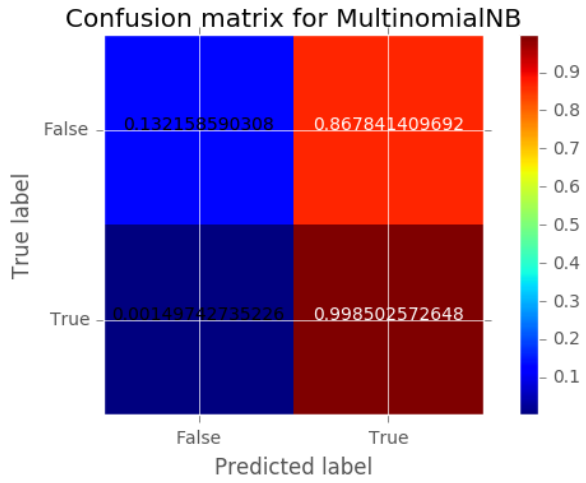


Figure 6: Confusion Matrix for Naive Bayes

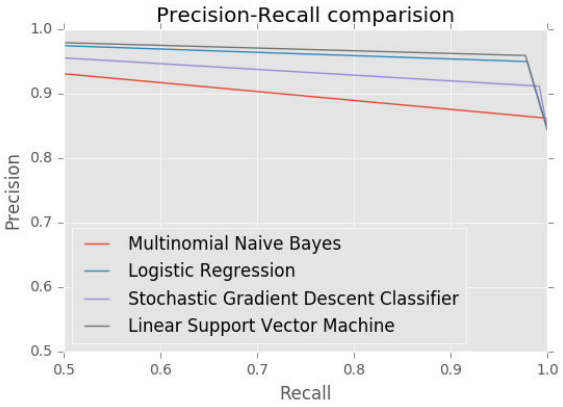


Figure 7: Precision-Recall Graph

Each review is first predicted as either as a good review or bad review, with the predicted label the words are checked for the distance between its position and the cluster centroid. The words which are closest to the centroid are the aspects of the products described in hat particular review. For a product a complete word matrix of all reviews is created individually and the less meaningful words are omitted, then the words score are summed together to get a collective total score for each word in the matrix. Now this word is checked for the distance between the centroid and its location using Eucilidian distance. All the words which lies closest to the centroid are extracted from the matrix and displayed to the user as the aspects of the product which are described the most in the user reviews[7]

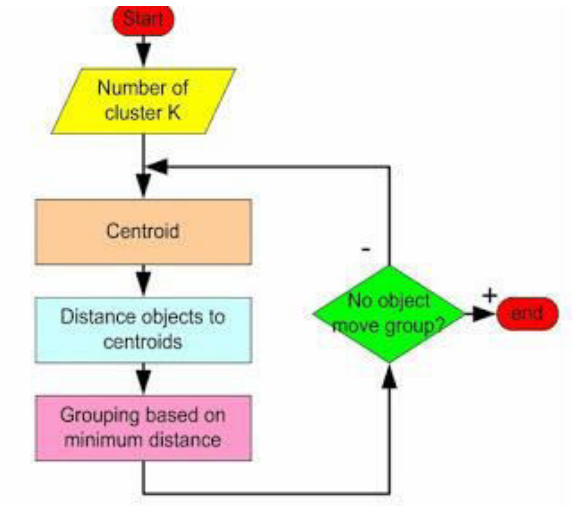


Figure 8: K-Means Flow chart

6. Examples from work

```
~~~~~Positive Review~~~~~
I have been a K-coffee user for some time
and I especially like Timothy's. I ordered
this flavored coffee and the regular decaf
at a very good price with the Amazon 1-day
sale.  All the K-cup varieties are good and
your K-pot makes a fresh cup in 30 sec.
Amazon delivered the coffee in just a couple
of days and I'm still enjoying it.  Thanks,
Amazon and Keurig.  When are you going to
have another sale?  I don't hesitate to
purchase at the regular price because I
really enjoy the coffee, but sales allow us
buy more!

~~~~~Extracted Aspects~~~~~
sale,sec,user,sales,hesitate
```

Figure 9:Example Aspect extraction in a Positive review

```
~~~~~Negative Review~~~~~
Nespresso makes GREAT coffee and GREAT
machines. I switched over to a Nespresso
machine 7 years ago and have never looked
back. I save a small fortune every year by
making my lattes at home.<br /><br />That
being said, the Nespresso capsule offers
posted here are from a third party who is
putting a large additional margin on their
price.<br /><br />You can order the same
products online from Nespresso for approx
$0.55 each (half the price here) AND you can
specify exactly what you want to buy rather
than taking a mixture of items, some of
which you may not like.

~~~~~ Extracted Aspects ~~~~~
nespresso,margin,specify,capsule,fortune
```

Figure 10:Example Aspect extraction in a Negative review

7. Conclusion

We developed a system with two machine learning algorithms to perform classification and clustering on the user product review data. First in order to process the data, we use sci-kit learn's TfidfVectorizer to convert the user reviews into word count matrix or simply called Bag of Words. The Linear SVM Classification algorithm is used to label the review which has 5 different star ratings into either a positive review or negative review then we use the K-means Clustering technique to group the positive and negative scoring words into clusters.

The words which are closest to the respective cluster centers are said to be the important words which decides the product review is either good or a bad review. So far we taken advantage of the large dataset to overcome the misleading reviews and we even omitted the 3 star reviews which are difficult to be labeled into a particular category. By using advanced concepts like neural nets, our system can be improved to provide great accuracy.

4.References

[1]. OnePlus One (Sandstone Black, 64GB) http://www.amazon.in/OnePlus-One-Sandstone-Black-64GB/dp/B00OK2ZW5W. Accessed November 11, 2015.

[2]. Amazon Fine Food Reviews Dataset https://www.kaggle.com/snap/amazon-fine-food-reviews

[3]. B. Liu and L. Zhang. A survey of opinion mining and sentiment analysis. In C. C. Aggarwal and C. Zhai, editors, *Mining Text Data*, pages 415–463. Springer US, 2012.

[4]. B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL 02 conference on Empirical methods in natural language processing-Volume 10*, pages 79–86. Association for Computational Linguistics, 2002.

[5] S. R. Das and M. Y. Chen. Yahoo! for amazon: Sentiment extraction from small talk on the web. *Management Science*, 53(9):1375–1388, 2007.

[6] Xing Fang and Justin Zhan. Sentiment analysis using product review data. *Journal of Big Data*.© Fang and Zhan; licensee Springer.Published:16 June 2015

[7] Zheng-Jun Zha, Jianxing Yu, Jinhui Tang, Meng Wang, Tat-Seng Chua, "Product Aspect Ranking and Its Applications", IEEE Transactions on Knowledge & Data Engineering vol. 26 no. 5, p. 1211-1224, , 2014