SENTIMENT ANALYSIS: AN APPROACH BASED ON TEXTUAL DATA AND PRODUCTS REVIEWS

Nisha VM¹, Dr Ashok Kumar R²

¹M. Tech Computer Network Engineering, Dept of ISE, BMS College of Engineering Bangalore, India
²Associate Professor, Dept of ISE, BMS College of Engineering Bangalore, India.

ABSTRACT

Sentiment analysis is one of the most important approaches used to analyze the data and identify sentiment expressed by the customers on the text. Opinion mining is also the name given for sentiment analysis. In the Internet large amount of text data is getting generated either in the form of suggestion, feedbacks, tweets or comments. The Amazon web services are generating lots of text data through customer reviews. Analyzing this data will help the online retailers to understand customer expectation and also provide shopping experience that can increase the sale. The Sentiment Analysis helps to identify the positive, negative and neutral information from the customer reviews. Most of the researchers have developed a technique for the sentiment analysis. Machine learning algorithms are used for analyzing text data. This work concentrates on the Amazon customer reviews on electronics devices such as kindle, fire TV stick, computer etc which includes basics product information like rating, reviews, etc. Using Machine Learning Algorithm based on text reviews the algorithm can predict positive or negative response given by the customers.

Keywords - Sentiment Analysis, Text Summarization, Information extraction, opinion mining, machine learning methods, and online reviews.

[1] INTRODUCTION

Sentiment Analysis is the platform where people express their reviews or opinions through text [11]. As there is a rapid growth in the e-commerce, millions of customers share their opinion about the products through discussion group, review website or through their personal blogs. Due to which the reviews that are available online in the Internet are increasing rapidly. The reviews that are available in the Internet are in the form of text that can be predicted the quality of the product. Sentiment Analysis can also be applied to the news article, blogs, stock market, movies reviews etc. Now a day’s people book the products, hotels, movies based on positive and negative reviews [11][13][14]. Analyzing this kind of data helps to understand the customer point of view towards the brand strategies. Machine Learning concepts help in analyzing the text data using some classification algorithms for training such as Navies Bayes and testing TF_IDF (Term Frequency and Inverse Document Frequency) count vectorization. There are basically three level of Sentiment Analysis as seen in figure1.

i. Document Level
ii. Sentence Level
iii. Aspect Level
Document Level: In Document level of Sentiment Analysis at a time the complete document has to be analyzed [7]. This level assumes that the review belongs to a single person. Document level supports for both supervised and unsupervised classification algorithm.

Sentence level: The document has to be broken into sentences in the Sentiment level Analysis. The breaking of sentences into a document leads to a subjectivity of classification. Each sentence is examined at a time [9]. The main aim of this approach is to find target of the sentence without the prediction of positive, negative and neutral of the sentence detected which is not useful [6].

Aspect Level: Aspect Level Sentiment Analysis mainly concentrates on the target entity attributes. It is mainly performed on the feedbacks, comments and compliments etc. Aspect level sentiment analysis application includes online store reviews, hotel reviews movie and more [6] [22].

[2] RELATED WORK

Satuluri Vanaja and Meena Belwal [1] has worked on Aspect Level Sentiment Analysis on E-Commerce Data. These authors considered the customer reviews and worked on parts of speech which includes noun, pronoun, verb, adverb, adjective, preposition etc. Using an A priori algorithm they extracted feature from the dataset and tried to identify emotions through the customer’s reviews.

Pankaj gupta and Ritu Tiwari [2] has worked on Sentiment Analysis using machine learning techniques. The main aim of this work was to differentiate the Sentiment Analysis of the traditional based approaches that included the sentence level classification algorithms. This work basically consists of two algorithms Naive Bayes Classifier and Support Vector Machine. The results obtained were more improved and precise when compared to earlier techniques.

ANH-DUNG VO and QUANG-PHUOC NGUYEN [3] has worked on opinion Aspect Relationship in Cognizing Feeling via Reviews which concentrates on Opinion Mining. These two authors introduces a new concept i.e. knowledge extraction and Sentiment Analysis [6]. Knowledge level extraction was performed using some natural language processing tools such as dependency parser (DP), Named Entity Recognition (NER).

Xiaojiang Lei and Xueming Qian [4] these were the two authors who worked on the Rating Prediction Based on Social Sentiment from Textual Reviews. These two authors mainly concentrated on the rating prediction task. The main aim of this paper was to find the effective information from the reviews and predict ratings. This was basically done using LDA (Linear Discriminate Analysis).

The Sentiment Analysis mainly concentrates on the opinion mining[7] where customers express their reviews on the particular products [10]. All the above related word is mainly focused on the features either it can be the parts of speech or usage through tools etc. but this paper works on the textual data.
and products reviews using classification algorithm such as Naïve Bayes and TF_IDF (Term Frequency and Inverse Document Frequency) count vectorization algorithms [13].

[3] DESCRIPTION OF DATASET

The dataset consist of Amazon products textual data and reviews on particular products such as computer, fire Stick TV, kindle etc [20]. On the particular products customers expressed their reviews either it can be positive or negative. The screenshot of the dataset can be viewed in table I It also consists of review_id, customer_id, url, address, brand, ratings, textual reviews etc. Table II describes about the details that are available in dataset. The file is in the CSV format. Table III focuses on the selected columns from Table II.

Table I. Screenshot of dataset

Table II

<table>
<thead>
<tr>
<th>Sl.no</th>
<th>Column name</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>User_id</td>
</tr>
<tr>
<td>2</td>
<td>name</td>
</tr>
<tr>
<td>3</td>
<td>brand</td>
</tr>
<tr>
<td>4</td>
<td>category</td>
</tr>
<tr>
<td>5</td>
<td>Reviewers_date</td>
</tr>
<tr>
<td>6</td>
<td>Reviewers_time</td>
</tr>
<tr>
<td>7</td>
<td>Reviewers_ratings</td>
</tr>
<tr>
<td>8</td>
<td>Reviewers_url</td>
</tr>
<tr>
<td>9</td>
<td>Reviewers_text</td>
</tr>
<tr>
<td>10</td>
<td>Reviewers_title</td>
</tr>
<tr>
<td>11</td>
<td>Reviewers_userCity</td>
</tr>
<tr>
<td>12</td>
<td>Reviewers_username</td>
</tr>
</tbody>
</table>

Table III

<table>
<thead>
<tr>
<th>Selected column</th>
</tr>
</thead>
<tbody>
<tr>
<td>1,2,5,6,7,9,12</td>
</tr>
</tbody>
</table>

[4] SYSTEM DESIGN AND IMPLEMENTATION

Sentiment Analysis is one of the approaches where customers express their emotions through text. In this paper the emotions that are expressed on Amazon electronics products that can be either a
positive or a negative sentiment. This Sentiment Analysis is associated with the text data i.e. the textual representation of the text data. As there is a large amount of text data available in the Internet lots of preprocessing has to be done before the actual classification. The preprocessing includes deletion of empty rows also the deletion of null values from the table so that the further processing becomes easier. Other preprocessing techniques include the tokenization which explains how many tokens are available in a particular text [20].

Figure 2 describes the pre-processing techniques that are included in this work. The pre-processing techniques include converting the text to the lowercase, Array of tokenization, removing of stop words and erase punctuation. Sentiment Analysis is the natural language processing techniques that include the emotions expressed by the customers either it can be positive or negative [23]. In order to identify the sentiments the text data has to undergo the pre-processing techniques.

![Processing Techniques](image)

**Figure 2. Pre-processing techniques involved in the text data**

The feature extraction includes the count of positive and negative reviews from the dataset. The graph is generated which shows exactly how many positive and negative reviews are associated in the dataset. Based on the review given on the particular product through textual data or through the rating the count of reviews is classified as positive and negative.

![Count of Reviews](image)

**Figure 3. Count of Positive and Negative Review.**
In order to extract the feature from the text i.e count of positive and negative review the condition is specified on the reviewers rating i.e it reviewers rating should be greater than or equal to 4 (reviews rating>=4). This is the sentiment that the customers have expressed on the particular product. If the customers have given the rating on particular electronic devices is greater than or equal to 4 then the review is positive else the review is negative. Figure3 describes completely based on the above concept.

![Classification Model of Text Data](image)

**Figure 4. Classification Model of text data**

Figure4 describes the detailed information and classification involved in the textual data.

The classification algorithm includes two concept training and testing. This step is usually used to classify based on opinion expressed by the customers either positive or negative. The classification algorithm used for training the model is Naive Bayes algorithm.

Naive Bayes algorithm is one of the machine learning algorithms used for training the dataset [13]. It is the probability based algorithm. Each Aspect level dataset is calculated using Naive Bayes algorithm. In natural language processing that is the Sentiment Analysis (opinion mining) describes the probability of the features.

The training set includes the information about the customers who have given their reviews on the multiple products. For example a customer with single id has given his opinion on multiple products it can be a kindle or computer or a fire stick TV. Different opinions on multiple products by single customers are considered on the training set. Detailed description of training set is designed in figure5.

![Training Phase Dataset](image)

**Figure 5. Example of training phase dataset.**
Once the data is classified into training set the Accuracy is calculated using Naïve Bayes classifier and the accuracy obtained is 58.97%. Further the most informative features were calculated which is from the training set and also gave the detailed information about the algorithm. The example for this feature and the Accuracy percentage can be viewed in the figure 6.

![Figure 6. Accuracy using Naive Bayes classifier](image)

The most informative features from the training classifier were “warning”, “deleted”, “rotate”, “bent”, “nope”. This Naïve Bayes classifier also explained about these features i.e. the probability of occurring of each feature from the training set. For example in the feature “warning” the probability of occurring the sentiment (positive/negative) is 51.3 is more as negative where as less in positive sentiment. The ratio or the probability is same with the other feature i.e. deleted, rotate, bent and nope. So from the training dataset using the Naïve Bayes classifier the work can be concluded that the most informative features obtained words are occurring as negative is more due to which the accuracy obtained is less.

In order to increase the accuracy from the training set we proceed with the testing phase of the classification. In natural language processing i.e the machine learning the sentiments that are expressed can be in the form of positive or negative. The performance vector is used in the testing phase i.e. the count vectorization. To increase the performance of the training set TF_IDF algorithm is used.

TF_IDF represents term frequency inverse document frequency. The TF-IDF is a machine learning algorithm used to represent the text data. The text data can be a positive or the negative sentiment expressed by the customers. TF-IDF is used for checking the probability of occurrence of each word in the text [8]. The probability can be calculated as increase the proportionality of word that are occurring in the document to the offset by number of documents occurring in the corpus dataset. The TF-IDF also used for recommendation system where it recommends the sentiments based on the words expressed by the customers.

The corpus dataset is divided into training and testing phase. The training phase is expressed in the above concepts. The testing phase in the corpus dataset includes the information about the customers reviews expressed on particular products. For example customer 2 and 3 have expressed their reviews on specific products i.e. single customer expressed his interest on a single product as described in the figure 7.
There are basically three fitting techniques that are used in the testing phase of dataset. All the three fitting techniques are basically used for predicting techniques.

**Multinomial Naïve Bayes (MNB):** This multinomial Naive Bayes algorithm is the machine learning algorithm used for predicting the sentiments expressed by the customers on the particular products. This prediction technique has to be applied on the discrete values like word count 1, 2, 3, 4 … is the underlying calculation to deal with it. When this fitting technique was applied on the testing data the accuracy was increased. This testing dataset with multinomial NB the obtained accuracy percentage of 93.29% which is an obviously more than the training dataset. The exact accuracy can be observed in figure 8.

**Figure 8. Multinomial Accuracy Percentage.**

**Bernoulli Naïve Bayes (BNB):** This Bernoulli Naive Bayes algorithm is also the machine learning algorithm used for predicting the sentiments expressed by the customers on the particular products. The Bernoulli NB is applied on a testing dataset and it should be applied on binary values or Boolean values like true/false or 0/1. The obtained results i.e the accuracy for this fitting technique is 92.04%. The exact accuracy can be viewed in the figure9.

**Figure 9: Accuracy results for Bernoulli NB.**
Logistic Regression (LNB): The Logistic Regression is the regression analysis that is used to examine the relationship between two or more variables. The logistic regression is also one of the machines learning prediction technique used for prediction of the text data. The Logistic Regression is the combination (average) of both Multinomial NB and the Bernoulli NB. The LNB predicts the average accuracy of the testing dataset that is obtained from the fitting techniques. The exact accuracy percentage can be observed from figure10.

From the above three fitting techniques the testing phase has obtained the improved accuracy from the testing phase. The aim of the classification techniques was to increase the accuracy. For training the dataset was divided and obtained the accuracy of around 58.97% and in order to increase the accuracy percentage the data was divided into testing where three fitting techniques were used and the average accuracy was obtained from the logistic regression which is 93.73%.

[5] PERFORMANCE METRIC

The metric used to analyze the performance of the trained model are described in this section. Performance metric is determined based on the following four measures:

**True Positive (TP):** Positive review on the electronic devices that are correctly classified as positive.

**False Positive (FP):** Negative review on the electronic devices that are wrongly classified as positive.

**False Negative (FN):** Positive review on the electronic devices that are wrongly classified as negative.

**True Negative (TN):** Negative review on the electronic devices that are correctly classified as negative.

The important metric are accuracy, precision, recall, specificity and F1 score. Recall and specificity together evaluates the accuracy measurement.

- **Precision(P):**
  Precision is defined as proportion of predicted positive instances that are correctly real positives. Hence, P is given by,
  \[ \text{Precision (P)} = \frac{TP}{(TP+FP)} \]  

- **Recall(R):**
  Recall is defined as proportion of real positive instances that are correctly predicted as positive. Hence, R is given by,
  \[ \text{Recall (R)} = \frac{TP}{(TP+FN)} \]  

- **F1 Score(F1):**
  F1 score can be defined as weighted mean of precision and recall. Hence, F1 is given by,
  \[ F1 = 2 \times \frac{\text{Precision}\times\text{Recall}}{(\text{Precision} + \text{Recall})} \]
Therefore the equations (1), (2), (3) are used to compute the precision, recall, and the F1 score respectively. The outputs of these equations are probabilistic values, where negative represents model is inaccurate and positive represents model is accurate.

[6] ANALYSIS
In this section the Receiver Operating Characteristic curve (ROC) graph is plotted which shows the performance of the classification model of all thresholds. The curve consists of two parameters;

- True Positive Rate
- False Positive Rate

**True Positive Rate (TPR)** is similar to Recall and is defined as shown below.

\[ \text{TRP} = \frac{TP}{(TP+FN)}; \]  

(4)

**False Positive Rate (FPR)** is similar to the precision and is defined as shown below,

\[ \text{FPR} = \frac{TP}{(TP+FP)}; \]  

(5)

[7] RESULTS
The results include different fitting techniques that were performed and the accuracy was calculated. The comparison is made between the different classifier i.e. TRP vs. FPR and the graph for different fitting techniques is show in figure11 (i.e multinomial, Bernoulli and logistic regression). The ROC [15] curve is obtained from the three fitting techniques i.e. the accuracy from the multinomial, Bernoulli and the logistic Regression. The Area under the Curve (AUC) provides the accurate measure of performance across the possible classification threshold. The average accuracy obtained is around 94% hence the failure rate is less. Therefore the obtained results are accurate and prediction of the reviews becomes more accurate and easier.

![Graph showing ROC curve with three fitting techniques](image)

**Figure11. TP vs. FP rate at different classification threshold**

Let’s see the precision, recall, and F1 score of different classifier

**Multinomial:**

<table>
<thead>
<tr>
<th></th>
<th>Precision</th>
<th>Recall</th>
<th>F1 Score</th>
<th>Support</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>464</td>
</tr>
<tr>
<td>Negative</td>
<td>0.93</td>
<td>1.00</td>
<td>0.97</td>
<td>6461</td>
</tr>
<tr>
<td>Avg/total</td>
<td>0.87</td>
<td>0.93</td>
<td>0.90</td>
<td>6925</td>
</tr>
</tbody>
</table>
Bernoulli’s:

<table>
<thead>
<tr>
<th></th>
<th>Precision</th>
<th>Recall</th>
<th>F1 Score</th>
<th>Support</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive</td>
<td>0.33</td>
<td>0.17</td>
<td>0.23</td>
<td>464</td>
</tr>
<tr>
<td>Negative</td>
<td>0.94</td>
<td>0.97</td>
<td>0.96</td>
<td>6461</td>
</tr>
<tr>
<td>Avg / total</td>
<td>0.90</td>
<td>0.92</td>
<td>0.91</td>
<td>6925</td>
</tr>
</tbody>
</table>

Bernoulli’s Naïve Bayes has to be used for a feature with binary values like True/False. The comparison is made between the TPR and FPR which gave the above results.

Logistic Regression:

<table>
<thead>
<tr>
<th></th>
<th>Precision</th>
<th>Recall</th>
<th>F1 Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive</td>
<td>0.56</td>
<td>0.33</td>
<td>0.47</td>
</tr>
<tr>
<td>Negative</td>
<td>0.95</td>
<td>0.98</td>
<td>0.97</td>
</tr>
<tr>
<td>Avg / total</td>
<td>0.93</td>
<td>0.94</td>
<td>0.97</td>
</tr>
</tbody>
</table>

The analysis results obtained is more accurate for all the three parameters i.e. precision, recall and F1-score [15] that can be viewed from the above tables for all the three fitting techniques Multinomial, Bernoulli and Logistic Regression.

Table IV describes the test for classifier of hand written examples
The table IV explains about the opinion expressed on a particular product the sample of the reviews can be viewed in the table IV.
For a particular sample text example given in this work can predict the given reviews are positive or negative. The probability of occurring a review in a particular context is positive or negative. This can be predicted based on the comment or the sentiment expressed by customers. The TF-IDF count vectorizer is used that predicts the probability of occurring that review in a textual data is positive or negative.

<table>
<thead>
<tr>
<th>Sample review</th>
<th>Sample estimated as</th>
<th>Probability of Positive Review</th>
<th>Probability of Negative Review</th>
</tr>
</thead>
<tbody>
<tr>
<td>The product was good and easy to use</td>
<td>Positive</td>
<td>1.000000</td>
<td>0.000000</td>
</tr>
<tr>
<td>The whole experience was horrible and product is worst</td>
<td>Negative</td>
<td>0.005560</td>
<td>0.994440</td>
</tr>
<tr>
<td>Product is not good</td>
<td>Negative</td>
<td>0.977007</td>
<td>0.22993</td>
</tr>
</tbody>
</table>
When a review is fed into the trained model the model will determine the polarity of the review. The Table IV shows the example where the model has predicted example reviews as either positive review or negative review.

The table V predicted values of classifiers for check on the basis of review text. This table V describes the information from the original dataset that is considered in this work. This work tried to predict the actual review expressed by the customers on the electronics devices. Based on the textual reviews expressed on devices these classification algorithms predict whether it is a positive review or the negative review.

<table>
<thead>
<tr>
<th>Reviewers_id</th>
<th>Reviews_text</th>
<th>Reviewers_username</th>
<th>Words</th>
<th>Multinomial</th>
<th>Bernoulli’s</th>
<th>logistic</th>
<th>Predicted Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>The Kindle is my first e-ink reader. I own an a book and it’s a good book</td>
<td>Jeffrey Stanley</td>
<td>[the, kindle, is, my, first, e, ink, reader,]</td>
<td>pos</td>
<td>neg</td>
<td>Pos</td>
<td>Positive</td>
</tr>
<tr>
<td>2</td>
<td>I’m a first-time Kindle owner, so I have nothing...</td>
<td>Matthew Coenen</td>
<td>[i, m, a, first, time, kindle, owner, so, i, h,]</td>
<td>pos</td>
<td>neg</td>
<td>pos</td>
<td>Positive</td>
</tr>
<tr>
<td>3</td>
<td>UPDATE NOVEMBER 2011: My review is now over a n...</td>
<td>Ron Cronovich</td>
<td>[update, november, my, review, is, now, over,]</td>
<td>pos</td>
<td>neg</td>
<td>pos</td>
<td>Positive</td>
</tr>
<tr>
<td>4</td>
<td>woke up to a nice surprise this morning: a n...</td>
<td>C. Tipton</td>
<td>[i, woke, up, to, a, nice, surprise, this, mor]</td>
<td>pos</td>
<td>neg</td>
<td>pos</td>
<td>Positive</td>
</tr>
<tr>
<td>5</td>
<td>All of them quit working. There's absolutely n...</td>
<td>M. Lansford Kindle fave</td>
<td>[all, of, them, quit, working, there, s, absol...]</td>
<td>pos</td>
<td>pos</td>
<td>pos</td>
<td>Positive</td>
</tr>
<tr>
<td>6</td>
<td>Great</td>
<td>Matthew Coenen</td>
<td>[great]</td>
<td>pos</td>
<td>pos</td>
<td>pos</td>
<td>Positive</td>
</tr>
<tr>
<td>7</td>
<td>I used to hate to read by now that I have my own k.</td>
<td>Jeffrey Stanley</td>
<td>[i, used, to, hate, to, read, by, now, that, i, have, my, own, k]</td>
<td>pos</td>
<td>pos</td>
<td>pos</td>
<td>Positive</td>
</tr>
<tr>
<td>8</td>
<td>Excellent product Easy to Use large screen makes watching movies and reading easier</td>
<td>Matthew Coenen</td>
<td>[Excellent, product, Easy, to, Use, large, screen, makes, watching, movies, and, reading, easier]</td>
<td>pos</td>
<td>neg</td>
<td>pos</td>
<td>Positive</td>
</tr>
<tr>
<td>9</td>
<td>Great for begineers. Brought as a giveand she loves it.</td>
<td>Jorge Truman</td>
<td>[Great, for, begineers, Brought, as, a, giveand, she, loves, it]</td>
<td>pos</td>
<td>neg</td>
<td>pos</td>
<td>Positive</td>
</tr>
<tr>
<td>10</td>
<td>All of them quit working. There's absolutely n...</td>
<td>D. Tatro</td>
<td>[all, of, them, quit, working, there, s, absol...]</td>
<td>pos</td>
<td>pos</td>
<td>pos</td>
<td>Positive</td>
</tr>
</tbody>
</table>

Another way of representation of text data is the word cloud. Word cloud is the pictorial representation and visualization of text data. Word cloud is an image composed of words used in a particular text or a subject in which the size of each word indicates its frequency and importance of
each word. This word cloud is represented for two sentiments a positive sentiment and a negative sentiment. Figure 12 describes about the complete word cloud from the textual data.

![Figure 12. Complete word cloud visualization of text data](image)

The next representation of the text data is for the positive words that are related to the text data. Figure 13 describes the negative words that are associated with the textual reviews.

![Figure 13. Word cloud with Positive Words](image)

![Figure 14. Word cloud with Negative Words](image)

The next representation of the text data is for the negative words that are related to the text data. Figure 14 describes the negative words that are associated with the textual reviews.

[8] CONCLUSION

In this paper, classification has been done using training and testing the model specifically with the Naïve Bayes and TF-IDF algorithm, which has powerful impact on predicting the polarity of the product review. The model is then trained using logistic regression machine learning classification algorithm, which is showing the accuracy of 94%. When a new set of reviews are feed into the model the model will predict the polarity of the products reviews.
Depending on the sentiment expressed by the people on the particular product the model predicts based on textual review as positive or the negative review.

Future Work

In future work, a mobile application can be developed to predict the reviews so that the model can recommend the customers based on the reviews. A model can be generated that can have more precise feature and also by applying different machine learning algorithm accuracy of the model can be improved.

REFERENCES


[4] Xiaojiang Lei, Xueming Qian, “Rating Prediction Based on Social Sentiment From Textual Reviews”, IEEE transactions on multimedia, vol. 18, no. 9, september 2016


