www.ijecs.in
*International Journal Of Engineering And Computer Science ISSN:2319-7242*
*Volume 4 Issue 5 May 2015, Page No. 11718-11717*

# Using Business Intelligence For Mining Online Reviews For Predicting Sales Performance

*Mohsin Nadaf*[1], *Sunil Lahane*[2], *Akshay Deshpande*[3,] *Sneha Tirth*[4]

Department of Information Technology, Trinity College of Engineering and Research,
Savitribai Phule Pune University, Pune

[1]mohsinanadaf@gmail.com

[2]sunilnlahane@gmail.com

[3]akkisayshello@gmail.com

[4]sneha.tirth@gmail.com

**Abstract:** ***Nowadays, posting reviews online has become very popular way for people to express their opinions and sentiments toward the products bought or the services received. Analyzing large volume of online reviews available would produce the useful actionable knowledge which could be of economic values to the vendors and the other interested parties. Here, we understood and tried to solve the problem of mining reviews for predicting the product sales performance. The reviews which are posted by consumers involve the sentiments. So, these sentiments expressed in reviews and the quality of the reviews has significant impact on the future sales performance of products or services. And these sentiments are hidden in the Document Corpus which is also known as a Comments Document. The document-level sentiment classification aims to automate task of classifying the textual review, which is given on single topic, as expressing either positive or negative sentiment. Hence by getting these sentiments the overall review or feedback about the particular product can be known in a summarized form which will help vendors to know the overall statistics and the future performance of their product.***

**Keywords:** Business Intelligence, Data Mining, Sentiment Mining, Reviews Mining, Opinion Mining, Sentiment Analysis, Prediction

## 1. Introduction

When consumer purchases any product, process of quality evaluation takes place naturally in his/her mind. So posting the online reviews has become an increasingly popular way for people to share with other users their thoughts, opinions and sentiments toward services and products. It has become a very common practice for e-commerce websites to provide the facilities for people to publish their reviews, with a prominent example being Flipkart (www.flipkart.com), Twitter. Reviews are also very common and frequent thing in blog posts, social networking websites like Facebook, Twitter as well as other dedicated review websites. So those online reviews present a wealth of information on the services and products, and if properly utilized, can provide the vendors highly valuable network intelligence and social intelligence to facilitate the improvement of their business. They will also help to support the strategic market decisions like what changes we need to do in our product or should we continue with this product or not.

As a result, review mining has recently received a great deal of attention. A growing number of recent studies have focused on the economic values of the reviews, exploring relationship between the sales performance of products and their reviews. Since what the general public thinks of a product can no doubt influence how well a particular service or product sells, then understanding the opinions and the sentiments expressed in the relevant reviews is of a high importance, because finally such reviews reflect what the general public think and thus can be very good indicator of the product's future sales performance.

Sentiment is nothing but determining an opinion about a product whether it is positive or negative or neutral. Sentiment classification is a special case of text categorization problem, where the classification is done on basis of attitude expressed by the consumers in discussion forums or the blogs etc. Sentiment analysis requires a deep understanding of the document under analysis because the concern here is how the sentiment is being communicated.

But simply classifying reviews as positive or negative, as the most current sentiment mining approaches are designed for, does not provide comprehensive understanding of the sentiments reflected in reviews. Hence in order to model the multifaceted nature of the sentiments, we will view sentiments which are embedded in reviews as an outcome of the joint contribution of a number of hidden factors. So here in our project we have fetched reviews from social networking website named Facebook. Then we did the analysis on those reviews got the result whether the comments are positive, negative or neutral and also predicted the Sales prediction of a particular product. Here the reason behind calling like a specific product is nothing but to state that our project is Product neutral, that is it can give result for any product or service. And that is what this whole process we are trying to explain in this paper.

## 2. Project Objective

The main objective of this system is to predict the Sales performance of a particular product. Generally the Vendor manufactures his product and sales in market. So it should be known to that vendor about his product feedback or performance in the market so that he can come to know the statistics of same thus will be able to take decisions based on that. So in general, if a consumer wants to analyze any product he can do such research on his own regarding same. But what if vendor wants to do the same regarding his products or services, because there are only few such products for vendors to know the statistics and sales prediction of his product. So when we understood this problem, we decided to develop such Vendor Oriented Product for the same which can

- Classify reviews in three different categories namely positive, negative and neutral
- Predict Sales Performance of a product and
- Give results which can help in making strategic business decisions

## 3. System Architecture

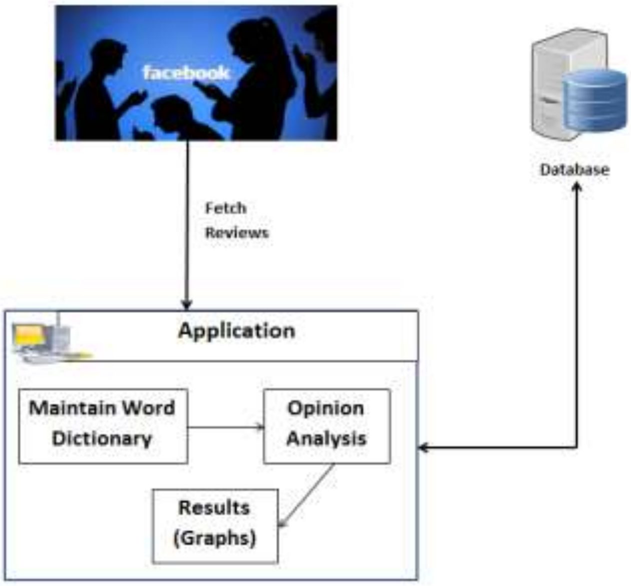


**Figure 1:** System Architecture

In our application we are fetching the live comments from Facebook. Firstly we need to specify the name of that product for which the reviews are to be fetched. Then the comments or reviews from the Facebook pages are dynamically fetched and responded to the system. Then to classify those reviews we already need to maintain a dictionary of words (positive, negative). And that words dictionary is stored and maintained in the database. So as soon as we get any words as positive or negative that we need to add in the dictionary specifying its proper category. Then after reviews are fetched this dictionary is compared to the same. And before we can compare, the fetched reviews are first broken down into the tokens. Then algorithms are applied to sort comments into positive, negative and neutral section. This will be calculated to generate the desired result. Then the results are displayed in the forms of bar graph specifying the number of positive, negative and neutral reviews. And then Sales Prediction is also calculated based on this processing which tells you, if there is a hike or decrease in sales and if so, then in how much percentage. Thus the opinion analysis is done and the result is displayed. The whole opinion mining flow is as follows:

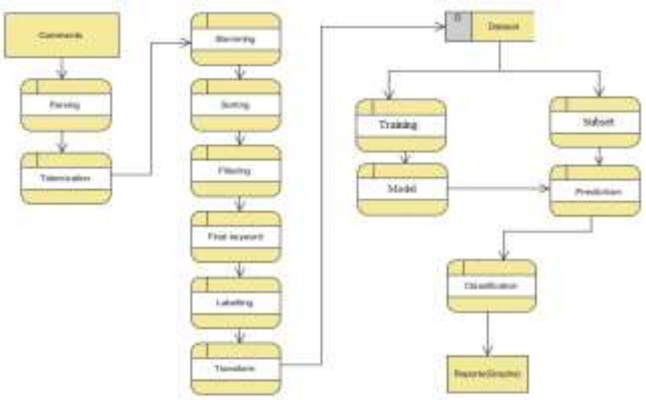


**Figure 2:** Opinion Mining Flow

- **Document Corpus**- It is a comment document. The fetched reviews are thus in the form of the documents which are taken dynamically from the website.
- **Parsing**- Breaking paragraph or data block into small parts like sentences, these sentences from the document which contains the comments or reviews fetched from the site. The whole paragraph is broken down to sentences.
- **Tokenization**- The sentences which are broken down from paragraph are further separately tokened into or broken into Keywords.
- **Stemming-** It is the technique to find base or root form of words which we have tokenized from sentences for example "user engineering" is converted into "use engineer". Stemming is very important in this process.
- **TF-IDF-** It stands for Term Frequency-Inverse Document Frequency is use to find the frequency of a word in a document this will help in improvement of the result.
- **Sorting**- The sorting is done on the basis of TF-IDF Scores. As according to the frequency of the words they are rated and then sorting is applied.
- **Filtering**- Then in the filtering process, important keywords are taken and then unnecessary words are filtered or rejected which have no use here.
- **Final Keyword**- List of final keywords that are tokenized, stemmed, sorted and filtered are ready to be compared with the dictionary which is maintained.
- **Labeling**- Labeling the words whether it is Positive or Negative after comparing the keywords with the dictionary.
- **Transform**- In this step the string is converted to integer for applying algorithms. So the words are converted into integer as "1" and "0" and then the algorithm is applied to get the desired result.
- **Final Data set**- All the above processing results in generation of final data set which is then ready for applying algorithms. Then the training to that is given, subsequently it goes through modeling, subset stages. And then after all this processing is done the next step which comes is of classification and prediction which means to classify whether the fetched reviews are positive or negative or neutral. And then this result is shown in terms of bar graph.

## 4. Modules

As explained above the whole process of our project, so in total there are three modules as follows:

1. **Manage Word Dictionary**
2. **Fetch and Manage Comments**
3. **Analysis of Reviews**

### 4.1 Manage Word Dictionary

This module is designed according to the project implementation path as the comments are being read by the machine dynamically we have to make the machine learn how to classify them by maintaining the dictionary. This is handled by the admin manually and is updated. As the system proceeds the comments are handled and maintained and classified in two class namely positive and negative words. These words are the set with which the comments will be matched. The comments which will be entered will be stemmed down and then will be stored in the dictionary and then will be compared with the operations performed in the next module.

### 4.2 Fetch and Manage Comments

The module 2 of this project has the operation of fetching online comments dynamically from the social networking website. Here in this module the comments are fetched dynamically, networking sites have many pages or reviews that a product can be decided by the user. Then these comments are parsed and then they are broken down into small tokens and then compared to the dictionary maintained in module 1. Then the comment is classified into positive and negative accordingly. Then the result will be in form of positive and negative and neutral sets. Here this module gives output of stemming like if you select any comment, then it will show its stemmed word. And will also show the category (positive or negative or neutral) of that Comment in Output box (Check the same in Section 6.1 Results)

### 4.3 Analysis of Reviews

In module 3 of this project we are analyzing all the online comments fetched from Facebook. This third module needs above two modules to be processed first and if it is so then only you can process this module. So with the help of the algorithms like Jaccard and Cosine similarity we generate exact calculations and this whole calculations result is presented in formats like bar graphs and charts. One more thing this module does is Sales Prediction. And this Sales Prediction is based on the results which are generated above. So it shows the sales prediction as in whether there will be hike or depletion in the sales of that product and if so, in how much percentage.

## 5. Algorithms

We have used two algorithms in our project, namely Jaccard and Cosine similarity. The comments are processed and result is given by both the Jaccard and Cosine similarity. Then as we know there is some difference in accuracy of every algorithm, so that is the reason why we have used two algorithms. So by using them we are able to get the accurate result of the reviews. So while processing we use these two algorithms and then afterword give final result as average of these two.

### 5.1 Jaccard

The Jaccard index, also known as Jaccard similarity coefficient is a statistic used for comparing the similarity and diversity of the sample sets. Jaccard coefficient measures similarity between the finite sample sets and is defined as the size of the intersection divided by size of the union of the sample sets. Formula is:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}. \qquad (1)$$

The Jaccard distance, which measures dissimilarity between sample sets, is the complementary to the Jaccard coefficient and is obtained by subtracting the Jaccard coefficient from 1, or, equivalently, by dividing difference of the sizes of the union and the intersection of two sets by the size of the union:

$$d_J(A,B) = 1 - J(A,B) = \frac{|A \cup B| - |A \cap B|}{|A \cup B|}. \qquad (2)$$

### 5.2 Cosine Similarity

The Cosine similarity is a measure of similarity between the two vectors of an inner product space that measures the cosine of the angle between them. Cosine of 0° is equal to 1, and it's less than 1 for any other angle. It is thus judgment of orientation and not magnitude: two vectors with the same orientation have a cosine similarity of 1, and two vectors at 90° have a similarity of 0, and the two vectors diametrically opposed have a similarity of -1, independent of their magnitude. The Cosine similarity is particularly used in positive space, where outcome is neatly bounded in [0,1]. The cosine of two vectors can be derived by using Euclidean dot product formula:

$$\mathbf{a} \cdot \mathbf{b} = \|\mathbf{a}\| \, \|\mathbf{b}\| \cos\theta \qquad (1)$$

Given two vectors of attributes, A and B, cosine similarity, cos(θ), is represented using the dot product and magnitude as

$$\text{similarity} = \cos(\theta) = \frac{A \cdot B}{\|A\| \|B\|} \qquad (2)$$

## 6. Results

### 6.1 Apply Mining

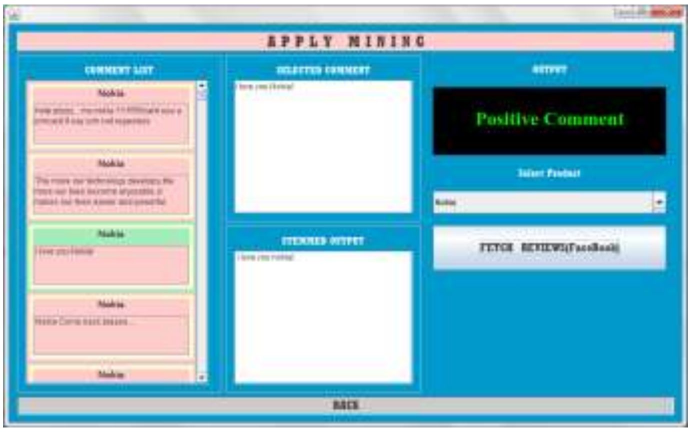
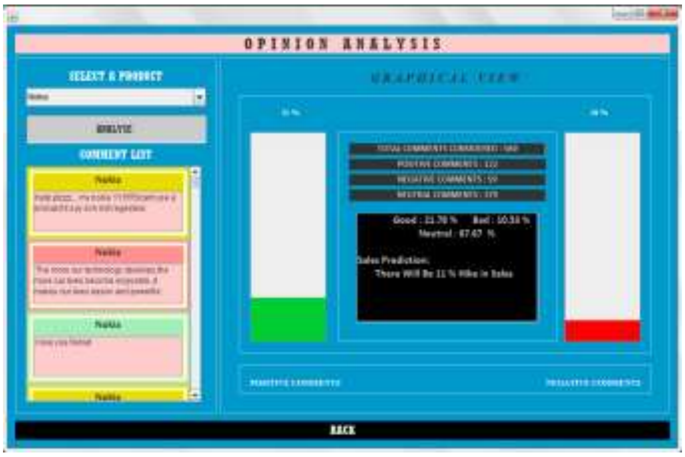


**Figure 3:** Apply Mining

Here for demo purpose we have taken example of product "Nokia". So in this module we fetch all the comments associated with the product "Nokia". So all the live comments from of "Nokia" are fetched dynamically from the social networking sites (in our case it is Facebook). So when comments are fetched they are parsed and then are broken down into tokens and stemmed and then compared to the dictionary maintained in module 1. The stemming results are also shown in this module. So if you select any comment it will show its stemmed output. The dictionary maintained is updated as per our need and is compared with the comments parsed, broken down into tokens.

### 6.2 Opinion Analysis



TOTAL COMMENTS CONSIDERED : [illegible]
POSITIVE COMMENTS : [illegible]
NEGATIVE COMMENTS : [illegible]
NEUTRAL COMMENTS : [illegible]

Good : 21.70 % Bad : 10.53 %
Neutral : 67.67 %

Sales Prediction:
There Will Be 11 % Hike in Sales

**Figure 4:** Opinion Analysis

So now all the ”Nokia” related comments which are already fetched in module 2 are analyzed here. The algorithms are applied to sort comments into positive, negative and neutral section. Cosine similarity and Jaccard algorithms are used to perform operations like comparing and classification of comments. These both algorithms are thus then used to calculate the average value of the comments. Then the Sales Performance (Prediction) is also calculated in terms of percentage of hike or decrease in Sales.

## 7. Conclusion and Future Scope

So here we have successfully applied the Jaccard and Cosine Similarity algorithm and got the result in terms of Positive and Negative comments and also predicted the Sales Performance as in rise and fall in sales. This helps vendor to know or predict the future performance of his product. Thus helping vendors and management people take strategic business decisions.

As of now we have taken reviews from Social Networking website like Facebook. In future, the reviews can be also collected from Twitter, YouTube or other similar kind of Websites. This will help in getting more kind of reviews on a particular product or many. This will move towards the accuracy of the result of a product. A product can be reviewed from many sites at a time as well. Languages can also be updated as required. Many languages can be introduced as well which the comments will be stored and processed accordingly.

## References

[1] Xiaohui Yu, Yang Liu, Jimmy Xiangji Huang, Aijun An, “Mining Online Reviews for Predicting Sales Performance: A Case Study in the Movie Domain”, jounal name, year, VOL. 24, NO. 4, APRIL 2012

[2] Rodrigo Moraes, Joao Francisco Valiati, Wilson P. Gaviao Neto, Document-level sentiment classification: An empirical comparison between SVM and ANN, 40 (2013) 621–633

[3] Ms. Neha S. Joshi, Prof. Mrs. Suhasini A. Itkar, A Feature Dependent Method for Sentiment Analysis to understand User Context in Web, cPGCON 2014

[4] S. ChandraKala and C. Sindhu, Opinion Mining And Sentiment Classification: A Survey, IJSC_Vol3_Iss1_Paper4_420_427

[5] Jianxing Yu, Zheng-Jun Zha, Mengwang, Tat-Seng Chua, Aspect Ranking: Identifying Important Product Aspects From Online Consumer Reviews, ACLWEB P11-1150

## Author Profile

**Mohsin Nadaf** is pursuing his Bachelors Degree in Information Technology from Savitribai Phule Pune University, Pune. He is Data Science enthusiast and till now he has published a research paper on “Data Mining In Telecommunication” in Journal IJACTE ISSN (Print):2 319 – 2526, Volume-2, Issue-3, 2013 www.irdindia.in/journal_ijacte/pdf/vol2_iss3/16.pdf and also published an online PowerPoint Presentation on Slideshare (http://www.slideshare.net/MohsinNadaf2/data-mining-in-telecommunication)

**Sunil Lahane** is pursuing his Bachelors Degree in Information Technology from Savitribai Phule Pune University, Pune. He is having knowledge about Li-Fi Technology (Light Fidelty) and has given the paper presentation on the same.

**Akshay Deshpande** too is pursuing his Bachelors Degree from Savitribai Phule Pune University, Pune, in Information Technology.

**Sneha Tirth** is an Assistant Professor in Department of Information Technology, Trinity College of Engineering & Research, Savitribai Phule Pune University, Pune. She has over 7 years experience in the field of teaching.