

Product Opinion Mining Using Sentiment Analysis on Smartphone Reviews

Shilpi Chawla, Gaurav Dubey, Ajay Rana

Amity University Uttar Pradesh, Noida, India

Shilpichawla13@gmail.com, gdubey@amity.edu, ajay_rana@amity.edu

Abstract: An unexpected increase in the application of social web sites has led to the need in development of a very robustful and trust worthy systems and varied machinery forms in order to analyze such large forms of data and their ware houses which is being collected from various industries. We are always eager and excited to know what people think what they feel and perceive about various aspects of living and non-living beings. In need to understand and analyze various traits of behavior and the varying personality there is a need of opinion mining. It is a collection of all the extracted information from all the available resources and possible instances which could be in the form of hidden emotions, paragraphs and could be text , urban language and various un-identified representations. It is not only related topics like sensex, politics, finance and other meaningful words , but represents many things in the wide area of application. Almost all the sites have provided the option to display and present various medias and their views on various real life events. They can share various aspects of philosophy too and many more. They can represent various opinions on various streams of life which has really become crucial in our daily life to analyze the pattern in which things are exponentially growing. This research paper provides you with sentimental analysis of various smart phone opinions on smart phones dividing them Positive, Negative and Neutral Behavior. This is basically being obtained by studying the various posts being posted by varied number of users considering their areas of interest categorizing the smart phones. Analysis of plenty of words coupled in a sentence represent various sentiments of users and the various experiences and impact that product has given them. This analysis compiles a structural modeling approach and Bayesian Interface system to identify the polarity of the opinion which subsequently classifies positive and negative opinions.

Keywords: Text Mining--Opinion mining--sentiment analysis--text analysis--Sentiment Mining.

I. INTRODUCTION:

1. Sentiment Classification

This type of classification also known as Polarity Classification . It is used to analyze in ample amount of text in which every sample is being labeled as either positive, negative or neutral sample depending on the overall response received that is being expressed in that particular text.

This type of classification can be carried out at various levels to gain surety over the produced opinion on the set of texts.

The higher level of classification in the pyramid will be the difficult level amongst all the levels as it would be having the opinions as well as the things.

1.1. Document Level Sentiment Classification:

At this level of classification, the sentiment is being analyzed in the whole document and is categorized as positive, negative or neutral or is just objective. In this level of classification , the utmost challenge is to extract that subjective test which could be inferred in the complete sentiment of the overall document.[2] Thus for polarity classification it is supposed that the document should be focused on a single object and should contain opinion from a single holder so as not to have any conflicts in the opinions of various opinion holders.

1.2. Sentence Level Sentiment Classification:

At this level of sentiment analysis the sentiments are being analyzed in the text appearing at the sentence level. It is a considered to fined grained classification as compared to the classification made at document level in which the polarity was expressed only in three forms either positive, negative or neutral format.[2] This type of sentence classification can be done in the below mentioned 2 ways- Grammatical Syntactic Approach or semantic approach . [33] The former approach considers the grammatical hierarchy or structure of the various sentences occurring in the text by taken into consideration various parts of speech. And the latter one counts the occurrences of positive as well as negative words in order to propose the polarity of the analyzed sentence.

1.3. Feature Level Sentiment Classification:

Product features or attributes referred to as components are being analyzed under this classification covering the sentence level as well as document level. Opinion is being derived from the already extracted features.

General flow of the complete process-

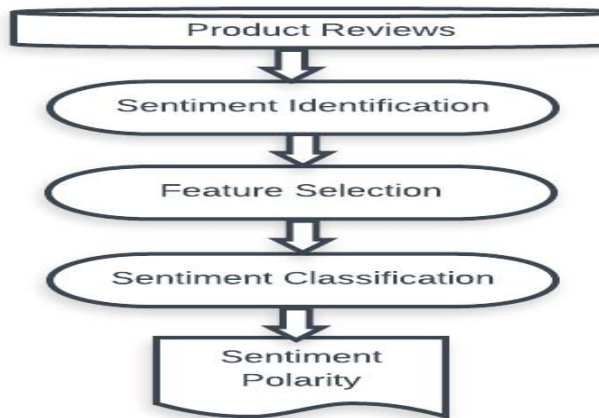


Figure 1: Sentiment Analysis process on product reviews.

Sentiment analysis has following features

□ **Term presence and their frequency:**

Includes UNI-grams or N-grams along with their frequency of the presence. In terms of Movie reviews, the concluded sentiment analysis by Pang et al. [14] was that uni-grams always give better performance results as compared to that than of bi-grams. Moreover, Dave et al. [16] concluded that bi-grams and tri-grams proved to give much higher polarity classification in case of product reviews.

□ **Part of speech information:**

Tagging specific words to their respective POS tags supports in disambiguating manner which gives help in feature selection [17]. Every word in a sentence is being assigned a label which represents the role or the position held by that word in the grammatical sense. Also, POS Tags could be helpful in identifying adjectives and various adverbs which can be rigorously used in sentiment indicators or classifiers [13].

□ **Negations:**

Negative words have the potential to reverse the selected sentiment so they should always be taken into consideration[17].

□ **Opinion words and phrases:**

Phrases or words(also known as opinion words) are those words which represents some positive or negative sentiments. In order to detect the semantic orientation of the mentioned words in the sentences , Lexicon and statistical based methods are the best ones. Taking an example of WordNet which was used by Hu and Liu et al. [4] to examine the polarity of the various extracted adjectives. Machine Learning Techniques widely uses Maximum Entropy , Support Vector Machines and Naïve Bayes Algorithms and have achieved success in the categorization of text.

Commonly used learning techniques in the natural language processing area are the K nearest

neighborhood, winnow classifier, centroid classifier, ID3, C5, and the N-gram model.

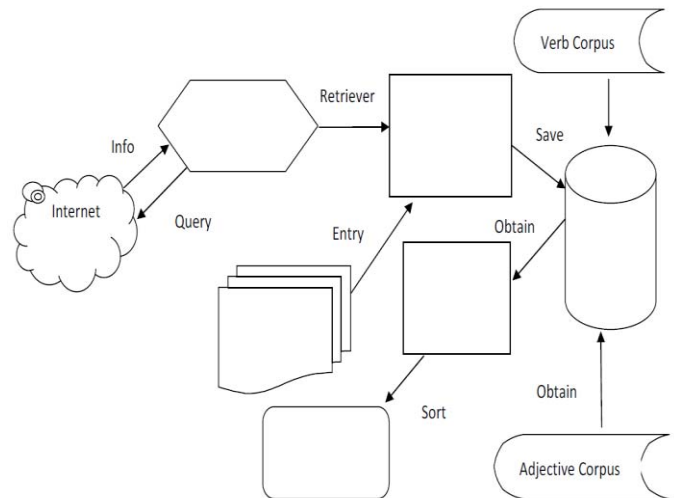


Figure 2: Opinion Mining Process.

II. RELATED WORK

The word “sentiment” which is used in context to the automated analyzing of the text which is being evaluated along with the conclusion or the judgment in terms of analytics that has already appeared in a paper published in 2001 by “Das” and “Chen” [2]. On the further evaluation and research , the same was opted and enhancement was made in it by “Turney” [3] and “Pang et al [4]. In the upcoming years the same concept was adopted by various other researchers as well ie “Nasukawa” & “Yi” [5] and “Yi et al.” [6]. Above mentioned events collaboratively can summarize the effect and importance of “the sentiment analysis” among various domains and communities which are basically focused on NLP(Natural Language Processing). A numerous number of papers have been published on the term “sentiment—analysis” which are specialized in classifying, collecting and categorizing the reviews collected by customer in two – either positive or negative. Sentiment analysis in all the published papers have been carried out at almost all the levels of categorization ie document, sentence, and attribute levels.

Supervised Learning Techniques can be applied to “Naïve Bayes”, “Support Vector Machine” and many more. As mentioned in [7] , the authors have tried to provide a comparison of “Naïve Bayes” and “Support Vector Machines” with the use of an Artificial Neural Network Methods in terms of document level classification of sentiment analysis.

The authors or the researchers contributed to the comparison of the sentiment – classification – literature survey-

- a) Dominant comparison survey along with an efficient approach which computationally proposed using an ANN methodology in the same respect (SVN and NB)
- b) a mixture of real life contexts including the positives and the negatives in un-equal.
- c) A dataset of new movie reviews was used to produce an evaluation based on performance using ANN approach [1].

Many supervised methodology was used in a classic manner for the selection of features and conducted various experiments manually using ANN [1], along with the reviews which were short listed from AMAZON web-site in Cameras, GPS and Books.

They have found that ANN has performed tremendously outstanding SVM specially when the input data set is highly un-balanced in respect to the accuracy of the classification of the obtained dataset of various smartphones reviews ((Pang & Lee, 2004).

When the balanced data set was referred that ANN was over SVM with approx 13 test cases, while SVM was over ANN in only 2 test cases.

Overall conclusion is that ANN has over achieved the accuracy in classification in all the data sets. SVM in the experiments has been graded as less effected by the Noisy data as compared to ANN when there is an increase in the imbalance.

“Pang et al. [4]” has carried out three commonly used Machine learning Methods ie Maximum Entrophy, Naïve Baiyesian and SVM on the data that is being collected from smartphones present on Internet.

The extracted ratings were being classified into 3 broad chunks: the positive ratings, the negative ratings, and the neutral ones too. Moreover the concentration was more on the pros and the cons. They accepted the standard chunks of the feature framework.

A far as performance is considered, worst results are been produced by Naïve Bayes Theorem while the SVM’s give the best results, more over the differences are not that large.

A Supervised Learning can too be converted into an Un – supervised Supervised learning method as reported in [3,8] by many researchers.

In the paper presented [3] authors have used an un-supervised learning algorithms to classify an opinion | result | review as being recommended or not.

Algorithm has been widely divided into 3 main broad categories –

Step-1:

Extracting of those phrases or idioms in the sentences that represent an adverb or adjective.

In many direction there may be some isolated form of adjectives which may indicate in a subjective manner and could be insufficient enough to find in an semantic manner.

Step-2:

Usage of PMI-R algorithm in order to estimate the orientation (semantic analysis) of the phases that were extracted in step-1

Step-3:

Last step is to find out the average of the various semantic orientation found out in the second step in the mentioned review and the same is classified as positive or negative review depending on the recommendation.

In the paper [8], the researchers proposed a learning approach in a depth with the aim to extract the way to extract a represent able format for each review using an un-supervised learning method.

Algorithm based on finding some intermediate format of representing the whole structure in an hierarchical manner. The data set has been taken from a very well – known famous site ie Amazon.com which proposed rules in lacs, out which around 50 products were reviewed and were being categorized as a positive or a negative recommendation.

III. DATA SET

A diverse data set was used to evaluate the results using Naïve Bayes Technique. The following set of data was obtained from a very reputed merchant like Amazon that act as a epic for opinion mining.

Collected various beneficial customer reviews about various smart phones including One Plus, Micromax, Samsung, Nokia, Lava and many more.

About 1000 reviews collected including positive negative and the rest of them were neutral ones. These reviews consists of approximately 3000 sentences

We have deleted some inconsistent data and annoying data and sentences from the set to be more precise and specific.

Table 1: depicts most of the typical features and the sentences number in the corpus.

| | Samsung | MI | OnePlus |
|-----------|---------|------|---------|
| Reviews | 200 | 500 | 300 |
| Sentences | 600 | 1500 | 900 |

Table1: Representing The count of reviews and corresponding sentences.

Total distribution of the data set in terms of positive and negative reviews can further be classified as-

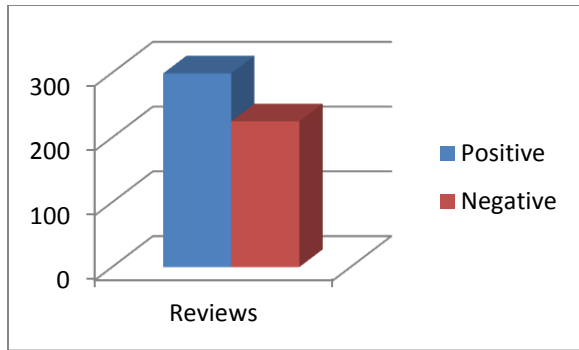


Figure-3 : Positive Vs Negative Reviews.

5. Implementation:

Flow chart of the process done can be summed by including various steps-

After installing all the necessary libraries and packages needed to execute the task following are the actions done-

Step1:

Set up the Current working directory in which R is being installed and activate the current session.

Read the CSV file which consists of all the reviews.

Step2:

Generate the factor type of the collected data that is categorized under positive and negative.

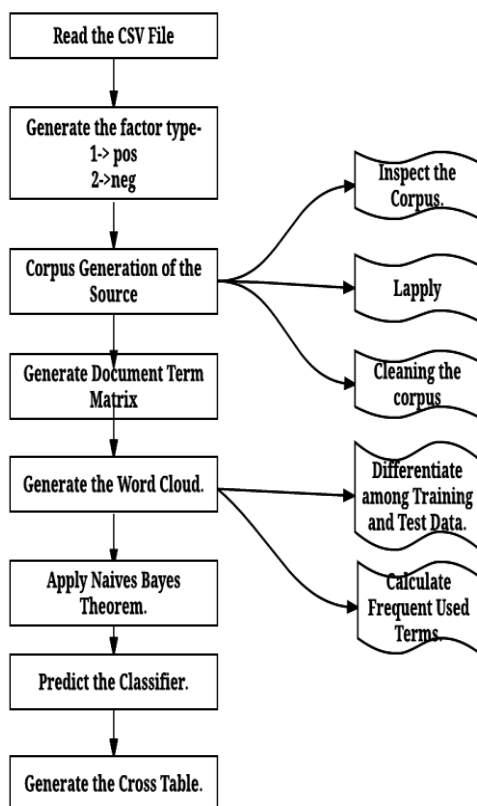


Figure 4: Flow of the process.

Step3:

Used Corpus based approach of Lexicon Tree to go through the sentiment analysis process.

- Inspecting the corpus
- Lapply to the corpus
- Cleaning the generated corpus(by removing punctuation marks, white spaces, duplicates and null values)

Step4:

Generate the Term Document Matrix which seems like as-

```
<<DocumentTermMatrix (documents: 524, terms: 1727)>>
```

```
Non-/sparse entries: 7380/897568
```

```
Sparsity : 99%
```

```
Maximal term length: 25
```

```
Weighting : term frequency (tf)
```

Step5:

Form the Word Cloud depending on the frequency of the words encountered in the textual data. Applying rainbow colors to the will give an attractive feel to the user.



Figure 4: Word Cloud from for Document Matrix.

Creating the subsets for Training Data and Testing data. The data set assigned training should be well equipped to evaluate to conclude the relevancy of the data set.

Accumulated Word cloud for Negative reviews with a minimum frequency of 6 is as-



Figure 5: Word Cloud for Negative Reviews

Word Cloud for Positive Reviews with a minimum frequency of 5 -



Figure 6: Word Cloud for Positive Reviews

| | | | | |
|----------|------------|---------------|--------------|-----------|
| "bad" | "get" | "good" | "now" | "oneplus" |
| "review" | "used" | "using" | "will" | "battery" |
| "charge" | "day" | "life" | "long" | "nice" |
| "phone" | "user" | "amazing" | "bought" | "got" |
| "time" | "via" | "always" | "android" | "buy" |
| "camera" | "can" | "design" | "experience" | "fast" |
| "first" | "front" | "give" | "great" | "happy" |
| "makes" | "money" | "one" | "product" | "provide" |
| "ram" | "software" | "worth" | "box" | "instead" |
| "budget" | "decent" | "even" | "light" | "like" |
| "look" | "perfect" | "performance" | "phones" | "price" |

Figure 7: Words with minimum frequency of 20 appearing in the Document Term Matrix

Step 6:

Application of Naïve Bayes theorem for textual classification.

It depends on the posterior probability of any class selected which designates the distribution of the words in the referred document.

$$P(\text{label}|\text{features}) = \frac{P(\text{label}) * P(\text{features}|\text{label})}{P(\text{features})}$$

$$P(\text{label}|\text{features}) = \frac{P(\text{label}) * P(f_1|\text{label}) * \dots * P(f_n|\text{label})}{P(\text{features})}$$

Step 7:

Generate the cross table to compare the actual values and predicted value producing the Row as well as column total.

| Predicted/ Actual | Negative | Positive | Row total |
|----------------------|----------------------|---------------------|--------------|
| Negative | 5 0.1 0.078 | 45 0.9 0.45 | 50 0.305 |
| positive | 59 0.518 0.922 | 55 0.482 0.55 | 114 0.695 |
| Column Total | 64 0.39 | 100 0.61 | 164 |

Figure 8: Cross table of predicted and actual values using Naïve Bayes Classifier

According to the shown classifier the model produced an accuracy of around 40% which is not that much satisfactory to mine a data with this kind of classifier.

In order to produce much more refine results , we have opted another approach to classify our unstructured data.

So SVM (Support Vector Machine) approach which is one step ahead the Nave Bayes Classifier and overcomes all the drawback of the existing classifier.

On the existing data , we have applied another classifier which works as follows-

Document Term Matrix Formed for SVM is-
 <<DocumentTermMatrix (documents: 524, terms: 1735)>>

Non-/sparse entries: 7656/901484

Sparsity : 99%

Maximal term length: 16

Weighting : term frequency (tf)

The training model this formed after having a ratio of around 70:30 against the training data as well as the test data is as follows-

Call:

```
svm.default(x = container@training_matrix, y = container@training_codes,
            kernel = kernel, cost = cost, cross = cross,
            probability = TRUE,
            method = method)
```

Parameters:

SVM-Type: C-classification

SVM-Kernel: linear
 cost: 1
 gamma: 0.0005763689

Number of Support Vectors: 172

| SVM_PROB | |
|----------|---------|
| Min. | :0.5160 |
| 1st Qu. | :0.8752 |
| Median | :0.9619 |
| Mean | :0.9068 |
| 3rd Qu. | :0.9799 |
| Max. | :1.0000 |

| SVM_LABEL | |
|-----------|-----|
| neg: | 51 |
| pos: | 174 |

Cross table thus generated between the actual and predicted values is as follows-

| Predicted/ Actual | Negative | Positive | Row total |
|----------------------|----------|----------|-----------|
| Negative | 49 | 5 | 54 |
| | 0.907 | 0.093 | 0.241 |
| | 0.662 | 0.033 | |
| positive | 25 | 145 | 170 |
| | 0.147 | 0.853 | 0.759 |
| | 0.338 | 0.967 | |
| Column Total | 74 | 150 | 224 |
| | 0.33 | 0.67 | |

Figure 9: Cross table of predicted and actual values using SVM Classifier

The accuracy of the model thus produced is 90% which is far better than that of Naïve Bayes Classifier.

IV. CONCLUSION

With this data set we have just tried to examine an unstructured Smartphone reviews that are being collected from a very well-known marketing site termed as Amazon. The data set was being classified amongst the classes that is the positive reviews and the negative reviews. The data set had been under gone through two well-known machine learning techniques ie Naïve Bayes and Support Vector machine Classifier. The overall accuracy of the classifier thus trained using Naïve Bayes Classification technique was around 40% which was quite un-satisfactory to deal with and to rely on. So in order to have a reliable trained model which is able to classify the data according to our needs we have opted another approach which is SVM. The same data set which when parsed under this approach [produces good result with an accuracy of around 90%. Hence the SVM

approach seems to be much better and reliable to work with the mining of data.

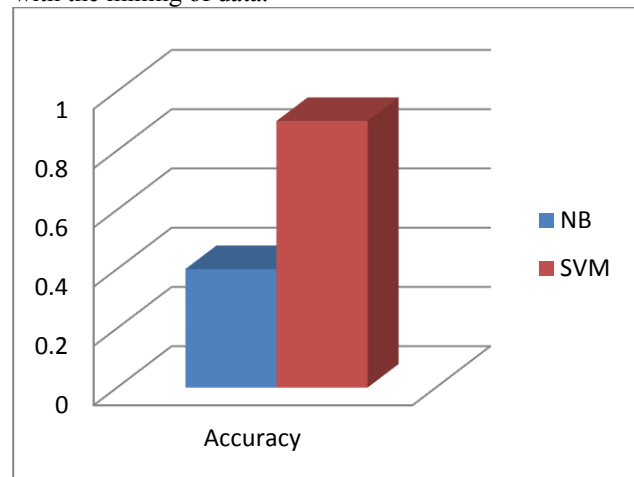


Figure 10: Comparison of the accuracy of Naïve Bayes Vs SVM

References

- [1] Luciano Barbosa and Junlan Feng. 2010. Robust sentiment detection on twitter from biased and noisy data. Proceedings of the 23rd International Conference on Computational Linguistics: Posters, pages 36–44.
- [2] Alec Go, Richa Bhayani, and Lei Huang. 2009. Twit-ter sentiment classification using distant supervision. Technical report, Stanford.
- [3] B.Pang and L. Lee. 2004. A sentimental education: Sentiment analysis using subjectivity analysis using subjectivity summarization based on minimum cuts. ACL.
- [4] Alexander Pak and Patrick Paroubek. 2010. Twitter as a corpus for sentiment analysis and opinion mining. Proceedings of LREC.
- [5] Bing Liu, Minqing Hu and Junsheng Cheng. "Opinion Observer: Analyzing and Comparing Opinions on the Web." Proceedings of the 14th International World Wide Web
- [6] B.Pang and L. Lee. 2004. A sentimental education: Sentiment analysis using subjectivity analysis using subjectivity summarization based on minimum cuts. ACL.
- [7] Bing Liu, Minqing Hu and Junsheng Cheng. "Opinion Observer: Analyzing and Comparing Opinions on the Web." Proceedings of the 14th International World Wide Web conference (WWW-2005), May 10-14, 2005, Chiba, Japan.

- [8] Zhongwu Zhai, Bing Liu, Hua Xu and PeifaJia. "Clustering Product Features for Opinion Mining." Proceedings of Fourth ACM International Conference on Web Search and Data Mining (WSDM-2011), Feb. 9-12, 2011, Hong Kong, China
- [9] Aue, A. and M. Gamon. Customizing sentiment classifiers to new domains: a case study. In Proceedings of Recent Advances in Natural Language Processing (RANLP-2005), 2005.
- [10] Blitzer, J., M. Dredze, and F. Pereira. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In Proceedings of Annual Meeting of the Association for Computational Linguistics (ACL-2007), 2007.
- [11] Bo Pang and Lillian Lee, Opinion Mining and Sentiment Analysis, Computer Science Department, Foundations and Trends in Information Retrieval Vol. 2, No 1-2 (2008) 1-135, 2008, Cornell University, Ithaca, NY 14853, U.S.A.
- [12] Daniel Lowd and Pedro Domingos, Naïve Bayes Model for Probability Estimation, Department of Computer Science and Engineering, University of Washington, Seattle, WA 98195-2350, USA.
- [13] Behrooz Parhami, IEEE Transactions on reliability, Voting Algorithms, Vol. 43, No. 4, 1994 December, University of California, Santa Barbara.
- [14] Simon Tong, Support Vector Machine Active Learning with Applications to Text Classification, Journal of Machine Learning Research (2001) 45-66, Computer Science Department Stanford University Stanford CA 94305-9010, USA
- [15] Vincent J. Della Pietra, Adam L. Berger, Stephen A. Della Pietra, A Maximum Entropy Approach to Natural Language Processing(1996).
- [16] Guo-Xun Yuan, Chia-Hua Ho and Chih-Jen Lin, An Improved GLMNET for L1-regularized Logistic Regression(2011).
- [17] S. B. Kotsiantis, Supervised Machine Learning: A Review of Classification Techniques, Informatica 31 (2007) 249-268, Department of Computer Science and Technology.