2017 International Conference on Intelligent Computing and Control (I2C2)

Sentiment Analysis of Customer Product Reviews Using Machine Learning

Zeenia Singla, Sukhchandan Randhawa, Sushma Jain
Computer Science and Engineering Department
Thapar University
Patiala, India
zeenia_77@yahoo.com

***Abstract*—Today, digital reviews play a pivotal role in enhancing global communications among consumers and influencing consumer buying patterns. E-commerce giants like Amazon, Flipkart, etc. provide a platform to consumers to share their experience and provide real insights about the performance of the product to future buyers. In order to extract valuable insights from a large set of reviews, classification of reviews into positive and negative sentiment is required. Sentiment Analysis is a computational study to extract subjective information from the text. In the proposed work, over 4,000,00 reviews have been classified into positive and negative sentiments using Sentiment Analysis. Out of the various classification models, Naïve Bayes, Support Vector Machine (SVM) and Decision Tree have been employed for classification of reviews. The evaluation of models is done using 10 Fold Cross Validation.**

***Index Terms*—Big data, text mining, text classification, sentiment analysis, machine learning.**

I. INTRODUCTION

With an ever increasing demand of smart phones, the mobile phone market is expanding at an exponential pace. With such a boom in the smart-phone industry, there is a need to realize the holistic review of the brand and the model of phone. There are numerous brands present in the market, out of which some are dominant and occupy quite a big part of the industry. For instance, Samsung, Apple, etc. are names associated with brands which are famous throughout the world. Electronic commerce plays a vital role in increasing the sales of the mobile phones and influencing consumer buying patterns. Reviews available on such e-commerce platforms act as a guiding tool for the consumers to make informed decisions. Retail websites like Amazon.com offer different options to the reviewers for writing their reviews. For instance, the consumer can provide numerical rating from 1 to 5 or write comments about the product.

As there are innumerable products manufactured by many different brands, so providing relevant reviews to the consumers is the need of hour. Number of reviews associated with a product or a brand is increasing at an alarming rate, which is no less than handling the big data. Classifying the reviews on the basis of sentiment of customers into positive and negative sentiment provides sentiment orientation of the review, hence results in better judgement. Segregation of reviews on the basis of their sentiment can help future buyers to evaluate positive and negative feedback constructively and reach at better decisions as per their requirements. This evaluation acts as a testimony to the users who are looking to know the details and specifications of the smartphones; thereby increasing user credibility.

In this research, unstructured data of Mobile Phone Reviews have been extracted from Amazon.com. It has been filtered to remove noisy data and has been pre-processed to evaluate sentiment of the reviews using supervised learning. The reviews have been classified using machine learning classification models like Naïve Bayes, Support Vector Machine (SVM) and Decision Tree and have been cross validated to find the best classifier for this purpose.

II. RELATED WORK

Data analytics has enabled to unravel the hidden patterns in data. Volume, Velocity and Variety define Big Data [1]. Veracity and Value are two more Vs that play an important role in Big Data. The volume and the relentless rapidity at which data are being generated every day are exceeding the computing capacity of many IT departments. E-commerce websites are loaded with a large set of diverse reviews for various products. These reviews can be used to determine consumer behavior and make informed decisions. Reviews can be both structured and unstructured. Valuable business insights can be fetched by filtering the irrelevant data. Big Data has enabled businesses to flourish and improvise on the basis of evidence rather than intuition. It aids in gaining insights on better targeted social influencer marketing, segmentation of customer base, recognition of sales and marketing opportunities, detection of fraud, quantification of risks, better planning and forecasting, understanding consumer behavior, etc. [2].

Sentiment analysis implies identifying sentiment of reviews on the basis of positive, negative and neutral connotations. Sentiment analysis can be performed at three levels, viz.

document level, sentence level and phrase level [3]. A lot of prior research has been done in this field where words and phrases have been classified with prior positive or negative polarity [4]. This prior classification is helpful in many cases but when contextual polarity comes into the picture, the meaning derived from positive or negative polarity can be entirely different. The contextual polarity of the phrases was taken into consideration and ambiguity was removed [5]. Also, a refined method has been devised to establish contextual polarity of phrases by using subjective detection that compressed reviews while still maintaining the intended polarity [6].

Delineated study has been conducted on tweets available on Twitter, movie reviews to build the grounds on sentiment analysis and opinion mining. A sentiment classifier has been built to categorize positive, negative and neutral sentiments not only in English but also for other languages using corpus from Twitter [7]. The polarity of smartphone product reviews has been found only on the basis of positive and negative orientation of the review [8]. A system has been built using support vector machine where sentiment analysis is carried out by taking into consideration sarcasm, grammatical errors and spam detection [9]. An enhanced Naïve Bayes model by combining methods like effective negation handling, word n-grams and feature selection has been utilized to conduct sentiment analysis [10].

Sentiment analysis is not only confined to the English language but has been implemented for various languages. Sentiment analysis of Chinese text by implementing four feature selection methods and five classifiers viz. Centroid classifier, K-nearest neighbor, Window classifier, Naïve Bayes and SVM has been done [11]. Through this learning paradigm it was concluded that SVM outperforms all the other learning methods in terms of sentiment classification. Sentiment analysis on travel reviews using three machine learning models namely, Naïve Bayes, SVM and character based N-gram model has been performed in which SVM and N-gram approaches have better performance than Naïve Bayes [12]. It has been observed that in maximum number of cases SVM showcases best performance in comparison to other classification models.

Sentiment Analysis is not limited just to reviews or Twitter data but is also applicable on stock markets ([13], [14]), news articles [15] or political debates [16]. Sentiment analysis can be used to flourish consumer products related business [17]. It uses rule-based approach for sentiment analysis to extract topic words of negative opinion sentences and thus promote the competitors of the products receiving negative feedback. Similarly, relevant ads based on a person's liking or disliking are displayed on various blogging sites targeting bloggers. Blogger-Centric Contextual Advertising Framework has been concocted to determine users' personal interests and display those ads that intersect with them [18].

The organization of paper is as follows: Section 3 explicates the dataset used and the approach followed to conduct analysis. Section 4 demonstrates the sentiment of the text, i.e., positive or negative. Section 5 shows the three classification models used to classify reviews. Section 6 explains the technique employed for data balancing. Section 7 presents the experimental results of the models and cross validation of predictive accuracy of models. Finally, section 8 concludes the proposed work and describes its future scope.

III. FRAMEWORK

The proposed framework of the research work is conducted in three different modules as shown in Fig. 1.

A. Dataset and its Features

The first module includes data collection and pre-processing of data. A large sample of online reviews is collected from the e-commerce giant Amazon.com. The data set consists of over 400,000 reviews for approximately 4500 mobile phones. It includes six features as explained in table 1.

B. Approach

The approach followed by the proposed framework is described in Fig. 1. Initially, the experimental data is collected from an e-commerce website Amazon.com. Each data set is in the Comma Separated Values (CSV) file format and available as supplement. In the second step, data are pre-processed to remove stop words, punctuation marks, whitespaces, digits and special symbols. 'tm' package [19] is employed for text mining. In the third step, feature selection is performed to extract relevant features from the data set. In the given data set out of the six features, only three features, i.e, Product Name, Brand Name and Reviews have been considered. In the fourth step, sentiment orientation of the reviews is determined. In the fifth step, 'Pos/Neg' tags are appended to the dataset to corresponding to each review to conduct supervised learning. The sixth step involves training and testing the classified data using Naïve Bayes, SVM and Decision Tree models. The accuracy so obtained is validated using 10-fold cross validation.

IV. SENTIMENT ANALYSIS

An inbuilt package named 'Syuzhet' [20] has been used to conduct Sentiment Analysis. This package encompasses three sentiment dictionaries. NRC sentiment dictionary is used to extract eight different emotions and their corresponding valence in the text including all the reviews. Ten different emotions represented are: anger, anticipation, disgust, fear, joy, sadness, surprise, trust, positive and negative.

TABLE 1: FEATURES INCLUDED IN THE DATA SET

Feature	Description
Product Name	Model name of mobile phone
Brand Name	Manufacturing brand
Price	Price of the mobile in dollars
Rating	User rating between 1 to 5
Reviews	User reviews provided for every mobile phone
Review Votes	Number of people who found the review helpful

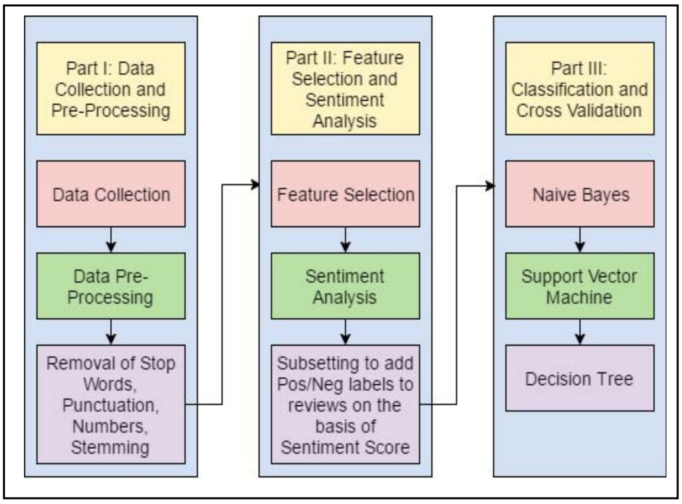


Fig. 1. Proposed Framework

In this research, only positive and negative sentiment orientation has been considered for classification of reviews. Figure 2 depicts the percentage of positive and negative reviews present in the dataset using bar graph.

V. CLASSIFICATION

Classification is the process of classifying reviews on the basis of their sentiment into two classes: Positive and Negative. The sentiment score established for each positive and negative polarity using NRC sentiment dictionary is then added to the reviews dataset. This individual score is used to calculate the overall polarity as given by Eq. 1 of the sentiment as shown in Fig. 3.

$$\text{Polarity} = \text{PositiveScore} - \text{NegativeScore} \quad (1)$$

After calculating the polarity corresponding to each review, different subsets of positive and negative sentiment are formed having polarity 0 to 10 and -1 to -10 respectively. The subsets so obtained are combined together in a CSV file along with

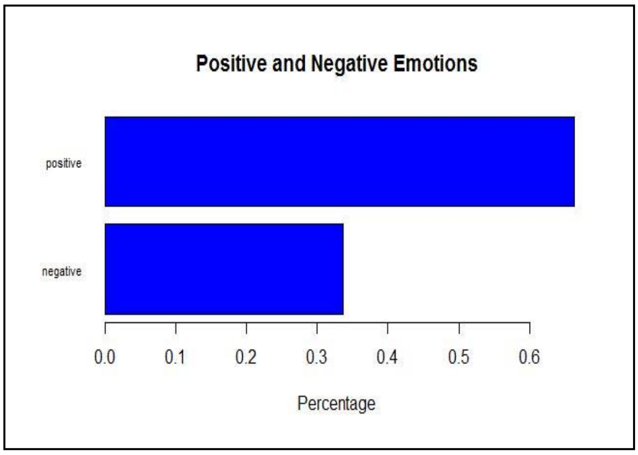
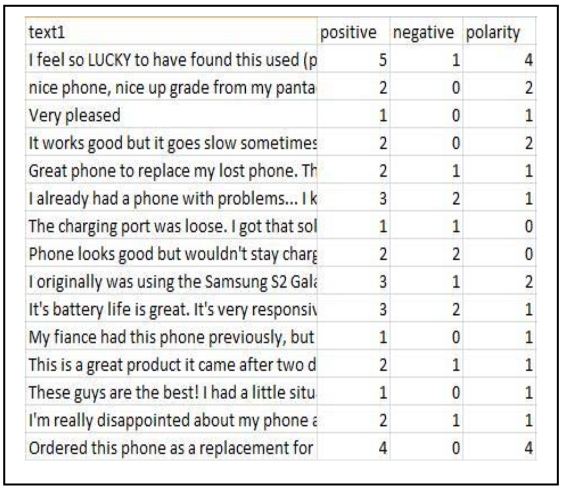


Fig. 2. Percentage of Positive and Negative Emotions



text1	positive	negative	polarity
I feel so LUCKY to have found this used (p	5	1	4
nice phone, nice up grade from my panta	2	0	2
Very pleased	1	0	1
It works good but it goes slow sometimes	2	0	2
Great phone to replace my lost phone. Th	2	1	1
I already had a phone with problems... I k	3	2	1
The charging port was loose. I got that sol	1	1	0
Phone looks good but wouldn't stay charg	2	2	0
I originally was using the Samsung S2 Gal	3	1	2
It's battery life is great. It's very responsi	3	2	1
My fiance had this phone previously, but	1	0	1
This is a great product it came after two d	2	1	1
These guys are the best! I had a little situ	1	0	1
I'm really disappointed about my phone a	2	1	1
Ordered this phone as a replacement for	4	0	4

Fig. 3. Overall Polarity of the Review

the Pos/Neg tags. The classified data is shown in Fig. 4.

A. Classifiers

The classification models selected for categorization of text are: Naïve Bayesian, Support Vector Machine and Decision Tree.

1) Naïve Bayesian Classifier

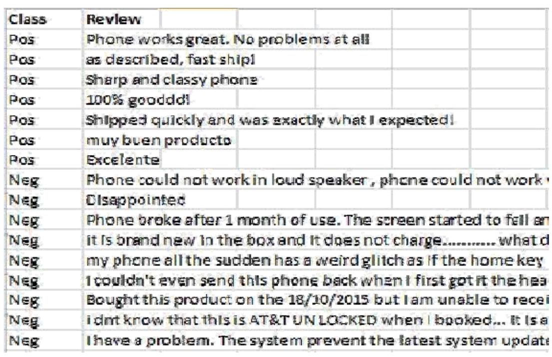
A statistical classifier that maps input feature vectors to output class labels [21]. For a set of training data D, each row is represented by an n-dimensional feature vector, $X = x_1, x_2, \ldots, x_n$. There are K classes, $K_1, K_2, \ldots, K_m$ in the output class label. For every tuple X, the classifier will predict 2 as given by Eq. 2 that X belongs to K_i if and only if: $P(K_i|X) > P(K_j|X)$, where $i, j \in [1,m]$ and $i \neq j$.

$$P(K_i|X)=\Pi^{n}_{a=1}P(x_k|K_i) \quad (2)$$

2) Support Vector Machine (SVM)

SVM is used for a labelled training data that categorizes testing dataset using an optimal hyperplane [22]. A hyperplane is separates data of one class from another which is defined as given in Eq. 3.

$$W - X+b=0 \quad (3)$$



Class	Review
Pos	Phone works great. No problems at all
Pos	as described, fast ship!
Pos	Sharp and classy phone
Pos	100% goodddd!
Pos	Shipped quickly and was exactly what I expected!
Pos	muy buen producto
Pos	Excelente
Neg	Phone could not work in loud speaker , phone could not work
Neg	Disappointed
Neg	Phone broke after 1 month of use. The screen started to fall ar
Neg	it is brand new in the box and it does not charge........... what d
Neg	my phone all the sudden has a weird glitch as if the home key
Neg	I couldn't even send this phone back when I first got it the hea
Neg	Bought this product on the 18/10/2015 but I am unable to recei
Neg	i dnt know that this is AT&T UN LOCKED when I booked... It is a
Neg	I have a problem. The system prevent the latest system updat

Fig. 4. Classified Data

3) Decision Tree

A hierarchical tree structure encompassing decision nodes for representing attributes and edges for denoting attribute values. This representation in the form of a tree allows to construct decision rules that classify new instances of the data [23].

VI. DATA BALANCING

Figure 2 represents that the number of positive reviews is more than double the negative reviews. This implies that the data are imbalanced as the target variable has imbalanced proportion of classes. Therefore, running machine learning models for classification would yield biased predictions and misleading accuracies. To avoid such scenarios, data balancing is employed. There are different methods that can be used to transform imbalanced data into balanced data like undersampling and oversampling [26].

For treating imbalanced data, undersampling techinque has been used. Undersampling means to reduce the number of observations from majority class to balance the data set. The balanced data so obtained has almost equal number of positive and negative reviews.

VII. MODEL EVALUATION AND RESULTS

After appending the data with a class having positive or negative tags and removing imbalanced proportion of classes, a random sample of 3000 reviews is taken to train and test the dataset on three classifiers.

A. *Cross Validation*

Cross Validation is a model evaluation parameter that demonstrates the ability of the system to make new predictions accurately. In the proposed work, K-fold cross validation has been implemented to determine the efficiency of the models. In K-fold cross validation, the dataset is divided into k subsets which is repeated k times. For every iteration, k subset is used as the training sample and k-1 subsets are used for testing. The three models are cross validated 10 times. Table 2 shows the cross validation of the three models for ten runs. It is observed that the accuracy of all the three models in all the iterations varies in the range of ±10.

The scatter plot in Fig. 5 elucidates that the SVM model reaches the highest accuracy mark of 81.75 among all the models for a number of iterations. Naïve Bayes model has the lowest accuracy of 64.57 among the three models. The graph clearly depicts that SVM model has the best accuracy out of the three models and Naïve Bayes model has the least predictive accuracy.

TABLE 2: CROSS VALIDATION

Runs	Naïve Bayes	SVM	Decision Tree
1	67.11	80.12	74.31
2	64.57	79.44	77.13
3	67.57	78.96	70.80
4	66.77	82.25	78.00
5	67.57	80.52	72.63
6	64.57	78.92	76.22
7	67.77	77.70	71.84
8	66.71	79.72	77.62
9	68.31	78.00	67.68
10	68.57	81.75	81.25

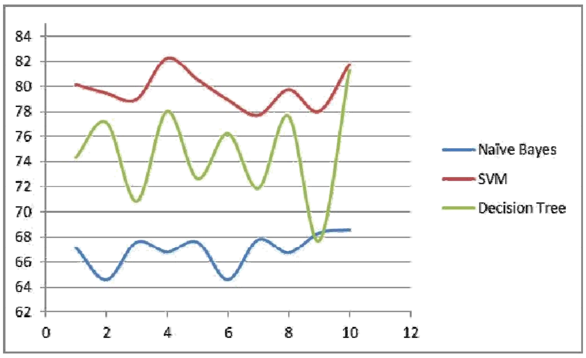


Fig. 5. Scatter Plot

B. *Performance Comparison*

After testing and training the dataset, their predictive accuracy is determined to find out which model is the best classifier for classifying reviews. As shown in table 3, SVM model has the best predictive accuracy among the three models and Naïve Bayes has the worst predictive accuracy.

TABLE 3: PREDICTIVE ACCURACY OF MODELS

Model Name	Accuracy
Naïve Bayes	66.95
SVM	81.77
Decision Tree	74.75

VIII. CONCLUSION AND FUTURE SCOPE

An evolutionary shift from offline markets to digital markets has increased the dependency of customers on online reviews to a great extent. Online reviews have become a platform for building trust and influencing consumer buying patterns. With such dependency there is a need to handle such large volume of reviews and present credible reviews before the consumer. Our research is aiming to achieve this by conducting sentiment analysis of mobile phone reviews and classifying the reviews into positive and negative sentiment. After balancing the data with almost equal ratio of positive and negative reviews, three classification models have been used to classify reviews. Out of the three classifiers, i.e., Naïve Bayes, SVM and Decision Tree, predictive accuracy of SVM is found to be the best. The accuracy results have been cross validated and the highest value of accuracy achieved was 81.75% for SVM among the three models.

In future, the work can be extended to perform multi-class classification of reviews which will provide delineated nature of review to the consumer, hence better judgement of the product. It can also be used to predict rating of a product from the review. This will provide users with reliable rating because sometimes the rating received by the product and the sentiment of the review do not provide justice to each other. The proposed extension of work will be very beneficial for the e-commerce industry as it will augment user satisfaction and trust.

REFERENCES

[1] S. Erevelles, N. Fukawa, and L. Swayne, "Big data consumer analytics and the transformation of marketing," Journal of Business Research, vol. 69, no. 2, pp. 897–904, 2016.

[2] P. Russom et al., "Big data analytics," TDWI best practices report, fourth quarter, pp. 1–35, 2011.

[3] S. Erevelles, N. Fukawa, and L. Swayne, "Big data consumer analytics and the transformation of marketing," Journal of Business Research, vol. 69, no. 2, pp. 897–904, 2016.

[4] V. Hatzivassiloglou and K. R. McKeown, "Predicting the semantic orientation of adjectives," in Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics. Association for Computational Linguistics, 1997, pp. 174–181.

[5] T. Wilson, J. Wiebe, and P. Hoffmann, "Recognizing contextual polarity in phrase-level sentiment analysis," in Proceedings of the conference on human language technology and empirical methods in natural language processing. Association for Computational Linguistics, 2005, pp. 347–354.

[6] B. Pang and L. Lee, "A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts," in Proceedings of the 42nd annual meeting on Association for Computational Linguistics. Association for Computational Linguistics, 2004, p. 271.

[7] A. Pak and P. Paroubek, "Twitter as a corpus for sentiment analysis and opinion mining." in LREc, vol. 10, no. 2010, 2010.

[8] M. WAHYUDI and D. A. KRISTIYANTI, "Sentiment analysis of smartphone product review using support vector machine algorithm-based particle swarm optimization." Journal of Theoretical & Applied Information Technology, vol. 91, no. 1, 2016.

[9] D. N. Devi, C. K. Kumar, and S. Prasad, "A feature based approach for sentiment analysis by using support vector machine," in Advanced Computing (IACC), 2016 IEEE 6th Interna-tional Conference on. IEEE, 2016, pp. 3–8.

[10] V. Narayanan, I. Arora, and A. Bhatia, "Fast and accurate sentiment classification using an enhanced naive bayes model," in International Conference on Intelligent Data Engineering and Automated Learning. Springer, 2013, pp. 194–201.

[11] S. Tan and J. Zhang, "An empirical study of sentiment analysis for chinese documents," Expert Systems with applications, vol. 34, no. 4, pp. 2622–2629, 2008.

[12] Q. Ye, Z. Zhang, and R. Law, "Sentiment classification of online reviews to travel destinations by supervised machine learning approaches," Expert Systems with Applications, vol. 36, no. 3, pp. 6527–6535, 2009.

[13] L.-C. Yu, J.-L. Wu, P.-C. Chang, and H.-S. Chu, "Using a contextual entropy model to expand emotion words and their intensity for the sentiment classification of stock market news," Knowledge-Based Systems, vol. 41, pp. 89–97, 2013.

[14] M. Hagenau, M. Liebmann, and D. Neumann, "Automated news reading: Stock price prediction based on financial news using context-capturing features," Decision Support Systems, vol. 55, no. 3, pp. 685–697, 2013.

[15] T. Xu, Q. Peng, and Y. Cheng, "Identifying the semantic orientation of terms using s-hal for sentiment analysis," Knowledge-Based Systems, vol. 35, pp. 279–289, 2012.

[16] I. Maks and P. Vossen, "A lexicon model for deep sentiment analysis and opinion mining applications," Decision Support Systems, vol. 53, no. 4, pp. 680–688, 2012.

[17] G. Qiu, X. He, F. Zhang, Y. Shi, J. Bu, and C. Chen, "Dasa: dissatisfaction-oriented advertising based on sentiment analysis," Expert Systems with Applications, vol. 37, no. 9, pp. 6182–6191, 2010.

[18] T.-K. Fan and C.-H. Chang, "Blogger-centric contextual advertising," Expert systems with applications, vol. 38, no. 3, pp. 1777–1788, 2011.

[19] I. Feinerer, "Introduction to the tm package text mining in r," 2015.

[20] M. Jockers, "Package âĂŸsyuzhetâĂŹ," 2016.

[21] K. M. Leung, "Naive bayesian classifier," Polytechnic University Department of Computer Science/Finance and Risk Engineering, 2007.

[22] S. Tong and D. Koller, "Support vector machine active learning with applications to text classification," Journal of machine learning research, vol. 2, no. Nov, pp. 45–66, 2001.

[23] I. Jenhani, N. B. Amor, and Z. Elouedi, "Decision Trees as possibilistic classifiers," International Journal of Approximate Reasoning, vol. 48, no.3, pp. 784-807, 2008.