Analysis of Customer Reviews Using Opinion Mining

Khine Zar Oo
University of Computer Studies, Magway
Magway, Myanmar
khinezaroodagayin@gmail.com

Ei Ei Moe Tun
Faculty of Information Sciences
University of Computer Studies, Magway
eieimoetun@gmail.com

***Abstract*— Since 2012, internet usage in the world has grown from a mere 1.8% to 59% in 2020. The numbers of online shopping webs in the world are becoming greater than ever before. People use the internet to access or update reviews. Many reviews are long and a few sentences contain an opinion on the products. Each product can get hundreds or thousands of customer reviews. Opinion mining is widely used in reviews and social media for a variety of applications, ranging from marketing to customer service. This paper is a system of opinion mining or sentiment analysis on the level of user satisfaction of product reviews. Opinion mining works for finding and classifying opinions in customer reviews on any products as either positive or negative. Support Vector Machine algorithm with RBF (Radial Basis Function) kernel is used to classify reviews; this is one of the supervised opinion mining techniques. The classification result obtained 87% accuracy using the camera product reviews dataset from amazon.com.**

Keywords*— *Opinion mining, Classification, Support Vector Machine; RBF

I. INTRODUCTION

Electronic commerce is becoming increasingly popular, many products are added on the web and people are also buying a product online. Online shopping market sizes are also expanding. Over the past year, Amazon has become the largest online marketplace in the world. For this reason, the goal of this paper is to perform an accurate sentiment analysis on Amazon product reviews. Also, the customers are becoming more comfortable buying the products on the internet web because they can save their time consumption. The online websites provide comment boxes to be reviewed their products by the customers. Customers can review and express opinions on the products which they have purchased. So analyzing the data from those customers' reviews to make data more dynamic is an essential field nowadays. The objectives of this paper are to categorize the positive and negative feedbacks of the customers over different products. The feedback from the customers is very important for the organization. By following the positive feedbacks, the organization can know what product should produce how much or how many quantity. And they can also know what product should make quality control. They can determine the needs, wants, and interests of the target markets. Customers want to buy and to interest in the products that got good comments. When seeing other user's comments or reviews, the customers can determine what product should be bought/not. Therefore, opinion mining will prove to know the quality of a product using a machine learning technique for buyers around the world.

This paper consists of the following components: Session II provides the related work. System work flow and methodology gives in Session III. Session IV shows the accurate results and conclusion of this paper describes in Session V.

II. RELATED WORK

Much of the research papers related to product reviews, sentiment analysis, or opinion mining has been done recently.

In [5] Mohan Kamal Hassan, Sana Prasanth Shakthi, and R Sasikala sentiment from the reviews and analyze the result to build up a business model. The commonly used programming language was R. They mainly used Multinomial Naïve Bayesian (MNB) as their main classifier.

In paper [11] Fuzzy clustering, K- mean, Mean Shift, Dragonfly Algorithm is used and describes the most effective web-based shopping sites and how they are carried on.

In the work [4], the authors examine opinion mining in various domains with the SVM method and compare the corpus in science research. The summary of this study states the SVM is better than Naïve Bayes.

In this paper, propose to use a Support Vector Machine from one of the classification methods for opinion mining to provide better accuracy. Many of the SVM kernel functions use the RBF kernel. We commonly used the python programming language.

III. METHODOLOGY

This paper aims to perform sentiment analysis or opinion mining tasks by applying a machine learning algorithm (supervised method); all the methodology required is described as follows. The first step, gather the reviews from the dataset. And then cut the text (reviews) into the collection of words text through preprocessing techniques such as transfer case, stemming, stop word removal, and tokenization. The next step is feature extraction because the reviews or comments need to convert text to vectors. Term Frequency-Inverse Document Frequency method will be used in this process. Finally, classify the reviews or comments with the Support Vector Machine (SVM) method. The overall operation of the system is performed as shown in Figure 1.

Support Vector Machine (SVM) is one of the classification methods that predict classes based on the pattern from the result of the training process. It performs classification by constructing a hyperplane to separate different classes in a high dimensional space. The data must be separated by a linear/nonlinear hyperplane. So, we need a linear/nonlinear dividing hyperplane to separate the classes for classification. To separate data according to a nonlinear

hyperplane, SVM uses the kernel function to assign data to two classes. The most commonly used nonlinear kernel function is as follows:

1. Polynomial kernel

   $K(x, y) = (kx^T y + 1)^d$

2. Radial basis function (RBF) kernel

   $K(x, y) = exp(-\gamma \parallel x - xi \parallel^2), \gamma > 0$

3. Sigmoid kernel

   $K(x, y) = \tan(kx^T y + \theta)$

Where'd', 'γ', and 'θ' are user-defined parameters. We will use an SVM classifier based on the class of hyperplane, $f(x) = w.x + b$, $w \epsilon R$, $b \epsilon R$. And then the decision function is $f(x) = a^T y K(x, y) + b$. The overview of the system is as follows.

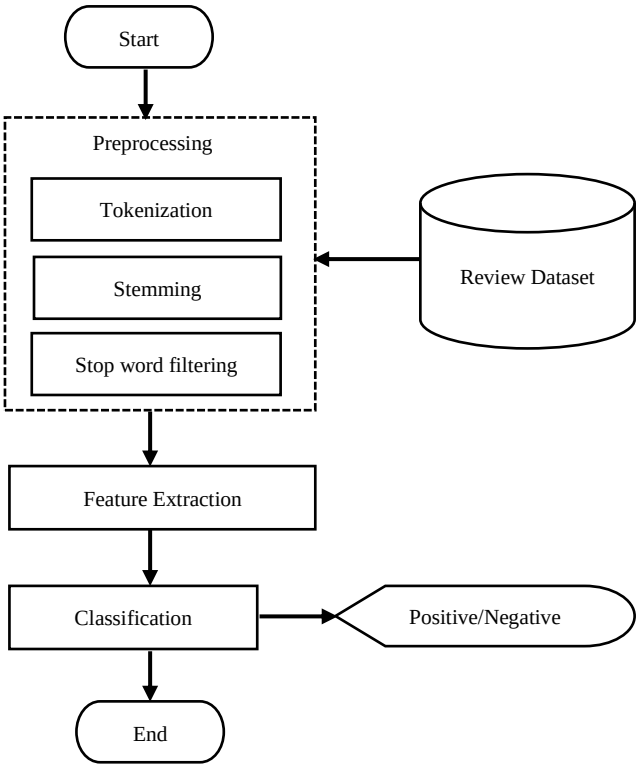

Fig. 1. Flow Diagram for Opinion Mining

A. Domain area

In this paper, the camera products data set will be used as the domain area. This data set is taken from s3.amazonaws.com that has been filtered based on many keywords from products. If the data contains errors that may distort our model they must be corrected. For example, some values from the product reviews in the dataset were empty so these errors must be removed. Dataset is many attributes but mainly used Review_body attribute that meaning the user's comments/reviews and Star_rating attribute. A star rating is the rating question that lets people rate a product with several stars. In this dataset, users or customers can give one to five stars. So, we assumed one star and two stars are negative and three stars, four stars, and five stars are positive. In Table I, the first review was labeled positive and the second review was negative because the user gave the first review five stars and the second review one star.

TABLE I. REVIEW HAS BEEN LABELED

Star_rating	Review_body
Positive	soft and comfortable on my neck
Negative	korner is unequalled

B. Preprocessing

Preprocessing removes unnecessary raw data such as stop words, missing attributes, punctuations, and HTML tags from user's comments or reviews. It improves the accuracy of the process of opinion mining. In this paper, preprocessing techniques are tokenization, stemming, transfer case, and stop word removal. The steps for preprocessing are shown in Figure 2.

- Tokenization: Tokenization is the process of splitting up a larger body of text into pieces called tokens. The reviews are first tokenized into tokens. It becomes the input for another process like stemming and stops word removal.
- Stemming: Stemming is the process of reducing inflected words to their word stem based on the root. For example, the three words- sleep, slept, sleeping has the same root word sleep.
- Stop word filtering: Stop words are the English word which does not add much meaning to a sentence. So, need to remove stop word such as a, an, the, in, at, he, she, they, these, I, and, on, it, is, that, etc.

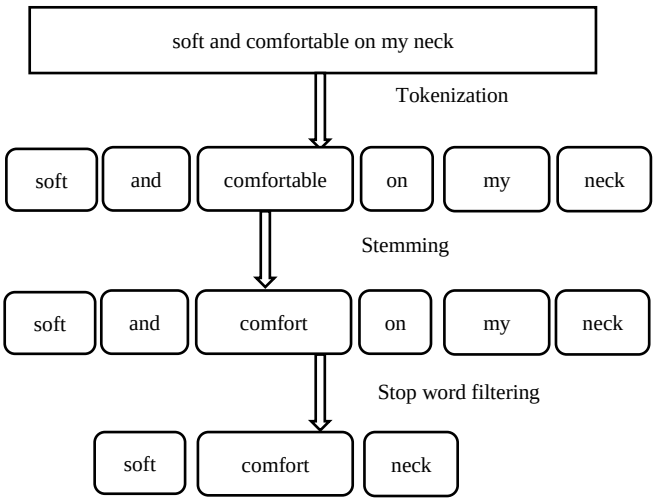

Fig. 2. Preprocessing steps

After the preprocessing step, there are two documents with labels positive and negative. One of the two examples reviews below will be set to D1 (document1) and the other to D2 (document2) as shown in Table II.

TABLE II. PREPROCESSED REVIEWS

Label	Review_body	Initial
Positive	'soft' 'comfort' 'neck'	D1
Negative	'korner' 'unequal'	D2

C. Feature Extraction

After preprocessing, the documents must be presented as a vector so every word is converted into a number. Feature extraction work on extracting product features from the text of the opinion contained in the dataset and converts the

document into binary numbers (0 and 1). This step will give you the weight of each term in every document. In this paper, the Term Frequency-Inverse Document Frequency (TF_IDF) method will be used for feature extraction. This is a technique to extract information that weights a term's frequency and its inverse document frequency. For example, a term i in document j is calculated using the following equations.

$$TF - IDF_{term,doc} = TF_{term,doc} * IDF_{term} \quad (1)$$

TF = Numbers of time term t in a document
DF = Numbers of the document that term t appear in
$IDF = log(N/DF)$,
N = Total number of documents
Table III. Shows the TF_IDF values of each term

TABLE III. CALCULATE OF TF_IDF VALUES

Term	***TF***		***DF***	$TF * (\log(\frac{N}{DF}))$	
	D1	***D2***		***D1***	***D2***
soft	1	0	1	0.3010	0
comfort	1	0	1	0.3010	0
neck	1	0	1	0.3010	0
korner	0	1	1	0	0.3010
unequal	0	1	1	0	0.3010

D. Classification with Support Vector Machine

Support Vector Machine (SVM) algorithm came out as an important learning technique for solving classification problems in various fields. So, this method with nonlinear kernel RBF (Radial Basis Function) will be used as a classification method. The preprocessed data will be separated by a non-linear hyperplane because it maps samples into a higher dimensional space unlike to linear kernel. To classify data, we need to find the nonlinear hyperplane $f(x) = w.x + b$. Where w is a weight and b is a bias. x is the input feature vector that is a matrix n*d (n is the number of data and d is the number of features). To find the input vector x, we used the RBF kernel formula as follows [7]:

$$K(x, y) = exp(-\gamma \parallel x - xi \parallel^2), \gamma > 0 \quad (2)$$

In this paper, we used the γ value is 0.5. The weight value of each term in D1 (document1) is set to x_1, and the value in D2 is set to x_2. According to that setup, training data $\{x_1, x_2\}$, and their labels are $\{y_1, y_2\}$ where $y_1 = 1$ and $y_2 = -1$. First of all, calculate the distance between two documents (x_1 and, x_2) following in Table IV.

TABLE IV. RESULT OF x-x_i

x_1-x_1	x_1-x_2
0	0.3010
0	0.3010
0	0.3010
0	−0.3010
0	−0.3010

And next, calculate the length between these two documents (x_1 and, x_2) in Table V.

TABLE V. RESULT OF $\parallel x - x_i \parallel$

$\parallel x - x_i \parallel$	$\sqrt{x^2 + y^2}$	**Length**
$\parallel x_1 - x_1 \parallel$	$\sqrt{0^2}$	0
$\parallel x_1 - x_2 \parallel$	$\sqrt{3(0.3010)^2 + 2(-0.3010)^2}$	0.6731

Calculate exp $(-\gamma \parallel x - xi \parallel^2)$ as follows:

$$K(1,1) = \exp(-\gamma \parallel x - xi \parallel^2)$$
$$= \exp(-0.5(0)^2)$$
$$= 1$$
$$K(1,2) = \exp(-0.5(0.6731)^2)$$
$$= 0.7973$$

We got the K kernel matrix as follows:

$$K = \begin{bmatrix} 1 & 0.7973 \\ 0.7973 & 1 \end{bmatrix}$$

Next, calculate y (labels) in Table VI.

TABLE VI. SCORE OF Y

y_1	y_2
1	−1
y_1y_1	y_1y_2
1	−1
y_2y_1	y_2y_2
−1	1

The support vectors are points on the borders of the margins. These are the only points needed to calculate the margin and get the best hyperplane. In this paper, calculate the support vectors by using the Langrage Multiplier equation as follows [1]:

$$min\, Ld = \tfrac{1}{2}\textstyle\sum_{i,j} a_i a_j y_i y_j x_i x_j - \sum_{i=1}^{N} a_i \quad (3)$$
$$0 \le a_i \le C, (\, i = 1,2,3, \dots, n)$$
$$\min Ld = \frac{1}{2}\sum_{i,j} a_i a_j \begin{bmatrix} 1 & 0.7973 \\ 0.7973 & 1 \end{bmatrix} * \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} - a_1 + a_2$$
$$= \left(0.2027{a_1}^2 - 0.2027{a_2}^2 - a_1 + a_2\right)$$
$$a_1 = 0.7, \qquad a_2 = 0.7$$

Find the value of w (weight) and b (biased) by the following methods [6]:

$$w = \textstyle\sum_{i=1}^{m} a_i y_i x_i, \quad (4)$$
$$w = \begin{bmatrix} 0.1419 \\ -0.1419 \end{bmatrix}$$
$$b_i = 1 - yi(w.xi), \quad (5)$$
$$b = 0.97$$

According to the above calculation, the value of $w, x,$ and b have been obtained, resulting in a nonlinear hyperplane $f(x) = w.x + b$. After the hyperplane come out, perform the SVM decision function for the new data is as follows:

$$f(x) = a^T yK\left(X_{training}, X_{testing}\right) + b \quad (6)$$
$$= \begin{bmatrix}0.7\\0.7\end{bmatrix}\begin{bmatrix}1 & -1\\-1 & 1\end{bmatrix} exp\left(-0.5 \parallel X_{training}, X_{testing} \parallel^2\right) + 0.97$$

The value obtained from that decision function is positive if it is greater than the value of hyperplane $w.x + b$ and negative if it is smaller.

IV. ACCURACY

In our basic dataset consist of 10000 camera product reviews, 5000 labeled positive and 5000 labeled negative. Accuracy is a method for measuring how accurate it is. If the test is 10 times, 8 times are true then the accuracy of this dataset is 80%. Accuracy is calculated by:

$$Accuracy = \frac{(all\ correct\ measurements)}{(total\ data)}$$

a) Accuracy with the values of C and Gamma

There are two parameters for an RBF kernel: C and γ [12]. We used the values of C are 1, 10, 100, 1000 and Gamma (γ) are 0.1, 0.2, 0.5, 1, and 2. Answer the comparison of accuracy with values of C and gamma in Table VII.

TABLE VII. ACCURACY OBTAINED BY C AND GAMMA VALUES

C	Gamma (γ)	Accuracy
1	0.1	84.8%
1	0.2	86.0%
1	0.5	85.6%
1	1	86.4%
1	2	86.6%
10	0.1	85.1%
10	0.2	85.2%
10	0.5	86.1%
10	1	86.1%
10	2	86.8%
100	0.1	83.2%
100	0.2	84.4%
100	0.5	85.8%
100	1	86.1%
100	2	86.8%
1000	0.1	83.0%
1000	0.2	84.4%
1000	0.5	85.8%
1000	1	86.1%
1000	2	86.8%

Testing the values of C and Gamma, the best values are C=10 and Gamma=2, C=100 and Gamma=2, and C=1000 and Gamma=2.

b) Accuracy with K-fold cross-validation

We will use the grid search method in cross-validation to select the best parameter set. Table VIII shows the best cross-validation rate with K-fold cross-validation. The values of K = 5, 10, 15, and 20 are used.

TABLE VIII. ACCURACY WITH K-FOLD CROSS-VALIDATION

No.	K-fold	Accuracy
1	5	85.9%
2	10	86.2%
3	15	86.2%
4	20	86.5%

The values K=20 or 20-fold gives the best classification performance to build the model.

c) Accuracy with cutting the data size

Finally, inserted the best parameter accuracy is calculated by changing the ratio of training and testing.

TABLE IX. ACCURACY OF CLASSIFIER USING SUPPORT VECTOR MACHINE

Testing size (%)	Accuracy (%)	Positive			Negative		
		Precision (%)	*Recall (%)*	*F1-score (%)*	*Precision (%)*	*Recall (%)*	*F1-score (%)*
90:10	87	89	84	86	85	90	88
80:20	86	89	82	85	84	90	87
70:30	86	89	82	85	83	90	86
60:40	85	88	82	85	83	89	86
50:50	85	88	82	85	83	89	86

When testing 5 times, the best data structure is 90%:10% (training: testing) with 87% accuracy, 89% precision, 84% recall, and 86% f1-score.

V. CONCLUSION

In this paper, have shown work on opinion mining of online customer reviews of camera products on amazon. We developed a product review using preprocessing techniques. To increase the accuracy of the learned model by extracting features from processed data used feature extraction (TF-IDF) method. Finally, we performed opinion mining on the reviews using a machine learning algorithm Support Vector Machine with RBF kernel because it gives the best accuracy when compared to other methods. This paper aims to help the user in selecting the best products they need. The main aim of all this is to collect beneficial information from the thousands of reviews about products.

REFERENCES

[1] Budi. Santosa, Tutorial Support Vector Machine, no.x. Surabaya, 1995.

[2] Chandra Kala S and Sindhu C 2012 'Opinion mining and sentiment classification: a survey' ICTACT J. Soft Comput. pp420-427.

[3] Chinsha T C PG Scholar Dept. of Computer Science, Shibily Joseph Asst. Professor, "Aspect based Opinion Mining from Restaurant Reviews ", 2014.

[4] M. R. Saleh, M. T. Martín-Valdivia, A. Montejo-Ráez, and L. A. UreñaLópez, "Experiments with SVM to classify opinions in

different domains," SINAI Research Group, Department of Computer Science, University of Jaén, Campus Las Lagunillas, Spain, 2011

[5] Mohan Kamal Hassan, Sana Prasanth Shakthi, and R Sasikala. 'Sentimental analysis of Amazon reviews using naïve Bayes on laptop products with MongoDB and R". November IOP Conf. Series: Materials Science and Engineering 263 (2017) 042090 doi:10.1088/1757-899X/263/4/042090

[6] N. Guenther and M. Schonlau, "Support vector machines," Stata J., vol. 16, no. 4, pp. 917–937, 2016.

[7] N. Rahmansyah, "Analisa Algoritma Support Vector Machine (Svm) Dalam Memprediksi Nasabah Yang Berpeluang Kredit Macet," J. KomTekInfo, vol. 3, no. 1, pp. 67–77, 2016.

[8] Rimba Nuzulul Chory, Muhammad Nasrun, Casi Setianingsih, "Sentiment Analysis On User Satisfaction Level Of Mobile Data Services Using Support Vector Machine (SVM) Algorithm", 2018.

[9] S. Mujilahwati, "Pre-Processing Text Mining Pada Data Twitter," Semin. Nas. Teknol. Inf. dan Komun., vol. 2016, no. Sentika, pp. 2089–9815, 2016.

[10] Sindhu Chandra Sekharan, SRM Institute of Science and Technology, "Sentiment Analysis Based Product Rating Using Textual Reviews", October 2017.

[11] S.K.Lakshmanaprabu, Dr. K Shankar, Deepak Gupta, Ashish Khanna, "Ranking Analysis for Online Customer Reviews of Products Using Opinion Mining with Clustering", September 2018.

[12] Vasileios Apostolidis-Afentoulis, University of Macedonia, "SVM Classification with Linear and RBF Kernels", July 2015

[13] 06 JUNE 2017, 'Online shopping popular with Myanmar youth', multiverseadvertising.com.

[14] Dec 17, 2018, 'Data Driven Investor', medium.com.