SENTIMENT ANALYSIS OF TWEETS FOR E-COMMERCE SITES USING SVM AND MAXIMUM ENTROPY

Gagan Tyagi[1]
M.Tech, Computer Science & Engineering R.D. Engineering College, Ghaziabad, India

Prof. Dr. Manish Kumar[2]
Professor, Dept. of Computer Science & Engineering R.D. Engineering College, Ghaziabad, India

Received: 14 March 2020 Revised and Accepted: 8 July 2020

Abstract
Amazon, Flipkart, SnapDeal are the best internet business applications and sites in India. It is significant for organizations or association to thinking about how general society reacts about their items or administrations which they offer. Conclusions, Reviews, Views, and Opinions can influence how the manner in which individuals settles on certain choices in their decisions and prerequisites. It is certain that feeling which originates from people, in general, can influence the growth and potential of an association or organization. Nonetheless, checking and arranging general supposition isn't a simple work to do. The measure of supposition communicated in internet-based life is a lot to be handled physically. In this manner, an extraordinary strategy or procedure is expected to classify the surveys consequently, regardless of whether it is sure or negative. In this manner, Data got from the twitter micro blog website is additionally marked and investigated utilizing the Support Vector Machine (SVM) and Maximum Entropy (Maxent) technique to group information survey. From the marking results that have been done at that point will be seen text relationship on each class of opinion to discover a reality and data that is viewed as significant and can be helpful for decision making. With the SVM along with the Maxent technique demonstrated under the scheme and accuracy of 91,05% is achieved. Moreover, positive classes including related to commodities, dealings, features, product and services, sales orders, logistics and delivery, reaction, e-shopping, needs and installments. While the negative classes that are often complained include merchandise, updates, servers, chat, electronic mail, materialistic-transactions, uploads, promotions, gift-vouchers, brochure and upgrades.

***Keywords*—**Sentiment Analysis, Machine Learning, Support Vector Machine, Maximum Entropy.

I. INTRODUCTION

The world of technology is currently growing increasingly rapidly towards the all-digital direction. The digital age has made humans enter a new lifestyle that cannot be separated from the all-electronic devices. Technology becomes a tool that helps human needs, with technology; anything can be done more easily. It was this important role of technology that began to bring civilization into the digital age. The increasing need for data and information encourages people to develop new technologies so that data and information processing can be done easily and quickly [1]. Advances in technology, computers and telecommunications have supported the development of internet technology. Internet users continue to grow every year.

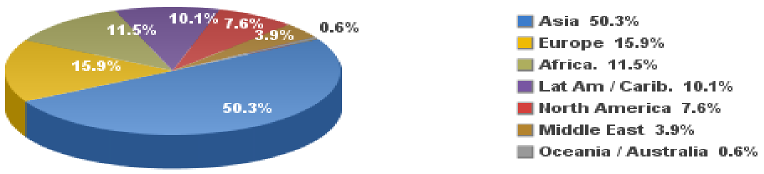


Figure 1: Growths of Internet Users

A survey conducted by the Internet Service Providers Association of India (ISPAI) [2] showed that the number of internet users in India increased each year until the end of 2021 to reach 601.6 million people, out of a total population of 1,387.58 million people. These developments have an impact on various fields. One of them is the rampant shopping activities through the internet media. According to a survey conducted by ISPAI 62% of internet users in India know the internet as a place for buying and selling goods and services, and 63.5% have done online transactions. In addition, 62% often visited online shop commercial content.

India has become the largest e-commerce market in Southeast Asia. In 2019, Euromonitor [3] noted, India's online sales reached US $ 129.1 billion, higher than USA, UK and Australia and second after China. However, if compared to the total retail trade, e-commerce sales in India only contribute 2.7 percent. This means that India's e-commerce market has the opportunity to grow even greater. Especially with the population and the level of gross domestic product (GDP), the largest in Asia. Euromonitor estimates the average annual growth of India online sales during 2017-2019 at 38% (India Brand Equity Foundation –IBEF.org).

One of the most popular e-commerce sites most visited in India according to Alexa data as of October 2018-January 2019 is amazon, flipkart, snapdeal and jabong site it came from a search engine of 34.30% and the total connected to the site was 131,110,000 (Alexa, 2019). These all are one of the Business-To Customer (B2C) e-commerce models. This model is the most widely applied and found on e-commerce sites in India. The Business-To Customer (B2C) market is currently still dominant in India's online retail market. Data collected from Euromonitor estimates that the B2C market contributed 13% of the retail market in India in 2019, while the B2C market contributed 1.7% (Euromonitor, 2019). In addition, Indian people prefer to shop online on the B2C model because there are many product choices. That is why this research will analyze the B2C model.

Google Play is Google's digital content service that consists of online stores of products such as clothes, households, electronic items, songs, books etc. or cloud-based media players. Google Play can be accessed via the web, android applications (Play Store), and Google TV. Google Play is equipped with a feature that contains reviews from users that can be used to view reviews from application users.

User reviews are often used as an effective and efficient tool in finding information about a product or service. According [4-6], that recent research found nearly 50% of internet users rely on word-of-mouth recommendations before using a product, because reviews from other users can provide the latest information about the product This is based on the perspective of other users who are already using the product. Customers or clients who are not satisfied with the services or products offered by a company will usually write their complaints on social media. On the other hand, there are also satisfied customers who express their positive attitude towards a product on social media. Whether we realize it or not, the opinions of customers written on social media, little or much, will influence prospective customers. However, monitoring and organizing opinions from the public is also not easy. Opinions posted on social media are too numerous to be processed securely. For this reason, we need a special method or technique that is able to categorize these reviews automatically, whether positive or negative, based on a property. The amount of E-Commerce application user review data that enters the Google Play site continues to grow over time, this makes it difficult for the company to obtain overall information from all reviews, because it will take a long time to read one by one every review that comes on the site page Google Play.

Many user reviews on Google Play regarding the E-Commerce app. A good brand image will form a good opinion of consumers about a product / service, and is expected to drive the buying process by consumers, and vice versa. Various kinds of responses on the Google Play site will certainly affect the image of E-Commerce Apps. Negative and positive responses from users may be influenced by a number of things that have not been addressed by E-Commerce Apps. This might have happened because of several factors that had to be corrected and were not yet known by E-Commerce Apps. By using text mining can be seen what talks are often discussed by users. One analysis of text mining is that sentiment analysis can be applied to companies that issue a product or service and provide services to receive opinions (feedback) from consumers for the product. Sentiment analysis is applied to classify positive, negative, and neutral feedback from consumers so as to speed up and simplify the company's task to review their product deficiencies. If negative sentiment is found, the company can quickly take action to overcome it.

From the aforementioned series of backgrounds, the researcher feels it is necessary to conduct further analysis of E-Commerce Apps user reviews on Google Play to find out how the user's opinion of Apps. Researchers will classify E-Commerce application user reviews whether including positive or negative reviews for evaluation material from e -commerce Apps is commonly called sentiment analysis using the Support Vector Machine (SVM) and Maximum Entropy (Maxent) method. The classification process is done using the SVM algorithm because it has the highest level of accuracy in terms of text classification [7-9]. Whereas the Maximum Entropy method is able to find the distribution $p\ (a \mid b)$ which will give the maximum entropy value with the aim of getting the best probability distribution that is the closest to reality. Based on these reasons, the researchers chose to use the Support Vector Machine and Maximum Entropy methods to classify reviews of the E-Commerce application.

After doing the classification, the author tries to extract and explore the widest possible information that is in each classification of positive sentiments and negative sentiments that are considered important for various purposes. In the process of extraction and exploration of information, the authors use descriptive statistics and associations between words to find topics that are often discussed by users. The hope, this research is able to classify the text of the review well so that later the information contained in it can be extracted properly and the presentation of information from the observed data can provide useful information for various parties who need it.

II. Related Work

In a study conducted by [11] on sentiment classification using the Machine Learning technique. In this study using the Naïve Bayes Classification (NBC), Maximum Entropy (ME), and Support Vector Machine (SVM) methods to classify film reviews into positive and negative classes. The experimental results show the SVM method has the highest level of accuracy compared to other methods, which is equal to 82.7%.

In the paper [12] classifies opinions using the NBC and SVM methods. From the results of his research, Saraswati stated that the SVM method has a higher level of accuracy than the NBC method for testing positive opinion data, while the NBC method shows better results when used in testing negative opinion data.

In a study conducted by [13] relating to sentiment analysis to find out how customer satisfaction is compared to three e-commerce sites that are frequently visited in Indonesia, namely Bukalapak, Tokopedia and Elevenia. Data obtained from social media Twitter, in his research Naïve Bayes Classifier is used as a classification technique with TF-IDF weighting, while to validate and evaluate the Naïve Bayes text classification is done using K-fold cross validation and confusion matrix. The results showed that negative sentiments towards the three e-commerce sites were more dominant in social media, and e-commerce sites with the highest negative sentiment were Bukalapak then Tokopedia and Elevenia. Corpus-based methods are a main direction in Twitter sentiment classification. They typically rely on machine learning algorithms to train classifiers on sentiment annotated datasets. Such methods have a clear advantage, in that they produce substantial amounts of annotated data. However, they are dependent on the effectiveness of emoticons in highlighting representative training instances [14].

Many researchers have investigated sarcasm on the data collected from various sources such as tweets on Twitter, Amazon product reviews, website comments, Google books and online discussion forums with the help of various features such as lexical, pragmatics and hyperbole. Sarcasm detection can be classified into three categories, namely lexical, pragmatics and hyperbole [15]

Suchita V Wawre and Sachin N Deshmukh [14] took dataset of movie reviews and did sentiment classification and compared three supervised machine learning algorithms namely- SVM. Naïve bayes and KNN and observed that SVM performed the best with an accuracy of more than 80%. The SVM uses classification to find hyper- plane with maximum margin that separates the document vector in one class from the other with maximum margin. They are large margin rather than probabilistic, classification in contrast to naïve bayes.

Author of [17] collected tweets having direct sentiment in the tweets in the form of hash tags using Twitter API. These direct sentiments were expressed like #happy, #sad, #delighted, #grief, #cheerful, #disappointed, #sarcastic etc. Unimportant and insignificant parts of tweets like tweets in other languages, re-tweets, spam's, punctuations etc. were filtered out. Then the data was divided into 4 feature sets and 12 different algorithms were applied to know the most important feature sets. Accuracy of 79% was obtained using gradient descent. They gave the concept of SCUBA framework in which sarcasm was divided into various forms as described in above diagram.

I. METHODOLOGY

The proposed scheme is as under:-

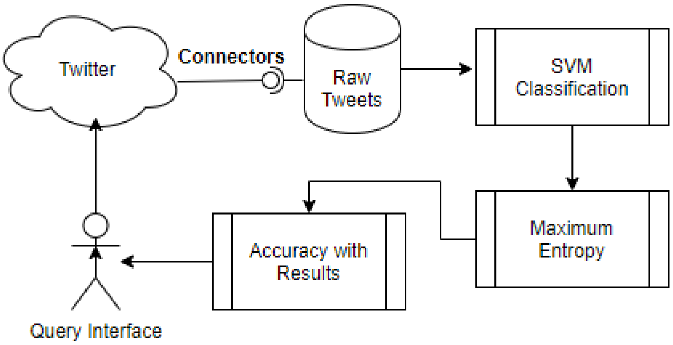


Figure 2: Proposed Work Flow

Support Vector Machine: Support Vector Machine was developed by Boser, Guyon, and Vapnik was first presented in 1992 at the Annual Workshop on Computational Learning Theory. The basic concept of SVM is actually a a combination of computational theories that existed decades before. The basic principle of SVM is a linear classifier, and so on developed so that it can work on non-linear problems, by including kernel trick in high-dimensional workspaces. Commonly used kernel functions usually that is linear, polynomial, radial basis function and sigmoid. This technique trying to find the optimal classifier function that can separating two data sets from two different classes. In the case of linear classification a separator function can be used defined as follows:-

$$g\ (x) \coloneqq\ sgn\left(f\ (x)\right)\ with\, f\ (x) =\ w^T x\ +\ b \qquad (eq.1)$$

Whereas in the case of non-linear classification to find an accurate hyper-plane separator for classifying two classes, is with using a nonlinear separator function (that is, one whose mapping function is $\emptyset$ is nonlinear mapping from the input space into several feature spaces). Defined as follows:

$$\phi\binom{x}{y} = \begin{cases} \begin{pmatrix} 6-x_1+(x_1-x_2)^2 \\ 6-x_2+(x_1-x_2)^2 \end{pmatrix} if\sqrt{x_1^2+x_2^2 \geq 2} \\ \qquad\qquad {}^{x_1}_{x_2}\, if\ not \end{cases} \qquad (eq.2)$$

Then calculate the value to find 3 parameters α_1, α_2, and α_3 based on the following linear equation:-

$$\alpha_1\tilde{S}_1.\tilde{S}_1 +\ \alpha_1\tilde{S}_1.\tilde{S}_1 +\ \alpha_1\tilde{S}_1.\tilde{S}_1 =\ +1\ (Positive)$$
$$\alpha_1\tilde{S}_1.\tilde{S}_1 +\ \alpha_1\tilde{S}_1.\tilde{S}_1 +\ \alpha_1\tilde{S}_1.\tilde{S}_1 =\ +1\ (Positive)$$
$$\alpha_1\tilde{S}_1.\tilde{S}_1 +\ \alpha_1\tilde{S}_1.\tilde{S}_1 +\ \alpha_1\tilde{S}_1.\tilde{S}_1 =\ +1\ (Negative)$$

After getting the value α, the next step is to search hyper-plane to separate positive classes and use negative classes' equation as under:-

$$\widetilde{W} = \sum_i \alpha_i\, \tilde{S}_1 \qquad (eq.3)$$

The following is an illustration of the use of the support vector machine method in separating relevant and irrelevant data.

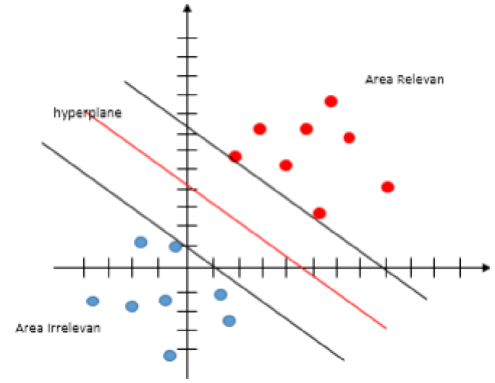


Figure 3: Separating relevant and irrelevant data

Maximum Entropy: The Maximum Entropy technique states that when no information from the data is known, the distribution is attempted to be uniform, that is, it has maximum entropy. In text classification, Maximum Entropy estimates the distribution of labels in documents. The document is represented by a set of word appearance features. In the Maximum Entropy method, the classification process is done by only using information on the appearance of a feature in a document. This relates to the user $fi \epsilon$ {0, 1}. Broadly speaking, the Maximum Entropy method looks for the most uniform probability distribution.

Classification Algorithm with Maximum Entropy

The following is a text classification algorithm using the method Maximum Entropy:
1. Identify specific words in the document (sentence).
2. Form a matrix containing the value of the appearance of specific words with the following index

$$f_j(a,b) = \begin{cases} 1; if\ f_j\ appears\ in\ document\ b\ of\ the\ class \\ 0; if\ f_j\ does\ not\ appears\ in\ document\ b\ of\ the\ class \end{cases}$$

3. Creating a Maximum Entropy model with training data that is calculating the value of α_j for each class with the GIS (Generalized Iterative Scaling) procedure $\alpha_j^{(0)} = 1$.

$$\alpha_j^{(n+1)} = \alpha_j^{(n)} \left[\frac{E_{\tilde{p}} f_j}{E^{(n)} f_j}\right]^{\frac{1}{2}} \quad (eq. 4)$$

Where
$E_{\tilde{p}} f_j = \Sigma_{x\epsilon\varepsilon} p(\mathrm{x})\ f_j(x)$
$E^{(n)} f_j = \Sigma_{x\epsilon\varepsilon} p(\mathrm{x})\ f_j(x)$
$p^{(n)}(x) = \pi \prod_{j=1}^{k} (\alpha_j^{(n)})^{fj(x)}$
$\forall\ x\ \epsilon\ \sum_{j=1}^{k} f_1(x) = C$

4. Look for joint probability p (a, b) for testing data a = {$positive$, $negative$}

$$\mathrm{p} * (\mathrm{a,b}) = \pi \prod_{j=1}^{k} \alpha_j^{f_j(a,b)} \quad (eq. 5)$$

5. Determination of the topic of the data testing document by looking at the value of $a*$ the greatest in a class

$$a * = \mathbf{argmax}\ p\ (a, b) \qquad \text{(eq. 6)}$$

$a\ \epsilon$ ($positive$, $negative$)

IV. Results and Discussion

Classification with Support Vector Machine: Support Vector Machine or SVM is one technique to predict which is good in classification and regression.. SVM is a method that can solve problems linearly as well as non-linear problems. In solving non-linear problems in use the concept of the kernel in the workspace dimension height, by finding hyper plane that can maximize margins between data classes. Hyper plane is useful in separating 2 class +1 (positive) and class –1 (negative) groups where each class has each pattern. In making decisions with SVM method is used the kernel function K (x_i, x^d). The kernel to use with the research shown in Equation below:

K (, x_d) = $(X_i^T\ X_j + C)^d$, $\gamma > 0$

Processing is done on training data Sequential training algorithm is used because it is a simple algorithm without takes a lot of time with calculation stages:

1. Initialization of various parameters,
like α_i, γ, C, and ε.
αi = alpha, to find support vector
γ = gamma constant to control speed
C = slack variable
ε = epsilon is used to find value error

2. Calculate the Hessian matrix obtained from multiplication between polynomial kernels and y
is vector 1 and -1.
The equation from the Hessian matrix is: $D_{ij} = y_i\, y^j\ (K\ (x_i,\ xj) + \lambda^2)$

3, Perform the following calculations until the interaction Data i to j:

a. $Ei = \sum_j^i aj\ Dij$

b. $\delta\alpha_i$ = min(max[γ(1 - E_i), α_i] , C - α_i

c. $\alpha i = \alpha i + \delta\alpha i$

4. Perform the three steps above in a manner repeat until it reaches the maximum limit
Iteration

5. Sequential learning process from stage 1 up to 4 will get value from support vector (SV), where the value SV = (αi> $th\,resh\ oldSV$). After that, it needs to be done calculation of the value of bias b obtained from Eq below:-

b= $-\frac{1}{2}\sum_{\mathrm{i=0}}^{\mathrm{n}} \alpha_i\, y_i\ (x_i,\ x\text{-}) + \sum_{\mathrm{i=0}}^{\mathrm{n}}\ \alpha_i\, y_i\ (x_i,\ x\text{+}))$

6. To find out the results of the tweet classification on a sarcastic sentiment class is then carried out the process of calculating the function f (x). If results from the function is sarcastic, then classified tweet on class sentiment (positive to negative). If the function value positive value, the tweet is classified on the positive sentiment class or negative class. The function f (x) is obtained in Equation below :

b= $-\frac{1}{2}\sum_{\mathrm{i=0}}^{\mathrm{n}} \alpha_i\, y_i\ (x_i,\ x\text{-}) + \sum_{\mathrm{i=0}}^{\mathrm{n}}\ \alpha_i\, y_i\ (x_i,\ x\text{+}))$

Given an example of testing tweet as: "amazon is awesome figure despite many other e-commerce sites excellent service"

Tweet testing the results of the pre-process as: “amazon awesome figure despite excellent service”

The classification process will determine the class of a tweet based the frequency of occurrence of words from the previous process. As for classification the steps are as follows:

1. Calculate the value of prior probability At this stage the training data will be calculated with prior probability using the formula:

$$P(c) = \frac{N_c}{N'}$$

Class C Document	Prior Probability
Love	2/10
Happy	2/10
Angry	2/10
Sad	2/10
Fear	2/10

Table 1: Example of Conditional Probability Calculation

2. Calculate the value of conditional probability At this stage the conditional probability of words in each will be calculated using Hyper-plane based on kernel function. The class uses the formula below is to avoid the zero value:

b= $-\frac{1}{2}\sum_{\mathrm{i=0}}^{\mathrm{n}} \alpha_i\, y_i\ (x_i,\ x\text{-}) + \sum_{\mathrm{i=0}}^{\mathrm{n}}\ \alpha_i\, y_i\ (x_i,\ x\text{+}))$

Word	Conditional Probability Term In the Class				
(Term)	Love	Happy	Angry	Sad	Fear
Amazon	10.01667	20.03175	30.01695	40.0137	40.0164

Proof	10.01667	20.01587	30.01695	40.0137	40.0328
Blur	10.01667	20.01587	30.01695	40.0274	40.0164
The gap	10.01667	20.01587	30.01695	40.0274	40.0164
Love	10.05	20.01587	30.01695	40.0137	40.0164
Awesome	10.01667	20.01587	30.01695	40.0274	40.0164
Sin	10.01667	20.01587	30.01695	40.0137	40.0164
Support	10.01667	20.03175	30.01695	40.0137	40.0164
Figure	10.01667	20.01587	30.0339	40.0274	0.01639
Failed	10.01667	20.01587	30.01695	40.0137	0.03279
Over	10.01667	20.01587	30.01695	40.0137	0.03279
Hoax	10.01667	20.01587	30.01695	40.0274	0.01639
Results	10.01667	20.01587	30.01695	40.0274	40.0164
Hoaks	10.01667	20.01587	30.0339	40.0137	40.0164
Sincere	410.03333	420.01587	30.01695	40.0137	0.01639
Street	10.01639	20.01587	30.01695	40.0137	0.01639
Bores	10.03333	20.03175	30.0339	40.0274	0.03279
Excellent	10.03333	20.01587	30.01695	40.0137	0.01639
Service	10.01667	20.01587	30.01695	40.0274	0.01639
Kite	10.03333	20.01587	30.01695	40.0137	0.01639
Dirty	10.01667	20.01587	30.01695	40.0274	0.01639
Behavior	10.01667	20.01587	30.01695	40.0274	440.016
Bores	10.01667	20.01587	30.0339	40.0137	40.0164
Mud	10.01667	20.01587	30.01695	40.0274	4440.02
Eat	10.01667	20.04762	30.01695	40.0137	440.016
Angry	10.01667	20.01587	30.05085	40.0137	40.0164
Let	10.05	20.03175	30.0339	40.0137	4440.02
Win	10.01667	20.03175	30.01695	40.0137	440.016
Enemy	10.01667	20.01587	30.01695	40.0274	40.0164
Hell	10.01667	20.01587	30.01695	40.0137	4440.03
Morning	10.01667	20.03175	30.01695	40.0137	440.016
Short	10.01667	20.01587	30.01695	40.0274	40.0164
Full	10.01667	20.01587	30.01695	40.0274	440.016
Period	10.01667	20.01587	30.01695	40.0137	40.0328
Exactly	10.01667	20.01587	30.01695	40.0274	40.0164
Prejudice	10.01667	20.01587	30.01695	40.0274	4440.02
Doubt	10.03333	20.01587	30.01695	40.0137	40.0164
People	10.03333	20.03175	30.01695	40.0137	40.0164

Restless	10.01667	20.01587	30.0339	40.0137	40.0328
Sibling	10.01667	20.03175	30.01695	40.0137	40.0164
Rice fields	10.01667	20.01587	30.01695	40.0274	40.0164
Sad	10.01667	20.01587	30.01695	40.0411	40.0164
Happy	10.01667	20.04762	30.01695	40.0137	40.0164
attack	10.01667	20.01587	30.01695	40.0274	440.016
Noon	10.01667	20.03175	30.01695	40.0137	40.0164
Figure	10.03333	20.01587	30.01695	40.0137	40.0164
Steadfast	10.01667	20.01587	30.0339	40.0137	40.0164
Afraid	10.01667	20.01587	30.01695	40.0137	40.082
Please	10.01667	20.01587	30.01695	40.0274	40.0164
Sincere	10.03333	20.01587	30.01695	40.0137	40.0164

Table 4.2 Example of Conditional Probability Calculation

3. Matching between the data in the model and testing data At this stage will look for matching results by checking the words that exist both in the model and testing

Words in Training Dictionary	It's on Data Testing?	
	Yes	Not
also		√
Proof		√
blur		√
the gap		√
love		√
village		√
sin		√
support		√
Awesome	√	
Failed		√
Over		√
used up		√
the results		√
Figure	√	
sincere		√
Street		√
Despite	√	
Family		√
Work		√
Kiwi		√

Dirty		√
behavior		√
Lamias		√
Mud		√
Eat		√
Angry		√
Let		√
Win		√
Enemy		√
morning		√
Short		√
Full		√
Period		√
Exactly		√
prejudice		√
People		√
Restless		√
rice fields		√
Sibling		√
Sad		√
Excellent	√	
Attack		√
Noon		√
Service	√	
Steadfast		√
Afraid		√
Please		√
Sincere		√

Table 4. Examples of Matching Term Results in Training and Testing Data

4. Get the value of conditional probability on the matching results. At this stage the conditional probability value of the word in the model will be put into words in testing if both words are the same.

5. Calculating probability based on kernel function This stage calculates the probability to determine the tweet class which has the multiplication value between prior probability and conditional probability using hyper-plane therefore, classes uses the formula as:-

$$K\ (\ , x_d) = (X_i^T\ X_j + C)^d\ , \gamma > 0$$

Class	*Probability based on Kernel Function*

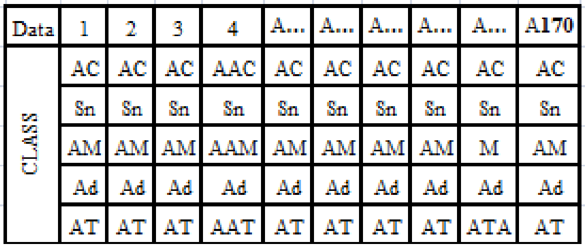
Love	= P (love) × P ("awesome" | love) × P ("figure" | love) × P ("despite" | love) * P "("excellent "| love) * P (" service "| love) = 0.2 × 0.0333 × 0.0333 × 0.01667 × 0.01667 × 0.01667 = 1. 027 × 10^{-9}
Happy	= P (awesome) × P ("despite" | happy) × P ("figure" | happy) × P ("service" | happy) * P "(" excellent"| happy) * P ("amazon"| happy) = (0.2) × 0.03175 × 0.01587 × 0.01587 × 0.01587 × 0.01587 = 4. 028 × 10^{-10}
Angry	P (angry) × P ("amazon" | angry) × P ("figure" | angry) × P ("excellent" | angry) * P "(" service "| angry) * P ("awesome"| angry) = (0.2) × 0.0339 × 0.01695 × 0.01695 × 0.01695 × 0.0339 = 1. 119 × 10^{-9}
Sad	= P (sad) × P ("amazon" | sad) × P ("figure" | sad) × P ("" | sad) * P "(" despite "| sad) * P (" service "| sad)) = (0.2) × 0.0274 × 0.0137 × 0.0274 × 0.0274 × 0.0274 = 1. 544 × 10^{-9}
Afraid	= P (scared) × P ("service" | scared) × P ("figure" | scared) × P ("amazon" | fear) * P "(" excellent"| fear) * P (" despite "| fear) = (0.2) × 0.03279 × 0.01639 × 0.01639 × 0.01639 × 0.01639 = 4.732 × 10^{-10}

Table 5 Example of Calculation of Probability based on SVM Kernel Function

It can be seen that the greatest posterior probability value belongs to the sad class with a value of 1.544 * 10-9 then the tweet will be classified into the sad class.

Maximum Entropy: This technique will calculate distribution probability and data accuracy both for data using Cross Validation comparing all labels. However, In the system has a total of 2,500 pieces of data tweets with a description of 500 class love tweets data, 500 happy class tweets, 500 angry class tweets, 500 sad class tweets, 500 classes tweets scared taken 170 data from each class as testing data so that testing data amounts to 850. The training data for 5 classes totalling 1650 pieces with the distribution of 330 data tweets per class.

Information:
'C' represents love
'Sn' represents Happy
'M' represents Angry
'Sd' represents Sad
'T' represents Fear



Data	1	2	3	4	A…	A…	A…	A…	A…	A170
CLASS	AC	AC	AC	AAC	AC	AC	AC	AC	AC	AC
	Sn	Sn	Sn	Sn	Sn	Sn	Sn	Sn	Sn	Sn
	AM	AM	AM	AAM	AM	AM	AM	AM	M	AM
	Ad	Ad	Ad	Ad	Ad	Ad	Ad	Ad	Ad	Ad
	AT	AT	AT	AAT	AT	AT	AT	AT	ATA	AT

Table 6 Distribution Probability of data using Maximum Entropy

Of the total data of 170 love testing tweets that entered the tweet system were classified into love classes as many as 106 pieces, 170 happy testing tweets that entered 160 tweets were classified into happy classes, out of 170 angry testing tweets 108 classified into angry classes, out of 170 tweets sad testing 121 is classified into sad class, out of 170 fear testing tweets 125 are classified into fear class, the classification results can be seen in the following confusion matrix:-

		Predictable class				
		C	**Sn**	**M**	**Sd**	**T**
CLASS INFACT	**C**	190				
	Sn		160			
	M			158		
	Sd				121	
	T					125

Table 7 Confusion Matrix System Test Results

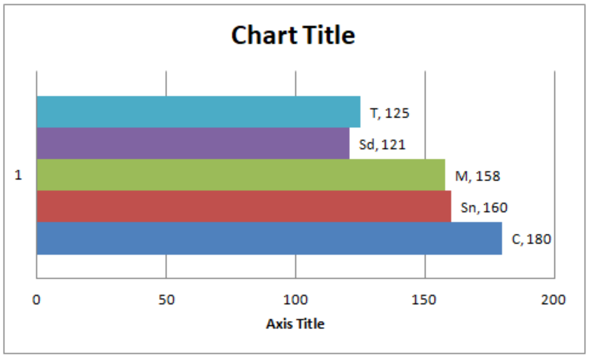


Figure 4. Graph Representation of Confusion Matrix

Accuracy Test Results: From the confusion matrix table above, we can calculate the accuracy value as follows:

$$\boldsymbol{Accuracy} = \frac{\mathbf{190 + 180 + 158 + 121 + 125}}{\mathbf{850}} * \mathbf{100\%}$$

$$\boldsymbol{Accuracy} = \frac{\mathbf{774}}{\mathbf{850}} * \mathbf{100\% = 91.05\%}$$

V. Conclusion and Future Scope

Conclusion: Based on the results of the analysis and discussion in the previous chapter, the author can draw some conclusions as follows:-

1. Based on the tweets it can be seen that the majority of e-commerce app users have a good rating or perception of the e-commerce. It is known that every month many users feel very satisfied by these e-commerce applications,

because there are indeed many features offered by these e-commerce apps to users who are different from other e-commerce sites. Whereas it is observed that, the sentiment labeling the number of positive reviews is more than the number of negative reviews. The number of positive reviews is 1,063 reviews or 55% while the rest are negative reviews.

2. By using a comparison analysis based on different applications the results of sentiment classification using the Support Vector Machine (SVM) along with Maximum Entropy method obtained an accuracy rate of 91.95% meaning that of the 385 review data tested, there were 354 reviews that were correctly classified by the SVM method. Whereas by using the Maximum Entropy (Maxent) method validates the of accuracy the results produced by SVM method.

3. Based on the results of the classification and association of texts carried out, it is generally known that the majority of e-commerce applications users discuss regarding goods and transactions because they always appear in both positive and negative sentiment classes. In general, the text association method used shows the results of information extraction on positive classes including related to commodities, dealings, features, product and services, sales orders, logistics and delivery, reaction, e-shopping, needs and instalments. While the negative classes that are often complained include merchandise, updates, servers, chat, electronic mail, materialistic-transactions, uploads, promotions, gift-vouchers, brochure and upgrades.

Future Scope: Based on the results of the analysis and conclusions, suggestions can be given as follows:

1. For e-commerce application, the results of information extraction from the reviews that have been given by users, especially negative reviews can be used as an evaluation material in increasing user satisfaction and providing services as much as possible, as well as for the development of further application and business updates.

2. The data used in this study is only one period the application is running on the Google Play system, so it is necessary to add data so that the results of sentiment classification are better.

3. This study only analyzes one e-commerce which is a C2C model, for subsequent research it can compare more than one e-commerce or analyze other e-commerce models such as B2C and C2B.

4. Sentiment class labelling system used in this study is limited to detecting sentiments between words using the lexicon dictionary, so negation words cannot be identified properly, for further research it is better to use a labeling system that has a higher level, which is able to detect sentiments on phrases and sentences.

REFERENCES

[1] Dadianova, Irina & Katasonova, Galiia. (2020). Information Technology. 10.31483/a-156.

[2] http://ispai.in/UI/index.php

[3] https://www.euromonitor.com/

[4] Albano, Roberto & Curzi, Ylenia & Fabbri, Tommaso. (2020). Information System. 10.4324/9781003057260-10.

[5] Mallach, Efrem. (2020). Information Systems. 10.1201/9780429061011.

[6] Sunyaev, Ali. (2020). Information Systems Architecture. 10.1007/978-3-030-34957-8_2.

[7] Yu, L.-Y & Zhang, Yong. (2015). Multiplier maximum entropy algorithm of support vector machines. 1195-1199. 10.1109/CCDC.2015.7162099.

[8] Cindo, Mona & Rini, Dian & Ermatita, Ermatita. (2020). Sentiment Analysis On Twitter By Using Maximum Entropy And Support Vector Machine Method. Sinergi. 24. 87. 10.22441/sinergi.2020.2.002.

[9] Wang, Ran & Kwong, Sam. (2010). Sample selection based on maximum entropy for support vector machines. 3. 1390-1395. 10.1109/ICMLC.2010.5580848.

[10] Akaichi, Jalel. (2017). Sentiment Classification. 10.4018/978-1-5225-1759-7.ch076.

[11] Tripathy, Abinash. (2016). Classification of Sentiment of Reviews using Supervised Machine Learning Techniques. International Journal of Rough Sets and Data Analysis (IJRSDA). 4. 56-74. 10.4018/IJRSDA.2017010104.

[12] D. A. Kristiyanti, A. H. Umam, M. Wahyudi, R. Amin and L. Marlinda, "Comparison of SVM & Naïve Bayes Algorithm for Sentiment Analysis Toward West Java Governor Candidate Period 2018-2023 Based on Public Opinion on Twitter," 2018 6th International Conference on Cyber and IT Service Management (CITSM), Parapat, Indonesia, 2018, pp. 1-6, doi: 10.1109/CITSM.2018.8674352.

[13] Alamsyah, Andry and Fatma Saviera. “A Comparison of Indonesia’s E-Commerce Sentiment Analysis for Marketing Intelligence Effort (case study of Bukalapak, Tokopedia and Elevenia).” (2018).

[14] Suchita V Wawre1, Sachin N Deshmukh2 “Sentiment Classification using Machine Learning Techniques”, International Journal of Science and Research (IJSR), 2018

[15] Santosh Kumar Bharti , Korra Sathya Babu , Sanjay Kumar Jena “Parsing-based Sarcasm Sentiment Recognition in Twitter Data” , IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. ISBN 978-1-4503-3854-7/15/08, 2018.

[16] Ravinder Ahuja, Shantanu Bansal, Shuvam Prakash, Karthik Venkataraman, and Alisha Banga “Comparative study of different Sarcasm Detection Algorithms based on Behavioral Approach ”, Procedia Computer Science 143, 2018.

[17] Sreelakshmi k, Rafeeque “An Effective Approach for Detection of Sarcasm in Tweets”, IEEE, 2018.

[18] Buntoro, Ghulam. (2016). Analisis Sentimen Hatespeech Pada Twitter Dengan Metode Naïve Bayes Classifier Dan Support Vector Machine. Jurnal Dinamika Informatika.

[19] Korovkinas, Konstantinas & Danėnas, Paulius & Garšva, Gintautas. (2017). SVM and Naïve Bayes Classification Ensemble Method for Sentiment Analysis. Baltic Journal of Modern Computing. 5. 10.22364/bjmc.2017.5.4.06.

[20] N. D. Putranti and E. Winarko, “Analisis Sentimen Twitter untuk Teks Berbahasa Indonesia dengan Maximum Entropy dan Support Vector Machine,” p. 10, Jan. 2014.

[21] Sayali D. Jadhav, H. P. Channe2Comparative Study of K-NN, Naive Bayes and Decision Tree Classification Techniques ,International Journal of Science and Research (IJSR) ,2013

[22] Ulwan, M. N. 2016. Pattern Recognition in Unstructured Text Data Using Support Vector Machines and Associations. Thesis Statistics Study Program, Faculty of Mathematics and Natural Sciences, Indonesian Islamic diversity.

[23] Fanani, F: Classification of Software Review on Google Play Using Sentiment Analysis Approach. Essay. Information Technology Study Program UGM Faculty of Engineering Yogyakarta (2017).

[24] Bernhard, Michael & Mühling, Thorsten. (2020). E-Commerce. 10.1007/978-3-658-29037-5_3.

[25] Prasad, Ramjee & Rohokale, Vandana. (2020). E-commerce. 10.1007/978-3-030-31703-4_12.