Sentiment Analysis using Naive Bayes Classifier and Information Gain Feature Selection over Twitter

Manjit Singh [#1], Swati Gupta [#2]

[#1] Student M.Tech. (CSE), Satya College of Engg. and Technology, J.C. BOSE University(YMCA) India
[#2] Assistant Professor Satya College of Engg. and Technology J.C. BOSE University(YMCA) India

Abstract **-** *The development of the internet today is growing very rapidly which indirectly encourages the creation of personal web content that involves sentiments such as blogs, tweets, web forums and other types of social media. Humans often make decisions based on input from friends, relatives, colleagues and others. Supported by the availability of growth and popularity of opinion-rich resources or sentiments such as online site reviews for e-commerce products and personal blogs For example, the expression of personal feelings that allows users to discuss everyday problems, exchange political views, evaluate services and products like Smartphone's Smart TV's etc. This research applies opinion mining method by using Naïve Bayes Classifier and Information Gain algorithm based on Feature Selection. Testing this method uses the E-Commerce based tweet dataset downloaded from the Twitter Cloud Repository. The purpose of this study is to improve the accuracy of the Naïve Bayes algorithm in classifying documents along with Information Gain methodology. Accuracy achieved in this study amounted to 88.80% which is appropriate to evaluate the sentiments.*

Keywords — *Machine Learning, Sentiment Analysis, Information Gain, Naïve Bayes Classifier..*

I. INTRODUCTION

As the development of internet technology in India, led to many online buying and selling sites or more popularly called e-commerce. Nowadays e-commerce is an online buying and selling place that is increasingly in demand by the people of the city [1]. This is because the convenience of transactions without having to come to a physical store, and every year the number of e-commerce users is increasing. According to data released by research firm Ernst & Young, together with Price Water House, India is the country with the largest e-commerce market growth in the world with an average growth of 17% every year[2]. From the increasing number of e-commerce users, of course crime or fraud in cyberspace is also increasing. In research conducted by Kaspersky Lab and B2B International [3], it was revealed that 48% of consumers were targeted by frauds designed to deceive and trick them into disclosing sensitive information and financial data for crime. Concerning, of the 24 countries surveyed, India occupies the highest position of 50% of consumers who have lost their money as a result of being targeted by online fraud [4]. Of the many e-commerce sites in India, there must be more and more online crimes going on, so that people's judgment on an e-commerce can be used as an analysis of the online market.

To overcome this problem, public opinion on an e-commerce can help other communities to be more careful in conducting online transactions. The author feels the need to conduct this research by creating an opinion analysis system or commonly called community sentiment analysis so that it can find out and help provide information about community e-commerce sentiment analysis. The e-commerce objects of this research are phones, smart-TV and smart watches. Reporting from articles written by online www.quora.com, the three e-commerce sites are among the five best e-commerce sites in India that are most frequently visited by consumers. The data is obtained from the Alexa site with the first rank is amazon, the second is flipkart and the third is snapdeal. Public opinion can be obtained from various print and electronic media. Urban society is now more often using social media to comment on a problem, including a product. One of the most popular social media Indian people is Twitter. Indians are known to be very active on social media. Country Business Head Twitter India Manish Maheshwari said, the number of Indian people singing during January to December 2019 reached 4.1 billion tweets. Although he did not mention the number of users in India, Manish said that the number of active Twitter users in India reached 77 percent of all users in the world.. Of course, the information contained in this tweet was very valuable as a tool for determining policy and this can be done with Text Mining..

Text Mining is one technique that can be used to classify documents where Text Mining is a variation of Data Mining that tries to find interesting patterns from a large collection of textual data. Sentiment Analysis or Opinion Mining [6-9] is a computational

study of people's opinions, sentiments and emotions through entities or attributes that are expressed in text form. Sentiment analysis will classify the polarity of the text in the sentence or document to find out the opinion expressed in the sentence or document whether it is positive, negative, or neutral[6-9]. Based on research [5] which examined the accuracy comparison between information gain, chi square, forward selection, backward selection obtained information gain was the best. So that in this study the author will create a Sentiment Analysis application using the Naïve Bayes method for classification with information gain features selection.

II. LITERATURE REVIEW

Research on the classification of sentiments has been carried out by Jockers, Matthew & Thalken, Rosamond [31-32]. In his journal, Bo Pang classified sentiments on film reviews using various machine learning techniques. The machine learning techniques used were Naïve Bayes, Maximum Entropy, and Support Vector Machines (SVM). In that study also used several approaches to extract features, namely unigram, unigram + bigram, unigram + Part of Speech (POS), adjective, and unigram + position. The results of the experiments conducted in this study found that SVM became the best method when combined with unigrams with an accuracy of 82.9%. In a study conducted by an opinion mining system was developed to analyze public opinion in tertiary institutions. In the document subjectivity and target detection sub processes used Part-of-Speech (POS) Tagging using Hidden Makov Model (HMM) [33]. In the results of the Portaging process, a rule is applied to find out whether a document is an opinion or not, and to find out which part of the sentence is an object that is the target of opinion. Documents that are recognized as opinions are then classified into negative and positive opinions (sub process opinion orientation) using Naïve Bayes Classifier (NBC). The test results obtained precision and recall values for the subjectivity document sub processes are 0.99 and 0.88 [34], for the target detection sub processes are 0.92 and 0.93, and for opinion orientation sub processes are 0.95 and 0.94.

Research on classification was also carried out by [35] conducting Data Mining techniques[35-37] that are used to visualize traffic congestion in a city. In this research the method used is Naïve Bayes by combining prior knowledge with new knowledge. From the trial results, the application shows that the smallest accuracy value of 78% is generated in testing with a sample of 100 and produces a high accuracy value of 91.60% in testing with a sample of 13106. The test results with Rapid Miner 5.1 software obtained the smallest accuracy value of 72% with a sample of 100 and the highest accuracy value of 93.58% with a sample of 13106 for the Naive Bayesian Classification method.

Other studies on the classification of film review sentiments were conducted by [38-39] comparing Machine Learning classification methods such as Naïve Bayes (NB), Support Vector Machines (SVM), and Artificial Neural Networks (ANN) and feature selection such as Information Gain, Chi Square, Forward Selection and Backward Elimination. [40] Explained that opinion mining is the process of extracting and evaluating customer feedback or reviews on a topic with text analysis. There are many cases found that opinions are hidden in blogs and forums, and it is very difficult to take sentences that are useful for retail and business. Plus the sentence is not arranged according to correct grammar. This research formed a new methodology to help new customers decide whether to buy or not the product by summarizing reviews of the product. The classification algorithm used is Support Vector Machine, and 4 phases of this research are data collection, preprocessed data, data processing and visual summarization review [41]. Algorithms use data mining techniques in classifying news on Twitter [42]. Gupta, Divya & Sharma, Aditi & Kumar, Mukesh argue that the information on Twitter needs to be organized, one of them is by classification. SVM and Naïve Bayes Classifier are the most popular classification techniques used in text classification. Theoretically, this proves that naïve bayes show faster performance than other classifiers with low error rates, but this is very dependent on the naïve bayes assumption that the feature of the sentence is independent. in a study entitled Twitter News Classification [47]: Theoretical and Practical comparison of SVM against Naive Bayes". In their research [43] also run the evaluation phase of the model to ensure that the model formed from a series of research activities can accurately predict positive and negative opinions. Quantitatively calculate his performance by randomly selecting 600 sentences [44-45] (200 sentences per product) that are parsed with data mining tools. From the results of the training, the best accuracy is to do sentence substring.

positive, negative and neutral opinions on a topic in textual form. Sentiment analysis has been studied and applied by several researchers in the last ten years, and Twitter has been widely used as a data source because it is considered one of the popular social media [47] This paper shows an innovative methodology for diverse texts. Modeling for data in the form of tweets must be different from data in the form of reviews from customers, this is because the form of text and information taken from tweets is very limited compared to customer reviews. The appearance of words in a text also helps research to determine the facts in a sentence. How to find the appearance of this word using TF-IDF calculation. This research also uses several sources to compare models from Facebook, Google+, Yelp, etc. Utilization of sentiment analysis in the industrial

world is one of the needs of companies to remain active for years to come. In a paper from Chavan, Somanath & Chavan, Yash. in 2019 entitled " Sentiment Classification of News Headlines on India in the US Newspaper: Semantic Orientation Approach vs Machine Learning". In this research a system will be made to classify whether the news is positive, negative or neutral news and classify the topic of the news. The case to be classified is a common case that appears on the news. Such as politics, crime, economics, sports, national, world, technology and others. Sentiment analysis using the Naive Bayes Classifier yields an accuracy of 74.2% for sentiment and 32% for the topic.

Twitter is a social media and micro blogging [50] service that allows users to send real-time messages. This message is popularly known as a tweet. A Tweet is a short message with a character length limited to 280 characters. Due to the limitations of the characters that can be written, a tweet often contains abbreviations, slang and spelling errors. From the start, Twitter was created as a mobile-based service that was designed according to the character limits of a text message (SMS), and to this day, Twitter can still be used on any cell phone that has the ability to send and receive text messages. Twitter was created to be a place to share experiences among fellow users without any barrier. By using it, users will be easy to follow trends, stories, information and news from all corners of the world. In addition, Twitter also helps users to always be connected with the people closest to them. When users send a tweet, the message is public and can be accessed by anyone, anywhere and anytime. In fact, for people who follow the Twitter account, the tweet will automatically appear in the person's timeline. The following are some of the terms known on Twitter:

1. Mention. Mention is mentioning or calling other Twitter users in a tweet. Mentions are done by writing ‘ ’ followed by another username

2. Hash tags. The hash tag is used to mark a topic of conversation on Twitter. Writing hash tags starts with '‘#’ followed by the topic being discussed. Hash tags are used to increase the visibility of user tweets.

3. Emoticons. Emoticons are facial expressions that are represented by a combination of letters, punctuation and numbers. Regular users use emoticons to express the mood they are feeling.

4. Trending Topics. If a hashtag is a way to mark a topic of conversation on Twitter, then trending topics is a collection of very popular topics on Twitter.

Information Gain is a feature selection technique that uses the scoring method for nominal or weighting continuous attributes that are discrete using maximum entropy. An entropy is used to define the value of Information Gain [57,58]. Entropy describes the amount of information needed to encode a class. Information Gain (IG) of a term is measured by counting the number of bits of information taken from the prediction of the category with the presence or absence of terms in a document Feature selection technique with information gain means to select feature vertices from a decision tree based on information gain values. The information gain value of a feature is measured by the effect of the feature on class uniformity in data that is broken down into sub data with a certain feature value. Class uniformity (entropy) is calculated in the data before it is broken with equation 1 and in the data after it is broken with equation below:-

$$Entropy(S) = \sum_{i=1}^{k} P_i \log_2 (P_i) \qquad (eq.1)$$

With the value of P_i is the proportion of data S with class i. K is the number of classes at output S.

$$Entropy(S,A) = \sum_{i=1}^{v} \left(\frac{Sv}{S} * Entropy(Sv)\right) \qquad (eq.2)$$

With the value v, all possible values of attribute A, S_v are the subset S where attribute A is worth v. Information gain value is calculated by equation as follows:

$$Gain(S,A) = Entropy(S) - Entropy(S,A) \qquad (eq.2)$$

With the value of Gain (S, A) is the value of information gain. Entropy (S) is the value of entropy before the separator. Entropy (S, A) is the value of entropy after the separator. The value of information gain indicates how much influence an attribute has on the classification of data.

Naïve Bayes Classifier is a popular method used for data mining purposes because of its ease of use [59] and its fast processing time, easy to implement with a fairly simple structure and structure a high level of effectiveness. With simpler language, Naïve Bayes Classifier assumes that the presence or absence of a feature in a class has nothing to do with the presence or absence of other features. For example, something that is red, round, and has a diameter of about 10 cm can be categorized as an apple . Although this feature depends on one feature with another feature. Naïve Bayes Classifier will continue to assume that these features are independent and have no influence on each other[30]. Depending on the probability model, the Naïve Bayes Classifier can be trained to carry out supervised learning very effectively. In a variety of applications, parameter estimation for the Naïve Bayes model uses the maximum likehood method, which means users can use the Naïve Bayes model without needing to trust Bayesian probabilities or without using the Bayesian method refers to the concept of conditional probability[30,45].. In general the Bayes theorem can be denoted in the following equation as under:-

$$P(A|B) = \frac{P(A|B).P(A)}{P(B)} \qquad (eq.4)$$

On Naive Bayes Classification each tweet is represented in attribute pairs (a_1, a_2, a_3, ..., a_n) where a1 is the first word a2 is the word second and so on, whereas V is a class set. At the time of classification, this method will produce the category / class with the highest probability (VMAP) by entering the attributes

(a_1, a_2, a_3, ..., a_n). The VMAP formula can seen in equation as follows:

$$VMAP = \underset{v_j \in v}{\arg max}\ P\left(v_j | a_1 a_2 a_3 \ldots . a_n\right) \qquad (eq.5)$$

By using the Bayes theorem, equation (2.5) can be written Becomes as :-

$$VMAP = \underset{v_j \in v}{\arg max}\ \frac{P\left(a_1 a_2 a_3 \ldots . V_j\right) P(V_j)}{P\left(v_j | a_1 a_2 a_3 \ldots . a_n\right)} \qquad (eq.6)$$

P (a_1, a_2, a_3, ..., an) values are constant for all v_j so that the equation (2.6) can also be stated as equation below:-

$$VMAP = \underset{v_j \in v}{\arg max}\ P\left(v_j | a_1 a_2 a_3 \ldots . a_n | V_j\right) P(V_j) \qquad (eq.7)$$

Naïve Bayes Classifier simplifies this by assuming that in each category, each attribute is free conditional on each other. In other words,

$$P\left(v_j | a_1 a_2 a_3 \ldots . a_n | V_j\right) = \prod_i P(a_i | v_j) \qquad (eq.8)$$

Then if equation (7) is substituted into equation (2.8), then will produce the following equation as under:-

$$VMAP = \underset{v_j \in v}{\arg max}\ P\left(V_j\right) * \prod_i P(a_i | v_j) \qquad (eq.9)$$

P (v_j) and the word a_i probability for each category P (a_i | v_j) are calculated at the time of the training which is formulated as follows :-

$$P(v_j) = \frac{docs_j}{training} \qquad (eq.10)$$

$$P(a_i | v_j) = \frac{n_j + 1}{n + vocabulary} \qquad (eq.11)$$

Where $docs_j$ is the number of documents in category j and training is the number of documents used in the training process. Whereas this is the number of occurrences of a_i words in the v_j category, n is the number of vocabularies that appear in the v_j category and the vocabulary is the number of unique words in all training data.

II(i) EVALUATION PERFORMANCE

Evaluation is conducted to test the results of the classification by measuring the performance value of the system that has been made. The test parameter used for evaluation is accuracy calculated from a confusion matrix table (classification matrix or contingency table). Table 1 shows a confusion matrix for classification into two classes.

Table 1: Confusion Matrix

		True Class	
		Positive	**Negative**
Predicted Class	**Positive**	True Positive (TP)	False Positive (FP)
	Negative	False Negative (FN)	True Negative (TN)

The matrix has four values that are used as a reference in the calculation, namely:

a) True Positive (TP), when the predicted class is positive and the facts are positive.

b) True Negative (TN), when the predicted class is negative, and the facts are negative.

c) False Positive (FP), when the predicted class is positive and the facts are negative.

d) False Negative(FN), when the predicted class is negative and the facts are positive.

The alternative in assessing a system is with accuracy. Accuracy is the accuracy of a system doing the correct classification. Calculation for accuracy can be calculated with the following equation 12:

$$Accuracy\,(A) = \frac{TP + TN}{(TP + TN + FP + FN)} \qquad (eq.12)$$

III. PROPOSED METHODOLOGY

The classification system for sentiment analysis being worked on has a design for how this system will work. General description of the system to be made as in Figure 1 as follows:-

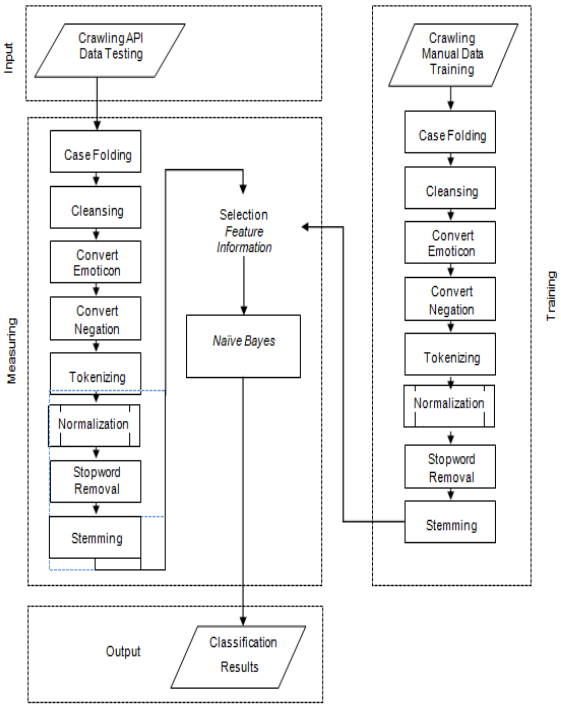


Figure 1 Proposed Schemes or Architecture

PSEUDO CODE OF NAÏVE BAYES CLASSIFIER

For each class c $\in C$ # Calculate P(c) terms
N_{doc}= number of documents/tweets in D
N_c = number of documents/tweets from D in class c
Logprior[c] ← $log \frac{N_c}{N_{doc}}$
V← vocabulary of D
Bigdoc[c] ←**append** (d) **for** d ∈ D **with** class c
For each word in V # Calculate P (w|c) terms
Count (w, c) ← # of occurrences of w in bigdoc[c]

Loglikelihood [w, c] ← $\log \frac{count(w,c)+1}{\sum_{w' in\ V}(count(w',c)+1)}$

Return logprior, loglikelihood, V

PSEUDO CODE OF INFORMATION GAIN

```
Segregate (attributearray, evalue):
    Outliste=e []
    foreie=e1etoelength(attribute array):
        ife (attributearray[i] e==evalue):
            outliste=e [outlist, i] e #e Appende"i" to outlist
    return eoutlist
# Assuminge label set akeevaluese1..M.

Compute Entropy(labels):
    entropye=e0
    for i =e1etoeM:
        probability_i =elength (segregate (labels, i)) e/e length (labels)
        entropy -=eprobability_ie*elog (probability_i)
    return entropy
# Find emoste frequent evalue. Assuminge label set akevaluese1..M

Most Frequently Occurring Value (labels):
    bestCount = -inf
    bestId = none
    for i = 1 to M:
        count_i = length (segregate (label, i))
        if (count_i > bestCount):
            bestCount = count_i
            bestId = i
    return bestId
```

IV. SIMULATION AND RESULTS

This chapter explains the results and discussion of the system that has been built. For this reason, testing of the system is carried out by making the information gain process, the classification process with Naïve Bayes, the accuracy of the system built, and making a data visualization. Here the authors give examples of calculations for data extracted from twitter for testing:-

Tweet	*Feature*	Category
Tweet7	thank you for ordering shanties	?
Tweet8	Please wait for the gift has been ordered	?
Tweet9	sorry can't send tomorrow	?

Table 2 Examples of Data Testing Cases

Weighting Information Gain : The process of weighting words using the information gain algorithm, the weight of each word will be used for selection of word features that have low weights. This process is done by calculating the weight of each word in the training data, then calculating the information gain value of each word will be ranked to get the best features using the threshold. In this study the authors used 20 different thresholds to determine the level of accuracy resulting from the change in the threshold. The best accuracy obtained will be used to select features in the training data. Information gain function can be seen in Table 3.

ID	Word	Word Entropy	Information Gain
103	not yet	0.876189627	0.051510373
102	why	0.899719421	0.279805791
79	already	0.912822731	0.014977726
100	message	0.913819075	0.013888092

Table 3. Results of Information Gain

Naïve Bayes Weighting: In the Naïve Bayes classification process weights each word in the training data. This weighting is done by calculating positive probability and negative probability of each word in the training document. The implementation of the calculation of the probability of each word is done by a function made like Figure 4.8 for positive probability or negative probability. The results of this weighting are stored in the training word database which will then be used as reference material for determining the weight of the testing data. The following result snippet is used for the process of weighting data word training vide function refers to equation 11 in section 2.

Stop Word	**Probability Positive**	**Probability Negative**
RT AnnieDreaXO Film on your iPhone and edit it in iMovie. Your iPhone camera works amaze with natural light. Back camera is better…	3.39805E-06	1.27805E-06
RT stickynickyyyy I need an indestructible iPhone charger	0.000655579	0.00032875

maybe I have not had a phone in a while but this iPhone quality is pretty nice lol pet approved	0.000722322	2.33342E-05
Can someone please explain to me how if I am lock out of my **** iPhone I can get a verification code that is	0.000332233	0.000533443
RT chubiei Top 5 Phones best camera	0.343	0.234
TRIVIA What year was the very first model of the iPhone release NBSYouthVoice NBSUpdates https //t.co/dqtAbemOHH	0.000636637	0.000344422
Whats your best travel photo so far Smile Face With Heart-Eyes iPhone 11 Pro Max	0.03344422	0.000447474

Table 4: Process Analyzing Probability (Positive/Negative) using Naïve Bayes

Classification of the Naïve Bayes Method : Sentiment determination is done by calculating the probability of testing data documents by referring to the probability of said data training. This process uses the naïve Bayes classifier algorithm. For implementation, it can be seen in Table 5. This sentiment classification is done automatically by implementing the Naïve Bayes Classification algorithm.

This process is implemented in the sentiment classification function by comparing the weight of each word in the testing data with the word in the training data, if the word is not found in the training data, the weight is assessed. The result of each training document is the positive and negative probability word weight. Furthermore, the document weight is compared, if the positive probability document weight is greater than the sentiment result is positive, and if the negative probability weight is greater than the sentiment result is negative.

Stop Word	Classification
RT AnnieDreaXO Film on your iPhone and edit it in iMovie. Your iPhone camera works amaze with natural light. Back camera is better…	Positive
RT stickynickyyyy I need an indestructible iPhone charger	Positive
maybe I have not had a phone in a while but this iPhone quality is pretty nice lol pet approved	Positive
Can someone please explain to me how if I am lock out of my **** iPhone I can get a verification code that is	Negative
RT chubiei Top 5 Phones best camera	Positive
TRIVIA What year was the very first model of the iPhone release NBSYouthVoice NBSUpdates https //t.co/dqtAbemOHH	Positive
Whats your best travel photo so far Smile Face With Heart-Eyes iPhone 11 Pro Max	Positive

Table 5: Process Naïve Bayes for Classification

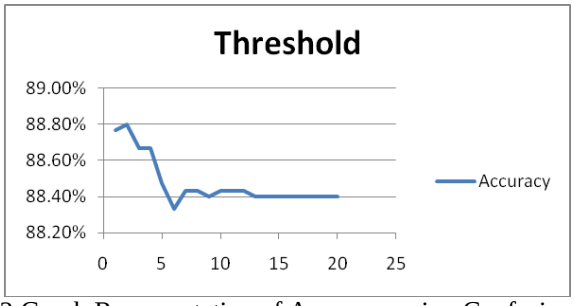
System Testing and Discussion : This section will explain the analysis of system test results that serves to determine the performance of the program in the classification process. This test is done by using all training data used as data testing to test the accuracy of the naïve Bayes method with information gain selection. Evaluation of system accuracy testing is done by giving epoch = 20, so the results of the accuracy of this system are 20 experiments. In this research, each time epoch is increased, a threshold is given by reducing 100 words for each word of training data. The reduced word is taken from the lowest information gain weight value, leaving the word with the best weight.

The results of testing this system can be seen in the following table 6:-

Epoch	Threshold	TP	TN	FP	FN	Accuracy	Time(s)
1	4100	733	1930	41	296	88.77%	651.8943
2	4000	731	1933	38	298	88.80%	672.8518
3	3900	728	1932	39	301	88.67%	669.4441
4	3800	725	1935	36	304	88.67%	646.3554
5	3700	720	1934	37	309	88.47%	661.5116
6	3600	714	1936	35	315	88.33%	629.6979
7	3500	716	1937	34	313	88.43%	636.2398
8	3400	716	1937	34	313	88.43%	644.808
9	3300	716	1936	35	313	88.40%	629.2616
10	3200	716	1937	34	313	88.43%	608.5906
11	3100	716	1937	34	313	88.43%	599.5477
12	3000	716	1937	34	313	88.43%	587.7442
13	2900	716	1936	35	313	88.40%	591.497
14	2800	716	1936	35	313	88.40%	564.8959
15	2700	716	1936	35	313	88.40%	575.5903
16	2600	716	1936	35	313	88.40%	568.4411
17	2500	716	1936	35	313	88.40%	555.9261
18	2400	716	1936	35	313	88.40%	536.6505
19	2300	716	1936	35	313	88.40%	536.9626
20	2200	716	1936	35	313	88.40%	514.0036

Table 6 Confusion Matrix with Accuracy

The system test results in table 6 can be seen each threshold reduction makes the accuracy and processing time change. At the 4000 threshold the highest accuracy is 88.80%, but the processing time is longer than before. In the next threshold reduction the relative processing time decreases, but also the accuracy decreases when the threshold decreases. That is because the more the threshold is reduced, the fewer words are processed so that time can be faster. In Epoch 13 to 20 the accuracy is stagnant at 88.40%. To more easily understand the picture between accuracy and speed, the writer makes a graphical visualization like in Figure 2



2:Graph Representation of Accuracy using Confusion Matrix

In Figure 2 the accuracy graph is depicted with a blue line. It can be concluded that the reduced feature gain information speeds faster processing, and accuracy decreases only slightly, so if this system is needed for speed then the 20th epoch can be selected. In this study the scheme chose the highest accuracy for the training data material so that the 2nd epoch was chosen.

V. CONCLUSIONS AND FUTURE SCOPE

Conclusion : Based on the research that has been carried out, it can be concluded that:

1. The Naïve Bayes method with Information Gain feature selection is proven to be able to analyze sentiments automatically. The trial is conducted using real-time testing data, each data is classified with positive or negative sentiment.
2. The performance of the combination of the Naïve Bayes method with the Information Gain feature selection has increased in the 2200 threshold limit the processing time to 514 seconds is faster than before the addition of Information Gain. Sentiment analysis system if coupled with the Information Gain feature selection can increase accuracy up to 88.8%.

Future Scope : Some suggestions for future research development are as follows:

1. In future studies the authors suggest adding the Adaboost method to reduce bias in order to significantly increase the accuracy of the Naïve Bayes algorithm.
2. Add a dictionary collection to the slang word database, because on social media Twitter is too many languages that are not standard.
3. Proposed scheme can be integrated with Huge Data Repositories like Big Data.

REFERENCES

[1] https://en.wikipedia.org/wiki/E-commerce_in_India

[2] https://www.slideshare.net/mailforveena/ernst-and young-rebirthof-ecommerce-in-india-report

[3] https://www.kaspersky.com/blog/security_risks_report_financial_impact

[4] https://economictimes.indiatimes.com/wealth/personal-finance-news/over-50-indians-fell-prey-to-discount-scams-tips-to-stay-safe-this-holiday-season/articleshow/72453319.cms

[5] Natarajan, Bhalaji & Kb, Sundharakumar & Selvaraj, Chithra. (2018). "*Empirical study of feature selection methods over classification algorithms*". International Journal of Intelligent Systems Technologies and Applications. 17. 98. 10.1504/IJISTA.2018.091590.

[6] Zhang, Lei & Liu, Bing. (2017). "*Sentiment Analysis and Opinion Mining"*. 10.1007/978-1-4899-7687-1_907.

[7] Zhang, Lei & Liu, Bing. (2016). "*Sentiment Analysis and Opinion Mining*". 1-10. 10.1007/978-1-4899-7502-7_907-1.

[8] Arya, Apoorva & Shukla, Vishal & Negi, Arvind & Gupta, Kapil. (2020). *"A Review: Sentiment Analysis and Opinion Mining*". SSRN Electronic Journal. 10.2139/ssrn.3602548.

[9] M.K., Sudha. (2020). "*Social Media Sentiment Analysis for Opinion Mining*". International Journal of Psychosocial Rehabilitation. 24. 3672-3679. 10.37200/IJPR/V24I5/PR202075.

[10] (2020). Machine Learning. 10.1007/978-981-15-2770-8_6.

[11] Suthaharan, Shan. (2016). "*Supervised Learning Models"*. 10.1007/978-1-4899-7641-3_7.

[12] Quinto, Butch. (2020). Unsupervised Learning. 10.1007/978-1-4842-5669-5_4.

[13] Sujatha, Christy. (2018). "*Building Predictive Model For Diabetics Data Using K Means Algorithm"*.

[14] Chang, Mark. (2020). Reinforcement Learning. 10.1201/9780429345159-11.

[15] Schuppert, A. & Ohrenberg, A.. (2020). data mining. 10.1002/9783527809080.cataz04524.

[16] Bramer, Max. (2020). "*Data for Data Mining"*. 10.1007/978-1-4471-7493-6_2.

[17] Wardle, Claire & Greason, Grace & Kerwin, Joe & Dias, Nic. (2019). "*Data Mining*". 10.32388/635643.

[18] Dawson, Catherine. (2019). "*Data mining*". 10.4324/9781351044677-13.

[19] Verma, Nishchal & Salour, Al. (2020). "*Feature Selection*". 10.1007/978-981-15-0512-6_5.

[20] Soltanian, Ali & Rabiei, Niloofar & Bahreini, Fatemeh. (2019). "*Feature Selection in Microarray Data Using Entropy Information*".10.15586/computationalbiology.2019.ch10.

[21] Bramer, Max. (2020). "*Decision Tree Induction: Using Entropy for Attribute Selection*". 10.1007/978-1-4471-7493-6_5.

[22] Hermanson, Eric. (2018). "*Claude Shannon Information Theory*".

[23] Strawn, George. (2014). "*Claude Shannon: Mastermind of Information Theory*". IT Professional. 16. 70-72. 10.1109/MITP.2014.87.

[24] Guizzo, Erico. (2003). "*The Essential Message: Claude Shannon and the Making of Information Theory"*.

[25] Nwanganga, Fred & Chapple, Mike. (2020). "*Naïve Bayes*". 251-275. 10.1002/9781119591542.ch7.

[26] Webb, Geoffrey. (2016). "*Naïve Bayes*." 10.1007/978-1-4899-7502-7_581-1.

[27] Janssen, Jürgen & Laatz, Wilfried. (2017). Naive Bayes. 10.1007/978-3-662-53477-9_25.

[28] Kaviani, Pouria. (2017). "*Naïve Bayes Algorithm".*

[29] Cichosz, Paweł. (2015). "*Naïve Bayes classifier*". 10.1002/9781118950951.ch4.

[30] Caraffini, Fabio. (2019). "*The Naive Bayes learning algorithm*". 10.13140/RG.2.2.18248.37120.

[31] Jockers, Matthew & Thalken, Rosamond. (2020). "*Sentiment Analysis*". 10.1007/978-3-030-39643-5_14.

[32] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. "*Thumbs up? Sentiment Classification using Machine Learning*" Techniques. Proceedings of EMNLP, 2002. Introduced polarity dataset v0.9.

[33] A. Paul, B. S. Purkayastha and S. Sarkar, *"Hidden Markov Model based Part of Speech Tagging for Nepali language*," 2015 International Symposium on Advanced Computing and Communication (ISACC), Silchar, 2015, pp. 149-156, doi: 10.1109/ISACC.2015.7377332.

[34] Rusydiana, Aam & Firmansyah, Irman & Marlina, Lina. (2018). "*SENTIMENT ANALYSIS OF MICROTAKAFUL INDUSTRY".* Vol 6, No 1 (2018). 10.15575/ijni.v6i1.3004.

[35] Yang, Xin-She. (2019). "*Data mining techniques".* 10.1016/B978-0-12-817216-2.00013-2.

[36] Roiger, Richard. (2017). "*Basic Data Mining Techniques*". 10.1201/9781315382586-3.

[37] Kaur, Harmeet & Kaur, Jasleen. (2018). *"Survey on Data Mining Technique*". International Journal of Computer Sciences and Engineering. 6. 915-920. 10.26438/ijcse/v6i8.915920.

[38] Gritta, Milan. (2019). *"A Comparison of Techniques for Sentiment Classification of Film Reviews".*

[39] Timor, Mehpare & Dincer, Hasan & Emir, Şenol. (2012). "*Performance comparison of artificial neural network (ANN) and support vector machines (SVM) models for the stock selection problem: An application on the Istanbul Stock Exchange (ISE)-30 index in Turkey*". African Journal of Business Management. 6. 1191-1198.

[40] C, Spoorthi & Ravikumar, Dr & M.J, Mr. (2019). "*Sentiment Analysis of Customer Feedback on Restaurant Reviews*". SSRN Electronic Journal. 10.2139/ssrn.3506637.

[41] Sui, Haiyang & Khoo, Christopher & Chan, Syin. (2003). "*Sentiment Classification of Product Reviews Using SVM and Decision Tree Induction*". 14. 10.7152/acro.v14i1.14113.

[42] Gupta, Divya & Sharma, Aditi & Kumar, Mukesh. (2020). "*TweetsDaily: Categorised News from Twitter*". 10.1007/978-981-15-0222-4_5.

[43] Rajvanshi, Nitin & Chowdhary, Prof. (2017). "*Comparison of SVM and Naïve Bayes Text Classification Algorithms using WEKA*". International Journal of Engineering Research and. V6. 10.17577/IJERTV6IS090084.

[44] Kesumawati, Ayundyah & Thalib, A.K.. (2018). "Hoax classification with Term Frequency - *Inverse Document Frequency using non-linear SVM and Naïve Bayes*". International Journal of Advances in Soft Computing and its Applications. 10. 116-128.

[45] Dey, Sanjay & Wasif, Sarhan & Tonmoy, Dhiman & Sultana, Subrina & Sarkar, Jayjeet & Dey, Monisha. (2020). "*A Comparative Study of Support Vector Machine and Naive Bayes Classifier for Sentiment Analysis on Amazon Product Reviews*". 217-220. 10.1109/IC3A48958.2020.233300.

[46] Garšva, Gintautas & Korovkinas, Konstantinas. (2018). "*SVM and Naïve Bayes Classification Ensemble Method for Sentiment Analysis".* Baltic J. Modern Computing.

[47] Dilrukshi, Inoshika & De Zoysa, Kasun & Caldera, Amitha. (2013). "*Twitter news classification using SVM".* Proceedings of the 8th International Conference on Computer Science and Education, ICCSE 2013. 287-291. 10.1109/ICCSE.2013.6553926.

[48] FARKHUND IQBAL, "*A Hybrid Framework for Sentiment Analysis Using Genetic Algorithm Based Feature Reduction",* Digital Object Identifier 10.1109/ACCESS.2019.2892852

[49] Chavan, Somanath & Chavan, Yash. (2019). "*Sentiment Classification of News Headlines on India in the US Newspaper: Semantic Orientation Approach vs Machine Learning*". 10.13140/RG.2.2.34008.75522.

[50] Duffy, Andrew. (2020). Twitter. 10.4324/9780429356612-7.

[51] Bell, Jason. (2020). "*The Twitter API Developer Application Configuration*". 10.1002/9781119642183.app2.