Analytical Study of Sentiments using Information Gain and Chi Square with Naïve Bayes Algorithm to enhance the performance of Sentiment Analysis

[1]Kapil Raj Dhaybhai, [2]Jitendra Singh Chauhan

[1]M.Tech Scholar, [2]Associate Professor
[1]Department of Computer Science & Engineering,
[1]Aravali Institute of Technical Studies, Umarda, Udaipur (Raj),India
Email – [1]kapilraj1412@rediffmail.com, [2]chauhan.jitendra@live.com

Abstract: *We are drowning in information yet starved for knowledge and with the development of web and web-based social networking inflow of information increments exponentially, to mine knowledge from this crude information. In this paper we take the Sentiment analysis for the Natural Language Processing to decide if the state of mind behind the content is Positive, negative or impartial. The main aim of this paper is to present Sentiment Analysis of Movie reviews by applying some machine learning strategies with associating feature selection technique: information gain and Chi Square to enhance the viability of the model. In our proposed procedure firstly the dataset utilized is of movie review. Furthermore, some pre-processing schemes are applied on the dataset. Thirdly, to acquire the outcomes with the expansion of viability of model we have use Naive Bayes algorithm with the combination of feature selection procedures. Model so constructed is optimized for better performance.*

Key Words:*Sentiment Analysis, Opinion Mining, Feature Selection, Chi Square, Information Gain, Naïve Bayes, Information retrieval, classifier, movie review*

1. INTRODUCTION:

1.1. Sentiment Analysis

Sentiment analysis (frequently utilized by NLP specialists) and opinion mining (received by the information recovery group) are regularly utilized interchangeably. In spite of the fact that, these examination fields are firmly related and are considered as one. Opinions, Emotions and Sentiments are the key ideas of the exploration, there is no record on which information to consider as an opinion or a sentiment. Opinions change from individual to individual. The objective of sentiment mining is to essentially distinguish subjective (opinions), objective (facts) and general sentiment of the text. Sentiment analysis can be arranged into three sections appeared by figure 1.1:

• **Document Level:** It considers the entire document as its central unit of data which is then arranged for positives and negatives. We have utilized document level sentiment analysis for our discoveries.

• **Sentence Level:** Analysis utilizes sentences for grouping of sentences. Not more than regularly document level analysis and sentence level analysis is utilized reciprocally as a document is a gathering of sentences.

• **Aspect Level:** Analysis manages a specific aspect or normal for the item. For instance a cell phone survey can have distinctive aspect, for example, battery life, camera, screen size and determination, RAM.

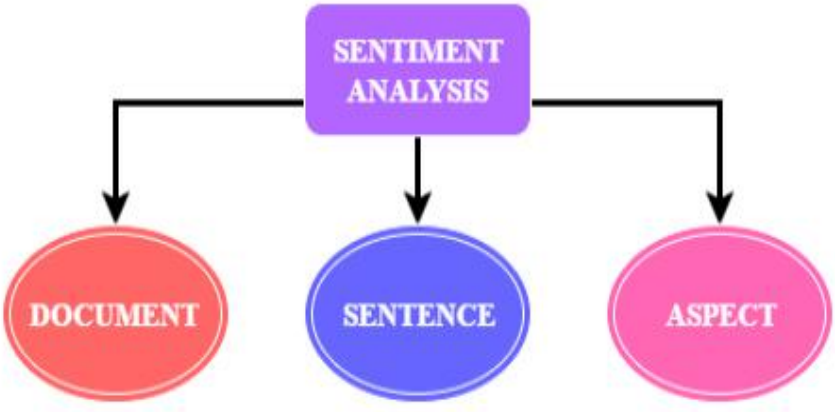

Fig. 1.1. Types of Sentiment Analysis

1.2. Pre-Processing

Pre-processing strategy is the initial phase in the content mining process and assumes an exceptionally essential part in content mining strategies and applications. The information in this present reality is inadequate, conflicting and uproarious. So to evacuate there inconsistencies information pre-processing is utilized. Pre-processing is utilized so that there is less information for the model to learn quicker and for high precision.

1.2.1 Stop Word Removal

In Information Retrieval and Text mining many every now and again utilized words in English are futile as they don't bestow any sentiment in the content and they make the content look heavier and less imperative for investigators. These words are known as Stop words and expelling stop words decrease the dimensionality of term space. Stop-words are frequent words like pronouns, prepositions, conjunctions that convey no data. In English dialect, there are around 400-500 Stop words (NLTK contains a rundown of 120 stop words). Cases of such words incorporate 'the', 'is', 'and', 'an'. The initial step amid pre-processing is to determine the issue of stop words, which has demonstrated as vital as they cut down general execution [1]. There are two sorts of strategies for the expulsion of stop words [2] that are the great strategy for utilizing stop list [3] and Tf-Idf .Tf-idf . The tf-idf weight is union of two terms:

• **TF: Term Frequency**, which evaluate how often a term happens in a record.

TF (t) = (No. of times term t appears in a document)/(Total no. of terms in the document).

• **IDF: Inverse Document Frequency**, which measures how essential a term is. While registering TF, all terms are supposed equally vital. However it is realized that certain terms, for example, "is", "on", and "of", "this", “are”, "that", may appear a considerable measure of times yet have little importance.

IDF (t) = log e (Total no. of documents/Number of documents with term t in it).

1.2.2 Stemming

Stemming procedures tries to discover the foundation of a word. Stemming change over words to their stems which consider dialect subordinate semantic information. As the words with a similar root for the most part portray same or generally close significance, these words can be conflated. For instance, the words, client, clients, utilized, utilizing all can be stemmed to the word 'Utilize'. There are two focuses to consider while utilizing a stemmer:

• Words having distinctive significance ought to be kept partitioned.
• Morphological types of a word are accepted to have a similar base significance and subsequently it ought to be mapped to a similar stem. The Porter Stemmer algorithm is the most ordinarily utilized algorithm in English. Watchmen stemming algorithm [2] [4] proposed in 1980 is a standout amongst the most well-known stemming algorithm.

1.3. Feature Selection

A "feature" (variable or attribute) alludes to the normal for the information. Features that might be discrete, continuous, or nominal are generally gathered before the features are determined or picked. Features can be:

• **Irrelevant:** Irrelevant features are those which don't impact the yield.
• **Relevant:** These are features which have an impact on the yield as they have an intrinsic significance which can't be accepted by the rest.
• **Redundant:** An excess exists at whatever point a feature can play the part of another.

The accompanying is the rundown of meanings of Feature Selection (FS) that are theoretically unique:

• **Idealized:** Find the insignificantly estimated feature subset that is fundamental and adequate to the objective idea [5].
• **Classical:** Select a subset of M features from an arrangement of N features, M <N, with the end goal that the estimation of a foundation work is improved over all subsets of size M [6].
• **Improving Prediction precision:** The point of feature selection is to pick a subset of features for enhancing prediction exactness or diminishing the span of the structure without fundamentally diminishing prediction precision of the classifier assembled utilizing just the chose features [7].
• **Approximating unique class appropriation:** the objective of feature selection is to choose a little subset to such an extent that the subsequent class dissemination, given just the qualities for the chose features, is as close as conceivable to the first class circulation given all feature esteems [7].

There are four fundamental strides in a commonplace feature selection strategy as appeared in figure 1.2:

- An era method produces next hopeful subset which holds enough information for better performance of the model.
- An assessment work assesses the hopeful subset;
- A halting foundation chooses when to end;
- An approval system approves the subset.

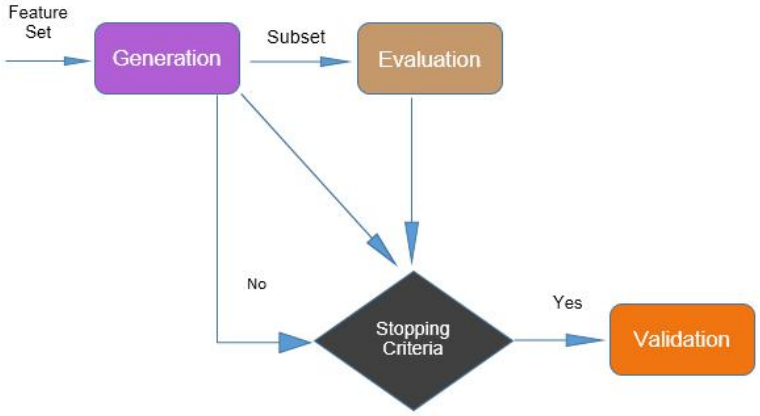

Fig.1.2 Steps of Feature Set Selection

The generation strategy is basically an inquiry method [8] [9] which fundamentally intend to produce subsets of features for assessment process. The generation work starts:

• With no features,
• With all features,
• With an irregular subset of features.

There are 3 general classes of feature selection calculations: filter method, wrapper method and embedded method.

• **Filter Method:** A scoring capacity is utilized by filter feature selection methods to dole out a score to the dataset which assistant chooses if the dataset merits keeping as it influences the execution of the model as appeared in Figure 1.3. The methods are frequently univariate and consider the feature freely, or with respect to the needy variable.

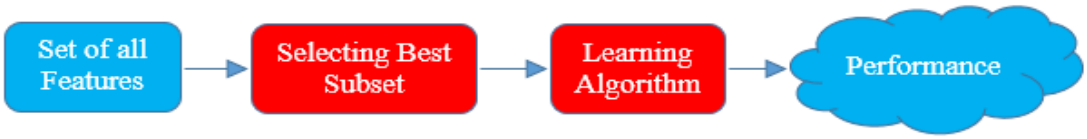

Fig.1.3. Filter Method

• **Wrapper Method:** Selection of features is finished utilizing a pursuit issue after which these chose set of features are assessed and contrasted and other blend and a score is relegated in light of model precision as appeared in figure 1.4. Features are included or evacuated utilizing stochastic hunt, for example, irregular slope climbing, methodical.

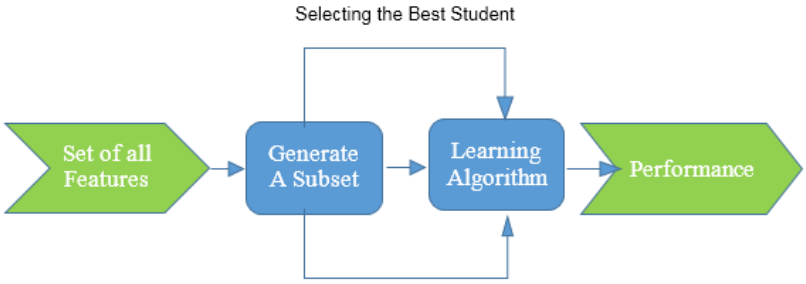

Fig.1.4. Wrapper Method

Embedded Method: Embedded methods absorb the information of the features (which features contributes the best to the precision) while the model is being made as appeared in figure 1.5.

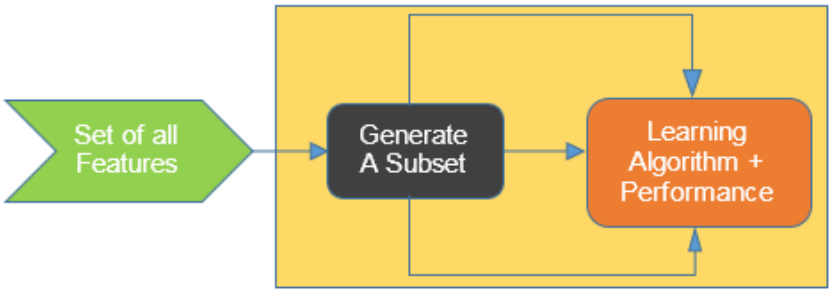

Fig.1.5 Embedded Method.

1.3.1 Chi Square Test

Chi Square Test is utilized as a part of the field of insights to test the freedom between two occasions. For the computation of chi square, we take the square of the distinction between the observed (o) and expected (e) qualities and afterward partition it by the expected esteem.

Chi Square measures the deviation between expected counts (e) and observed Count (o) as appeared in condition (i)

$$\chi 2 = \Sigma \frac{(o-e)^2}{e} \qquad \text{----------(i)}$$

1.3.2 Information Gain

Information gain measures the importance of a component for visualization of a class by knowing the nearness or nonappearance (recurrence) of a specific term in a record. To put it plainly, after the estimation of highlight is gotten the information gain measures the lessening in entropy of the class variable or we can likewise say that it gauges how visit an element is in one class when contrasted with different classes. Information gain is often utilized as a term-goodness standard in the field of machine learning [10] [11]. It gauges the quantity of bits of information got for class forecast by knowing the nearness or nonattendance of a term in an archive. Information gain is defined as:

$$I(w) = -\sum_{i=1}^{k} P_i \cdot \log(P_i) + F(w) \cdot \sum_{i=1}^{k} p_i(w) \cdot \log(p_i(w)) + \\ +(1 - F(w)) \cdot \sum_{i=1}^{k} (1 - p_i(w)) \cdot \log(1 - p_i(w)) \qquad \text{---- (ii)}$$

In the recipe in condition (ii) appears if the record contains the word w then Pi is the worldwide likelihood of class i, pi(w) is the likelihood of class I, F(w) is the bit of the reports that contain the word w. Information gain I(w) and biased energy of the word (w) are relative . The intricacy of the information gain for a corpus containing n records and d words is O(n·d· k)

1.4 Classification

Automatic Text classification has dependably been a vital application and research point since the beginning of advanced reports. Today, Text classification is a need because of the huge measure of content archives that we need to manage day by day.

Areas in which Text classification is regularly utilized are:

• News filtering and Organization: News administrations deliver huge measure of electronic information in the form new articles, recordings, realistic pictures. As it winds up plainly hard to organize this information physically Media organizations require mechanized techniques for news categorization [9] (text filtering) to stay aware of the enormous inflow of news.
• Document Organization and Retrieval: With web a wide range of information is available on the web. Be it advanced libraries, logical writing, web accumulations, articles or web-based social networking posts, all these should be organized progressively with the goal that it can be helpful to the clients and perusing and recovery is simple [12].
• Opinion Mining or Sentiment Analysis: Customer surveys or opinions are regularly short text archives which can be mined to decide helpful information from the audit [13].
• Email Classification and Spam Filtering: Classification of email is an exceptionally basic practice. It winds up plainly tedious for people to categorize every one of the emails [14] [15] [16] and furthermore to distinguish spam sends so a programmed strategy is expected to decide either the subject or to decide garbage email [17].

1.4.1 Naïve Bayes

In our proposed system we have utilize Naïve Bayes (NB) as classification algorithm which depends on bayes hypothesis. Bayes' Theorem relates restrictive probabilities and discovers its underlying foundations in likelihood hypothesis that demonstrates the impact of the event of one occasion on another. The critical terms in bayes hypothesis are the earlier likelihood which is the likelihood got at the outset before any extra data is acquired and back likelihood is the amended likelihood after some confirmation is gotten.

Bayes' Theorem can be composed as

$$P\left(\frac{A}{B}\right) = \frac{P\left(\frac{B}{A}\right)P(A)}{P(B)} \text{-------- (iii)}$$

Where,

- P(A) is the prior probability of A
- P(B) is the prior probability of B
- P(A|B) is the posterior probability of A given B
- P(B|A) is the posterior probability of B given A

2. PROPOSED METHODOLOGY:

Figure 2.1 demonstrates the definite stream outline for development of the model proposed. In consequent area we will give definite data about the approaches and functionalities utilized as a part of each progression of model development for our proposed framework.

2.1 Phase 1 (Information securing)

In our proposed procedure the information utilized is of motion picture audit. It has a few surveys of motion pictures isolated into positive or negative subcategory.

2.2 Phase 2 (Pre-processing)

Pre-processing is done in our proposed system to evacuate the words which obstruct our procedure of conclusion examination by expanding the quantity of false positives or false negative. We need to invoke a framework which accurately identifies the class with no false prediction. In proposed demonstrate we utilized two pre-processing systems stopwords removal and stemming.

Stop-words removal: In our model stop words are evacuated utilizing Tf-idf. Term Frequency-Inverse Document Frequency is known to locate the vital and no so essential word in the record. We ascertained tf-idf utilizing scikit-learn library.

Stemming: Stemming algorithm endeavor to consequently expel additions (and now and again prefixes) keeping in mind the end goal to discover the root word or stem of a given word. NLTK gives a few stemmer interfaces. In our proposed technique we have utilized doorman stemmer to discover the root words.

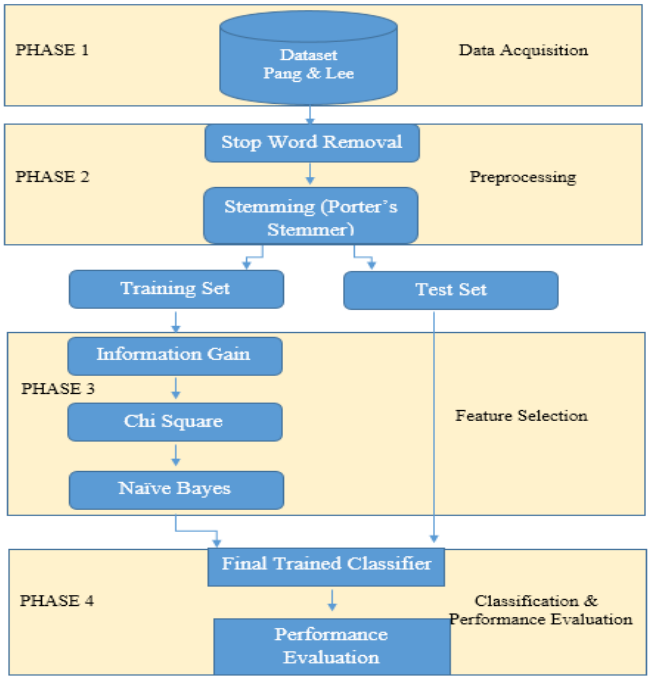

Fig 2.1. Proposed Methodology Design

2.3 Phase 3 (Feature selection)

In our approach we utilized two feature selection strategies chi square an information gain.

• **Chi square:** In our proposed system we utilized chi square as a scoring capacity with which we can discover if two terms are related to each other We at that point apply chi square capacity which gives the scoring capacity. Subsequent to applying chi square we learn whether the bigram or trigram happens as much of the time as every individual word.
• **Information gain:** It causes us in comprehension if a word is educational or not. On the off chance that a word for the most part happens in positive survey and once in a while in negative audits it would main be able to that the word is vital. So we discover how basic a word is in a specific class contrasted with different classes. For applying information gain in our model we have influenced utilization of chi to square scoring capacity.

2.4 Phase 4 (classification)

For our proposed procedure we have utilize guileless bayes baseline algorithm. To start with the information is isolated into training set (1500) and Test set (500).

2.5 Measure of optimal solution

Performance of our prepared model will be measured by exactness as well as by different measures which help in the better comprehension of model calculations. Confusion Matrix is appeared in table 2.1.

TABLE 2.1 Confusion Matrix

		Predicted	
Observed		**Yes**	**No**
	Yes	**True positive**	**False negative**
	No	**False positive**	**True positive**

• True Positive (TP) values are the effectively anticipated positive esteems which mean the estimation of both real and anticipated class is yes.
• True Negatives (TN) values are the accurately anticipated negative esteems which implies the estimation of both genuine and anticipated class is no.
• False positives and false negatives happen when the real class repudiates with the anticipated class.
• False Positives (FP) happens when real class is no and anticipated class is yes.
• False Negatives (FN) happens when genuine class is yes however anticipated class in no.

Figure 2.2 demonstrates the dissemination of positive and negative items. The dashed line in the center demonstrates the choice edge of classifier. The regions set apart as FN and FP speaks to the inaccurately grouped items.

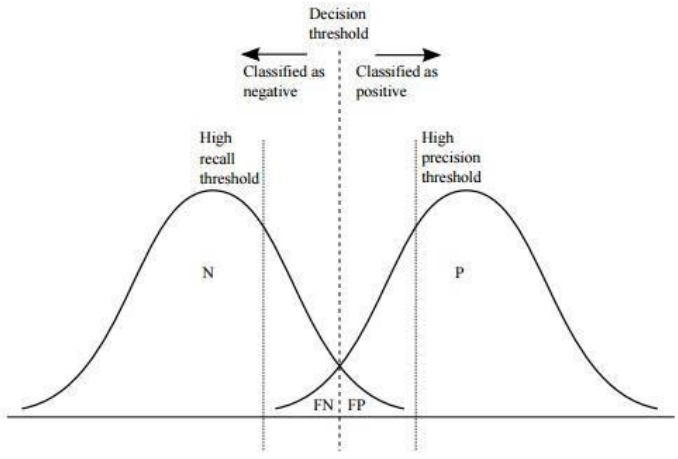

Fig 2.2. Precision and Recall

Precision and recall for the most part go up against each other and the spotted lines inside the bend speak to the choice limit for high recall or high precision separately. On the off chance that the choice limit is moved to one side, there will be a greater amount of FN objects and less FP objects, bringing about low recall and higher precision.

Accuracy: Accuracy is the most widely recognized execution measure and it is a proportion of effectively anticipated perception to the aggregate perceptions. Accuracy is formalized:

$$Accuracy\ (a) = \frac{TP + TN}{TP + FP + FN + TN}$$

Precision (Completeness): Precision is the proportion of accurately anticipated positive perceptions to the aggregate anticipated positive perceptions. On the off chance that precision is high there will be low false positives. It is frequently restricting to recall as it is intuitive that lower recall give higher precision.

$$Precision\ (p) = \frac{TP}{TP + FP}$$

Recall (Sensitivity): Recall is the proportion of accurately anticipated positive perceptions to the all perceptions in real class. Higher recalls relates to less false negatives as in condition

$$Recall\ (r) = \frac{TP}{TP + FN}$$

F-Measure: In factual examination, the F-measure (likewise F1 score or F-score) is a measure of the exactness. It considers both the precision (the quantity of right positive outcomes isolated by the quantity of every single positive outcome) and the recall (is the quantity of right positive outcomes separated by the quantity of positive outcomes that ought to have been returned) of the test to register the score.

$$F1 = 2\frac{Precision * Recall}{Precision + Recall}$$

3. RESULTS AND DISCUSSIONS:

3.1 Experimental Results andAnalysis

The dataset required to assess proposed research of conclusion investigation was first presented by Bo Pang and Lillian Lee [18]. The dataset is accessible under the name Sentiment Polarity Dataset Version 2.0. It is a User-Created movie reviews documented on the Internet Movie Database or IMDB (http://reviews.imdb.com/Reviews). This dataset is accessible on the connection (http://www.cs.cornell.edu/People/pabo/film survey information). The dataset is partitioned into two orders positive and negative. This dataset contains aggregate of 2000 printed reviews as .txt augmentation and is additionally arranged into Positive (pos) and Negative (neg) reviews.

3.2 Experimental Setup

In our proposed sentiment analysis method Classifier was readied using more than 1000 reviews. After pre-preprocessing of dataset using tf-idf for stop word removal and porter stemmer for stemming of words to its root word, we need to pick the extent of training set and test set with the objective that our proposed classifier is arranged properly. A self-assertive extent can be picked yet the extent of getting preparing set should be all the more by then test set as it urges the classifier to learn better so we pick 75% (1500) as training set and 25% (500) as test set. Both the training and the test sets are kept differently and are not introduced to each other at the period of model improvement and after that portrayal (Naive Bayes) is associated on getting training data to assemble the proposed appear

3.3. Performance of Proposed Model with Feature Selection

Underneath diagram demonstrates the accuracy of the model after test set is connected. Accuracy isn't the main measure to characterize the performance of the model. There are different Measures which accurately characterize the status of the model.

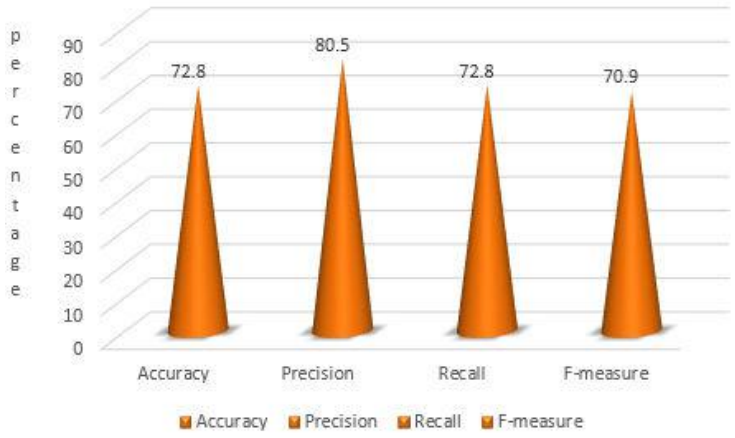

Fig 3.1. Performance metrics of our proposed methodology without feature selection.

These measures are Precision, Recall and F-measure. Figure 3.1 demonstrates the performance measure of the model without feature selection. We watch that the accuracy is low at 72.8%. Likewise precision and recall are additionally low at 80.5% and72.8%.

Feature Selection chooses features that will upgrade the performance of the model or that contains the most extreme data that will help in better comprehension of extremity of report. So to upgrade our model we apply Chi Square which fills in as a scoring capacity that aides in finding the autonomies between two words. This aides in finding the collocation (Words that co-happen more frequently than anticipated). Discovering Collocations can help in better understanding the extremity of the archive.Figure 3.2 demonstrates that when we connected feature selection (chi square) there is noteworthy increment in the performance of our proposed show.

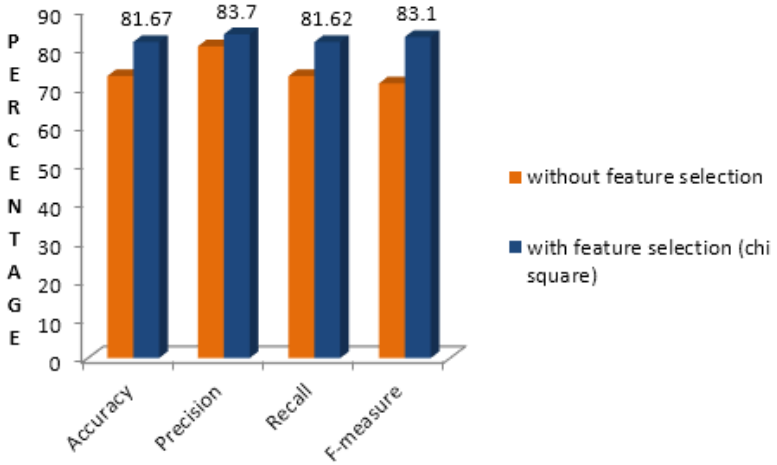

Fig 3.2. Performance Measure comparison of our proposed methodology with and without chi square

Accuracy expanded by 8.87%, Precision expanded by 3.2%, Recall expanded by 8.82% and F-measure expanded by 12.2%. The outcome demonstrates that including chi square can expand classifiers adequacy. At the point when classification model has a great many features of many (if not most) of those features are low in information. Therefore a need emerge to expel these features as they diminish the performance and causes overfitting and revile of dimensionality. Low information features can be evacuated utilizing information gain, otherwise known as mutual information.

Figure 3.3 demonstrates the performance measure of model in the wake of applying information gain at the features. Figure additionally contrasts the aftereffect of model and without information gain. We conclude that there is 20.25% expansion in accuracy; Precision is expanded by 12.6% while Recall expanded 21.2% and F-measure expanded 22.52%.

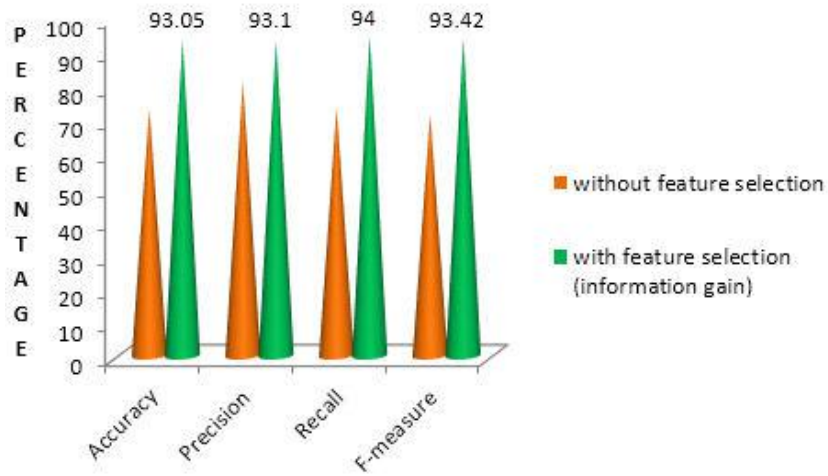

Fig 3.3. Performance Measure comparison of our proposed methodology with and without information gain

Figure 3.4 shows correlation between the performances got from applying chi square and information gain. Performance increments on applying information gain as there is 11.38% expansion in accuracy, 9.4% expansion in precision, 12.38% increment in recall and 10.32% expansion in f-measure. Along these lines we can state that use of significant features to classifier can build the performance evaluation of our proposed demonstrate.

It can be built up that our proposed approach of sentiment analysis has a tendency to perform better when feature selection is performed. We proposed our model with two feature selection procedures (chi square and information gain).

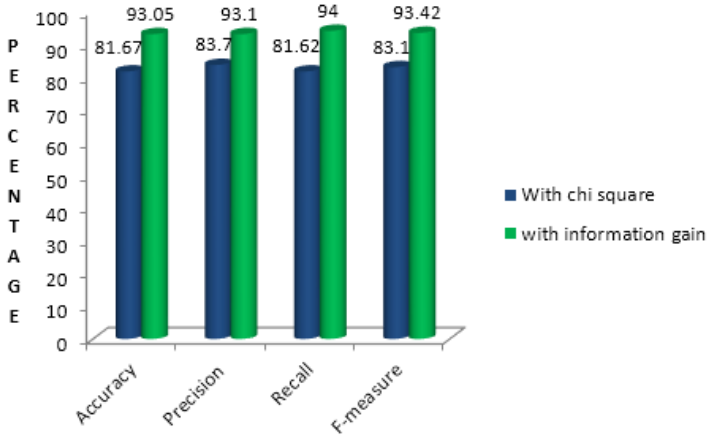

Fig 3.4. Performance comparison of chi square and information gain applied on our proposed methodology for model construction

On development of model and investigations are done with these diverse methods we induce that model built with information gain gives better outcomes in contrast with the model with chi square.

4. CONCLUSION:

In this work, We have done sentiment analysis on supervised classifier like Naïve Bayes utilizing highlight for the two grouping classifications (positive/negative, obstinate/real). In the first place, we separated information of movie reviews for sentiment prediction from Bo Pang and Lillian Lee Corpora. We have gathered 1000 positive reviews and 1000 negative reviews and utilized 75% for training set and 25% for testing set. At that point in pre-processing, we clean the information and select the highlights. Results demonstrate that after utilization of different component determination techniques the execution criteria increments in the long run. Utilization of Chi Square adds 9% development to the accuracy alone while use of Information Gain adds 20.25% development to precision. Not simply accuracy but rather Precision, Recall and F-measure additionally increment drastically so we can suggest that better feature selection choice strongly affect assurance of extremity of the movie reviews.

REFERENCES:

1. Xue, X. B., &Zhou, Z. H. 2009. Distributionalfeatures for text categorization.*IEEE Transactions on Knowledgeand Data Engineering*, 21: 428-442.
2. Porter, M. F. 1980. Analgorithm forsuffixstripping.*Program*, 14 : 130-137.
3. Jivani,A.G.2011.Acomparativestudyofstemmingalgorithms.*Int.J.Comp.Tech.Appl*, 2: 1930-1938
4. Porter, M. F. 2001. Snowball: A language for stemmingalgorithms.
5. Kira,K.,&Rendell,L.A.1992,July.Thefeatureselectionproblem:Traditionalmethods and anewalgorithm. In *AAAI*2 : pp. 129-134.
6. Narendra,P.M.andFukunaga,K.1977,September. Abranchandboundalgorithmfor featureselection.*IEEE Transactions on Computers*, 26 : 917–922.
7. Koller, D.,&Sahami, M. 1996.*Toward optimal feature selection*. StanfordInfoLab.
8. Siedlecki,W.andSklansky,J.1988.Onautomaticfeatureselection.*InternationalJournalof Pattern Recognition and Artificial Intelligence*, 2 : 197–220.
9. Lang,K.1995,July.Newsweeder:Learningtofilternetnews.In *Proceedingsofthe12th internationalconferenceon machine learning*pp.331-339.
10. Quinlan, J. R. 1986. Induction ofdecision trees.*Machine learning*, 1 :81-106. R.
11. Tom Mitchell. MachineLearning. McCrawHill, 1996.
12. Chakrabarti,S.,Dom,B., Agrawal,R.,&Raghavan,P.1997,August.Usingtaxonomy, discriminants,andsignaturesfornavigatingintextdatabases. In*VLDB* 97:pp.446-455.
13. Liu, B., & Zhang, L. 2012. A survey of opinion mining and sentiment analysis. In Mining text data pp. 415-463.
14. Carvalho,V.R.,& Cohen,W.W.2005,August.Onthecollectiveclassificationofemail speechacts.In *Proceedingsofthe28thannualinternationalACMSIGIRconferenceon Researchand development in information retrieval*pp. 345-352.