An Iterative Approach For Processing The Large Scale Datasets

Dr. S. CHITRA

Assistant Professor, Department of Computer Science, Srimad Andavan Arts and Science College (Autonomous) (Affiliated to Bharathidasan University, Tiruchirappalli), Tamilnadu, India.

ABSTRACT

With the introduction of Web 2.0, the amount of emotive content available on the Internet has increased. Movie or product reviews, user comments, testimonials, remarks in discussion forums, and other forms of such content are frequently seen on social networking websites. The benefits of timely discovery of emotive or opinionated web content are several, the most important of which is monetization. Understanding the feelings of the general public toward various entities and products allows for better contextual advertising, recommendation systems, and market trend analysis. The goal of this study is to develop a sentiment-focused web crawling framework that will make it easier to find and analyse emotive material in movie and hotel reviews. Statistical methods are utilised in this study to capture aspects of subjective style and sentence polarity. The research compares and contrasts the overall accuracy, precisions, and recall values of two supervised machine learning algorithms: K-Nearest Neighbour(K-NN) and Nave Bayes'. It was discovered that while Nave Bayes performed significantly better than K-NN in terms of movie reviews, these algorithms performed similarly poorly in terms of hotel evaluations..

KEYWORDS: Iterative Classification, K-Nearest Neighbour, Naïve Bayes, Web Content Mining, Sentiment analysis

1. INTRODUCTION

Data mining is the process of extracting useful information from a big quantity of data. For sentiment analysis, a variety of data mining analysis algorithms (such as clustering, classification, regression, and so on) can be utilised [8]. [9]. Sentiment mining is an important part of data mining that allows crucial data to be mined depending on the positive or negative meanings of the data. Sentiment Analysis, often known as Opinion Mining, is the technique of identifying and extracting subjective information from source materials using natural language processing, text analysis, and computational linguistics. The source materials in this case refer to several social networking sites' opinions, reviews, and comments [1]. Sentiment in comments, feedback, and critiques serves as a valuable indicator for a variety of purposes and can be classified according to polarity [2]. We use polarity to determine if a review is overall good or negative. Consider the following scenario: 1) Positive Sentiment in a Subjective Sentence: "I liked the movie Mary Kom"—From the sentiment threshold value of the word "loved," we can deduce that this sentence expresses positive sentiment regarding the movie Mary Kom. As a result, the numerical threshold value for the word "loved" is positive. 2) Negative sentiment in subjective sentences: The specified sentence "Phata poster nikla hero is a flop movie" expresses negative sentiment regarding the film "Phata poster nikla hero," as determined by the sentiment threshold value of the word "flop." As a result, the numerical threshold value for the word "flip" is negative. There are three types of sentiment analysis: document level, sentence level, and entity level. However, we're looking into sentiment analysis at the phrase level. Traditional text mining focuses on fact analysis, whereas opinion mining focuses on attitudes [3] [10] [11] [12] [13] [14] [15].

Sentiment classification, feature-based sentiment classification, and opinion summarising are the key study areas. The application of sentiment analysis in a commercial setting is becoming more common. The growing number of brand tracking and marketing organisations that provide this service demonstrates this. - Tracking user and non-user views and ratings on products and services are just a few of the options available. - Keeping an eye on the company's problems in order to prevent viral spread. - Analyzing market sentiment, competitive activity, and customer trends, fads, and fashion. - Measuring public reaction to a company-related activity or issue [4]. In this research, we calculate the accuracies, precisions (of positive and negative corpuses), and recall values using two Supervised Machine Learning algorithms: Nave Bayes' and K-Nearest Neighbor (of positive and negative corpuses). The difficulties in Sentiment Analysis include that an opinion word that is considered positive in one context may be considered bad in another. The degree of optimism or negativity has a big influence on people's opinions. For example, the terms "excellent" and "very good" cannot be used interchangeably. [2] Although classical text processing states that a minor alteration in two bits of text has no effect on the meaning of the sentences, this is not the case. However, the most recent text mining technology allows for advanced analysis that measures the word's intensity. This is the point at which the accuracy and efficiency of certain algorithms can be scaled [4] [5] [6] [7] [8] [9] [16] [17] [18].

2. PROPOSED ITERATIVE APPROACH FOR LARGE SCALE DATA SETS

The research's major purpose is to examine the data from the surveys and determine whether it is suitable for analysis using the data mining methods outlined. Figure 1 shows a graphical representation of the processes involved in sentiment analysis.

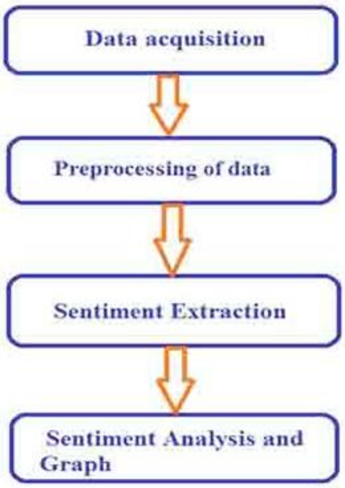

Figure 1: Flowchart of Iterative Classification Scheme for Large Scale Data sets

3.1 Chi-Square Test

- Initialize an empty frequency distribution.
- Initialize an empty conditional frequency distribution (based on words being positive and negative).
- This work fills out the frequency distributions, incrementing the counter of each word within the appropriate distribution.
- It finds the highest-information features is the count of words in positive reviews, words in negative reviews, and total words.
- This work use a chi-squared test (also from NLTK) to score the words. It find each word’s positive information score and negative information score, add them up, and fill up a dictionary correlating the words and scores, which we then return out of the function.

3.1 Naïve Bayes Classification

A prominent supervised classification paradigm is Bayesian network classifiers. The Nave Bayes' classifier is a probabilistic classifier based on the Bayes' theorem that considers the Nave (Strong) independence condition. It is a well-known Bayesian network classifier. It was brought into the text retrieval community under a different name and is still a popular(baseline) method for text categorization, the problem of evaluating documents as belonging to one of two categories using word frequencies as the criterion. The fact that Nave Bayes only takes a little quantity of training data to estimate the parameters required for classification is an advantage. In its most basic form, the Nave Bayes model is a conditional probability model. The nave Bayes' classifier has been shown to operate satisfactorily in a variety of areas, despite its simplicity and heavy assumptions. Prior knowledge and observed data can be merged in Bayesian classification, which enables practical learning techniques. The core idea behind the Nave Bayes technique is to use the joint probabilities of words and categories to find the probabilities of categories given a text document. It is predicated on the idea of word independence.

The starting point is the Bayes' theorem for conditional probability, stating that, for a given data point x and class C:

$P\ (C\ /\ x) = P(x/C)/P(x)$ ------------ (1)

Furthermore, by making the assumption that for a data point

x = {x1,x2,...xj}, the probability of each of its attributes occurring in a given class is independent, we can estimate the probability of x as follows:

$P(C/x)=P(C).\prod P(xi/C)$ ----------- (2)

Algorithm

Input: a document d

A fixed set of classes C={c1,c2,…,cj}

Output: a predicted class c∈C

Steps: 1. Pre-processing:

i. About 10,000 reviews were crawled from www.imdb.com / Opin Rank Review Dataset ii. Positive reviews and negative reviews were kept in two files pos.txt and neg.txt

iii. 2 empty lists were taken, one for positive and one for negative reviews.

iv. Sentences of the positive and negative reviews were broken and 'pos' and 'neg' were appended to each accordingly and were stored in the 2 empty lists created.

v. ¾ of these sentences were kept in the dictionary for training while the ¼ were kept for testing.

Step 2. Using chi squared test:

we calculated the score of each of the remaining words and instead of using all of those words we only used the best 10,000.

Step 3. The classifier was trained using the dataset just prepared.

Step 4. Labelled sentences were kept correctly in reference sets and the predicatively labeled version in test sets.

Step 5. Metrics were calculated accordingly.

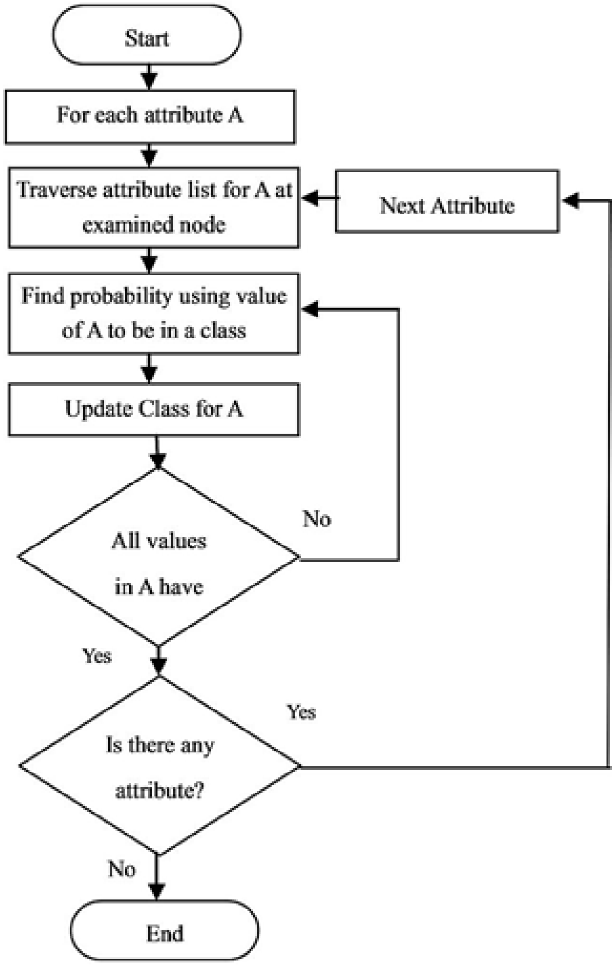

Figure 2: Flowchart of Naïve Bayes Classification Method

3.3 K-Nearest Neighbor Classification

K-NN is a sort of instance-based learning, often known as lazy learning, in which the function is only approximated locally and all computation is postponed until after classification. It is a non-parametric classification or regression method. If the result is class membership (the most common cluster may be returned), the item is classified by a majority vote of its neighbours, with the object being allocated to the most common class among its k nearest neighbours. During learning, this rule simply keeps the complete training set and gives a class to each query based on the majority label of its k-nearest neighbours in the training set.

When K = 1, the Nearest Neighbour rule (NN) is the simplest form of K-NN. All the distances between an unknown sample and all the samples in a training set can be computed given an unknown sample and a training set. The sample in the training set closest to the unknown sample correlates to the distance with the least value. As a result, the classification of the unknown sample can be based on the classification of its nearest neighbour. The K-NN is a simple method to comprehend and apply, as well as a strong tool for sentiment analysis. KNN is powerful because it makes no assumptions about the data other than that a distance measure between two instances can be determined reliably. As a result, it's referred to as non-parametric or non-linear because it doesn't take on a functional shape. Figure 3 shows the k-nn classifier's flowchart.

1. Pre-processing:

i). About 10,000 reviews were crawled from www.imdb.com/OpinRank Review Dataset

ii. Positive reviews and negative reviews were kept in two files pos.txt and neg.txt

iii. 2 empty lists were taken, one for positive and one for negative reviews.

iv. Sentences of the positive and negative reviews were broken and ‘pos’ and ‘neg’ were appended to each accordingly and were stored in the 2 empty lists created

v. ¾ of these sentences were kept in the dictionary for training while the ¼ were kept for testing.

2. Training:

i. Using chi squared test we calculated the score of each of the words occurring in the training dataset.

ii. An empty list is created, the dictionary in which the words from training dataset are stored followed by each of their scores thus calculated.

iii. for each word

iv. If it exists in the word score list, add its score to review score

v. Else find the word in word score list with minimum jaccard index to the unknown word and add its score to the review score.

vi. End for at step 3

vii. End for at step 4

viii. Find metrics accordingly.

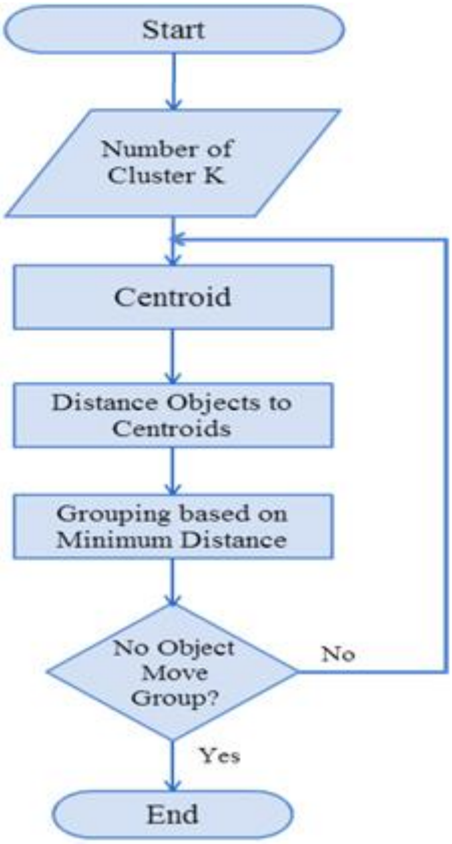
Start
Number of Cluster K
Centroid
Distance Objects to Centroids
Grouping based on Minimum Distance
No Object Move Group?
No
Yes
End

Figure 3: Flowchart of K-Nearest Neighbor

4. RESULT AND DISCUSSION

4.1 Data Source and Dataset

To conduct the research, two datasets are considered here - Movie Reviews & Hotel Reviews.

- All the movie reviews have been scanned from www.imdb.com.
- All the hotel reviews have been downloaded from Opin Rank Review Dataset (http://archive.ics.uci.edu/ml/datasets/OpinRank +Review+Data set)

The data set has been prepared by taking 5000 positive and 5000 negative reviews from each of the mentioned sites.

4.2 Performance Metrics

Accuracy, Precision and recall are method used for evaluating the performance of opinion mining. Here accuracy is the overall accuracy of certain sentiment models. Recall (Pos) and Precision (Pos) are the ratio and precision ratio for true positive reviews. Recall (Neg) and Precision (Neg) are the ratio and precision ratio for true negative reviews. In an ideal scenario, all the experimental results are measured according to the Table 1.and accuracy, Precision and recall as explained below [9].

$$\mathrm{Accuracy} = \frac{\mathrm{a+d}}{\mathrm{a+b+c+d}}$$

$$\mathrm{Recall\ (Pos)} = \frac{\mathrm{a}}{\mathrm{a+c}}$$

$$\mathrm{Recall\ (Neg)} = \frac{\mathrm{d}}{\mathrm{b+d}}$$

$$\mathrm{Precision(Pos)} = \frac{\mathrm{a}}{\mathrm{a+b}}$$

$$\mathrm{Precision\ (Neg)} = \frac{\mathrm{d}}{\mathrm{c+d}}$$

Table 1: A Confusion Matrix table

	True Positive Reviews	True Negative Reviews
Predict Positive reviews	a	b
Predict negative reviews	c	d

4.3 Analysis

The overall accuracies of the two algorithms in 10 rounds of experiments are indicated in Table 2.

Table 2: Accuracy comparison on Test Data sets

Number of Experiments	Number of reviews in the training set	Accuracy			
		Movie reviews dataset		Hotel Review dataset	
		Naïve Bayes	K-Nearest Neighbor	Naïve Bayes	K-Nearest Neighbor
1	100	56.78	47.64	43.11	45.35
2	200	64.29	55.07	41.26	40.97
3	500	70.06	58.44	42.56	41.42
4	1000	73.81	61.48	44.64	41.18
5	1500	77.23	64.21	48.21	42.01
6	2000	79.14	66.02	51.28	46.57
7	2500	79.82	67.89	52.03	47.04
8	3000	80.27	68.58	52.64	47.03
9	4000	82.11	69.03	53.92	49.75
10	4500	82.43	69.81	55.09	52.14

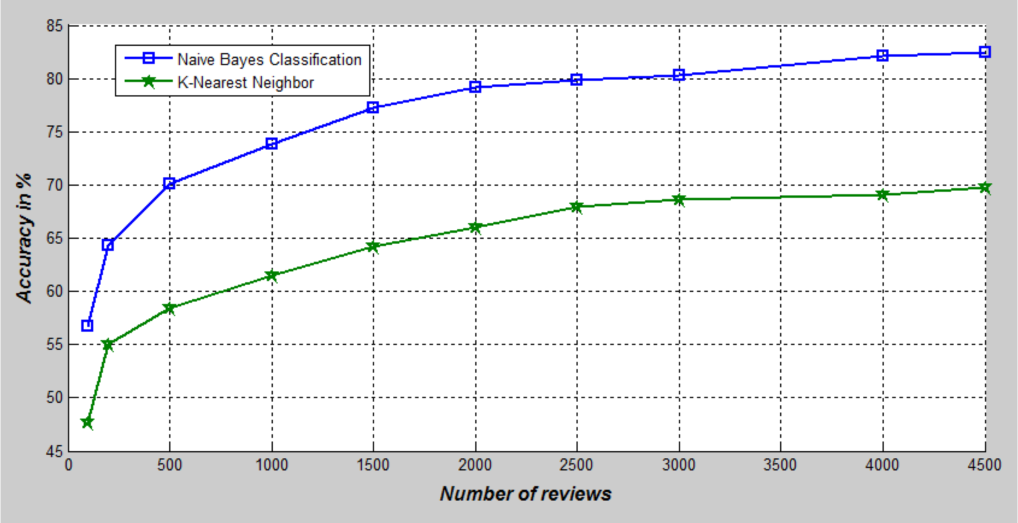

Figure 4a: Performance analysis of Naïve Bayes and KNN for Movie Review data set

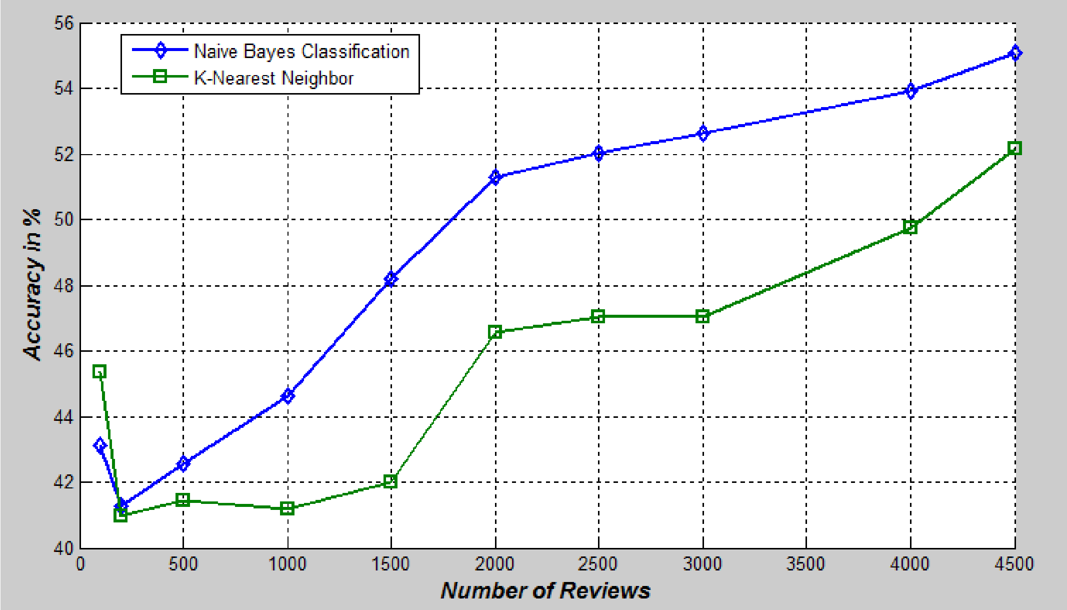

Figure 4b: Performance analysis of Naïve Bayes and KNN for Hotel Review data set

Table 3: Result of accuracies with maximum number of reviews

Total number of reviews	Classifier used	Review dataset used	Correct sample	Incorrect sample
1500	Naive Bayes	Movie	1237	263
		Hotel	827	673
	K-Nearest Neighbor	Movies	1047	453
		Hotel	782	718

Table 4: Precision comparison for Positive Corpus on Test Data sets

Number of Experiments	Number of reviews in the training set	Precision for Positive Corpus			
		Movie reviews dataset		Hotel Review dataset	
		Naïve Bayes	K-Nearest Neighbor	Naïve Bayes	K-Nearest Neighbor
1	100	59.04	41.35	42.11	44.51
2	200	64.96	50.97	40.26	40.86
3	500	69.56	54.42	41.56	5041
4	1000	73.64	58.18	43.64	42.21
5	1500	77.21	62.01	47.21	42.12
6	2000	80.28	65.57	50.28	45.36
7	2500	81.03	66.04	51.03	46.14
8	3000	81.64	67.03	51.64	47.13
9	4000	82.92	67.75	52.92	47.57
10	4500	84.09	68.14	54.09	48.21

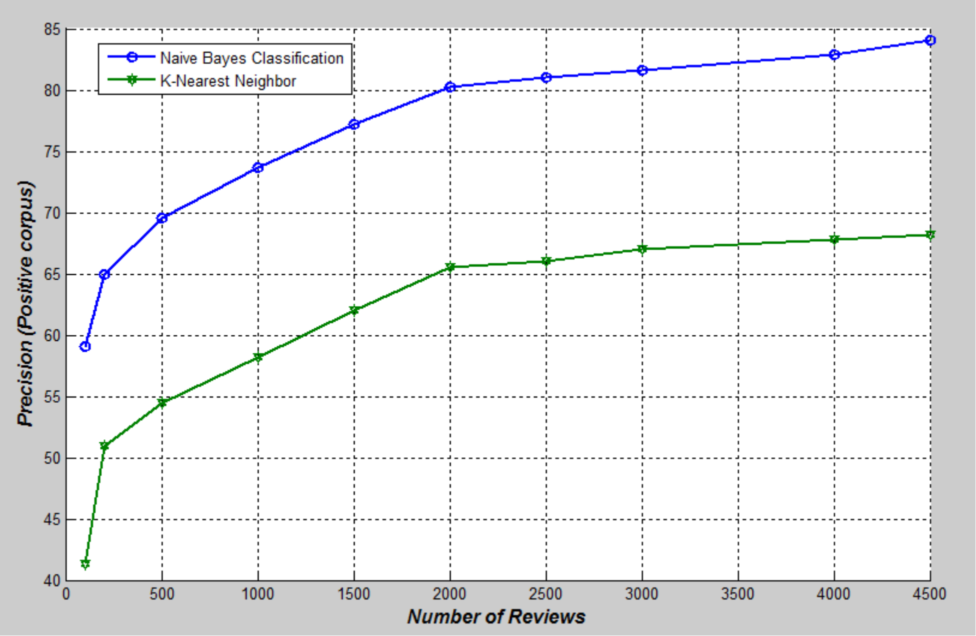

Figure 5a: Performance analysis on precision (positive corpus) of Naïve Bayes and KNN for Movie Review data set

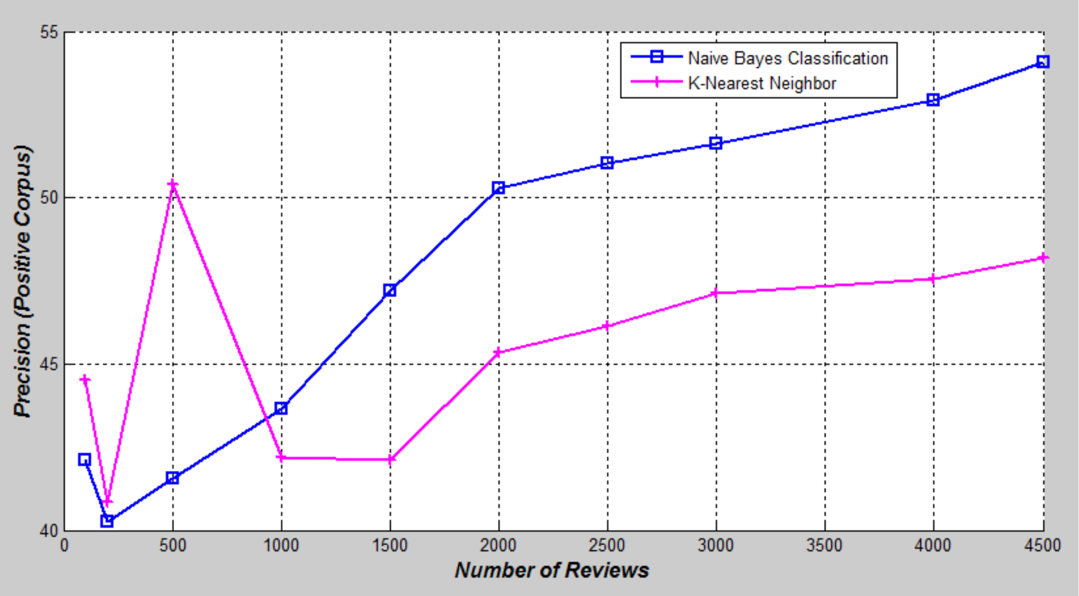

Figure 5a: Performance analysis on precision (positive corpus) of Naïve Bayes and KNN for Hotel Review data set

Table 5: Precision comparison for Negative Corpus on Test Data sets

Number of Experiments	Number of reviews in the training set	Precision for Negative Corpus			
		Movie reviews dataset		Hotel Review dataset	
		Naïve Bayes	K-Nearest Neighbor	Naïve Bayes	K-Nearest Neighbor
1	100	55.43	38.12	48.39	46.21
2	200	63.67	49.56	42.61	41.63
3	500	70.59	57.25	50.62	47.32
4	1000	73.99	62.12	53.81	52.15
5	1500	77.25	64.48	57.31	54.43
6	2000	78.09	65.73	58.11	55.69
7	2500	78.70	66.23	58.4	56.32
8	3000	79.00	66.47	59.91	56.51
9	4000	81.33	66.62	61.29	56.66
10	4500	81.01	66.73	61.11	52.14

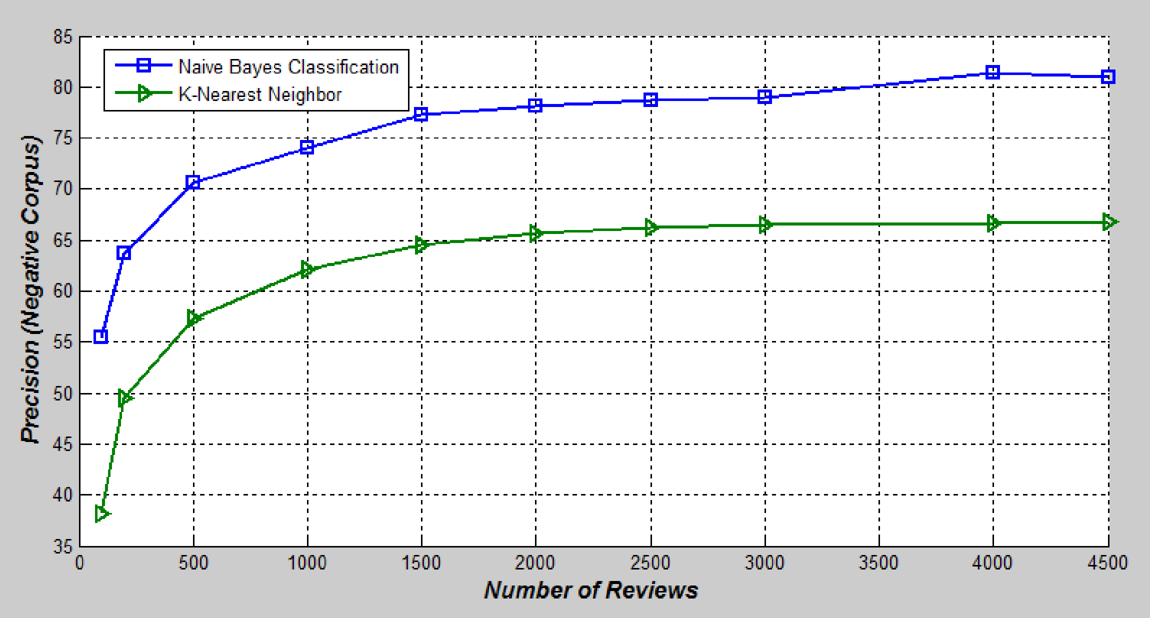

Figure 6a: Performance analysis on precision (negative corpus) of Naïve Bayes and KNN for Movie Review data set

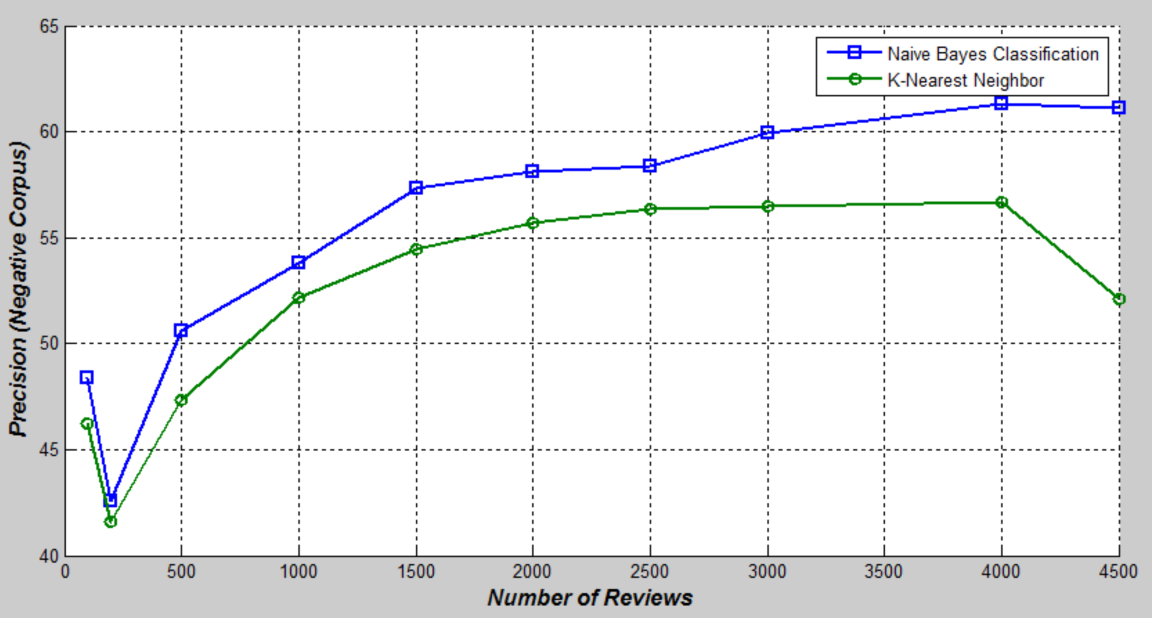

Figure 6b: Performance analysis on Precision (Negative corpus) of Naïve Bayes and KNN for Hotel Review data set

Table 6: Recall comparison for Positive Corpus on Test Data sets

Number of Experiments	Number of reviews in the training set	Recall for Positive Corpus			
		Movie reviews dataset		Hotel Review dataset	
		Naïve Bayes	K-Nearest Neighbor	Naïve Bayes	K-Nearest Neighbor
1	100	44.33	31.12	32.24	30.35
2	200	62.04	45.37	43.54	42.41

3	500	71.34	52.24	41.79	41.86
4	1000	74.19	56.31	47.44	42.21
5	1500	77.26	58.24	49.19	44.72
6	2000	77.26	60.02	50.02	45.03
7	2500	77.89	61.12	51.77	46.01
8	3000	78.09	61.53	51.44	46.52
9	4000	80.87	61.72	51.34	46.25
10	4500	80.12	61.81	51.84	46.31

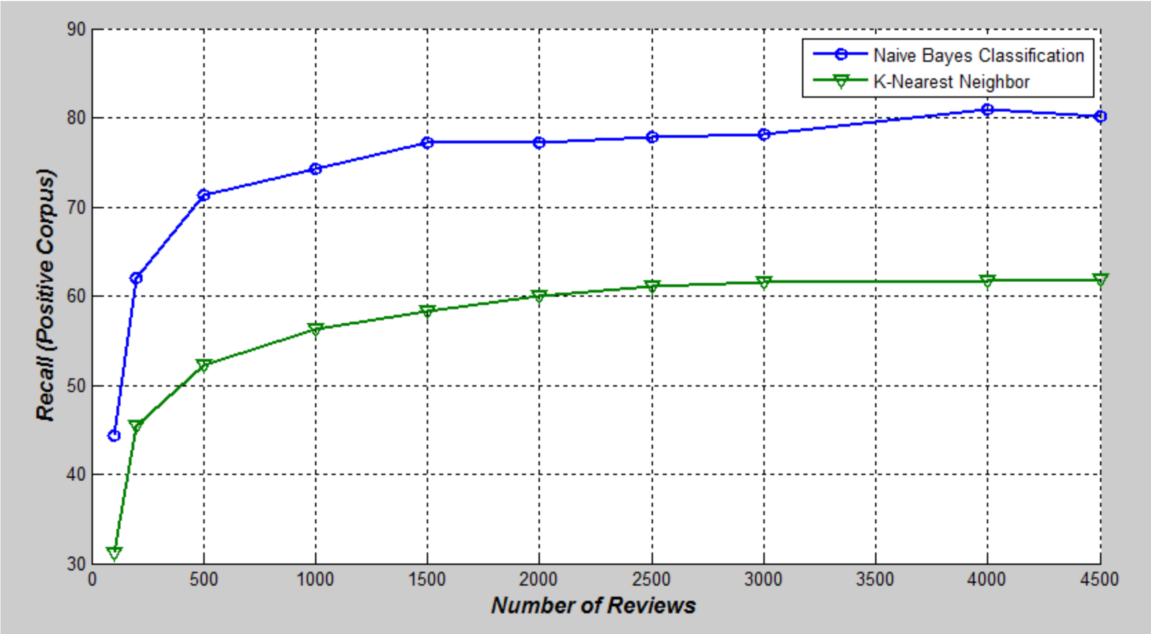

Figure 7a: Performance analysis on Recall (positive corpus) of Naïve Bayes and KNN for Movie Review data set

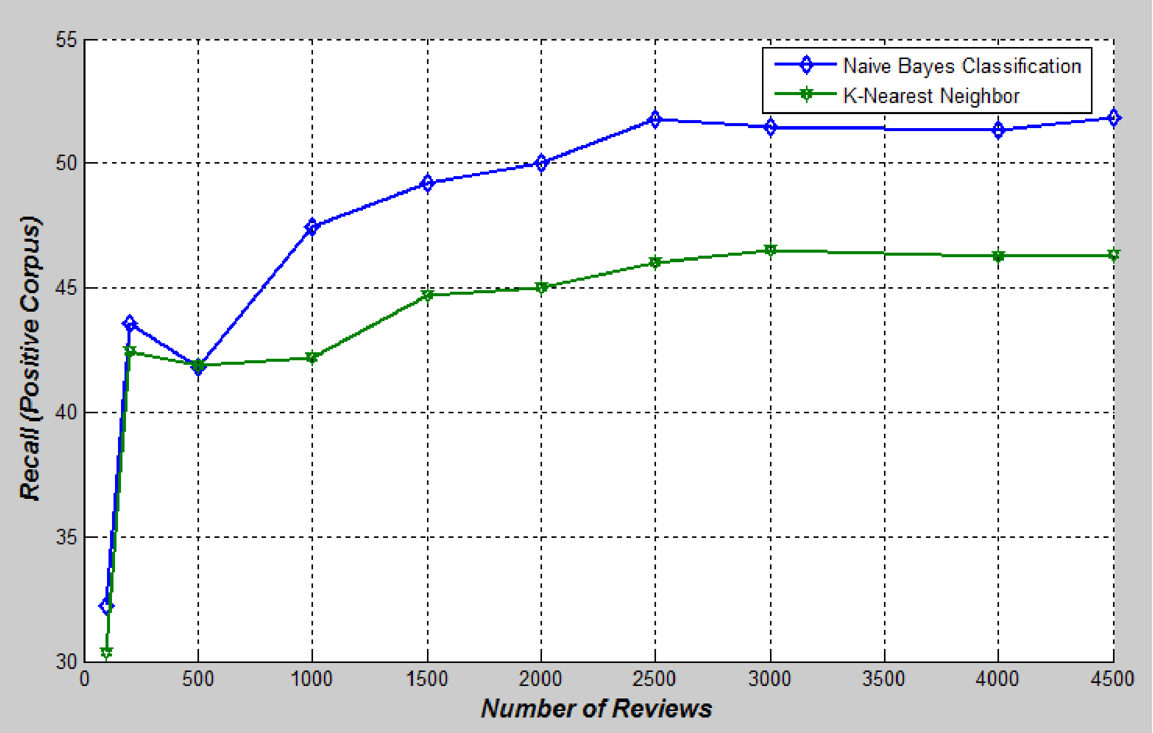

Figure 7b: Performance analysis on Recall (positive corpus) of Naïve Bayes and KNN for Hotel Review data set

Table 7: Recall comparison for Negative Corpus on Test Data sets

Number of Experiments	Number of reviews in the training set	Recall for Negative Corpus			
		Movie reviews dataset		Hotel Review dataset	
		Naïve Bayes	K-Nearest Neighbor	Naïve Bayes	K-Nearest Neighbor
1	100	69.24	39.25	62.33	60.35
2	200	66.54	55.12	53.51	52.41
3	500	68.79	53.86	51.81	51.89
4	1000	73.44	60.21	57.52	52.19
5	1500	77.19	63.72	59.24	54.77
6	2000	81.02	65.03	60.11	55.13
7	2500	81.77	66.01	61.83	56.11
8	3000	82.44	66.52	61.49	56.32
9	4000	83.34	66.25	61.37	56.35
10	4500	84.84	66.31	61.88	56.41

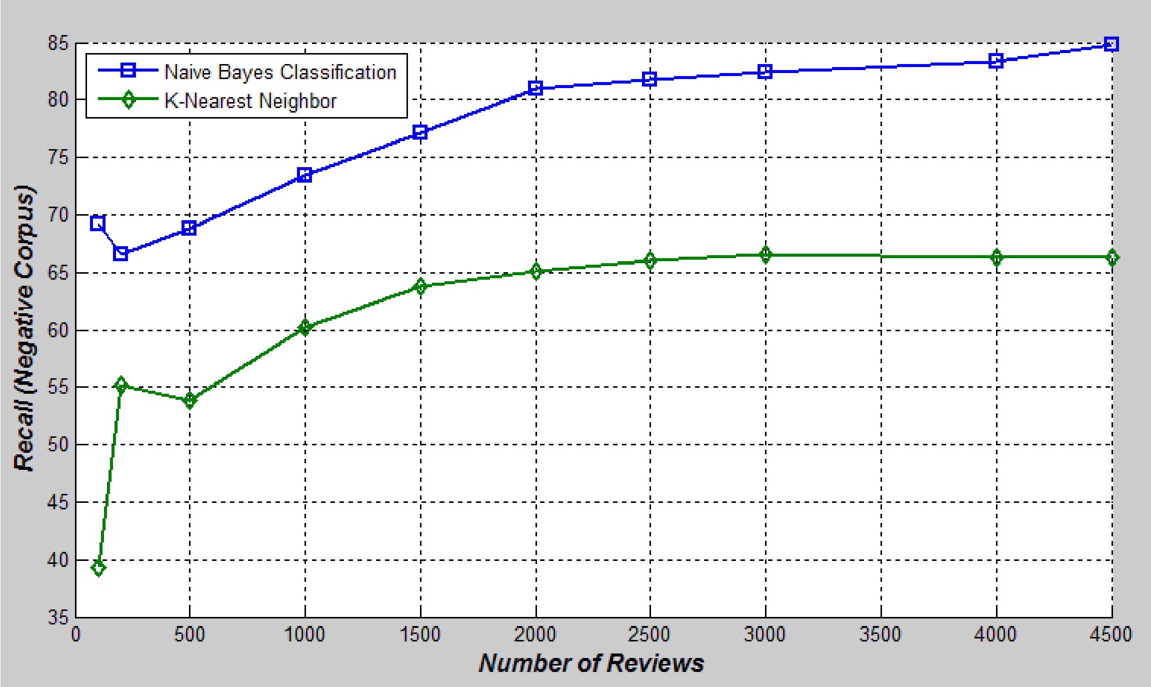

Figure 8a: Performance analysis on Recall (negative corpus) of Naïve Bayes and KNN for Movie Review data set

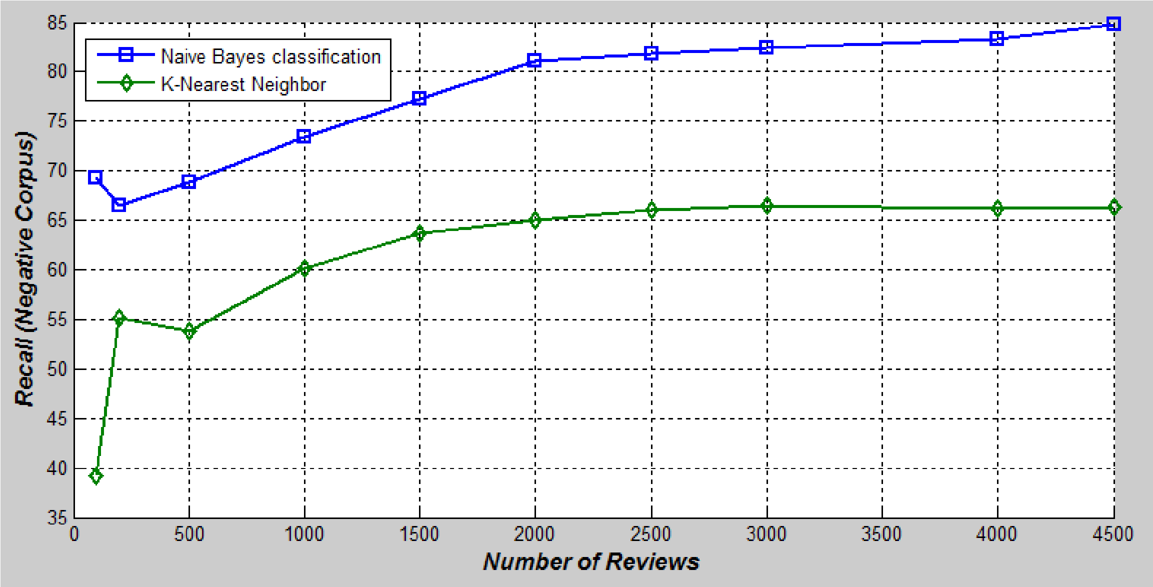

Figure 8b: Performance analysis on Recall (negative corpus) of Naïve Bayes and KNN for Hotel Review data set

From the above tables and figures, it is clear that Naïve Bayes classification method performs well than the existing technique like K- Nearest Neighbour algorithm

5. CONCLUSION

The goal of this study is to assess sentiment categorization performance in terms of accuracy, precision, and recall. The comparison of two supervised machine learning algorithms, Nave Bayes' and KNN, for sentiment categorization of movie and hotel reviews is presented in this study. The experimental results reveal that the classifiers performed better for movie reviews, with the Nave Bayes strategy surpassing the k-NN approach with accuracies of over 80%. The accuracies of hotel reviews, on the other hand, are substantially lower, and both classifiers produced similar results. As a result, it is possible to conclude that Nave Bayes' classifier can be successfully applied to the analysis of movie reviews..

REFERENCES

[1] Hand, D, Mannila, H & Smyth, P 2008, Principles of Data Mining, PHI.

[2] Digby, PG & Kempton, RA 1987, Multivariate Analysis of Ecological Communities, Chapman and Hill, London.

[3] Becker, RA, Ecik, SG & Wilks, AR 1995, 'Visualizing network data', IEEE transactions on Visualization and Computer Graphics, vol. 1, no.1, pp. 16-28.

[4] Fayyad, UM, Djorgovski, SG & Weir, N 1996, 'Automating the analysis and cataloging of sky surveys', Advances in Knowledge Discovery and Data Mining, pp 471-493.

[5] Cortes, C & Pregibon, D 1996, 'Giga Mining', proceedings of the fourth international conference on Knowledge Discovery and Data Mining, pp. 174-178.

[6] Bhandari, I, Colet, E, Parker, J, Pines, Z, Pratap, R & Ramanujam, K 1997, 'Advanced Scout: data mining and knowledge discovery in NBA data', Data mining and knowledge discovery, vol. 1, no.1, pp. 121-125.

[7] Fawcett, T & Provost, F 1997, 'Adaptive fraud detection', Data mining and knowledge discovery, vol. 1, no.3, pp. 291-316.

[8] Powell, M J D 1981, Approximation Theory and Methods, Cambridge University Press

[9] Hill, W, Stead, L, Rosenstein, M & Furnas, G 1995,'Recommending and evaluating choices in a virtual community of use', proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp.194-198.

[10] Subhashini, M., & Gopinath, R., Mapreduce Methodology for Elliptical Curve Discrete Logarithmic Problems – Securing Telecom Networks, International Journal of Electrical Engineering and Technology, 11(9), 261-273 (2020).

[11] Upendran, V., & Gopinath, R., Feature Selection based on Multicriteria Decision Making for Intrusion Detection System, International Journal of Electrical Engineering and Technology, 11(5), 217-226 (2020).

[12] Upendran, V., & Gopinath, R., Optimization based Classification Technique for Intrusion Detection System, International Journal of Advanced Research in Engineering and Technology, 11(9), 1255-1262 (2020).

[13] Subhashini, M., & Gopinath, R., Employee Attrition Prediction in Industry using Machine Learning Techniques, International Journal of Advanced Research in Engineering and Technology, 11(12), 3329-3341 (2020).

[14] Rethinavalli, S., & Gopinath, R., Classification Approach based Sybil Node Detection in Mobile Ad Hoc Networks, International Journal of Advanced Research in Engineering and Technology, 11(12), 3348-3356 (2020).

[15] Rethinavalli, S., & Gopinath, R., Botnet Attack Detection in Internet of Things using Optimization Techniques, International Journal of Electrical Engineering and Technology, 11(10), 412-420 (2020).

[16] Poornappriya. T.S., & Gopinath, R., Application of Machine Learning Techniques for Improving Learning Disabilities, International Journal of Electrical Engineering and Technology (IJEET), 11(10), 403-411, 2020.

[17] Priyadharshini, D., Gopinath, R., & Poornappriya. T.S., A fuzzy MCDM approach for measuring the business impact of employee selection, International Journal of Management (IJM), 11(7), 1769-1775, 2020.

[18] Poornappriya. T.S., & Gopinath, R., Rice plant Disease Identification using Artificial Intelligence Approaches, International Journal of Electrical Engineering and Technology (IJEET), 11(10), 392-402, 2020.