ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Vol. 6, Issue 1, January 2018

Plagiarism Detection Framework using Monte – Carlo Based Artificial Neural Network for Nepali Language

Rakesh Kumar Bachchan* and Arun Kumar Timalsina

Department of Computer Engineering, BVM Engineering College, V.V. Nagar, Gujarat, India

E-mail: rakeshbachchan01@yahoo.com[1]

Abstract: This research work develops two frameworks for detecting plagiarism of Nepali language literatures incorporating Monte Carlo based Artificial Neural Network (MCANN) and Backpropagation (BP) neural network, which was applied for the plagiarism detection on certain document type segment. Neural Network training is considered using Monte Carlo based family of algorithms as of these algorithms superiority and robustness. Both the frameworks are tested on two different datasets and results were analyzed and discussed. Convergence of MCANN is faster in comparison to traditional BP algorithm. MCANN algorithm achieved a convergence in the range of 10−2 to 10−7 for the training error in 40 epochs while general BP algorithm is unable to achieve such a convergence even in 400 epochs. Also, the mean accuracy of BP and MCANN are respectively found to be in the range of 98.657 and 99.864 during paragraph based and line based comparison of the documents. Thus, MCANN is efficient for plagiarism detection in comparison to BP for Nepali language documents.

Keywords: Plagiarism; Monte Carlo method; Artificial neural network; Back propagation

I. INTRODUCTION

Using documents of others without any reference or violating the copyright rules making the document as our own, is said to be plagiarism. Plagiarism detection is the act of finding the originality of a document i.e. whether a document or idea is of the same person who is claiming about it. Because of avalanche of electronic documents over the internet, contents about any topic could be easily found which the main reason behind plagiarism is. Plagiarism not only means using other's document but using ideas, concepts, thought of others without their consent. In this research work, Artificial Neural Network which is the most promising model simulating the biological neural network is combined with one of the most famous class of randomized algorithm, Monte Carlo Method, and is then trained for stepping towards detecting plagiarism.

A lot of research work has been carried out for detecting plagiarism in English documents and some other language documents like Arabic, Chinese and others. No any research work for detecting plagiarism in Nepali language documents is ever found. This research work focuses on detecting plagiarism in Nepali language documents. Also, several works using Monte Carlo method and artificial neural network have been carried but none of the works related to plagiarism is found using artificial neural network based Monte Carlo method.

II. LITERATURE REVIEW

There are lots of plagiarism checker tools (e.g., Turnitin, Eve2, CopyCathGold, etc.) still plagiarism detection is a difficult task because of huge amount of information available online [1]. None of the available tools checks plagiarism in Nepali language based documents. In the study done by Lukashenko et al. [2], different ways of reducing plagiarism along with widely used detection tools is discussed.

Two types of plagiarism detection method have been investigated in literatures: Intrinsic and External Plagiarism detection. In Intrinsic plagiarism detection method, identification of the document is done by checking its writing pattern, i.e., whether a document is written by a single author or not, if not which part of it is plagiarised. It is not compared with another document. In External plagiarism detection method document is compared with other documents for checking the document similarity.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Vol. 6, Issue 1, January 2018

Dara Curran [3] combined genetic algorithm with neural network for intrinsic plagiarism detection. The plagiarism detection classifier is capable of evolving both the weight and structure of the neural network. Salunkhe and Gawali [4] have used Temporal Difference (TD) algorithm of reinforcement learning for detecting plagiarism among documents. It improves data retrieval speed from database and plagiarism detection accuracy. Salha Mohammed Alzahrani and Naomie Salim [5], proposed statement based approach for detecting plagiarism in Arabic scripts using Fuzzy set information retrieval method. Here fuzzy-set IR model is adapted and used with Arabic language for detecting plagiarized statements based on the degree of membership between words. Shanmuga sundaram Hariharan [1], carried out plagiarism detection using similarity analysis where similarity is estimated using several measures like cosine, dice, jaccard, hellinger and harmonic. In this paper solution for "copy paste" and "paraphrasing" type of plagiarism is identified. In the research work considered by Efstathios Stamatatos [6], Plagiarism detection is done without removing the stop words. This method is based on structural information rather than content information. Stopword n-grams are able to capture syntactic similarities between suspicious and original documents and they can be used to detect the plagiarized passage boundaries is shown. Freitas et al. [7] train neural network using sequential Monte Carlo methods where they have used sampling techniques and illustrate their performance on the problem of pricing option contracts, traded in financial markets. A new algorithm named Hybrid SIR (hybrid gradient descent/sampling importance resampling algorithm) was also proposed in the same work. Man Yan Miranda Chong [8] shows that combining natural language processing and deep learning techniques improves the classification of plagiarised texts by reducing the number of false negatives. PAN plagiarism workshop is promoting research related to plagiarism detection since 2007 [8].

III. NEURAL NETWORK

Since learning is the result of communication between several neurons which is actually because of interconnection of a large number of neurons. Because of the highly inter-connected neurons, learning seems to be feasible in human. Neural Network although does not completely mimic the biological neural architecture but it resembles with the biological neural network to some extent. Also, it is an attempt to mirror the biological neural network; hence it is used for detecting plagiarism during the work. The back propagation algorithm which uses gradient descent method for minimizing the error was used for network training.

3.1 Backpropagation Neural Network

Backpropagation Neural Network was used for training purpose. Input to the neural network is cosine similarity and Jaccard similarity scores. The output from the network is either 0 or 1, where 0 represents the plagiarised case and 1 represents the non-plagiarised cases. The threshold value taken for indicating the document as plagiarised was ten percent. The equations for Backpropagation algorithm used in different phases during the training are discussed in [9].

IV. MONTE CARLO METHOD

Monte Carlo Method, a randomized algorithm, was used for updating the weights during the network training. For the purpose, some samples are drawn from the posterior distribution of cosine and jaccard similarity vectors. Generally, the method is used for generating samples from the state space in such a way that the samples resemble the target distribution. The posterior is calculated using NUTS sampler as discussed in [10]. The random samples are drawn from the posterior distribution of parameters during "learning phase".

V. DATASET

The corpus for Nepali document consists of different political, educational, biography, sports, stock exchange news from various daily, weekly and monthly Nepali newspapers [11-16]. The dataset statistic is given in Table 1.
Another corpus of Nepali language consists of 11 different theses collected from the Central Library Database. The statistic for the corpus is shown in Table 2. Both dataset consists of copy-paste and paraphrasing type of plagiarism.

Filename	No. of Paragraphs	No. of Words
Train1.txt	3	1225
Train2.txt	3	661
Train3.txt	7	2416
Train4.txt	4	775
Train5.txt	9	4177

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Vol. 6, Issue 1, January 2018

Train6.txt	6	2728
Train7.txt	10	1480
Test1.txt	6	301
Test2.txt	27	1426
Test3.txt	72	8294
Test4.txt	36	7404
Test5.txt	36	10858
Test6.txt	76	10519
Test7.txt	63	14114
Test8.txt	6	5571
Test9.txt	15	6194
Test10.txt	3	12092

Table 1: Statistics of dataset by Bam.

Filename	No. of Paragraphs	No. of Words
लमजुङ भोर्लेटार क्षेत्रमा प्रचलित लोकगीतको सङ्कलन वर्गीकरण र विश्लेषण	668	14113
नेपाली उपन्यासको सङ्कलनमा त्रिभुवन विश्वविद्यालय केन्द्रीय पुस्तकालयको योगदान	1453	27329
शान्तिकुमारी राईको जीवनी, व्यक्तित्व र कृतित्व	1059	37673
फणीन्द्रराज खेतालाको जीवनी, व्यक्तित्व र कृतित्व	856	21319
सेतो बाघ उपन्यासको पात्रविधान	507	13498
नेपाली नाट्यविधाका कृति र पत्रपत्रिकाको संरक्षणमा त्रिभुवन विश्वविद्यालय केन्द्रीय पुस्तकालयको योगदान	1839	19750
पश्चिमाञ्चल क्षेत्रमा प्रचलित नेपाली लोकगीतहरूको अध्ययन	3670	152085
माधवप्रसाद पोखरेलः जीवनी, व्यक्तित्व र कृतित्वको अध्ययन	957	35957
नुवाकोट जिल्लामा प्रचलित लोकगीतको अध्ययन	1190	28148
पुण्य निरौलाका उपन्यासमा पात्र विधान	1136	27286
सल्यानको पूर्वीक्षेत्रमा प्रचलित लोकगीतहरूको अध्ययन	1387	27610

Table 2: Statistics of Nepali language thesis dataset.

VI. DATA PREPROCESSING

Preprocessing includes paragraph segmentation (splits text into paragraphs), Punctuation removal (removes punctuation symbols), lowercasing (replaces uppercase letters with corresponding lowercase characters), number removal (removes number from the text), and stopword removal (removes stopwords from the text).

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Vol. 6, Issue 1, January 2018

The stop words of Nepali used are छ, र, पनि, छन्, लागि, भएको, गरेको, भने, गर्न, गर्ने, हो, तथा, यो, रहेको, उनले, थियो, हुने, गरेका, थिए, गर्दै, तर, नै, को, मा, हुन्, भन्ने, हुन, गरी, त, हुन्छ, अब, के, रहेका, गरेर, छैन, दिए, भए, यस, ले, गर्नु, औं, सो, त्यो, कि, जुन, यी, का, गरि, ती, न, छु, छौं, लाई, and नि.

The punctuation marks used in Nepali language are same as that used in English language except one additional "।" which is used for terminating the sentence.

Original Text	२०१६ सालमा त्रिभुवन विश्वविद्यालय केन्द्रीय पुस्तकालयको स्थापना काठमाडौको त्रिपुरेश्वरमा भएको हो । स्थापनाकालमा जम्मा १२ सय पुस्तकबाट सेवा दिन थालेको यस पुस्तकालयमा हाल विभिन्न भाषामा लेखिएका विभिन्न विषयमा केन्द्रित गरी जम्मा २ लाख ८० हजार पुस्तक र १ लाख थान विदेशी जर्नल र नेपाली अखवार तथा पत्रिकाहरु छन् ।
Paragraph Segmentation	Paragraph (1) २०१६ सालमा त्रिभुवन विश्वविद्यालय केन्द्रीय पुस्तकालयको स्थापना काठमाडौको त्रिपुरेश्वरमा भएको हो । Paragraph (2) स्थापनाकालमा जम्मा १२ सय पुस्तकबाट सेवा दिन थालेको यस पुस्तकालयमा हाल विभिन्न भाषामा लेखिएका विभिन्न विषयमा केन्द्रित गरी जम्मा २ लाख ८० हजार पुस्तक र १ लाख थान विदेशी जर्नल र नेपाली अखवार तथा पत्रिकाहरु छन् ।
Punctuation Removal	२०१६ सालमा त्रिभुवन विश्वविद्यालय केन्द्रीय पुस्तकालयको स्थापना काठमाडौको त्रिपुरेश्वरमा भएको हो स्थापनाकालमा जम्मा १२ सय पुस्तकबाट सेवा दिन थालेको यस पुस्तकालयमा हाल विभिन्न भाषामा लेखिएका विभिन्न विषयमा केन्द्रित गरी जम्मा २ लाख ८० हजार पुस्तक र १ लाख थान विदेशी जर्नल र नेपाली अखवार तथा पत्रिकाहरु छन्
Number Replacement	[#] सालमा त्रिभुवन विश्वविद्यालय केन्द्रीय पुस्तकालयको स्थापना काठमाडौको त्रिपुरेश्वरमा भएको हो । स्थापनाकालमा जम्मा [#] सय पुस्तकबाट सेवा दिन थालेको यस पुस्तकालयमा हाल विभिन्न भाषामा लेखिएका विभिन्न विषयमा केन्द्रित गरी जम्मा [#] लाख [#] हजार पुस्तक र [#] लाख थान विदेशी जर्नल र नेपाली अखवार तथा पत्रिकाहरु छन् ।
Stopword Removal	२०१६ साल त्रिभुवन विश्वविद्यालय केन्द्रीय पुस्तकालय स्थापना काठमाडौ त्रिपुरेश्वर। स्थापनाकाल जम्मा १२ सय पुस्तकबाट सेवा दिन थाले पुस्तकालय हाल विभिन्न भाषा लेखिए विभिन्न विषय केन्द्रित जम्मा २ लाख ८० हजार पुस्तक १ लाख थान विदेशी जर्नल नेपाली अखवार पत्रिका ।

Table 3: Example of several preprocessing tasks on an example Nepali text.

VII. VECTOR PROCESSING AND DIMENSIONALITY REDUCTION

The data from preprocessing stage was vectorized using Term Frequency - Inverse Document frequency (TF- IDF) [16]. After vectorizing the document, its dimensionality was reduced using Principal Component Analysis (PCA)

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Vol. 6, Issue 1, January 2018

discussed in [12] for reducing the processing complexity. Then, Cosine and Jaccard Similarity between each paragraph vector from the source and suspicious data were calculated as in [1].

Similarity Calculation

Cosine Similarity and Jaccard Similarity between each paragraph vector from the source data and suspicious data was then calculated. Cosine similarity is given by:

$$\cos\theta = \frac{\vec{a}.\vec{b}}{\|\vec{a}\|\|\vec{b}\|}$$

Where, a and b are vectors of suspicious and source paragraph respectively. Similarly, Jaccard Similarity is given by:

$$J(A,B) = \|Intersection(A,B)\|/\|Union(A,B)\|$$

Where, A and B are are vectors of suspicious and source paragraph respectively.

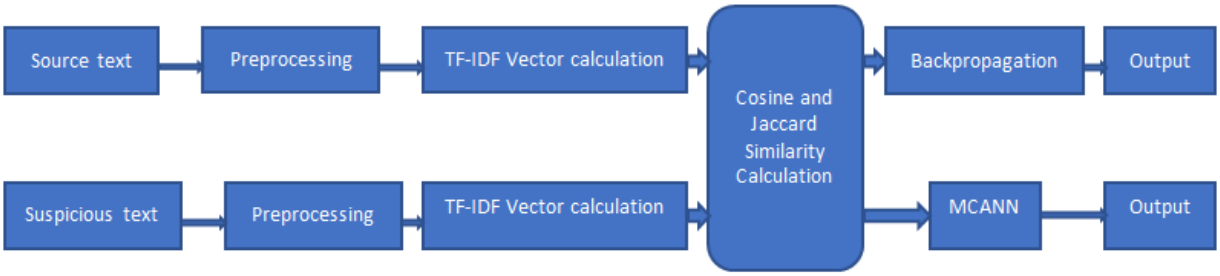

Figure 1: General processing framework.

VIII. RESULTS AND ANALYSIS

In this work, ANN model and MCANN model were developed for detecting the plagiarism of Nepali documents. Both the models were tested on several dissertations carried out in Nepali. Eleven dissertations of Nepali language were collected for the research. Similarly, testing was also carried out on Bam data [11].

8.1 Results of Paragraph Based Comparison

8.1.1 Experiment with Nepali thesis using back propagation

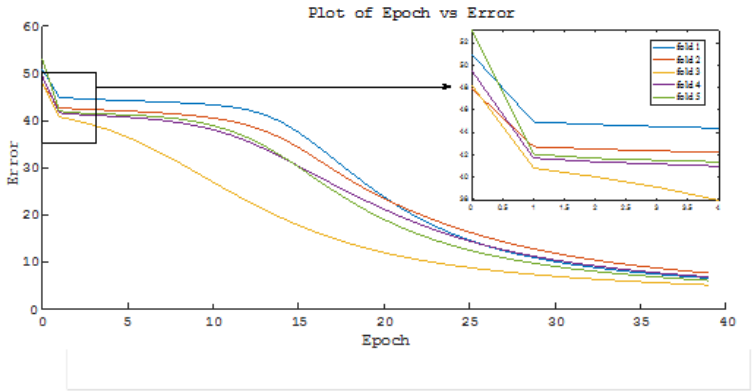

Figure 2: Error vs. Epoch for Nepali thesis for 40 Epochs. It is the case of 5-fold cross validation.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Vol. 6, Issue 1, January 2018

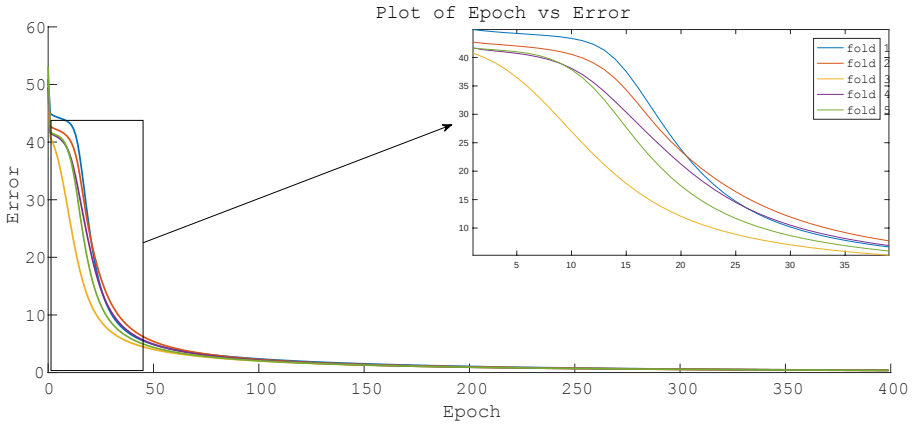

Figure 3: Error vs. Epoch for Nepali thesis using BP for 400 Epochs. It is the case of 5-fold cross validation.

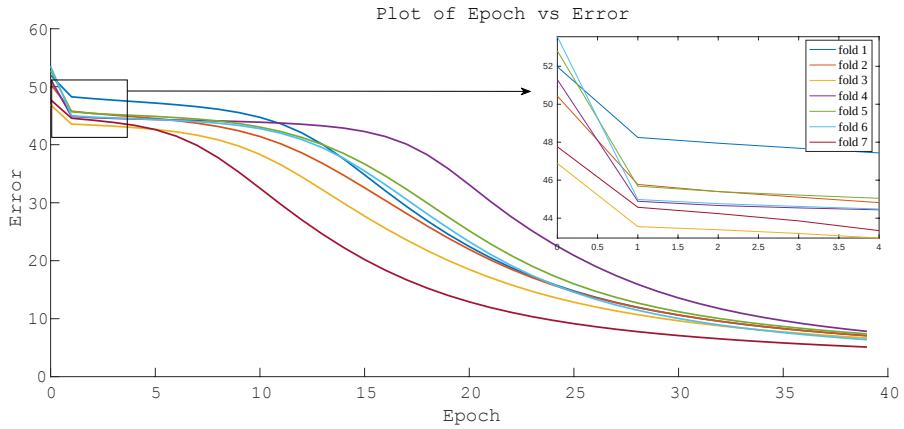

Figure 4: Error vs. Epoch for Nepali thesis using BP for 40 Epochs. It is the case of 7-fold cross validation.

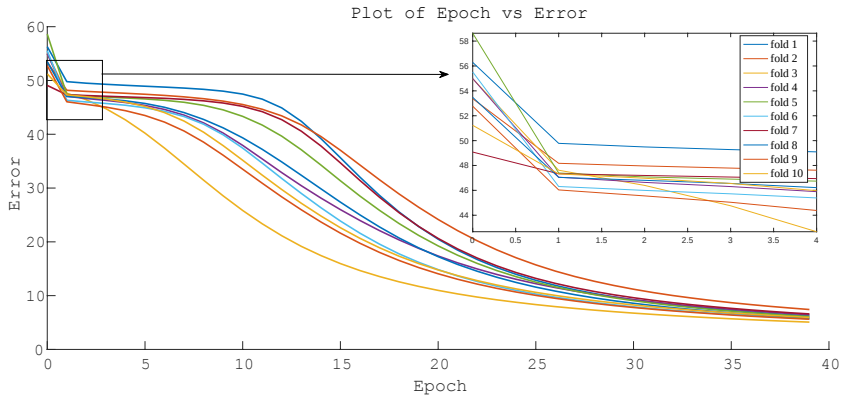

Figure 5: Error vs. Epoch for Nepali thesis using BP for 40 epochs. It is the case of 10-fold cross validation.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Vol. 6, Issue 1, January 2018

Figures 1-5 represent the plot of error against number of epochs when BP with two hidden layers were used for detecting the similarity of several thesis of Nepali languages. Results of training error obtained are summarized in Tables 3 and 4.

Dataset	Algorithm applied	Error obtained on different experiments		
		5-fold cross validation	7-fold cross validation	10-fold cross validation
Nepali Thesis	BP	6.613	5.111	5.952
Bam data	BP	335.854	350.929	360.370

Table 4: Result of backpropagation on Nepali thesis and Bam data during paragraph based comparison.

8.1.2 Experiment with Nepali thesis using MCANN

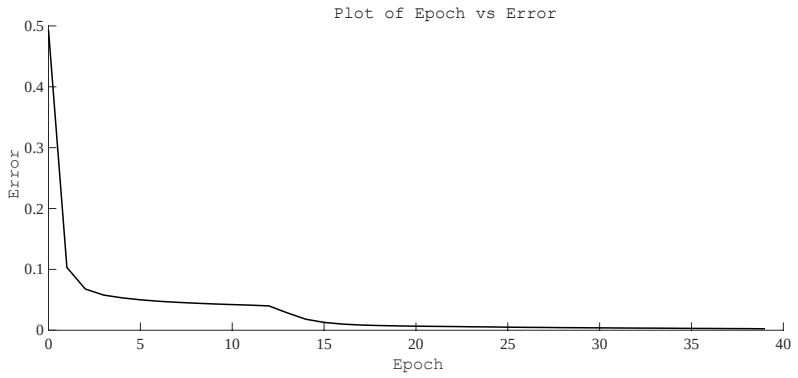

Figure 6: Error vs. Epoch for Nepali thesis using MCANN in 40 epochs. Ninety percent data was used as training data and ten percent as test data.

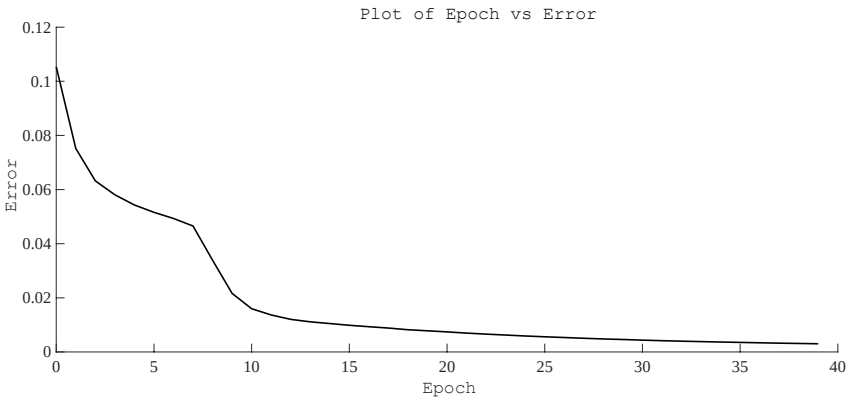

Figure 7: Error vs. Epoch for Nepali thesis using MCANN in 40 epochs. Eighty percent of data was used as train data and twenty percent data as test data.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Vol. 6, Issue 1, January 2018

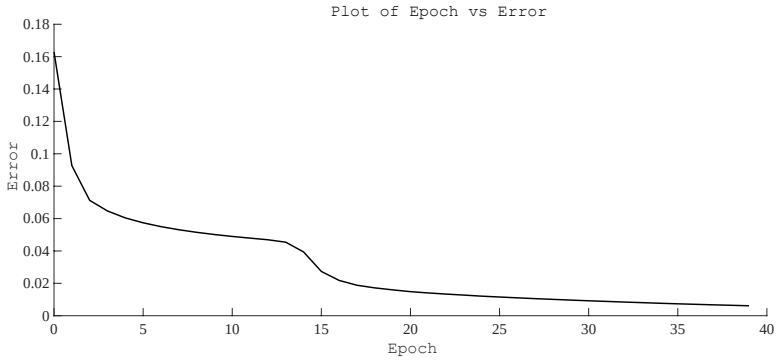

Figure 8: Error vs. Epoch for Nepali thesis using MCANN in 40 epochs. Sixty percent of data was used as train data and forty percent data as test data.

The error obtained on experimenting with Nepali thesis using MCANN using different training and testing data is shown in Figures 6-8 respectively. The results are summarized in Table 5.

Dataset	Algorithm applied	Error obtained on different experiments		
		60% train and 40% test data	80% train and 20% test data	90% train and 10% test data
Nepali Thesis	MCANN	6.1455e-03	3.0096e-03	2.4219e-03
Bam data	MCANN	3.0948e-04	2.1130e-04	8.1471e-05

Table 5: Result of MCANN on Nepali thesis and Bam data during paragraph based comparison.

8.1.3 Experiment with Bam data using BP

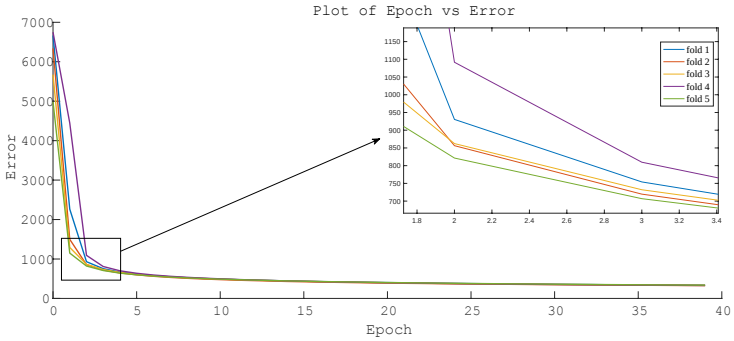

Figure 9: Error vs. Epoch for Bam \cite {bam2014named} data using BP (40 epochs). It is the case of 5-fold cross validation.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Vol. 6, Issue 1, January 2018

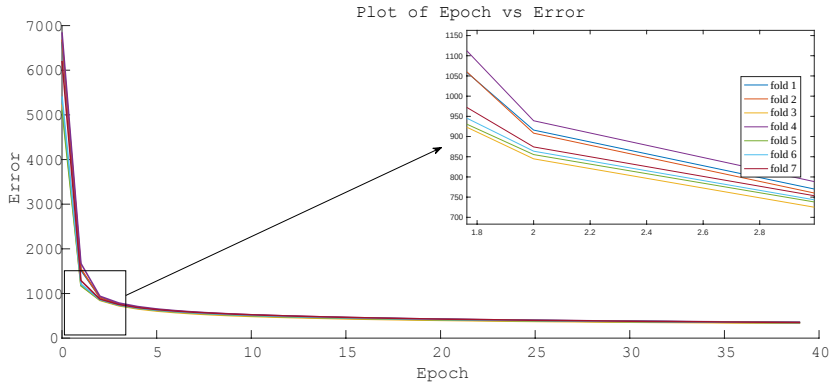

Figure 10: Error vs. Epoch for Bam \cite {bam2014named} data using BP (40 epochs). It is the case of 7-fold cross validation.

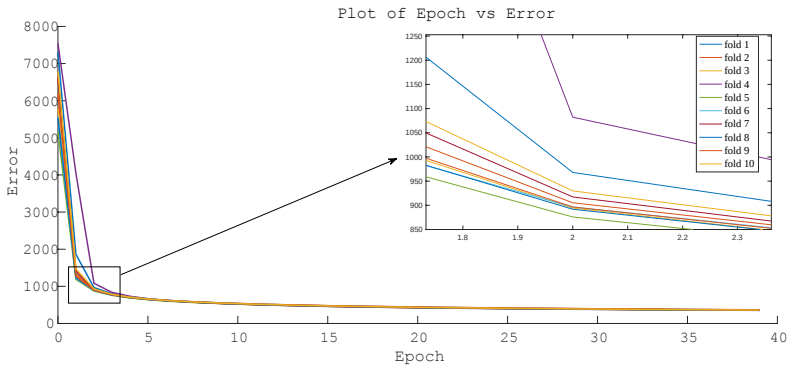

Figure 11: Error vs. Epoch for Bam \cite {bam2014named} data using BP (40 epochs). It is the case of 10-fold cross validation.

8.1.4 Experiment with Nepali data collected by Bam using MCANN

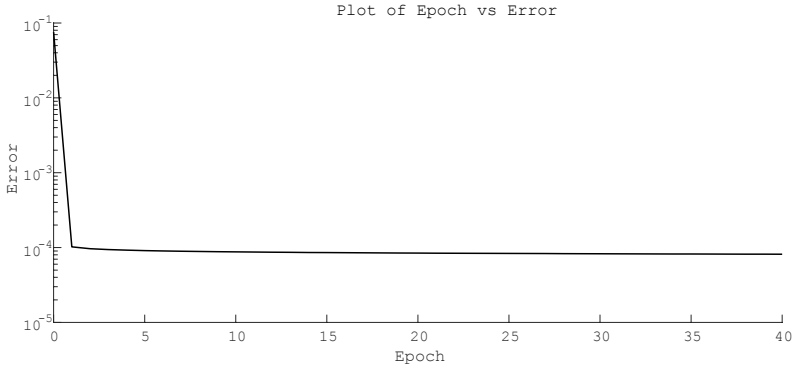

Figure 12: Error vs. Epoch for Bam \cite {bam2014named} using MCANN (40 epochs). Ninety percent data was used as training data and ten percent as test data.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Vol. 6, Issue 1, January 2018

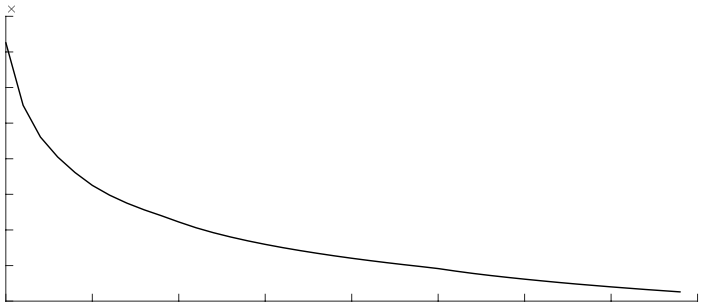

Figure 13: Error vs. Epoch for Bam \cite {bam2014named} using MCANN (40 epochs). Eighty percent data was used for training and twenty percent for testing.

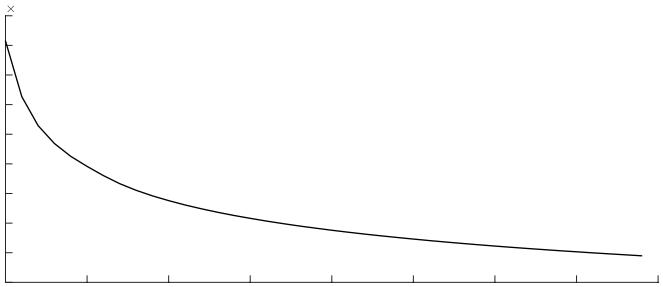

Figure 14: Error vs. Epoch for Bam \cite {bam2014named} using MCANN (40 epochs). Sixty percent data was used as training data and forty percent as test data.

Figures 9-11 represents the plot of error against number of epochs when BP with two hidden layers was used for detecting the similarity using Bam data [11]. Results of training error are summarized in Table 4.
The error obtained on experimenting with Bam data [11] using MCANN using different training and testing data is shown in Figures 12-14 respectively. The results are summarized in Table 5.

8.2 Results of Line Based Comparison

8.2.1 Experiment with Bam data using BP

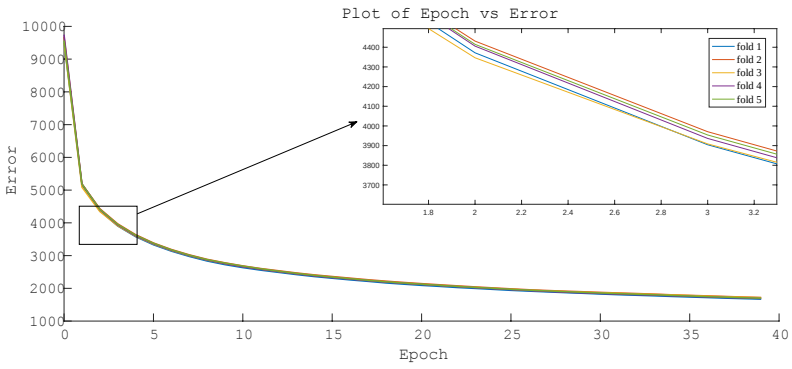

Figure 15: Error vs. Epoch for Bam \cite {bam2014named} data using BP (40 epochs). It is the case of 5-fold cross validation.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Vol. 6, Issue 1, January 2018

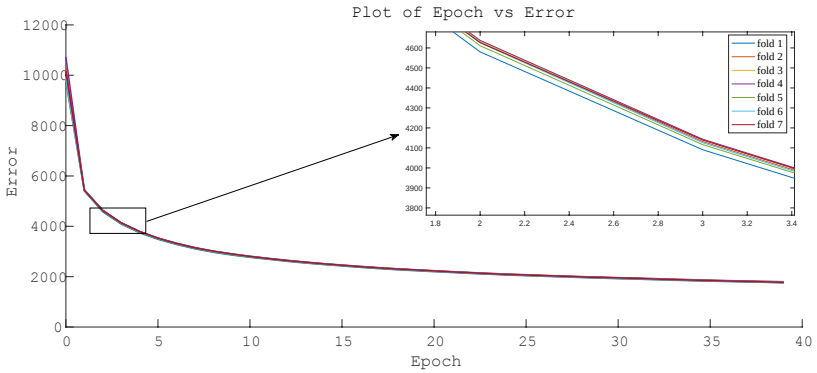

Figure 16: Error vs. Epoch for Bam \cite {bam2014named} data using BP (40 epochs). It is the case of 7-fold cross validation.

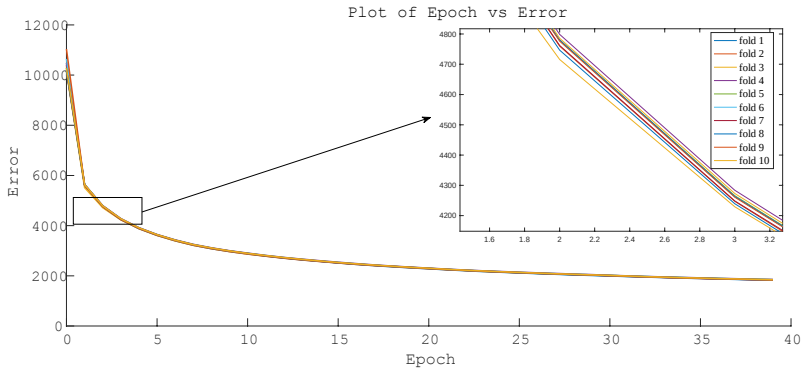

Figure 17: Error vs. Epoch for Bam \cite {bam2014named} data using BP (40 epochs). It is the case of 10-fold cross validation.

Figures 15-17 represents the plot of error against number of epochs when BP with two hidden layers was used for detecting the similarity of data in Nepali language by Bam [11]. The result of training error obtained is shown in Table 6.

8.2.2 Experiment with Nepali data collected by Bam using MCANN

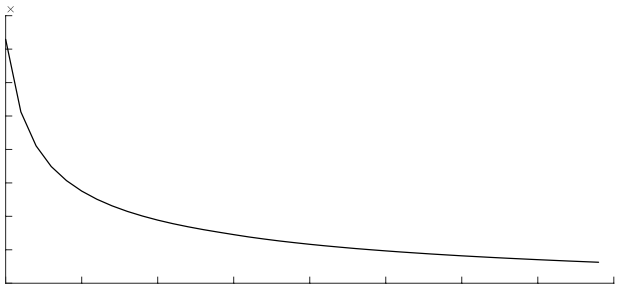

Figure 18: Error vs. Epoch for Bam \cite {bam2014named} using MCANN (40 epochs). Ninety percent data was used as training data and ten percent as test data.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Vol. 6, Issue 1, January 2018

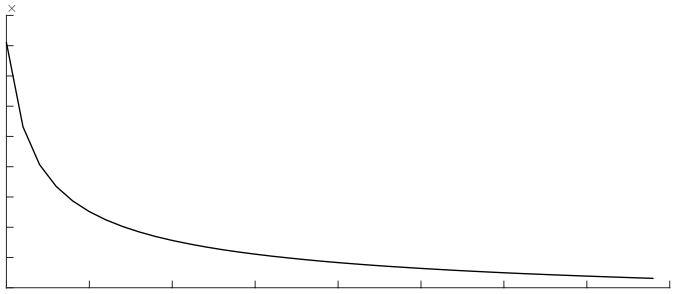

Figure 19: Error vs. Epoch for Bam \cite {bam2014named} using MCANN (40 epochs). Eighty percent data was used for training and twenty percent for testing.

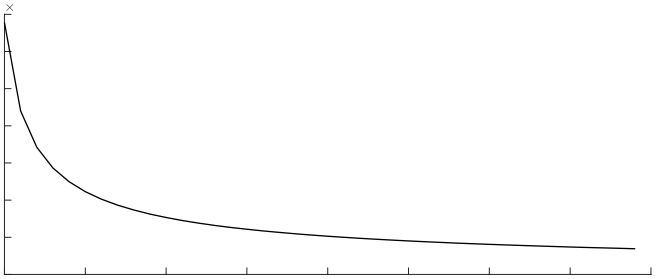

Figure 20: Error vs. Epoch for Bam \cite {bam2014named} using MCANN (40 epochs). Sixty percent data was used as training data and forty percent as test data.

The error obtained on experimenting with Bam data [11] using MCANN using different training and testing data is shown in Figures 18-20 respectively. The results are summarized in Table 7.

Dataset	**Algorithm applied**	**Error obtained on different experiments**		
		5-fold cross validation	**7-fold cross validation**	**10-fold cross validation**
Bam data	BP	1699.706	1775.384	1843.961

Table 6: Result of back propagation on Bam data during line based comparison.

Dataset	**Algorithm applied**	**Error obtained on different experiments**		
		60% train and 40% test data	80% train and 20% test data	90% train and 10% test data
Bam data	**MCANN**	3.4599e-05	1.3106e-04	3.2539e-04

Table 7: Result of MCANN on Bam data during line based comparison.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Vol. 6, Issue 1, January 2018

8.3 Results of Cluster based Analysis

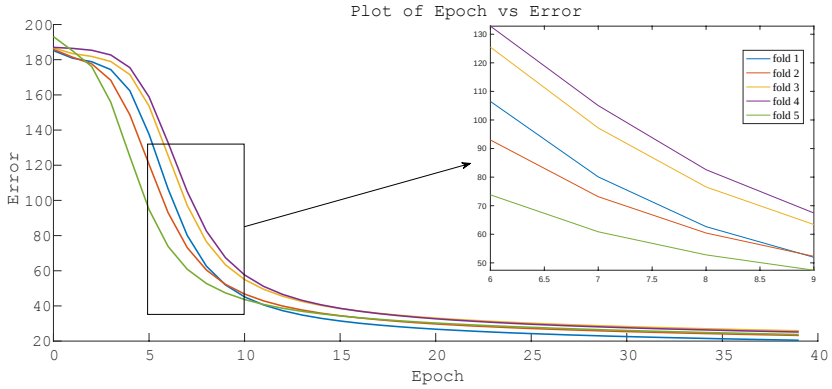

Figure 21: Error vs. Epoch for selected four documents using BP (40 epochs). It is the case of 5-fold cross validation.

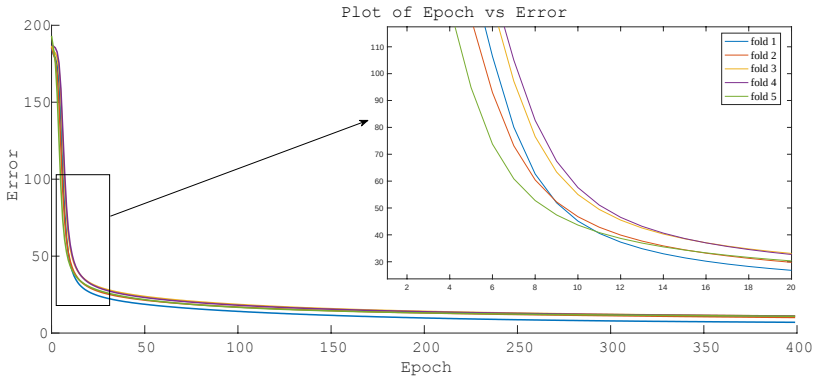

Figure 22: Error vs. Epoch for selected four documents using BP (400 epochs). It is the case of 5-fold cross validation.

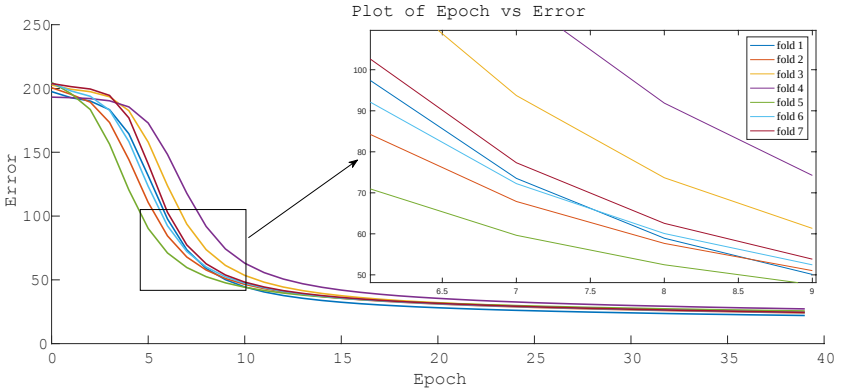

Figure 23: Error vs. Epoch for selected four documents using BP (40 epochs). It is the case of 7-fold cross validation.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Vol. 6, Issue 1, January 2018

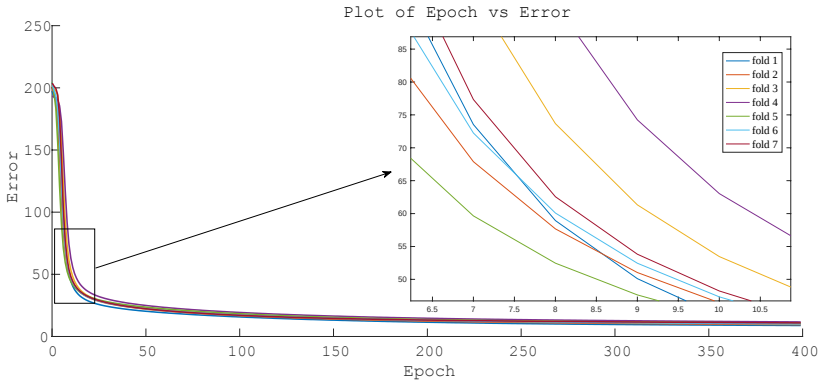

Figure 24: Error vs. Epoch for selected four documents using BP (400 epochs). It is the case of 7-fold cross validation.

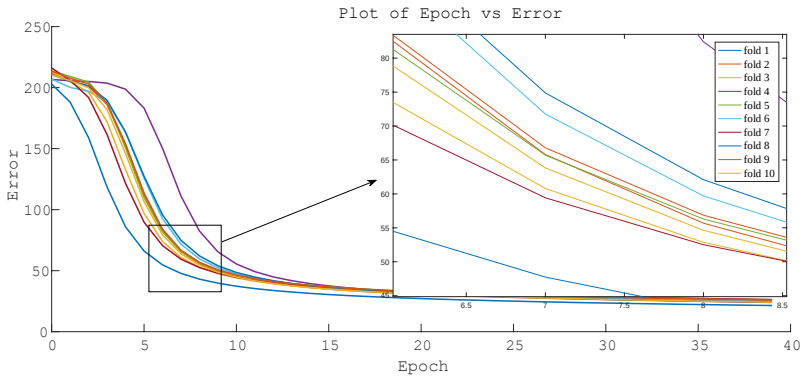

Figure 25: Error vs. Epoch for selected four documents using BP (40 epochs). It is the case of 10-fold cross validation.

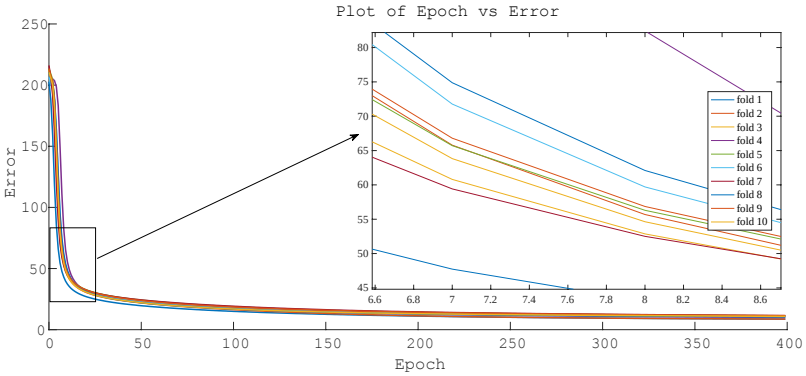

Figure 26: Error vs. Epoch for selected four documents using BP (400 epochs). It is the case of 10-fold cross validation.

Figures 21-26 represents the plot of error against number of epochs when BP with two hidden layers were used for detecting the similarity of selected Nepali thesis. The results are listed in Table 8.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Vol. 6, Issue 1, January 2018

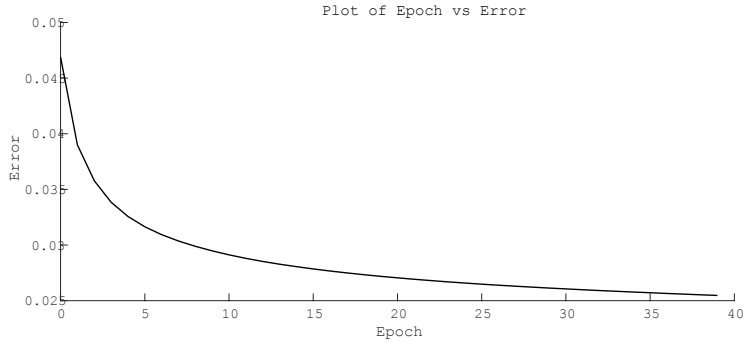

Figure 27: Error vs. Epoch for selected four documents using MCANN (40 epochs). Ninety percent data was used as training data and ten percent as test data.

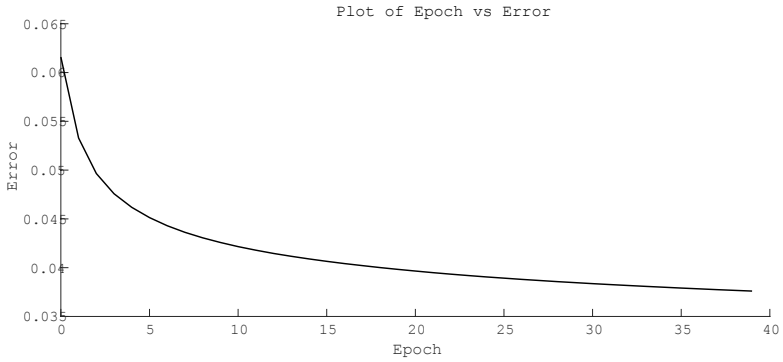

Figure 28: Error vs. Epoch for selected four documents using MCANN (40 epochs). Eighty percent data was used as training data and twenty percent as test data.

Figure 29: Error vs. Epoch for selected four documents using MCANN (40 epochs). Sixty percent data was used as training data and forty percent as test data.

The error obtained on experimenting with selected four documents using MCANN using different training and testing data is shown in Figures 27-29 respectively. The errors obtained are listed in Table 9.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Vol. 6, Issue 1, January 2018

Dataset	Algorithm applied	No. of Epoch	Error obtained on different experiments		
			5-fold cross validation	7-fold cross validation	10-fold cross validation
Selected Nepali Thesis	BP	40	23.743	24.496	24.137
Selected Nepali Thesis	BP	400	10.925	10.926	10.880

Table 8: Result of back propagation on selected Nepali thesis.

Dataset	Algorithm applied	Error obtained on different experiments		
		60% train and 40% test data	80% train and 20% test data	90% train and 10% test data
Selected Nepali Thesis	MCANN	1.5065e-03	3.7599e-02	2.5471e-02

Table 9: Result of MCANN on selected Nepali thesis.

8.4 Results of Experiments Carried Out with Selected Portion of Selected Nepali Thesis

Results of paragraph based experiment carried out on theory section of four documents.

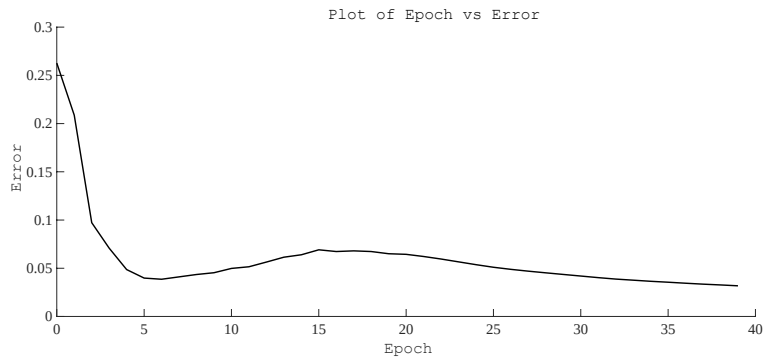

Figure 30: Error vs. Epoch for theory section of documents using MCANN (40 epochs). Ninety percent data was used as training data and ten percent as test data.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Vol. 6, Issue 1, January 2018

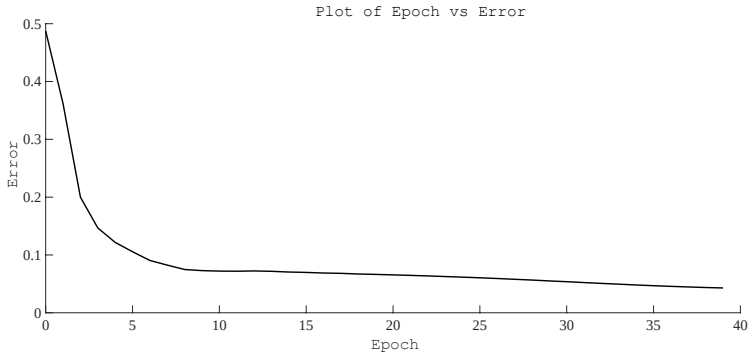

Figure 31: Error vs. Epoch for theory section of documents using MCANN (40 epochs). Eighty percent data was used as training data and twenty percent as test data.

Figure 32: Error vs. Epoch for theory section of documents using MCANN (40 epochs). Sixty percent data was used as training data and forty percent as test data.

The error obtained for above experiment with MCANN using 90% train and 10% test data was 3.1812e-02, 80% train and 20% test data was 4.2914e-02 and 60% train and 40% test data was 2.8589e-02 (in 40 iterations) as shown in Figures 30-32 respectively.

8.5 Results of Line based Experiment Carried out on Theory Section of Four Documents

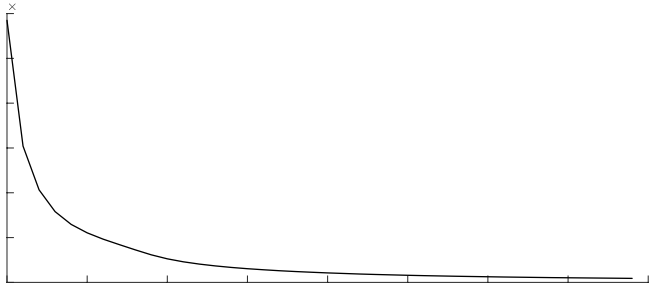

Figure 33: Error vs. Epoch for theory section of documents using MCANN (40 epochs). Ninety percent data was used as training data and ten percent as test data.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Vol. 6, Issue 1, January 2018

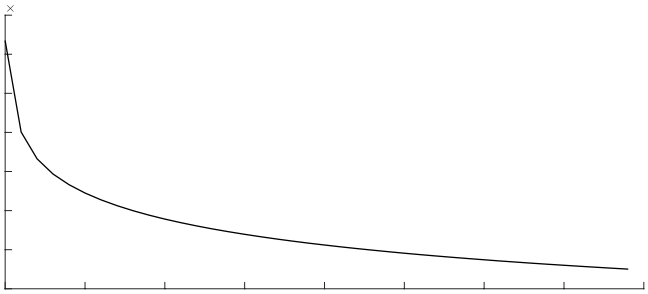

Figure 34: Error vs. Epoch for theory section of documents using MCANN (40 epochs). Eighty percent data was used as training data and twenty percent as test data.

Figure 35: Error vs. Epoch for theory section of documents using MCANN (40 epochs). Sixty percent data was used as training data and forty percent as test data.

The error obtained for above experiment with MCANN using 90% train and 10% test data was 4.4805e-07, 80% train and 20% test data was 1.5503e-04 and 60% train and 40% test data was 2.2957e-02 (in 40 iterations) as shown in Figures 33-35 respectively.

8.6 Results of Paragraph based Experiment Carried out on Result Section of Four Documents

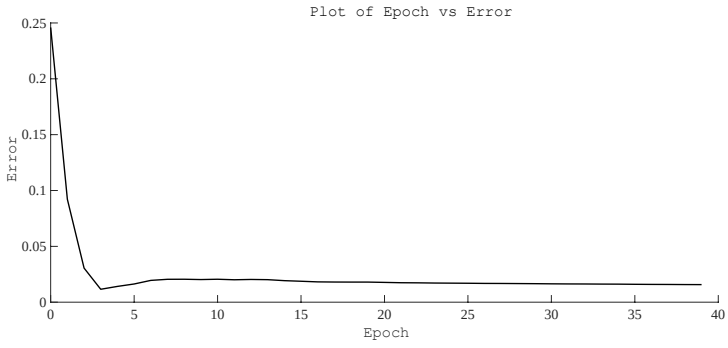

Figure 36: Error vs. Epoch for result section of documents using MCANN (40 epochs). Ninety percent data was used as training data and Ten percent as test data.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Vol. 6, Issue 1, January 2018

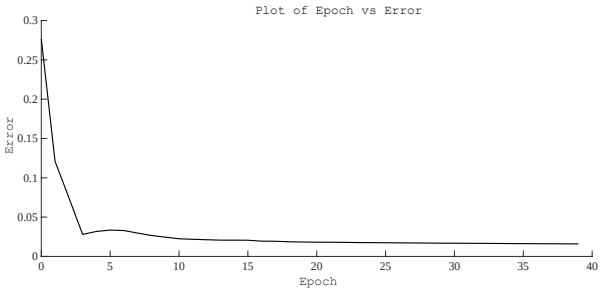

Figure 37: Error vs. Epoch for result section of documents using MCANN (40 epochs). Eighty percent data was used as training data and twenty percent as test data.

Figure 38: Error vs. Epoch for result section of documents using MCANN (40 epochs). Sixty percent data was used as training data and forty percent as test data.

The error obtained for above experiment with MCANN using 90% train and 10% test data was 1.5713e-02, 80% train and 20% test data was 1.5928e-02 and 60% train and 40% test data was 2.0076e-02 (in 40 iterations) as shown in Figures 36-38 and Tables 10-13 respectively.

IX. RESULTS SUMMARY

Dataset	**Algorithm applied**	**Analysis approach**	**No. of epoch**	**Error obtained on different experiments**		
				60% train and 40% test data	**80% train and 20% test data**	**90% train and 10% test data**
Bam data	MCANN	Paragraph based	40	3.0948e-04	2.1130e-04	8.1471e-05
Bam data	MCANN	Line based	40	3.4599e-05	1.3106e-04	3.2539e-04
				5-fold cross validation	7-fold cross validation	10-fold cross validation
Bam data	BP	Paragraph based	40	335.854	350.929	360.370
Bam Data	BP	Line based	40	1699.706	1775.384	1843.961

Table 10: Comparison of result of MCANN and BP model on Bam data.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Vol. 6, Issue 1, January 2018

Dataset	**Algorithm applied**	**Analysis approach**	**No. of epoch**	**Error obtained on different experiments**		
				60% train and 40% test data	**80% train and 20% test data**	**90% train and 10% test data**
Nepali Thesis (11 documents)	MCANN	Paragraph based	40	6.1455e-03	3.0096e-03	2.4219e-03
				5-fold cross validation	7-fold cross validation	10-fold cross validation
Nepali Thesis (11 documents)	BP	Paragraph based	40	6.163	5.111	5.952
Nepali Thesis (11 documents)	BP	Line based	400	0.385	0.231	0.131

Table 11: Lists the result of MCANN and BP on all eleven Nepali theses.

Dataset	**Algorithm applied**	**Analysis approach**	**No. of epoch**	**Error obtained on different experiments**		
				60% train and 40% test data	**80% train and 20% test data**	**90% train and 10% test data**
Selected Nepali Thesis (4 documents)	MCANN	Paragraph based	40	1.5065e-03	3.7599e-02	2.5471e-02
				5-fold cross validation	7-fold cross validation	10-fold cross validation
Selected Nepali Thesis (4 documents)	BP	Paragraph based	40	23.743	24.496	24.137
Selected Nepali Thesis (4 documents)	BP	Line based	400	10.925	10.926	10.880

Table 12: Lists the result of MCANN and BP on selected four Nepali thesis documents.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Vol. 6, Issue 1, January 2018

Dataset (Selected Nepali Thesis)	Algorithm applied	Analysis approach	No. of epoch	Error obtained on different experiments		
				60% train and 40% test data	80% train and 20% test data	90% train and 10% test data
Theory Section	MCANN	Paragraph based	40	2.8589e-02	4.2914e-02	3.1812e-02
Theory Section	MCANN	Line based	40	2.2957e-02	1.5503e-04	4.4805e-07
Result Section	MCANN	Paragraph based	40	2.0076e-02	1.5928e-02	1.5713e-02

Table 13: Lists the result of MCANN algorithm on different portion of selected four Nepali theses.

X. CONCLUSION

Nepali languages documents collected from different sources are passed in the framework for results. Obtained results are then analyzed for their accuracy. MCANN algorithm achieves a convergence in the range of 10−2 to 10−7 for the training error in 40 epochs while general BP algorithm is unable to achieve such a convergence even in 400 epochs. Also, the mean accuracy of BP and MCANN are respectively found to be in the range of 98.657 and 99.864 during paragraph based and line based comparison of the documents.

From the results obtained it is concluded that neural network trained with Monte Carlo method performs better than traditional backpropagation method. Thus, Monte Carlo based Artificial Neural Network is beneficial over general artificial neural network trained using backpropagation learning method for problems related to similarity detection, in particular for Nepalese language texts. When the data size is less (BAM data), the results are not consistent, whereas the Thesis data results (being large in size) are consistent.

XI. FUTURE ENHANCEMENT

This research focuses on extrinsic plagiarism detection of Nepali language based documents. It could be further extended for cross lingual plagiarism detection task. Similarly, performance could be increased by increasing more similarity measures as features. Better analysis could be carried out with datasets of different varieties collected from different fields. Also, effect of Evolutionary algorithms could be studied for detecting the plagiarism on Nepali language documents. Also, this research could be augmented for intrinsic plagiarism detection.

XII. REFERENCES

1. H Shanmugasundaram, Automatic plagiarism detection using similarity analysis. International Arab Journal of Information Technology 2012; 9: 322-326.
2. L Romans, G Vita, et al. Computer-based plagiarism detection methods and tools: an overview. International conference on Computer systems and technologies 2007.
3. C. Dara, An evolutionary neural network approach to intrinsic plagiarism detection. Artificial Intelligence and Cognitive Science 2010; 33-40.
4. SD Sudhir, SZ Gawali, A plagiarism detection mechanism using reinforcement learning," International Journal of Advance Research in Computer Science and Management Studies 2013; 1: 125-129.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Vol. 6, Issue 1, January 2018

5. AM Salha, S Naomie, Plagiarism detection in Arabic scripts using fuzzy information retrieval. Conference on Research and Development 2008.
6. S Efstathios, Plagiarism detection using stopword n-grams. Journal of the American Society for Information Science and Technology 2011; 62: 2512-2527.
7. JF de Freitas, M Niranjan, et al. Sequential Monte Carlo methods to train neural network models. Neural computation 2000; 12: 955-993.
8. MYM Chong, A study on plagiarism detection and plagiarism direction identification using natural language processing techniques. 2013.
9. SN Sivanandam, SN Deepa, Introduction to neural networks using Matlab 6.0. Tata McGraw-Hill Education 2006.
10. HD Matthew, G Andrew, The no-U-turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. Journal of Machine Learning Research 2014; 15: 1593-1623.
11. BB. Surya, SB Tej, Named entity recognition for Nepali text using support vector machines. Intelligent Information Management 2014; 6: 21-29.
12. M Stephen, Machine learning: an algorithmic perspective. CRC press 2015.
13. SS Richard, GB Andrew, Reinforcement Learning: An Introduction, Computer Science And Intelligent Systems The MIT Press 2012.
14. P Martin, E Andreas, et al. Overview of the 3rd International Competition on Plagiarism Detection. Working Notes Papers of the CLEF Evaluation Labs 2011.
15. P Martin, S. Benno, et al. An Evaluation Framework for Plagiarism Detection. in International Conference on Computational Linguistics. Beijing, China: Association for Computational Linguistics 2010.
16. DM Christopher, R Pravbhakar, et al. Introduction to information retrieval. Cambridge University Press 2008.