International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056
Volume: 03 Issue: 10 | Oct -2016 www.irjet.net p-ISSN: 2395-0072

Online news classification using Deep Learning Technique

Sandeep Kaur, Navdeep Kaur Khiva

M.tech Student , Dept. Of CSE at GZSCCET Bathinda, Punjab, India.
Assistant Professor Dept. Of CSE at GZSCCET Bathinda, Punjab, India.

Abstract - Abstract - *A news classification task begins with a data set in which the class assignments are known. Classification are discrete and do not imply order. The goal of classification is to accurately predict the target class for each case in the data. Due to the Web expansion, the prediction of online news popularity is becoming a trendy research topic. Many researches have been done on this topic but the best result was provided by a Random Forest with a discrimination power of 73% accuracy. So, in this paper, main aim is to increase accuracy in predicting the popularity of online news. Thus, Neural Network will be implemented to acquire better results.*

Key Words*: *Online news Classification, Neural Network, Precision rate, Recall rate, Accuracy

1. INTRODUCTION

There exists a large amount of information being stored in the electronic format. With such data, it has become a necessity of such means that could interpret and analyze such data and extract such facts that could help in decision-making [1, 2, 3]. Data mining which is used for extracting hidden information from huge databases is a very powerful tool that is used for this purpose. News information was not easily and quickly available until the beginning of last decade. But now, news is easily accessible via content providers such as online news services [4, 5].

A huge amount of information exists in form of text in various diverse areas whose analysis can be beneficial in several areas [6]. Classification is quite a challenging field in text mining as it requires prepossessing steps to convert unstructured data to structured information. With the increase in the number of news it has got difficult for users to access news of interest which makes it a necessity to categories news so that it could be easily accessed. Categorization refers to grouping that allows easier navigation among articles. Internet news needs to be divided into categories. This will help users to access the news of their interest in real-time without wasting any time [7, 8, 9]. When it comes to news it is much difficult to classify as news are continuously appearing that need to be processed and those news could be never-seen-before and could fall in a new category.

In this paper, a review of news classification based on its contents and headlines is presented. A variety of classification has been performed in past that are:

i. To evaluate a model, which first use data-driven models for predicting what is more likely to happen in the future, and then use modern optimization methods to search for the best possible solution given what can be currently known and predicted [10, 11, and 12].
ii. To implement neural network for classification of news based on features extracted.
iii. To evaluate the performance using various parameters like Precision rate, Recall rate and accuracy.

Definition: 1

Recall rate (r) is the positives that has been detected by algorithm

Definition: 2

Precision rate (p) is the negatives that have been recognized by the algorithm.

Definition: 3

The accuracy is the exactness of the true values obtained by implementation of proposed algorithm.

2. NEURAL NETWORK

Below Figures describes the process of NN working in proposed work [12]. Training has been done using newff function in MATLAB. In this work. neural network loop will be run for 5 times so that desired output can be obtained effectively.

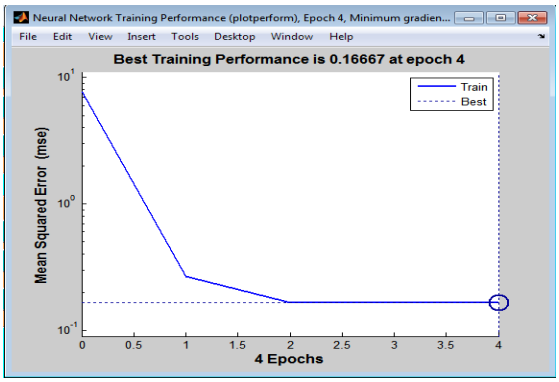

Fig 1. Mean Square Error

The figure Above shows the value of obtained mean square error for 4 epochs having values 10 $^{-1.}$

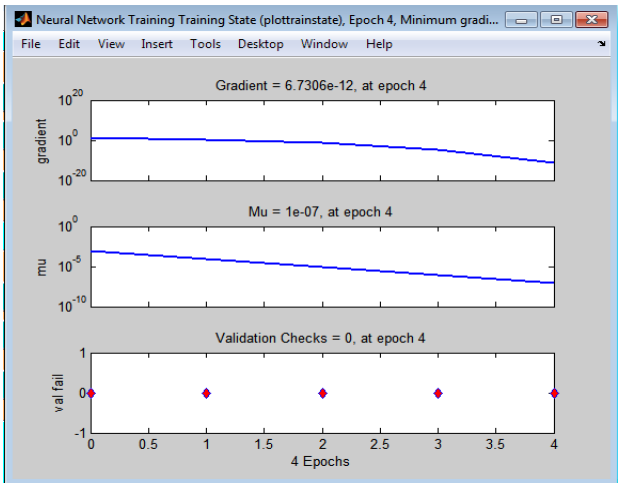

Fig 2. Neural Network Functions

Above figure shows the various function values based on neural network like gradient = 6.730, Mu= 1e-.7 and validation checks= 0 for 4 epochs.

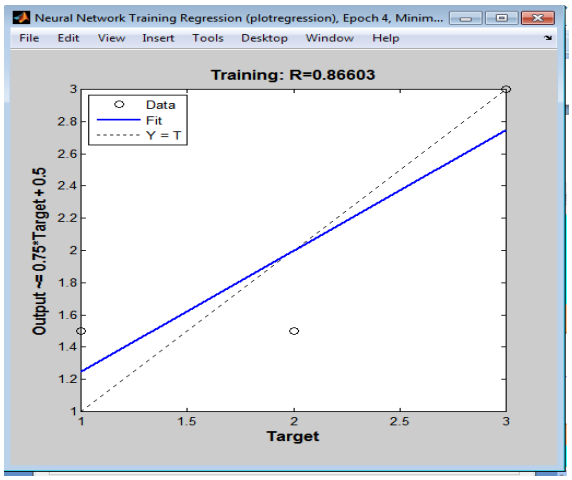

Figure 3. Neural Network Output

Above figure shows the training output value of neural network and the obtained value = .866.

3. A GLANCE OF EXISTING TECHNIQUES

Various algorithms has been proposed by authors, but in this literature survey only recently used methods has been presented [13- 25].

Table.1 Literature Survey

Author(year)	Title	Technology
Vishawnath et.al (2014)	Usage for Machine learning technique for text and document mining	KNN algorithm
Yun Lin et.al (2014)	Study on text classification utilizing SVM-KNN	SVM-KNN algorithm
Zaghoul et.al (2013)	Usage of Arabic Text Classification Based on Reduction of various Features Using ANN	artificial neural network (ANN)
Anuradha (2013)	Text Classification using Relevance Factor as Term Weighing Method with NN Approach	multilayer feed forward networks
Nidhi and Gupta, (2012)	A novel techniq for Punjabi text classification	Naive Bayes and Ontology Based Classification)
La Lei et.al (2012),	Categorization Text utilizing SVM with exponent weighted ACO	Ant Colony Optimization
Saha et.al (2012),	Web Text Classification Using a Neural Network	Machine Learning Algorithm
Aurangzeb et.al (2010)	Ontology based text categorization - telugu documents	survey of various techniques
Luo et.al (2010),	Feature selection for text classification using OR+SVM-RFE	on
Harrag et.al (2009)	Neural Network for Arabic text classification	SVD and ANN
Ali and Ijaz et.al (2009)	Urdu text classification	SVM and Naive Bayes classifier

Mohammad Abdul et.al (2009)	Text Classification Using Machine Learning	feature vector
Ikomonios et.al (2005)	Using Machine Learning Techniques for Text Classification	machine learning algorithms
Mitra et.al (2005)	Text classification using neuro-SVM model with latent semantic indexing	Recurrent neural network (RNN) and a least squares support vector machine (LS-SVM).
Lam et.al (1999)	Text categorization using neural network for Feature reduction	PCA
Frank and Bouckaert	text classification using naive bayes with unbalanced classes	Multino mial naive Bayes (MNB).

4. NEWS CLASSIFICATION WORKFLOW

The main objective of text classification is the classification of documents into a settled amount of predefined classes. Every single archive could be in no class, several, or precisely one by any means. Our main goal is to utilize machine learning algorithm to understand NN classifier.

Algorithm 1: Classification using NN

Input: Training of D Documents

Output: Classified Documents

Split the training set T into training T and validation V

Compute the accuracy on V of the classifier built on T

Compute P using the documents in T

For each document di ∈ T

For each term j

di,j = P1/δ j di,j

Scale di so that ||di||2 = 1

Compute the accuracy on V of the classifier built on T

If accuracy does not decrease

Accuracy will be measured using following classifier

{Neural Network

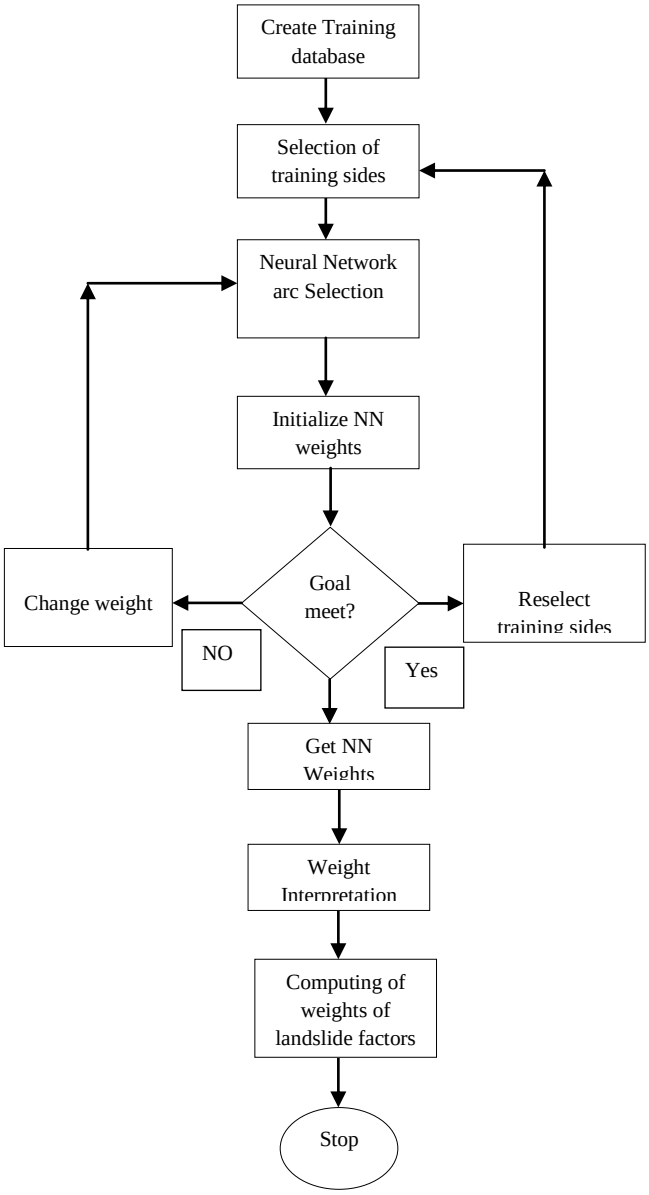

Fig 4. Proposed Architecture

Training set of American

Training set of Indian

Training set of German

Compute newftt function to train

{net=newff(Trainingset',Target,10);

net.trainParam.epochs=50;

net =train(net,Trainingset',Target)}

Save neural network

Load neural network for testing

Display results

End}

5. RESULT EVALUATION

The whole simulation for proposed work has been done in MATLAB 2010a using various parameters like precision rate, recall rate as well as using accuracy.

Table 2. NN Values

Iterations	Precision rate	Recall Rate	Accuracy
2	0.78	0.047	99.86
4	0.73	0.078	98.85
6	0.76	0.085	98.88
8	0.75	0.043	99.83
10	0.71	0.079	98.82
12	.79	.026	99.34
14	.81	.034	99.45
16	.76	.038	99.25

Table 2 shows the different values shown by neural network classifier of Precision rate, recall rate and accuracy. The values are according to the different iterations. Average of precision rate is 0.76125; Recall rate has an average of 0.05375 while the accuracy has an average of 99.285.

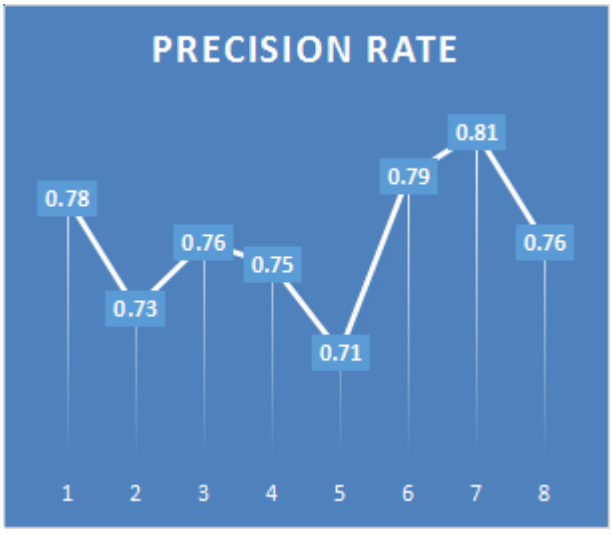

Fig 5 Precision rate

Precision rate values must be low to give good results and in proposed work, precision rate is having good rate values.

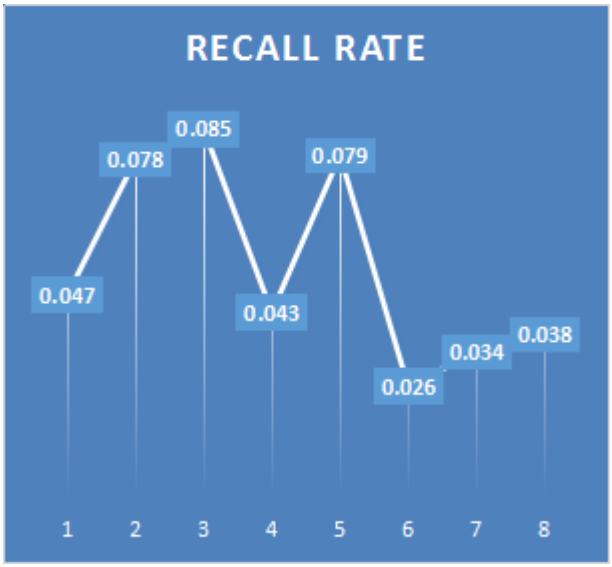

Fig 6 Recall rate

Also, the recall rate values must be low to give good results for proposed model and in proposed work, thisrate is having good rate values ranging from .026 to .079.

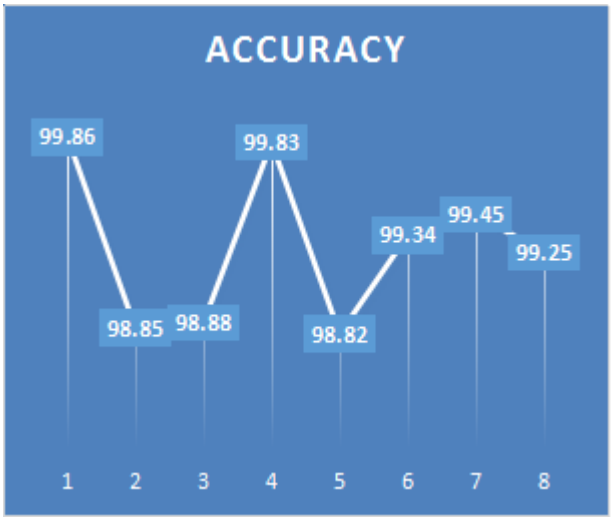

Fig 7 Accuracy rate

Accuracy is the best measure to evaluate the effectiveness of the system and in proposed work the accuracy values is nearby 99.46%.

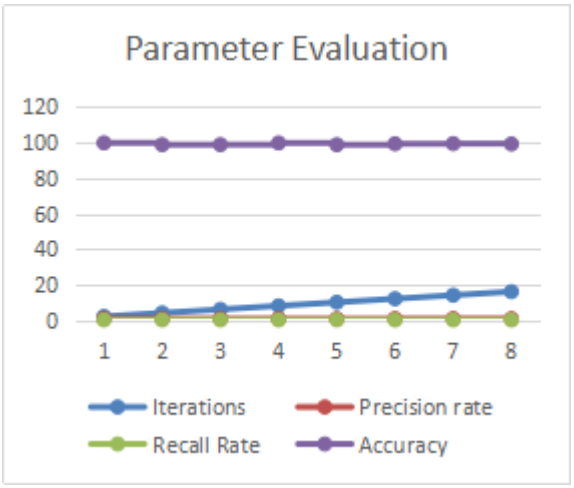

Fig 8. Neural Network Graph

The above figure shows the false acceptance rate and false rejection rate using neural network with respect to the number of itearations and shows that these performance parameters is having less measure which is having high accuracy on the basis of false acceptance rate and false rejection rate and shows that neural network classifier performance for the classification process of news.

6. CONCLUSION

It has been concluded that results showed that there are four types of categories that has been proposed like politics, financial and sports.Also the classification process has been implemented using neural network classifier. From simulation result it has been concluded that NN with accuracy 99.93 has been obtained and has provided good results w.r.t traditional methods.

REFERENCES

[1] P. D. Turney, "Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews," in Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, 2002, pp. 417–424.

[2] T. Wilson, J. Wiebe, and P. Hoffmann, "Recognizing contextual polarity: An exploration of features for phrase-level sentiment analysis," Computational Linguistics, vol. 35, no. 3, pp. 399–433, 2009.

[3] C. Quan and F. Ren, "Construction of a blog emotion corpus for chinese emotional expression analysis," in Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, vol. 3, 2009, pp. 1446–1454.

[4] D. Das and S. Bandyopadhyay, "Word to sentence level emotion tagging for bengali blogs," in Proceedings of the ACL-IJCNLP 2009 Conference, 2009, pp. 149–152.

[5] Y. Chen, S. Y. M. Lee, S. Li, and C.-R. Huang, "Emotion cause detection with linguistic constructions," in Proceedings of the 23rd International Conference on Computational Linguistics, 2010, pp. 179–187.

[6] M. Purver and S. Battersby, "Experimenting with distant supervision for emotion classification," in Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, 2012, pp. 482–491.

[7] K. H.-Y. Lin, C. Yang, and H.-H. Chen, "What emotions do news articles trigger in their readers?" in Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 2007, pp. 733–734.

[8] ——, "Emotion classification of online news articles from the reader's perspective," in Proceedings of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, vol. 1, 2008, pp. 220–226.

[9] Y.-j. Tang and H.-H. Chen, "Mining sentiment words from microblogs for predicting writer-reader emotion transition." in Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC), 2012, pp. 1226–1229.

[10] Y. Bengio, R. Ducharme, P. Vincent, and C. Janvin, "A neural probabilistic language model," The Journal of Machine Learning Research, vol. 3, pp. 1137–1155, 2003.

[11] F. Morin and Y. Bengio, "Hierarchical probabilistic neural network language model," in Proceedings of the international workshop on artificial intelligence and statistics, 2005, pp. 246–252.

[12] R. Collobert and J. Weston, "A unified architecture for natural language processing: Deep neural networks with multitask learning," in Proceedings of the 25th international conference on Machine learning. ACM, 2008, pp. 160–167.

[13] Vishwanath (2014), "Machine learning approach for text and document mining", Vol.7, Issue.1, pp.41-48.

[14] Yun Lin, (2014), "Research on text classification based on SVM-KNN", IEEE, Software Engineering and Service Science (ICSESS), 2014 5th IEEE International Conference, pp. 842 – 844.

[15] Zaghoul, (2013), "Arabic Text Classification Based on Features Reduction Using Artificial Neural Networks", IEEE, Computer Modelling and Simulation (UKSim), 2013 UKSim 15th International Conference, pp. 485 – 490.

[16] Anuradha (2013), "Neural Network Approach for Text Classification using Relevance Factor as Term Weighing Method", IJCA Journal, Vol. 68, Issue.17, pp-37-41.

[17] Nidhi and V. Gupta (2012), "Algorithm for punjabi text classification", International Journal of Computer Applications (0975 – 8887), Vol.37, Issue.11, pp. 30-35.

[18] Luo, (2010), "Feature selection for text classification using OR+SVM-RFE", IEEE, Control and Decision Conference (CCDC), pp. 1648 – 1652.

[19] Harrag, (2009), "Neural Network for Arabic text classification", IEEE, Applications of Digital Information and Web Technologies, 2009. ICADIWT '09. Second International Conference, pp. 778 – 783.

[20] A. R. Ali and M. Ijaz (2009)," Urdu text classification", ACM, Proceedings of the 7th International Conference on Frontiers of Information Technology, Vol. 80, Issue 3, pp.765-769.

[21] Michal Toman, Roman Tesar, and Karel Jezek (2006), "Influence of word normalization on text classification", Proceedings of InSciT, pp. 354–358.

[22] I. Ikonamakis (2005), "Text Classification Using Machine Learning Techniques", WSEAS TRANSACTIONS on COMPUTERS, Vol.4, Issue.8, pp. 966-974.

[23] Mitra, (2005), "A neuro-SVM model for text classification using latent semantic indexing", IEEE, Neural Networks, 2005. IJCNN '05. Proceedings. 2005 IEEE International Joint Conference, Vol.1, pp. 564 – 569

[24] Lam, (1999), "Feature reduction for neural network based text categorization", IEEE Database Systems for Advanced Applications, 1999. Proceedings. 6th International Conference, pp. 195 – 202.

[25] E. Frank1 and R. R. Bouckaert (2006), "naive bayes for text classification with unbalanced classes", ACM, Proceedings of the 10th European conference on Principle and Practice of Knowledge Discovery in Databases, pp.503-510