Proceedings of 9th IOE Graduate Conference
Peer Reviewed
ISSN: 2350-8914 (Online), 2350-8906 (Print)
Year: 2021 Month: March Volume: 9

Nepali News Document Classification using Global Vectors and Long Short Term Memory

Santosh Kumar Thapa [a], Sitaram Pokhrel [b]

[a, b] *Department of Electronics and Computer Engineering, Pashchimanchal Campus, IOE, TU, Nepal*
Corresponding Email: [a] santosh.752534@pasc.tu.edu.np, [b] sitaram.pokhrel@gmail.com

Abstract
Separating the news documents into different groups according to the predefined categories is called News Classification. The easy availability and accessibility of the News Articles has increased the popularity of Online News Portals and attracted thousands of online News audience. The increasing online Nepali news portals and rapidly growing news documents has an effective influence on the readers and their daily life routine. Thus, this research aims to upgrade Nepali News document classification based on Long Short Term Memory Recurrent Neural Network and Global Vectors for Word Representation to label the news into its relevant category. For this research, the total 116736 data in 14 different categories is collected through web scraping and embedded using Global vectors for word representation. The LSTM model implemented for News classification achieved the accuracy of 95.36% leading the accuracy of CNN with 93.97% and DNN with 90.75%.

Keywords
Convolutional Neural Network (CNN), Dense Neural Network (DNN), Global Vectors (GloVe), Long Short Term Memory (LSTM)

1. Introduction

The popularity of the Internet and rapidly growing online surfers has increased the demand of organized and easily accessible articles of their interest. The competitive online Nepali news portals and other media on the Internet produce a huge amount of news articles on regular basis. The basic need is to provide organized contents to the e-readers. It helps them to access and share particular news of their interest. This is possible when a machine learning algorithm is implemented for automatic text classification to classify the huge amount of news document to its proper category.

The complex morphology of Nepali text due to various forms of existing 36 consonants and 12 vowel letters makes feature extraction in Nepali document classification difficult than in English language. Feature extraction with TF-IDF method, based on word frequency count and word2vec word embedding technique is used mostly by the Nepali text researchers. This paper has implemented the GloVe for word embedding in Nepali plain text news documents which focuses on the globally co-occurring words in the text corpus and vectorize the words. These vectors represents the probability of two words appearing together.

Labeling the huge amount of text from heterogeneous source for differentiating the categories of the document and preserving its essence is difficult with the traditional machine learning algorithms and frequency based feature extraction methods. Deep neural networks have the tenacity to take large amount of information, learn from them and provide the best output. Due to the better potential and improved performance than the traditional machine learning algorithms, the aim of this research is to build a new dataset for 14 categories based on Nepali News, analyze the combined efforts of GloVe and LSTM model for Nepali News document classification and compare the results with CNN and DNN models.

2. Related Works

The experimental analysis of Shahi and Pant [1] for Nepali news classification based on SVM, Naïve Bayes and Neural Network shows SVM as the best

model. The paper discusses about the lower accuracy of 75% due to the feature extraction with TF-IDF method. The words in the corpus is context dependent, conflicting with the TF-IDF theory which assumes the independence of the words in the corpus. Shahi and Pant debate about the effect of diversity and noise of data on the classification accuracy.

Basnet [2] experimented the use of LSTM and SVM with word2vec for the improvement of Nepali news recommendation with 64,925 data for 8 classes. The LSTM model has exceeded the performance of SVM with 3.22%. Nepali SMS classification task is examined with Naïve Bayes and support vector machine approach. In contrast to the expression of the full text, the length of the SMS is very short, so the various factors such as SMS headings, words, and their frequency have been used to outline the SMS. To classify SMS into either spam or not spam, the SVM and Naïve Bayes based classification techniques were applied. The Naïve Bayes outperform the SVM with 5% (92% accuracy for Naïve Bayes and 87% accuracy for SVM) due to the few numbers of the feature taken in consideration [3].

Zaiying Wang et al.[4] researched on the performance of BiGRU and attention mechanism with the traditional common deep learning model for hot news classification. The performance of traditional deep learning model is outperformed by BiGRU + attention deep learning model.

To optimize the vector representation of the text, Kaushal Kafle et al. [5] has experimented the usage of neural networks in the word2vec. The TF-IDF method performance is surpassed by 1.6 percent using word2vec model where Support vector Machine was used for classification.

Pennington et al. developed a model that utilizes the advantage of count data while simultaneously capturing the predominant significant linear substructures in recent log-bilinear prediction-based techniques such as word2vec. Authors of paper [6], presents GloVe as a new log-bilinear regression model that dominates other models on word analogy, word similarity, and recognition task of named entities.

Niharetal. [7] exhibited better performance of LSTM in document classification in contrast to other machine learning algorithms. LSTM has achieved an accuracy of up to 93% in document classification. However, this approach combined with Global vectors for word embedding has not been tested with Nepali language, as the language itself has a different set of vocabularies and complexities as compared to English language.

Jo, Taeho [8] argues about the difference between document classification and text categorization. The large dimensionality and the infrequent distribution are the two major problems while embedding documents into numerical vectors. A string vector helps to avoid these problems. The string vectors represent documents quite transparently than numerical vectors. Thus, the author proposed that the documents should be encoded into string vectors.

As stated by Xue and Li [9], the generalization error is affected by the number of trees in the random forest. It relies on the strength as well as the degree of dependency between the decision trees. The experimental analysis of the random forest-based text categorization model shows that the number of decision trees in the random forest and the feature dimension have some effect on the categorization model's performance.

According to [10], LSI based classification is less accurate than LDA and SVM, as it aims to explore the most representative features for the text classification, rather than the most discriminatory features. Wang and Qian has used three steps for text classification. The first step used was LDA for dimensionality reduction to obtain text subspace. Then, the new text to be classified is projected into the text subspace. In the end, the text is classified by using the SVM classifier.

3. Methodology

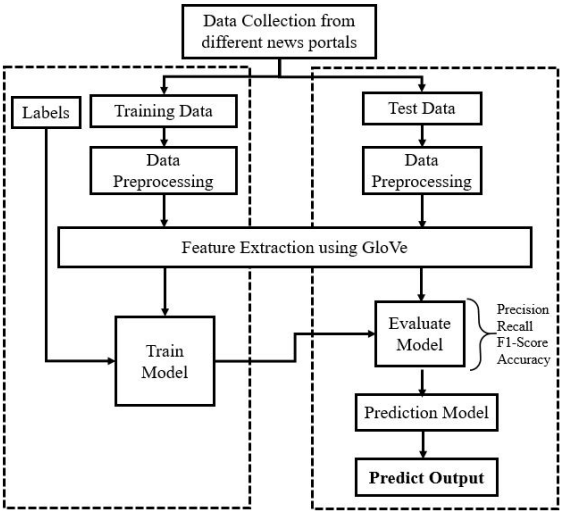

Figure 1: News Classification System Methodology

3.1 Data Collection

Collection of data is the crucial part of machine learning to train the model. It plays decisive role in training the models build based on machine learning algorithms. Among the most popular News portals in Nepal, six websites are scraped using Beautiful Soup, python scraping tool. News from the last five years till April 2020 is collected to train and test the model. Some of the well-known News portals used for data collection are:

1. Online Khabar (onlinekhabar.com)
2. Rato Pati (ratopati.com)
3. Image Khabar (imagekhabar.com)
4. Gorkhapatra Online (gorkhapatraonline.com)
5. DC Nepal (dcnepal.com)
6. Ujyalo Online (ujyaloonline.com)

Data collected in different categories are as follows:

Table 1: Data Collection in 14 categories

S.No.	News Categories	News Articles Collected
1	Automobiles	6513
2	Diaspora	6224
3	Economy	11705
4	Employment	6021
5	Entertainment	11865
6	Health	6000
7	International	13610
8	National	6000
9	Opinion	6604
10	Politics	9952
11	Society	6000
12	Sports	13890
13	Technology	6019
14	Tourism	6333
	Total	**116736**

3.2 Data Preprocessing

Data preprocessing is the process of cleaning and preparing the text for feature extraction and classification. This process involves removal of the noise and uninformative text. Data collected from different sources are tokenized. The white space, numbers and special symbol like [! ” # $ % &’ ( )*+,- . / : ; ¡ = ¿ ? @ [ ] ˆ_ ‘ —] are removed. Stop words do not hold any meaning useful for the analysis rather introduce noise in the machine learning model. Such empty words are removed for the better analysis and modeling process.

Dataset contains Nepali plain text news articles of various length along with their headlines. The maximum sequence length of the articles is fixed to 200 by padding and truncation before feeding it to the deep learning models.

3.3 Feature Extraction

Words needs to be converted into numeric form so that it can be implemented for machine learning. Here, Glove vectorizer is used to represent the words in the form of vectors. Mathematical equations of GloVe are:

$$P_{ij} = P(j/i) = X_{ij}/X$$

$$F(w_i, w_j, w_k) = P_{ik}/P_j k$$

$$F((w_i - w_j)^T w_k) = (F(w_i^T, w_k))/P_{jk}$$

$$F(w_i^T w_k) = P_{ik} = X_{ik}/X_i$$

$$J = \sum_{i,j=0} f(X_{ij})(w_i^T w_j + b_i + b_j^- - logX_{ij})^2$$

$$f(X_{ij}) = \begin{cases} (x/x_{max})^\alpha & \text{for } x < x_{max} \\ 1 & \text{otherwise} \end{cases}$$

Where, X_{ij} denotes the number of times word ’j’ occurs in the context of word ’i’, X_i is the total number of times word ’i’ appeared in the corpus, X denotes word-word co-occurrence matrix counts, P_{ij} is the probability of word ’j’ appearing in the context of ’i’ and F is the function with unknown parameters. Similarly, v is the vocabulary size, w_i is the vector for target word, w_j is the vector for context word, w_i^T is the transpose of w_i, b_i and b_j are the scalar biases for the target and context word. J denotes the cost function, $f(X_{ij})$ is a weighting function that acts to reduce the impact of frequent co-occurring terms with α=3/4.

Glove vectors deals with the context words and their relation to the target word along with their global co-occurrences. The co-occurrence matrix is calculated with the help of vocab count file. On the basis of co-occurrence matrix Glove is trained to produce word embedding of size 50,100 and 200. The word embedding preserves the meaning of the words in the sentence, it is not necessarily important to stem the words in preprocessing phase. One of the major disadvantage of word stemming is that the words gets morphed and may lose their actual meaning.

3.4 Algorithms

3.4.1 Dense Neural Network

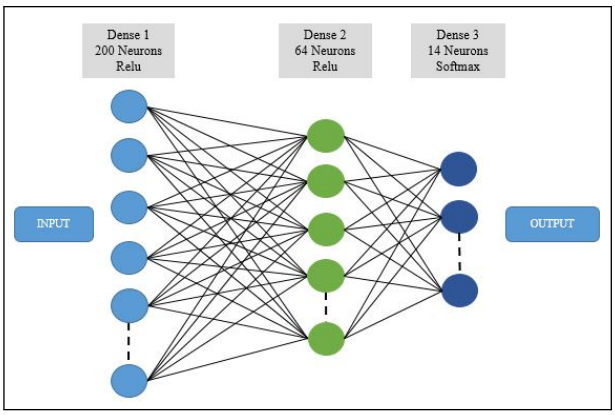

Figure 2: Dense Neural Network

A DNN is a network of fully connected layers linking each neurons in a layer to each neurons in the other layer. DNN consists of an embedding layer as an input layer. The input layer receives GloVe word vectors of news articles. The flatten layer is applied to convert the data into one dimensional array before feeding it to the dense layer. The first dense layer consist of 128/200 neurons, the second hidden layer of 25/64 neurons and an output layer of 14 neurons. All the neurons are interconnected with each other in each layer.The previous layer provides learning features as an input to the present layer's neurons. The two dense layers are provided with a Rectified Linear Unit (ReLU) activation function while the final output layer is provided with the Softmax activation function.

3.4.2 Convolutional Neural Network

A convolutional kernel looks at the embedding for multiple words and slides a window in the case of text classification. The window helps to look at several word embedding in a sequence. The kernels is large rectangle with dimension 3x200 (with an embedding length of 200). In 1D convolution, the kernel moves in only one direction. The kernels slides down through a list of word embedding, to process an entire sequence of words. The maxpooling layer interprets all the high level features to generate the sentence representations. This layer discards the less relevant, locational information regardless of their location in a sequence and forces the network to maintain only maximum value in a feature vector that is the most useful local feature.

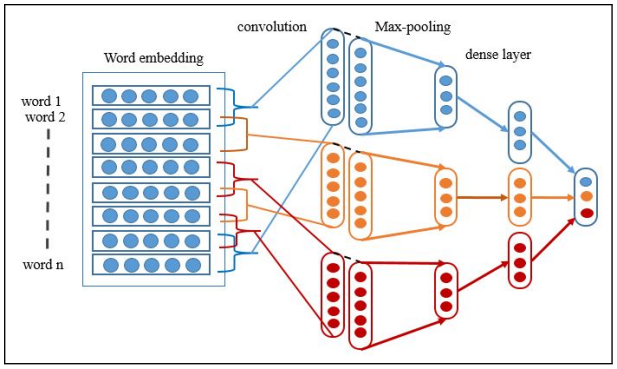

Figure 3: Convolutional Neural Network

3.4.3 Long Short Term Memory

A special form of RNN, capable of learning long-term dependencies, is LSTM networks. Forget gate, input gate and output gates are the internal mechanism in LSTM. These three gates helps to regulate the flow of information and learn which data in a sequence is important to remember/retain and which are irrelevant and can be forgotten/dropped. By this basic method, LSTM learns to use relevant information to make predictions.

$$f_t = \sigma(W_f * S_{t-1} + W_f * X_t)$$
$$i_t = \sigma(W_i * S_{t-1} + W_i * X_t)$$
$$o_t = \sigma(W_o * S_{t-1} + W_o * X_t)$$
$$c_t^{'} = \tanh(W_c * S_{t-1} + W_c * X_t)$$
$$c_t = (i_t * c_t^{'}) + (f_t * c_{t-1})$$
$$c_{(t+1)} = o_t * \tanh(c_t)$$

Where, σ is the sigmoid function, f_t is the forget gate, i_t is the input gate, o_t denotes the output gate, W_f is forget weight, W_i input weight, W_o output weight, X_t indicates the input, S_{t-1} is the previous state, C_t stands for the cell state and C_t' represents the intermediate cell state.

3.5 Model Evaluation

Model is evaluated on the basis of Accuracy, Precision, Recall and F1-Score. The model is validated using confusion matrix. The evaluation matrix used are:

$$\text{Precision} = \frac{TP}{(TP+FP)}$$
$$\text{Recall} = \frac{TP}{(TP+FN)}$$
$$\text{F1-Score} = \frac{2 * Recall * Precision}{(Recall + Precision)}$$
$$\text{Accuracy} = \frac{(TP+TN)}{(TP+TN+FP+FN)}$$

4. Result and Analysis

Three different experiments were conducted in this research. The first experiment was performed to find the best performance of the proposed model. Different LSTM models like, Vanilla LSTM, Stacked LSTM and Bidirectional LSTM are experimented to find the optimal one. The second experiment was carried out to compare the proposed model with Convolutional Neural Network and Dense Neural Network. The third experiment was based on the datasize.

4.1 Observation with different LSTM Models

Among different LSTM Models, Bidirectional LSTM is found to be the optimal model for Nepali news document classification. Bidirectional LSTM helps to get information of the future words. This is achieved by splitting the LSTM into two layers. The first layer is feed forward and the direction of second layer is reversed. BiLSTM is trained with 64 hidden units, which counts a total of 128 units considering 64 each for two individual LSTM layers propagating in two opposite direction. Table 2, illustrates the performance matrices of LSTM models.

Table 2: Experiment based on LSTM Models

Performance Measures	Vanilla LSTM	BiDirectional LSTM	Stacked LSTM
Accuracy	93.49%	**93.94%**	93.57%
Precision	93%	93%	93%
F1-Score	93%	93%	93%

4.2 Observation with Embedding Size

GloVe is trained by taking minimum vocab count 5 with window size 15 to produce the word embedding of 50, 100 and 200 dimension.The experiment conducted on these embedding dimension with DNN, CNN and LSTM model comparison is shown in table 3. The table clearly shows an improvement in accuracy as the embedding size increases in each deep learning model. BiLSTM has the best performance among DNN and CNN models with 93.94% accuracy in the small balanced dataset training.

Table 3: Experiment based on Embedding Size

Models	50	100	200
DNN	89.41%	89.58%	90.17%
CNN	87.91%	91.67%	92.20%
BiLSTM	**90.76%**	**92.83%**	**93.94%**

4.3 Observation with Small and large Dataset

The amount of data has the great impact in deep learning model. It is the basic thing from which model learn to distinguish the class and their related data. Two different experiments has been conducted to observe the performance of DNN, CNN and LSTM on the basis of dataset. First, the models were trained on balanced dataset of 84000, that is 6000 dataset for 14 each categories with 514189 vocab count. This was helpful to find the optimal LSTM model for news document classification. Secondly, total unbalanced dataset of 116736 was used, which contains uneven amount of data for 14 categories with total vocab count 561992. Deep learning models trained in large dataset has performed comparatively better than in small balanced dataset. Table 4 shows that the increment in datasize helps the models to learn sufficient vocabularies and give better performance.

Table 4: Experiment based on Dataset

Models	Small Balanced Dataset	Large Unbalanced Dataset
DNN	90.17%	**90.75%**
CNN	92.20%	**93.97%**
BiLSTM	93.94%	**95.36%**

Table 5: Classification Report of LSTM Model

Categories	Precision	Recall	F1-Score	Support
Automobiles	0.9969	0.9807	0.9887	1971
Diaspora	0.8999	0.9279	0.9137	1832
Economy	0.8920	0.9319	0.9115	3511
Employment	0.9960	0.9703	0.9830	1819
Entertainment	0.9577	0.9696	0.9636	3620
Health	0.9249	0.9586	0.9414	1811
International	0.9779	0.9671	0.9724	4067
National	0.9815	0.9470	0.9640	1793
Opinion	0.9602	0.9358	0.9478	1932
Politics	0.9185	0.9453	0.9317	3015
Society	0.8686	0.8019	0.8339	1772
Sports	0.9873	0.9857	0.9865	4190
Technology	0.9961	0.9983	0.9972	1781
Tourism	0.9968	0.9769	0.9867	1905
Avg/Total	0.9539	0.9498	0.9516	35019

The LSTM model evaluation results are illustrated in the form of Classification Report and accuracy/loss graph. The model initially run for 200 epochs regularized with 0.5 dropout has stopped at 162_{th} epoch due to early stopping with the patience of 6. Table 5, shows the Automobiles with highest precision 99.69% among 14 categories, followed by

Tourism 99.68% and Technology 99.61% whereas Society is found to be classified with lowest precision 86.86%.

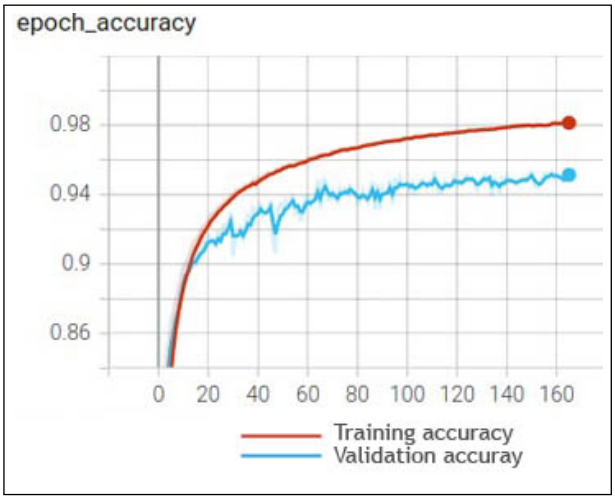

Figure 4: Accuracy graph of LSTM model

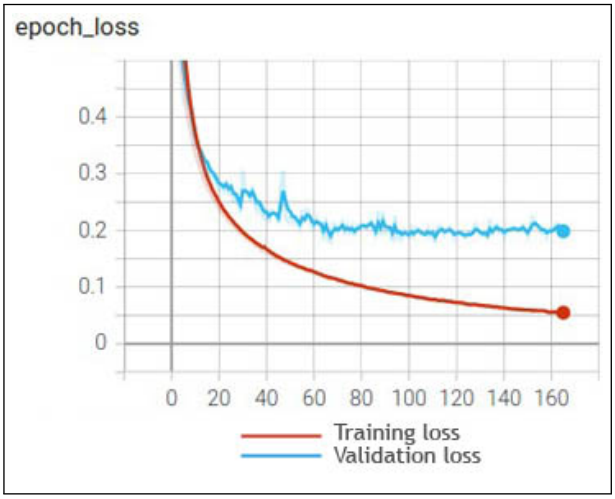

Figure 5: Loss graph of LSTM model

The CNN model consist of 128 filters, kernel size 3 and stride =1 is trained on large dataset with 200 embedding dimension. It is regularized with l1 and l2 regularizer with 0.01 value. Table 6 illustrates the classification report of CNN model. Employment has highest precision 98.97% of the 14 categories, followed by Sports, Tourism and Automobiles with 98.70%, 98.65% and 98.35% precision respectively. Society has the lowest precision 81.19%.

Table 6: Classification Report of CNN Model

Categories	Precision	Recall	F1-Score	Support
Automobiles	0.9835	0.9995	0.9914	1971
Diaspora	0.8714	0.9023	0.8866	1832
Economy	0.8739	0.9040	0.8887	3511
Employment	0.9897	0.9989	0.9943	1819
Entertainment	0.9575	0.9395	0.9484	3620
Health	0.8910	0.9481	0.9187	1811
International	0.9625	0.9469	0.9546	4067
National	0.9558	0.9398	0.9477	1793
Opinion	0.9206	0.9601	0.9400	1932
Politics	0.9400	0.8829	0.9106	3015
Society	0.8119	0.7551	0.7825	1772
Sports	0.9870	0.9790	0.9830	4190
Technology	0.9727	1.0000	0.9862	1781
Tourism	0.9865	0.9974	0.9919	1905
Avg/Total	0.9360	0.9395	0.9375	35019

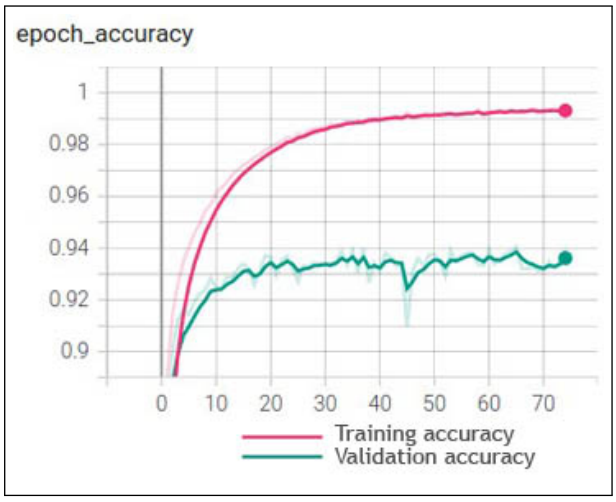

Figure 6: Accuracy graph of CNN model

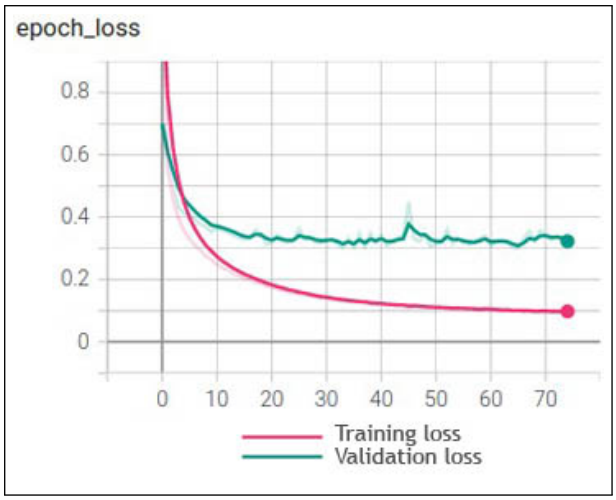

Figure 7: Loss graph of CNN model

Table 7: Classification Report of DNN Model

Categories	Precision	Recall	F1-Score	Support
Automobiles	0.9494	0.9995	0.9738	1971
Diaspora	0.8718	0.8614	0.8666	1832
Economy	0.8717	0.8089	0.8391	3511
Employment	0.9802	0.9549	0.9674	1819
Entertainment	0.9555	0.9066	0.9304	3620
Health	0.9102	0.8901	0.9001	1811
International	0.9510	0.9063	0.9281	4067
National	0.9518	0.9247	0.9380	1793
Opinion	0.9028	0.9136	0.9082	1932
Politics	0.8216	0.8949	0.8566	3015
Society	0.6256	0.6913	0.6568	1772
Sports	0.9636	0.9735	0.9685	4190
Technology	0.9766	0.9843	0.9804	1781
Tourism	0.9236	0.9963	0.9586	1905
Avg/Total	0.9040	0.9076	0.9052	35019

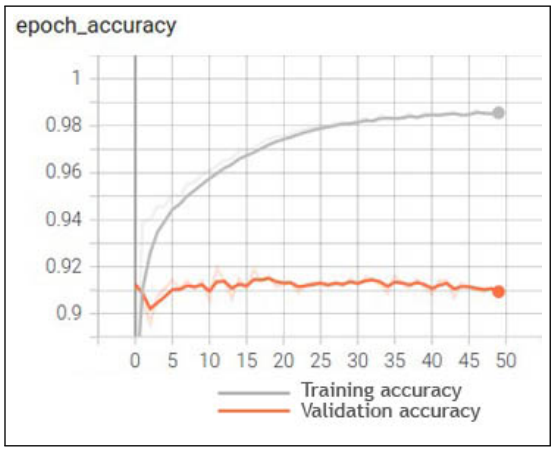

Figure 8: Accuracy graph of DNN model

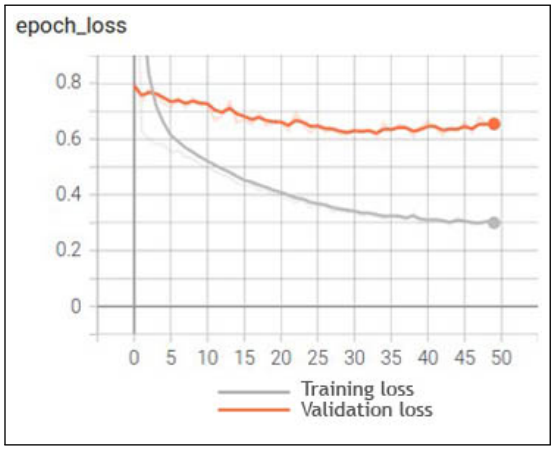

Figure 9: Loss graph of DNN model

Dense neural network model evaluation results are illustrated in the form of Classification Report and accuracy/loss curve. According to the table 7, Employment class is found to be with highest precision 98.02%, Technology with 97.66% precision and Sports falling little behind with 96.36% precision. The precision of Society 62.56% is the lowest among all the other categories like in LSTM and CNN models.

As we compare the table 5, 6 and 7, society has the lowest accuracy, precision, recall and f1-score as compared to the 13 other categories. The articles in the Society contains the general context which includes sports organized in communities, crimes in society and even the political programs organized in local levels. Such context makes the article general. In such articles, little change of the keywords completely alters the category.

Finally, the models built are compared on the basis of evaluation matrices. Figure 10, demonstrates LSTM as a best model with 95.36% accuracy, 95.39% precision, 94.98% recall and 95.16% f1-score. CNN model lags a little behind with 93.97% accuracy, 93.60% precision, 93.95% recall, 93.75% f1-score. DNN model is behind LSTM and CNN with the accuracy of 90.75%, 90.40% precision, 90.76% recall and 90.52% f1-score.

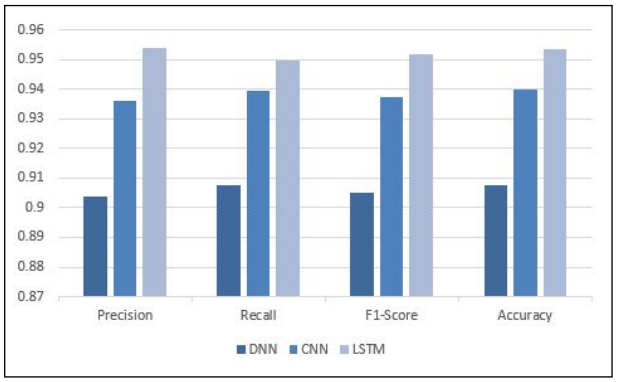

Figure 10: Summarized Evaluation Results

5. Conclusion

Vanilla LSTM, stacked LSTM, BiLSTM models along with CNN and DNN were experimented for this research. BiLSTM models showed efficient results in comparison to CNN and DNN. These models with 50, 100 and 200 features were experimented with 50, 100 and 200 epochs and tested eventually. Six common Nepalese websites have been used for data collection. Increase in news categories decreases the keyword's strength. Since, keywords are loosely related to multiple classes, it degrades the accuracy of the classification model. Experiment was focused on 14 different categories, each containing 6000 data.

Irregular distribution of data can lead to under and over training of different classes. To avoid this possibility, fourteen each categories were trained on equal amount of data. Later, available remaining data were added to different categories among 14 classes. Models were trained and tested with total 116736 data with imbalanced categories.

Accuracy, precision, recall and f1-score evaluation matrices were used to compare the performance. The LSTM model achieved the accuracy of 95.36% which is greater than the accuracy of CNN with 93.97% and DNN with 90.75%.

The LSTM model with Global Vectors for word embedding presented in this paper has outperformed the LSTM model presented by Basnet [2] with the accuracy of 84.63% using word2vec method for feature extraction.

6. Future Works

Deep learning requires huge amount of data for model training. This research is based on fourteen different news categories with total 116736 data. The uneven publication of news categories with uneven distribution of news articles are main reasons to limit model training with less than ten thousand data for each category. This work can be further extended by increasing the amount of data. More online news portals can be used for data scrapping in more categories.

Furthermore, performance evaluation of the GloVe for Nepali text can also be the important topic to research. This research is limited to single layered, stacked and bidirectional LSTM models, among them bidirectional model performed slightly better. In spite of that, the combined efforts of Convolutional Neural Network (CNN) and Long Short Term Memory (LSTM) can be experimented to achieve higher accuracy.

Acknowledgments

Authors would like to express sincere gratitude to Mr. Ashok Kumar Pant for his words of encouragement and genuine advices for the fulfillment of this research work.

References

[1] Tej Bahadur Shahi and Ashok Kumar Pant. Nepali news classification using naive bayes, support vector machines and neural networks. In *2018 International Conference on Communication information and Computing Technology (ICCICT)*, pages 1–5. IEEE, 2018.

[2] Ashok Basnet and Arun K Timalsina. Improving nepali news recommendation using classification based on lstm recurrent neural networks. In *2018 IEEE 3rd International Conference on Computing, Communication and Security (ICCCS)*, pages 138–142. IEEE, 2018.

[3] Tej Bahadur Shahi, Abhimanu Yadav, et al. Mobile sms spam filtering for nepali text using naïve bayesian and support vector machine. *International Journal of Intelligence Science*, 4(1):24–28, 2014.

[4] Zaiying Wang and Bohao Song. Research on hot news classification algorithm based on deep learning. In *2019 IEEE 3rd Information Technology, Networking, Electronic and Automation Control Conference (ITNEC)*, pages 2376–2380. IEEE, 2019.

[5] Kaushal Kafle, Diwas Sharma, Aayush Subedi, and Arun Kr Timalsina. Improving nepali document classification by neural network. In *Proceedings of IOE Graduate Conference*, pages 317–322, 2016.

[6] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[7] Mr Nihar M Ranjan, YR Ghorpade, GR Kanthale, AR Ghorpade, and AS Dubey. Document classification using lstm neural network. *Journal of Data Mining and Management*, 2(2):1–9, 2017.

[8] Taeho Jo. Categorization of news articles using neural text categorizer. In *2009 IEEE International Conference on Fuzzy Systems*, pages 19–22. IEEE, 2009.

[9] Dashen Xue and Fengxin Li. Research of text categorization model based on random forests. In *2015 IEEE International Conference on Computational Intelligence & Communication Technology*, pages 173–176. IEEE, 2015.

[10] Ziqiang Wang and Xu Qian. Text categorization based on lda and svm. In *2008 International Conference on Computer Science and Software Engineering*, volume 1, pages 674–677. IEEE, 2008.