Journal of Science and Technology
ISSN: 2362-1443 (Print) 3059-9490 (Online) 2026, Vol. 5, No. 1

Improving Nepali News Recommendation using Classification based on LSTM Recurrent Neural Networks

Ashok Basnet
Nepal College of Information Technology
Pokhara University, Nepal
mail@ashokbasnet.com.np

Arun K. Timalsina*
Pulchowk Campus, Institute of Engineering
Tribhuwan University, Nepal
t.arun@ioe.edu.np

*Corresponding author

***Abstract*— News classification is the process of grouping news documents into some predefined categories. Due to the increasing volume of the Nepali news content being generated everyday by thousands of online news portals, appropriate classification of these news items has become a necessity for the news readers. This paper was targeted to improve the Nepali news classification based on Recurrent Neural Networks, that uses deep layers of neural networks to classify the news to appropriate category. In this research paper, five popular news portals website across eight different categories was used for the purpose of data gathering and their classification accuracies was compared among these websites as well as overall accuracy was measured. The model was compared with the Support Vector Machine based on the parameters Accuracy, Precision, Recall and F1 Score. The use of Long Short Term Memory Recurrent Neural Network has improved the precision with the use of word2vec model. The presented model in the thesis have achieved a good accuracy of 84.63% and precision of 89% in compared to the SVM where the accuracy was 81.41% and precision 85% Among five news websites compared onlinekhabar.com was found to have good classification of the news where as ratopati.com was the least. Based on the categories of the news, sports news was classified more accurately by the model and economy was least accurately classified.**

Keywords*— *text classification, neural networks, recurrent neural networks, long short term memory

I. Introduction

The news content in Nepal is growing in a rapid pace. With so many news portals being started every year, there has been a massive growth in the content being produced. This generation of content is not being served right to the people who are in search of a particular category of news that they are interested in. The overcrowded social media content gives people with less choice and more of the sponsored content.

Traditional machine learning algorithms and neural network did not consider the understanding of the previous content. The persistence of the information allows the neural networks to get the best performance. The classification of the online articles has been a complex problem and there are numerous researches done in this field. In recent years, the deep learning frameworks are showing more potential than the machine learning algorithms that were developed earlier. The outcome of the deep learning networks is showing the signs of the great improvement over those algorithms.

A large number of news content is being generated and consumed via different media such as Facebook, twitter and other social media. Plenty of Nepali news are being produced in daily basis from popular news sites, so users cannot go through all the articles and miss their interested category news. Furthermore, many news portals are using manual way of recommending content to user only based on that particular news they are reading. The classification of the news document has become an important concern for growing online Nepali users, in order to stay updated on the concerned news category. The thesis collects large volume of available Nepali news documents and uses the deep learning mechanism to classify these news documents into 8 different categories from five most popular portals of Nepal. This thesis was focused on improving the performance and accuracy of news document classification based on the Long Short Term Memory recurrent neural networks.

II. Related Works

There have been a few works in Nepali news classification. However, there are good works done for English articles using Recurrent Neural Networks and they have achieved good performance using Long Short Term Memory networks. The research for the Nepali news classification is done for the conventional model of Machine Learning which focus on the SVM and Naïve Bayes mainly.

N. M. Ranjan et al. [1] demonstrated outstanding performance of Long Short Term Memory in document classification in comparison to other machine learning algorithms. The 20 newsgroups data set was used for the experiment with around 20,000 documents. They used TFIDF for feature extraction and used maximum features as 5000 for the documents. K. Kafle et al. [2] successfully tested the use of neural networks in the word2vec for word embedding to improve the vector representation of the text. The use of word2vec model outperforms the TF-IDF method by 1.6 percent. The classification is carried out with Support Vector Machine. They classified news articles into 12 different categories and used SVM to train the model. The machine learning model was able to classify the categories with the use of word2vec.

S. Kaur and N. K. Khiva [3] presented “Online news classification using Deep Learning Technique”, where they classified four different categories of news articles using neural network classifier, where they achieved up to 81% precision. C. Zhou et al. [4] worked on “A C-LSTM Neural Network for Text Classification” where they used unified

DOI : https://doi.org/10.3126/jost.v5i1.74563
Article History: Received: 12 Nov. 2025, Revised: 11 Jan. 2026, Accepted: 26 Feb. 2026.

model of Convolutional Neural Network and Recurrent Neural Network called as C-LSTM. C- LSTM was able to learn phrase-level features through a convolutional layer; sequences of such higher- level representations are then fed into the LSTM to learn long-term dependencies. They evaluated the learned semantic sentence representations on sentiment classification and question type classification tasks with very satisfactory results. F. Al. Zaghoul and S. Al. Dhaheri [5] worked on "Arabic Text Classification Based on Features Reduction Using Artificial Neural Networks" where the experiments on an Arabic corpus have demonstrated that the ANN model is effective in representing and classifying Arabic documents.

C. Chan et al. [6], on their research paper entitled "Automated Online News Classification with Personalization", designed and implemented a preliminary version of news classification system based on the SVM classification method. The system was capable of both general classification and personalized classification.

III. System Overview

This research was focused on the quantitative research. The data was collected from five different online news portals of Nepal under eight different categories. The data collected from the website are not in the right format, so the data cleaning process was done which removed the HTML tags, extra whitespaces. The data was selected for the training purpose. The selected data is processed through stop word removal process which looked up at NLTK standard data source for Nepali stop words collection.

Once the data was cleaned and selected, the feature extraction is applied using word2vec. The word2vec model created a vector matrix for the input word sequence. Then the data was split into test and train sets, where training set was further split into training set and validation sets. The model was developed using LSTM and trained using training set and validated against validation set. The final model was valuated against the test set. Fig1 describes the methodology flow of the research work.

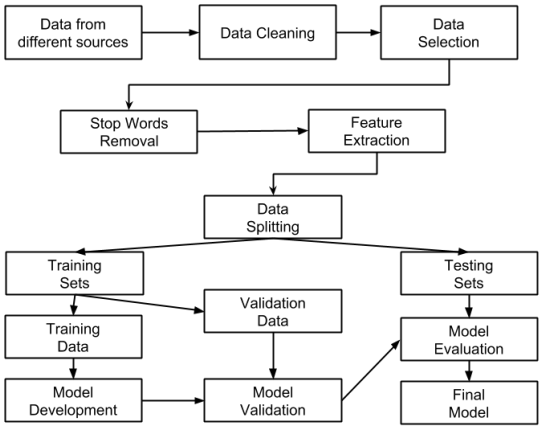

Fig. 1. Research Methodology Flow Diagram

A. Data Collection

Data from five popular sites were collected using python scraping tool, scrapy. The data were collected from around last 3 years till April 2018. Following sites were used for data collection.

TABLE I: ONLINE SOURCES FOR NEWS GATHERING

S.No.	Site Name	Site URL	No. of articles
1	DC Nepal	Dcnepal.com	12,872
2	Image Khabar	imagekhabar.com	19,296
3	Online Khabar	onlinekhabar.com	20,504
4	Ratopati	ratopati.com	7,855
5	Ujyaalo Online	ujyaaloonline.com	11,476
	Total		**72,003**

Different categories such as diaspora, economy, entertainment, health, international, opinion, politics, society, sports, technology. Since the combined dataset from all the websites didn't distribute the data over all categories fairly, the research was limited to eight categories only. The news gathered from different sources as mentioned above was categorized under eight different categories as below: -

TABLE II: CLASSIFICATION CATEGORIES

S.No.	Category Name	No. of Articles
1	Diaspora	6,224
2	Economy	14,067
3	Entertainment	7,588
4	Health	3,122
5	International	9,879
6	Opinion	2,675
7	Politics	8,352
8	Sports	13,018
	Total	**64,925**

B. Data Preprocessing

The first step in Data Preprocessing is the HTML tag removal from the news document. As news data contains different html tags such as div, paragraphs, headings, links, images etc., these tags need to be removed before the other preprocessing steps can be taken into consideration. The white space and other special symbol was removed so that the text document is clean from special symbol. As white space has no meaning in the text representation, the removal of the white space provides us with the clean data out of the news document text. The next step is the stemming which was used to convert the words into their base form, so that the word having different forms was taken as different words but was to be treated as one. Finally, news document contained certain words that won't have any impact on the news representation. These words, known as stop words was taken out from standard stop word vocabulary for Nepali language.

C. Model Building

For model building, Recurrent neural network was used to train the model. As per the training sets generated in the data selection phase, recurrent neural with few hidden units was used and the prediction was done. The performance was evaluated and the number of hidden units was increased to test out the accuracy and other model evaluation criteria. Here the Recurrent Neural Network with long short term memory was used.

The LSTM model was build using the Word2Vec model with 100,200 and 300 features. LSTM model was used with 3 hidden layers where hidden layer 1 contained 25/50/150 hidden units variations, layer 2 contained 25/50/100 hidden units variations and third hidden layer was used as Dense layer. The model was tested for 5 and 10 epochs with dropout of 0.2. The softmax activation was used in the output layer and categorical cross entropy as loss function of the model evaluation.

D. Model Validation

Confusion Matrix was used for validation of the model with the precision, recall, accuracy and f measure. Table 3 shows the evaluation metrics.

TABLE III: TEXT CLASSIFICATION EVALUATION METRICS

Evaluation metric	Formula
Precision	TP / (TP + FP)
Recall	TP / (TP + FN)
Accuracy	(TP + TN)/N
F Measure	2*Recall*Precision/(Recall + Precision)

IV. Long Short Term Memory

RNN uses sequential information. In a traditional neural network, it is assumed that all inputs (and outputs) are independent of each other. But for many tasks that's a very bad idea. RNNs are called recurrent because they perform the same task for every element of a sequence, with the output being depended on the previous computations. Another way to think about RNNs is that they have a "memory" which captures information about what has been calculated so far. In theory RNNs can make use of information in arbitrarily long sequences, but in practice they are limited to looking back only a few steps (more on this later). Here is what a typical RNN looks like:

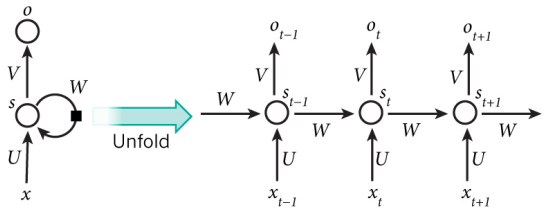

Fig. 2. A recurrent neural network and the unfolding in time of the computation involved in its forward computation

LSTM networks are a special kind of RNN, capable of learning long-term dependencies. The repeating module in a standard RNN contains a single layer.

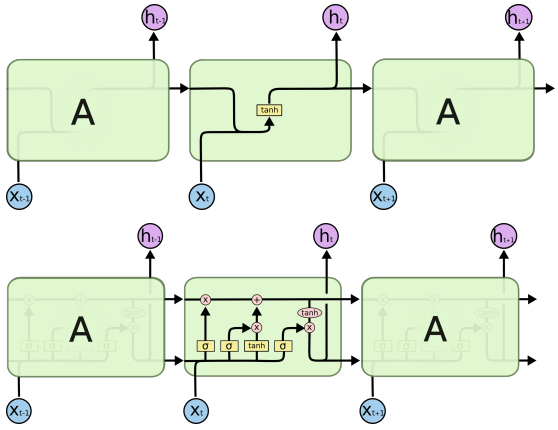

Fig. 3. The repeating module in an LSTM contains four interacting layers.

V. Experiment And Result

A. Experiment 1: LSTM Model Evaluation

Table 4 describes the experimental results for the classification of the overall data of five website using SVM model, where diaspora was found to be classified with 98% precision and economy news was classified with 66% precision.

TABLE IV:
OVERALL WEBSITE DATA CLASSIFICATION RESULT USING SVM

category	precision	recall	f1-score	support
diaspora	0.98	0.54	0.69	1309
economy	0.66	0.98	0.79	2774
entertainment	0.90	0.76	0.82	1543
health	0.98	0.25	0.40	618
international	0.83	0.91	0.87	1996
opinion	0.96	0.25	0.40	534
politics	0.82	0.85	0.84	1648
sports	0.93	0.96	0.95	2564
avg / total	**0.85**	**0.81**	**0.80**	**12986**

B. Experiment 2: SVM Model Evaluation

LSTM model was run for all the websites and following results were obtained in the form of classification report and accuracy / loss graphs. The results from the LSTM model with the 300 features and 10 epochs is listed below. The comparison of various hyper parameters was done in the next experiments.

TABLE V:
OVERALL WEBSITE DATA CLASSIFICATION RESULT USING LSTM

category	precision	recall	f1-score	support
diaspora	0.89	0.72	0.79	1896
economy	0.86	0.86	0.86	4194
entertainment	0.84	0.79	0.81	2208
health	0.82	0.75	0.78	909
international	0.92	0.85	0.88	3019
opinion	0.84	0.90	0.87	815
politics	0.88	0.85	0.86	2501
sports	0.96	0.93	0.95	3936
avg / total	**0.89**	**0.85**	**0.87**	**19478**

Table 5 describes the experimental results for the classification of the overall data of five website using LSTM model, where sports was found to be classified with 96% precision and health news was classified with 80% precision. Overall, good precision was found for the category wise classification by the model. The accuracy and loss graph were seen as below: -

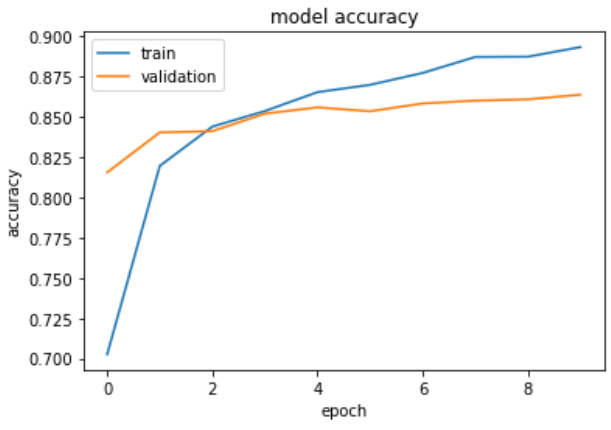

Fig. 4. Accuracy vs epoch graph for overall websites

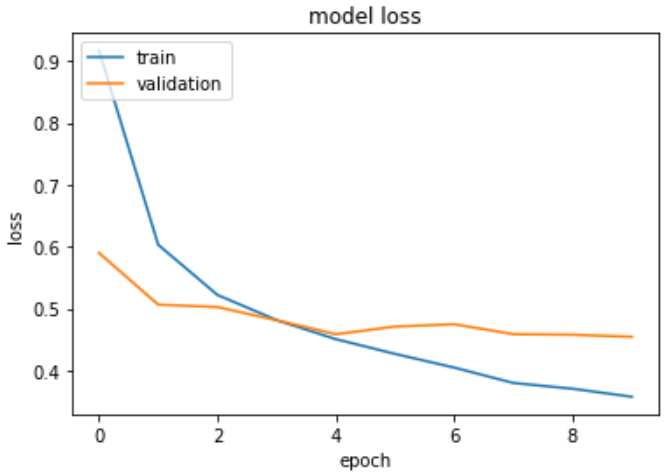

Fig. 5. Loss vs epoch graph for overall websites

C. Experiment 3: Comparison of models

1) Comparison of 5 websites using LSTM model

Table 6 describes the experimental results for the classification of the five different website data using LSTM model. Each website was compared against the four different evaluation metrics, accuracy, precision, recall and f1-score. It can be seen clearly that, onlinekhabar was found to be with highest accuracy and precision too, while the ratopati.com website was with the least accuracy as well as precision. The f1-score for imagekhabar and onlinekhabar was found to be similar with 90%.

TABLE VI: COMPARISION OF 5 WEBSITES USING LSTM MODEL

Website	Accuracy	Precision	Recall	f1-score
Dcnepal.com	81.45%	85%	81%	83%
Imagekhabar.com	88.83%	90%	89%	90%
Onlinekhabar.com	89.26%	91%	90%	90%
Ratopati.com	77.87%	83%	78%	80%
Ujyaaloonline.com	85.29%	89%	85%	87%

2) Comparison of category classification using LSTM

Fig 6 shows the comparison between eight different categories taken during the process of data collection. Among them, Sports category has the highest precision and accuracy. The other metrics also were compared such as recall, f1-score which shows the comparative analysis between the categories. The category diaspora has least precision among the categories.

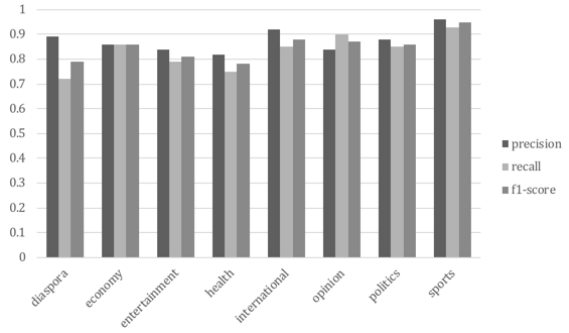

Fig. 6. Comparison of category classification using LSTM

3) Comparison of LSTM model with different hyper-parameters

Table 7 represents the experiment of the LSTM model with the variation of the hyper parameters of the model. The no. of features and no. of epochs was taken into consideration in the experiment. As the number of epoch grows, the classification accuracy was also found to increase whereas with the increase in the number of the features taken into consideration, the evaluation metrics were found to have increased.

TABLE VII:
COMPARISION OF LSTM MODEL WITH DIFFERENT HYPER-PARAMETERS

No. of features	No. of epochs	Accuracy	Precision	Recall	f1-score
100	5	70.43%	85%	70%	77%
100	10	73.22%	84%	73%	78%
200	5	79.19%	87%	79%	83%
200	10	80.95%	87%	81%	84%
300	5	83.05%	89%	83%	86%
300	10	84.63%	89%	85%	87%

4) Comparison of LSTM model based on number of hidden units for 1st and 2nd layer

The table 8 shows the comparison of LSTM models ran across different variations for the LSTM layer. Hidden layer 1 was varied from 25 to 300 units while hidden layer 2 was varied from 25 to 150 units. From the above experiment, we can see that there is good improvement when the hidden layers were increased. The accuracy was optimum for hidden layer 1 with 100 units and hidden layer 2 with 50 units which was found to be 85.6% and precision to be 88%. The precision was found to be better with lower number of hidden layers as seen in the table. The increase in hidden layers from 100 to 150 was not able to produce improvement over the accuracy which is due to the overfitting of the training data.

TABLE VIII:
COMPARISION OF LSTM MODEL BASED ON NO. OF HIDDEN LAYERS

Hidden Layer 1	Hidden Layer 2	Accuracy	Precision	Recall	f1-score
25	25	82.44%	89%	82%	85%
50	25	84.63%	89%	85%	87%
50	50	84.79%	89%	85%	87%
100	50	85.60%	88%	86%	87%
100	100	85.29%	88%	85%	87%
150	100	84.81%	87%	85%	86%
300	150	85.57%	88%	86%	86%

5) Comparison of SVM and LSTM models

The table 9 shows the comparison of the two models for the evaluation of the LSTM model over the SVM. The table shows that the LSTM model outperforms the SVM model by a good margin of 3% in accuracy and 4% in precision. F1-score for the LSTM model was way better in compared to SVM models for the same amount of data being fed to both the models.

TABLE IX: COMPARISION OF SVM AND LSTM MODELS

Model	Accuracy	Precision	Recall	f1-score
SVM	81.41%	85%	81%	80%
LSTM	84.63%	89%	85%	87%

VI. Conclusion

In conclusion, the use of Long Short Term Memory, Recurrent Neural Network showed expected improvement over the SVM where LSTM model was evaluated with word2vector implementation. LSTM model with 100 features and 300 features was experimented which was tested with 5 and 10 epochs. Data from five popular websites of Nepal were taken into consideration and their performance was compared with metrics viz. accuracy, precision, recall and f1-score. Onlinekhabar was found to have highest accuracy and precision with 89.26 % accuracy and 91% precision, whereas ratopati was found to be least effective with the categorization with 77.87% accuracy and 83% precision. The news was categorized into eight different categories where sports was found with 93% precision and economy category with 66% precision. The LSTM model was evaluated in overall data where the accuracy was found to be 84.63% compared to 81.41% accuracy of SVM and there was good improvement for precision of 89% in LSTM.

VII. Future Work

The collection of data was limited to eight categories and five websites, due to which the data needed for the deep learning module was limited to below 1 lakh. It can be improved by taking more websites into consideration and increasing the categories of the data will help to classify more news. The thesis uses rule-based stemming in the basic form, due to the unavailability of standard stemming library for Nepali language. The effective use of stemming can help to improve the overall accuracy of the classification. The thesis is limited to LSTM, where other technology like CNN can be used along with LSTM to achieve higher degree of accuracy and precision in the news categorization for Nepali news.

References

[1] N. M. Ranjan, Y. R .Ghorpade, G. R. Kanthale, A. R. Ghorpade and A. S. Dubey, "Document Classification using LSTM Neural Network", Journal of Data Mining and Management, Volume 2 Issue 2, 2017

[2] K. Kafle, D. Sharma, A. Subedi and A. K. Timalsina , "Improving Nepali Document Classification by Neural Network", IOE Graduate Conference, 2016, pp. 317–322m

[3] S. Kaur and N. K. Khiva, "Online news classification using Deep Learning Technique", International Research Journal of Engineering and Technology (IRJET), Volume: 03 Issue: 10, 2016

[4] C. Zhou, C. Sun, Z. Liu, and F. Lau., "A C-LSTM Neural Network for Text Classification", CoRR abs/1511.08630, 2015

[5] F. Al. Zaghoul and S. Al. Dhaheri "Arabic Text Classification Based on Features Reduction Using Artificial Neural Networks", UKSim 15th International Conference on Computer Modelling and Simulation, 2013

[6] C. Chan, A. Sun and E. Lim, "Automated Online News Classification with Personalization", Proceedings of the 4th International Conference of Asian Digital Library (ICADL2001), Pages 320-329, Bangalore, India, December, 2001