Nepali News Classification using Naïve Bayes, Support Vector Machines and Neural Networks

Tej Bahadur Shahi
Central Department of Computer Science and IT
Tribhuvan University, Kathmandu, Nepal
tejshahi@cdcsit.edu.np

Ashok Kumar Pant
Research Scholar
Central Department of Computer Science and IT
Tribhuvan University, Kathmandu, Nepal
asokpant@gmail.com

***Abstract*— Automated news classification is the task of categorizing news into some predefined category based on their content with the confidence learned from the training news dataset. This research evaluates some most widely used machine learning techniques, mainly Naive Bayes, SVM and Neural Networks, for automatic Nepali news classification problem. To experiment the system, a self-created Nepali News Corpus with *20* different categories and total *4964* documents, collected by crawling different online national news portals, is used. TF-IDF based features are extracted from the preprocessed documents to train and test the models. The average empirical results show that the SVM with RBF kernel is outperforming the other three algorithms with the classification accuracy of *74.65%*. Then follows the linear SVM with accuracy *74.62%*, Multilayer Perceptron Neural Networks with accuracy *72.99% and the* Naive Bayes with accuracy *68.31%*.**

Keywords—Natural Language Processing, Machine Learning, Nepali news classification, TF-IDF, Naive Bayes, Support Vector Machine, Neural Networks

I. INTRODUCTION

Online news portal and other media on the internet now produced the large amount of text, which is mostly unstructured in nature. When an individual wants to access or share particular news, it should be organized or classified in the proper class. Automatic classification of text is to assign a label or class to given text using a computer program. It is more important when a large number of texts come from the different source and needs to be reported into specific labels for further processing as it happens mostly in case of news. At present, as like in all other parts of the world, the most of the news now flashed out from the online media in Nepal. The news in print media becomes almost outdated or already known before it reached the hand of reader next day.

The online news portals classify their news into different categories such as "Political News", "Sports News", "Entertainment News" and so on. This task of manually labeling the news class becomes tedious when a large amount of news comes together from different sources. It is almost impossible to make this classification manually if some application tries to feed the trending news to the reader in real time. Hence it is necessary to develop an automatic tool that will be able to classify the Nepali news into relevant class. The technique developed to classify News text will be equally applicable to classify the other kind of Nepali text documents.

Since the Nepali language is morphologically rich and complex, the text classifier needs to consider the specific language features before classifying the text i.e. in training phase. In this work, the relevant features of Nepali text will be extracted using most popular feature representation technique TF-IDF. Three supervised machine learning algorithms, Naïve Bayes, Neural Network and Support Vector Machine with the same features are compared in terms of their accuracy.

II. RELATED WORK

Various techniques have been proposed to classify the text: Rule-based [1], Neural Network [2], Decision Trees [3], and Support Vector Machines [4]. Also, some tricks and deep learning based classifications can be found in[5, 6, 7]. The basic concept of these techniques is the classification of news type using the trained classifier that can automatically predict an incoming news type to some of the predefined classes.

In [4] the SVM is used to classify the Amharic text in hierarchical order with the data set collected from Ethiopian News agency. The flat classification was compared with hierarchical classification using SVM and claimed that when the number of features increased, the performance of SVM decreases in flat classification in comparison to hierarchical classification.

In [8], the medical text classification problem is addressed using the neural network. The convolutional neural network with medical data set consisting of several classes of health information is trained and tested with the accuracy of about 15% more than the existing approach in this field.

In the context of Nepali text, there is little literature available. The basic steps in natural language processing pipeline such as part of speech tagging [9, 10], steaming [11], named-entity recognition[12, 13] have been investigated separately. There is no result available collectively. Furthermore, the text classification problem for Nepali text is not even studied thoroughly. In [14], authors have proposed simple naïve Bayes classifier to address the problem. Naïve

978-1-5386-2051-9/18/$31.00 ©2018 IEEE

Bayes uses the concept of probability. The parameter in Naïve Bayes was learned from training the module with the Bayesian rule of probability. the representation of text document in the form of the bag of words where it is assumed that each word is independent of other, mainly degrade the performance of this approach. The simple Naïve Bayes was augmented with multinomial lexicon pooling [14]. Paper [15] applied some machine learning strategies for addressing the classification problem of the Nepali documents.

Nepali SMS classification task is discussed in [16] with Naïve Bayes and support vector machine approaches. Here the length of SMS is very limited in comparison to the length of the full text, the number of features such as SMS headings, words, and their frequency was taken to represent an SMS. The SVM and Naïve Bayes based classification technique were implemented to classify SMS into either spam or not spam. Due to the few number of the feature in consideration, the Naïve Bayes outperform the SVM with 5% (92% accuracy for NB and 87% accuracy for SVM)

III. METHODOLOGY

A. Dataset Preparations

The Nepali language belongs to one of the most common scripts, Devanagari, invented by Brahmins around the 11th century. It consists of 36 consonant symbols, 12 vowel symbols and 10 numeral symbols along with different modifiers and half forms. According to current census research, 17 million people worldwide speak the Nepali language. Nepali language character set is given in Table 1.

Table 1: Nepali character set

Numerals

०	१	२	३	४	५	६	७	८	९

Vowels

अ	आ	इ	ई	उ	ऊ	ए	ऐ	ओ	औ	अं	अः

Consonants

क	ख	ग	घ	ङ	च	छ	ज	झ	ञ	ट	ठ
ड	ढ	ण	त	थ	द	ध	न	प	फ	ब	भ
म	य	र	ल	व	श	ष	स	ह	क्ष	त्र	ज्ञ

A collection of Nepali news was collected from various online Nepali News portals using web crawler. The news portal namely ratopati.com, setopati.com, onlinekhabar.com, and ekantipur.com were used to gather text related to different news types. The distribution of news type in the Nepali news corpus is as shown in Table 2.

Table 2: Statistics of Nepali News Corpus

S.N.	News class	No. of documents
1	Agriculture	100
2	Automobile	95
3	Bank	417
4	Blog	209
5	Business	142
6	Economy	500
7	Education	85
8	Employment	154
9	Entertainment	500
10	Health	31
11	Interview	229
12	Literature	102
13	Migration	111
14	Opinion	500
15	Politics	500
16	Society	253
17	Sport	500
18	Technology	110
19	Tourism	214
20	World	212
	Total	**4,964**

B. Preprocessing

The text preprocessing cleans the text data to make it ready to use in training and testing of the machine learning model. Preprocessing is done to reduce the noise in the text that helps to improve the performance of the classifier and speed up the classification process, thus aiding in real time news classification. The main preprocessing techniques used are given below.

1. **Tokenization:** Breakdowns the text into sentences and then words. Vertical bar, question mark, and full stop are used to break down the sentences and while space and comma are used to break down the words.
2. **Special symbol and number removal:** Special symbols like !, :, ÷, ×, º, >, <, \, /, @, #, $, %, ^, &, *,), (, _, -, +, =, ~, ø, [,], ‘, ”, etc. and numbers, those

do not have much importance in classification, are removed.

3. **Stop word removal:** Stop words are high-frequency words that has not much influence in the text are removed to increase the performance of the classification. The list of 255 stop-words like "छ, म, हो, केही, हामी, मेरो, त्यो, हरु, फेरी, आफू, हुन्छ, राख, भयो, गर्नु, पनि, etc." were collected and removed from the text.
4. **Word Stemming:** Stemming is used to reduce the given word into its stem. Since the word stem reflects the meaning of a particular word, we have segmented the inflected word and derivational word into a stem word so that the dimension of vocabulary reduced in the significant manner[11]. Example:

नेपाल टेलिकमले दुर्गम हिमाली जिल्लाहरुमा सेवा उपलब्ध गराउनका लागि एनटीस्याट प्रविधिको प्रयोग गरिरहेको छ।

['नेपाल', ' टेलिकम', ' दुर्गम', 'हिमाल', 'जिल्ला', 'सेवा', 'उपलब्ध', 'गरा', 'लागि', 'एनटीस्याट', 'प्रविधि', 'प्रयोग', ' गर्', 'छ']

C. Feature Vector Construction

Feature Vector construction is the process of representing the news into a vector form. To represent Nepali news in vector form, the TF-IDF weighting value for each word in the text is taken as a dimensional value in a vector. It is calculated as,

$$W_{t,d,D} = tf_{t,d} * idf_{t,D}$$

Where,

$$tf_{t,d} = \frac{f_{t,d}}{max\{f_{t',d} : t' \in d\}}$$

$$idf_{t,D} = log\frac{N}{1+|\{d \in D : t \in d\}|}$$

tf = Term frequency.

idf = Inverse document frequency.

$f_{t,d}$ = Number of term t in document d.

$max\{f_{t',d} : t' \in d\}$ = Max occurring term t' in document d.

N = Total number of document in the corpus D.

$|\{d \in D : t \in d\}|$ = Number of documents where the term t appears.

D. Naïve Bayes Classifier

Naive Bayes Classifier is a simple probabilistic classifier based on Bayes Theorem with strong independence assumptions of feature space. Depending on the precise nature of the probability model, Naive Bayes classifier can be trained very efficiently in a supervised learning setting. In this research, we used multinomial Naive Bayes classifier with smoothing technique [17].

E. Support Vector Machine Classifier

Support Vector Machine [18] classifier is a hyperplane based discriminative classifier which finds a hyperplane to separate a multidimensional data into a two class. So these are basically binary classifier, but they can be extended into multi-class classifier using the techniques such as one-to-one and one-vs-rest. Linear and Radial Basis Function kernels are used for the decision making in the SVMs.

F. Neural Network Classifier

Neural network classifier is composed of a large number of highly interconnected processing elements working in unison to solve specific problems. In this research, Backpropagation Multilayer Perceptron[19, 20] with stochastic gradient descent optimization[21] is used for the classification learning and prediction.

IV. EXPERIMENTAL SETUP

The news classification learning and evaluation system pipeline is given in Figure 1. It consists of Preprocessing, Feature Extraction, Classification and Evaluation Phases. The complete news dataset that was used in the system training and evaluation was explicitly divided into training and testing sets. The five experiments were conducted to make the more accurate analysis of the outputs. The experimental parameters such as C and gamma (γ) for SVM and learning rate alpha (α) for the Neural network were determined for their optimal results/outputs. In each experiment, the different optimization parameters were analyzed to reach the optimal output.

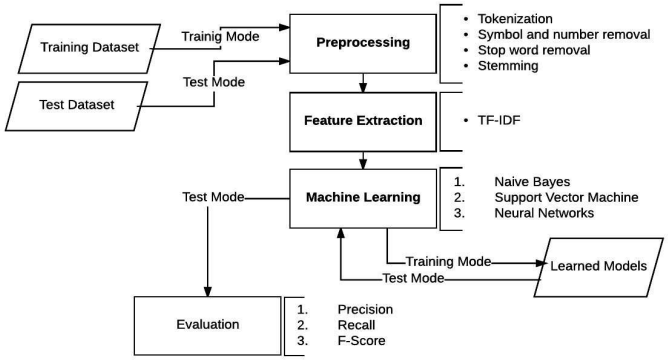

Figure 1: News classification system pipeline

V. EVALUATION AND RESULT

The experimental setup is done in five exclusive observations with variation in train and test splits. Experiments *1* to *5* (aka. Exp1 to Exp5) are respectively trained with *10%, 20%, 30%, 40%* and *50%* test splits of total training dataset described in Table 2. The number of features used in the experiments is *14332*, which is equal to the vocabulary size of the dataset.

The experimental result was analyzed for four evaluation parameters: Accuracy, Precision, Recall and F score [22].

A. Observations with Naive Bayes

Since the Naive Bayes learn the probabilities from the training corpus, the problem of zero probability of the individual word in a document should be considered. In this experiment, the optimal value of smoothing parameter alpha is determined using grid search. The best alpha value is found to be 0.03.

The result obtained from five experiments for Naïve Bayes is shown in Table 3.

Table 3: Experimental Result for Naïve Bayes

Exp.	Accuracy	Precision	Recall	F-Score
Exp1	68.2	69	68	67
Exp2	**68.98**	**70**	**69**	**68**
Exp3	68.78	69	69	68
Exp4	68.13	69	68	67
Exp5	67.45	69	67	66

B. Observations with Linear Kernel SVM

The value of regularization parameter C affects the accuracy of SVM. The optimal value of C should be determined before using SVM as a classifier. Here, the grid search is used to find the optimal value of C in each experiment automatically. For linear SVM, the best value of C is obtained as a 1.5.

The result obtained from five experiments for SVM with Linear kernel is described in Table 4.

Table 4: Experimental Result for SVM with Linear Kernel

Exp.	Accuracy	Precision	Recall	F-Score
Exp1	74.45	75	74	74
Exp2	74.62	76	75	74
Exp3	74.5	75	74	74
Exp4	**74.78**	**75**	**75**	**74**
Exp5	74.74	75	75	74

C. Observations with RBF Kernel SVM

The main optimization parameter in SVM classifier is regularization parameter C and kernel parameter gamma (γ) . The optimal values for the classifier parameters are selected using the grid search strategy. The optimal value of C is obtained as a 100 and of gamma as a 0.01.

Table 5: Experimental Result for SVM with RBF Kernel

Exp.	Accuracy	Precision	Recall	F-Score
Exp1	74.25	75	74	74
Exp2	**75.03**	**76**	**75**	**75**
Exp3	74.5	75	74	74
Exp4	74.58	75	75	74
Exp5	74.9	76	75	75

D. Observations with Neural Network

The result obtained from five experiments for Multilayer Perceptron (MLP) Neural Network is described in Table 5. The optimal parameters for the classifier are selected using grid search strategy and obtained as, learning rate = 0.0025, learning rate mode = adaptive, momentum = 0.9 and hidden layer neurons = 256.

Table 5: Experimental Result for MLP

Exp.	Accuracy	Precision	Recall	F-Score
Exp1	72.84	73	73	72
Exp2	**73.62**	**74**	**74**	**73**
Exp3	73.22	73	73	72
Exp4	72.76	73	73	72
Exp5	72.49	72	72	72

E. Summarized observations

The mean performance measure of three model taken from above five experiments are shown in Table 6 and corresponding visualization graph is shown in Figure 2.

Table 6: The mean performance measure for Four model

Measures	Naïve Bayes	SVM (Linear)	SVM (RBF)	MLP
Accuracy	68.31	74.62	**74.65**	72.99
Precision	69.2	75.2	**75.4**	73
Recall	68.2	74.6	**74.6**	73
F Score	67.2	74	**74.4**	72.2

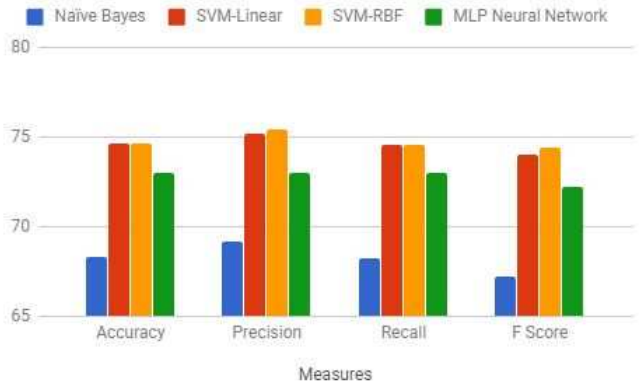

Figure 2: Summarized accuracy results

VI. CONCLUSION AND RECOMMENDATION

Automatic news classification has great importance in managing and targeting the crumbled news across different online sources. This paper introduced machine learning techniques for Nepali news classification task and presents a comprehensive comparison of the algorithms for classification accuracy. The classification pipeline includes the language-specific text preprocessing and feature extraction steps.

In the experimental setup, each algorithm has experimented with different optimization parameters and selected optimal from the grid search strategy. The experimental results show that the SVM best performs MLP neural networks and Naive Bayes. SVM with RBF kernel has the average accuracy of 74.65% and average F-score of 74.4% which is slightly higher than the SVM with a linear kernel. The accuracy then follows for MLP neural network with average F-score of 72.2% and then for Naive Bayes with average F-score of 67.2%.

In summary, it can be claimed that for the high dimension data such as the large volume of text, the support vector machine gives the best performance in comparison to Neural Network and Naive Bayes.

The accuracy obtained in this research is comparatively lower. It is because of the feature selection method-TF-IDF which assume the independence between words in the text corpus. However, in practice, the words in the corpus are context dependent- which violates the assumption of TF-IDF theory. Also, the classification accuracy is affected by the diversity of the classification data, noise in news documents, some preprocessing tasks like stemming and dataset size.

The future recommendations on the enhancement of this work are the extraction of semantic features from the text documents, finer stemmer, and increased experimentation dataset size. Also, deep learning based classification approaches like CNN and RNN(GRU, LSTM) can be applied to solve the Nepali news classification problem.

REFERENCES

[1] W. W. Cohen et al., "Learning rules that classify e-mail," in AAAI spring symposium on machine learning in information access, vol. 18. California, 1996, p. 25.

[2] I. Stuart, S.-H. Cha, and C. Tappert, "A neural network classifier for junk e-mail," Document Analysis Systems VI, pp. 442–450, 2004.

[3] X. Carreras and L. Marquez, "Boosting trees for anti-spam email filtering," arXiv preprint cs/0109015, 2001.

[4] A. K. Tegegnie, A. N. Tarekegn, and T. A. Alemu, "A comparative study of flat and hierarchical classification for amharic news text using svm," Culture, vol. 2007, p. 1, 2010.

[5] S. Lai, L. Xu, K. Liu, and J. Zhao, "Recurrent convolutional neural networks for text classification." in AAAI, vol. 333, 2015, pp. 2267-2273.

[6] A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov, "Bag of tricks for efficient text classification," arXiv preprint arXiv:1607.01759, 2016.

[7] K. Kowsari, D. E. Brown, M. Heidarysafa, K. J. Meimandi, M. S. Gerber, and L. E. Barnes, "Hdltex: Hierarchical deep learning for text classification," arXiv preprint arXiv:1709.08267, 2017.

[8] M. Hughes, I. Li, S. Kotoulas, and T. Suzumura, "Medical text classification using convolutional neural networks," arXiv preprint arXiv:1704.06841, 2017.

[9] T. B. Shahi, T. N. Dhamala, and B. Balami, "Support vector machines based part of speech tagging for nepali text," International Journal of Computer Applications, vol. 70, no. 24, 2013.

[10] A. Paul, B. S. Purkayastha, and S. Sarkar, "Hidden markov model based part of speech tagging for nepali language," in Advanced Computing and Communication (ISACC), 2015 International Symposium on. IEEE, 2015, pp. 149–156.

[11] B. K. Bal and P. Shrestha, "A morphological analyzer and a stemmer for Nepali," PAN Localization, Working Papers, vol. 2007, pp. 324–31, 2004.

[12] S. B. Bam and T. B. Shahi, "Named entity recognition for nepali text using support vector machines," Intelligent Information Management, vol. 6, no. 02, p. 21, 2014.

[13] A. Dey, A. Paul, and B. S. Purkayastha, "Named entity recognition for nepali language: A semi hybrid approach," International Journal of Engineering and Innovative Technology (IJEIT) Volume, vol. 3, pp. 21–25, 2014.

[14] S. Thakur and V. K. Singh, "A lexicon pool augmented naive bayes classifier for nepali text," in Contemporary Computing (IC3), 2014 Seventh International Conference on. IEEE, 2014, pp. 542–546.

[15] K. Kafle, D. Sharma, A. Subedi, and A. K. Timalsina, "Improving nepali document classification by neural network," in Proceedings of IOE Graduate Conference, 2016, pp. 317–322.

[16] T. B. Shahi and A. Yadav, "Mobile sms spam filtering for nepali text using naı¨ve bayesian and support vector machine," International Journal of Intelligence Science, vol. 4, no. 01, p. 24, 2013.

[17] C. D. Manning, P. Raghavan, and H. Schutze, "Introduction to information retrieval Cambridge University Press, 2008," Ch, vol. 20, pp. 405–416.

[18] C. Cortes and V. Vapnik, "Support-vector networks," Machine learning, vol. 20, no. 3, pp. 273–297, 1995.

[19] S. S. Haykin, Neural networks: a comprehensive foundation. Tsinghua University Press, 2001.

[20] D. E. Rumelhart, G. E. Hinton, R. J. Williams et al., "Learning representations by back-propagating errors," Cognitive modeling, vol. 5, no. 3, p. 1, 1988.

[21] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," arXiv preprint arXiv:1412.6980, 2014.

[22] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," Information Processing & Management, vol. 45, no. 4, pp. 427–437, 2009.