Automatic Bengali Document Categorization Based on Word Embedding and Statistical Learning Approaches

Conference Paper · February 2018

DOI: 10.1109/IC4ME2.2018.8465632

CITATIONS
3

READS
50

2 authors:

Rajib Hossain
Bangabandhu Sheikh Mujibur Rahman Science & Technology University
19 PUBLICATIONS **39** CITATIONS

SEE PROFILE

Moshiul Hoque
Chittagong University of Engineering & Technology
72 PUBLICATIONS **210** CITATIONS

SEE PROFILE

Automatic Bengali Document Categorization Based on Word Embedding and Statistical Learning Approaches

Md. Rajib Hossain
Dept. of Computer Science & Engineering
Chittagong University of Engineering & Technology
Chittagong, Bangladesh
e-mail: rajcsecuet@gmail.com

Mohammed Moshiul Hoque
Dept. of Computer Science & Engineering
Chittagong University of Engineering & Technology
Chittagong, Bangladesh
e-mail: moshiulh@yahoo.com

***Abstract*—The automated categorization of text documents into predetermined categories has witnessed a growing in the last few years, due to the huge availability of documents in digital form and the ensuing need to organize them. Automatic document categorization is the process of assigning one or more categories or classes to a document, making it easier to manipulate and sort. This paper proposes a Bengali document categorization technique based on word2vec word embedding model and stochastic gradient descent (SGD) statistical learning algorithm with multi-class svm. The semantic features of a document are extracting by Word2Vec and SGD improve the classification complexity with multi-class SVM that classify the unlabeled data. The experimental result with 10000 training and 4651 testing documents shows the 93.33% accuracy.**

Keywords—Bangla language prcessing, Documents categorization; Word embedding; Machine learing.

I. INTRODUCTION

In recent years, automatic document categorization has gained much attention by NLP researchers due to the availability of texts in digital form. Document categorization is the task of assigning a text document or a sequence of text documents into a single or multiple predefined categories. A Number of text documents in digital form have grown enormously day by day in size and variety. Therefore, an automatic document categorization system should develop to handle a large amount of text data to organize or sort it easily and quickly. Bangla is spoken by about 245 million people in Bangladesh and two states of India, with being 7th most spoken language in the world [1]. With the popularity of Unicode system and growing use of the Internet, Bangla text documents in digital domain have increased since last few years. Although there are few researches are conducted in the field of Bangla language processing such as syntax analysis, machine translation, optical character recognition and so on, an automatic text document categorization also an importance issue that need to solve. Bangla document categorization may be used in security agency to identify the suspected web contents or spam detection, the daily newspapers to organize by subject categories, library to classify papers or books medical to categorize patient reports from multiple aspects, using taxonomies of disease categories, and so on. There are many document categorization system developed for English language processing but there is no usable system is developed for Bangla texts.

In this work, we purpose is to design a framework to classify Bengali text documents using the word embedeing, SGD, and multi-class SVM. Convolutional neural network [2], Character-level [3] and recurrent neural network [4, 5] have achieved very good result in document classification but it required costly hardware and large dataset for training. Our propose system uses word embedding technique and this embedding text will use for categorization. Word embedding is a process of feature extraction, where each word extracts some feature depending on semantic and syntactic relations. Semantic vector space models of language represent each word with a real-valued vector. There are many famous statistical algorithms for word embedding but GloVe [6] and Word2Vec [7] are state-of-the-art or competitive algorithm. We tuned the hyperparameters of both algorithm and train a big amount of data for Bangla word embedding. Each word in the sentence is projected into embedding vector space by being multiplied with a weight matrix, forming a sequence of dense real-valued vectors. This sequence is then fed into the SGD which processes the word sequence that in turn SVM classify the text.

II. RELATED WORK

A number of significant researches has conducted on word embedding and document categorization in English language. In recent year, word embedding shown high performance for English and some European language text classification [6, 7]. However, no significant embedding technique is developed for classify Bangla texts. Very few attempts of word embedding techniques are found in Bangla language such as, TF-IDF [1], N-GRAM [8] and lexical [9] approaches. TF-IDF based word embedding technique used only word-count and it shown a low-performance due to lack of feature extraction capability. N-GRAM word embedding is a statistical process in which semantic relation depend on previous N words, as a result, the

current word embedding may be diverted to low performance. N-GRAM model is not represents semantic meaning of the whole sentence. Moreover, the lexical feature is not working properly for Bangla language due to its large inflectional diversity in verbs, tense, noun, etc. In a recent work Word2Vec word embedding technique is used for text classification [10], but it accuracy is not good due to the lack of Bangla corpus and hardware support.

Krendzelak et al. describe a text categorization system with machine learning and hierarchical structures which used a tree-based Naive Bayesian categorization process [11, 12]. It is a conventional machine learning system which performs low accuracy due to training feature extraction process and training techniques. An unsupervised technique with latent semantic feature and Gaussian mixture model is used for text categorization [13]. Most of the text documents contain a huge number of the sentence and the category name is the just summary of the document, for this reason, it is really hard to find the relation between text category name and document contents. Text categorization on Turkish language using SVM is proposed which is achieved good accuracy but time complexity is large due to the large feature dimensions [14]. A system for Arabic text categorization is developed using Naive Bays in control environment dataset with a resonable accuracy but it falls due to the unknown data set [15].

In the recent year few researches were conducted using machine learning techniques. Clustering based approach [10, 8] achieved the better result but there are lots of problem with a clustering-based solution. In cluster-based technique, the accuracy depends on a number of the clusters but there are no work is conducted to determine the optimize clusters. Data outlier is another problem of the cluster-based solution, a cluster center may huge change due to outliers, and as a result, the final train model should be overfitting. Bangla web documents categorization based on multiple supervised and unsupervised algorithm apply [1] in here but word embedding based on TI-IDF and classifier take more time for using multiple classifier algorithms. For this reason, its performance and accuracy is very low and cannot use in real time. In our work, we propose a word embedding with spec2vec with SGD based system for Bangla text categorization which expects to overcome the shortcomings of previous work.

III. METHODOLOGY

We proposed a documents categorization system which trains by Bengali text documents with a supervised algorithm and prepares a classifier model which projected by unladed documents and provide an expected category name. In this training module, we have a newspaper dataset which collects from a different newspaper. Let the training sets $X = \{x_1, x_2, x_3, .., x_n\}$ and its class $C = \{y_1, y_2, y_3, .., y_n\}$. Where n is the total number of training documents and c the total number of category or class. This module input as a text documents X with label C and output as a sequence of words list. A schematic representation of propose text documents classifier is shown in Fig. 1.

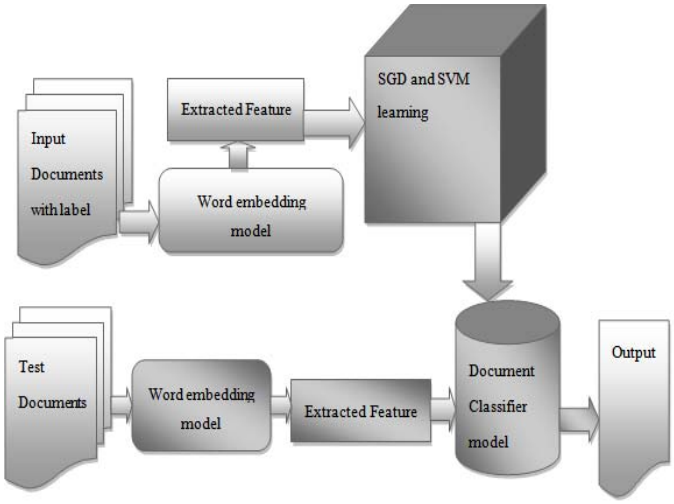

Fig. 1. Documents classifier training and projection module.

Let x_1 = সেডন পার্ক, হ্যামিল্টন। মাহমুদউল্লাহর জন্য মাঠটা বুঝি ভীষণ পয়া! টেস্টে একমাত্র সেঞ্চুরি এসেছিল এ মাঠেই।

Now, the document x_1 is projected with the classifier model and expected output will $y_1 = sports$.

A. *Word embedding model*

The main goal of the word embedding is to convert words into numeric value to manipulate an understanding of natural language. In this module, input takes as a one-hot vector for each word and output an embedding feature vector for that word.Word2Vec is a class of algorithm that learns in an unsupervised way to representations of the word vectors which captures semantic relations well. Now we will build a dense vector for each word so that it is easy to predict the other words appearing in the context. The output of this module is a dense vector ($Ww \bullet f$).Fig 2 shows the shallow 2-layer neural network word embedding module.

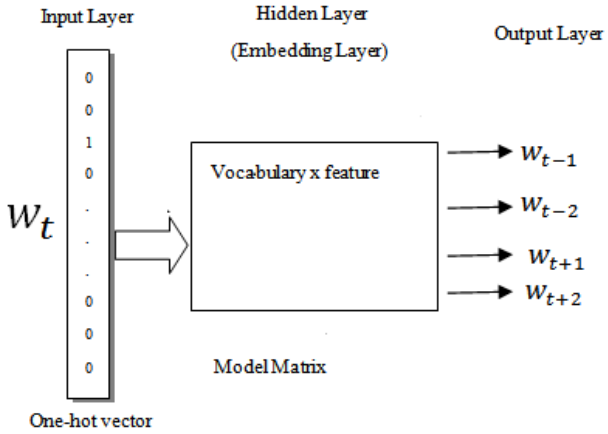

Fig. 2. Skip-gram model for the 2-layer shallow network.

We define a model that aims to predict between a centers words w_t and context words in terms of word vectors. We look at many positions t in a big Bengali language corpus. We keep adjusting the vector representations of words to minimize this loss. We have a large Bengali corpus and each word in the corpus $t = 1,...,T$ is predicted surrounding words in a window of "radius" m of every word.

Objective function $J'(\theta)$: The objective function Maximize the probability of any context word given the current center word.

$$J'(\theta) = \prod_{t=1}^{T} \prod_{-m \le j \le m;\, j \ne 0} p\left(w_{t+j} \mid w_t; \theta\right) \quad (1)$$

Where w_t the center word and w_{j+t} is the context word. The Log Likelihood objective function also optimizes the loss function and maximizes the context word probability.

$$J(\theta) = -\frac{1}{T} \sum_{t=1}^{T} \sum_{-m \le j \le m;\, j \ne 0} \log p\left(w_{t+j} \mid w_t\right) \quad (2)$$

$J(\theta)$ is the Log likelihood objective function which maximizes the context words probability. Now predict the surrounding or context words in a window of radius m of every word for $p(w_{t+j} \mid w_t)$ the simplest first formulation is:

$$p(o \mid c) = \frac{\exp\left(u_o^T \quad v_c\right)}{\sum_{w=1}^{v} \exp\left(u_w^T \quad v_c\right)} \quad (3)$$

$p(o \mid c)$ is the context word probability respect to the center word. Where o is the outside (or output) word index c is the center word index v_c and u_w center and outside vectors of indices c and o. Finally, using Softmax word c to obtain the probability of word o. The *Softmax function* is defined as in eq. (4).

$$p_i = \frac{e^{u_i}}{\sum_{j=1}^{c} e^{u_j}} \quad (4)$$

Where p_i is the i^{th} class soft probability numerator that denotes the actual probability and denominator denotes the score normalization. The probability maximization, Log likelihood, context word probability measurements and Softmax function are combined to design a shallow 2-layer neural network. The network input layer feed as one hot vector which projects into the hidden layer and output layer contains the semantic feature corresponding to the input word vector. Input feature function takes an embedding vector per-word and output as concatenate feature vector.

B. Extracted Feature

Feature extraction is a process of domain transformation. In documents categorization system input domain takes as raw ssstext documents and features extractor system process the raw text and output as a numeric value for each word with a fixed dimension. For each word lookup the embedding model $Ww \bullet f$ and concatenate the feature vector.

C. SGD and SVM Learning

The main objective of the SGD and SVM learning is to propagate the feature vector and collect the distinguishing feature for classifier model. The Multi-class SVM converts to Binary Classification problem using one vs. all (OVA) technique. Due to the large-scale dataset, we used stochastic gradient descent (SGD) for training that reduces the training time with gradient optimization technique.

Let the update parameters θ_j update with batch size N and each time update the parameters by eq. (5).

$$\theta_j = \theta_j - \alpha\left(h_\theta\left(x^i\right) - y^i\right) x^i_j \quad (5)$$

Where θ_j the updated gradient, α is the learning rate, x^i is the input data, y^i is the i[th] class label and h_θ is the hyperparameter. Now the SVM functions each time partition the input data and try to minimize the objective or error function.

$$L(y, \hat{y}) = -(y)\log(\hat{y}) + (1-y)\log(1-\hat{y}) \quad (6)$$

Here $L(y, \hat{y})$ is a loss function or objective function which reduces the error of SVM function, y is an actual label and $\hat{y}$ is predicted label.

D. Documents classifier model

Now the training algorithm generates a model which represent by $\theta_{k \bullet f}$ here k denotes the class number and f represent the trained weight feature dimension.Let x^i is the unlabeled feature and b is a bias term and projected by model matrix $\theta_{k \bullet f}$. The hyperparameter is given below:

$$h_\theta\left(x^i\right) = \theta_{k \bullet f} * \quad x^i + b \quad (7)$$

Now we got a score vector from equation (7) $\langle s_1, s_2, \cdots\cdots s_9 \rangle$ and get the maximum score obtained by the eq. (8). From equation (8) we determine the expected documents category class.

$$\max\left(h_\theta\left(x^i\right)\right) \quad (8)$$

IV. EXPERIMENTS

The whole system executes in GTX 1070 GPU with 32 GB physical memory and core i7 processor. We collected a number of Bengali documents for word embedding and documents classification from web, blogs, newspapers, online books. Table I summarize the statistics of data used for word embedding.

TABLE I. WORD EMBEDDING DATA SUMMARY

Number of documents	84000
Number of sentences	102096
Total unique words	850400
Word embedding dim	100

In order to categorize the text documents, we collected a handcraft dataset from different online newspaper [16-19]. Table II shows the summary of dataset used for classification.

TABLE II.HANDCRAFT CLASSIFIER DATASET SUMMARY

	Training	Testing
Number of class	9	9
Number of documents	10000	4651
Average word per documents	60	60
Feature-length per document	600	600
Padding with documents	Allowed	Allowed

In the Table II, the feature length of each document is considered as a floating point value that depends on file size and padding. We develop a documents categorization dataset with 10000 traing and 4651 testing documents. Table III illustrates the summary of data used for training and testing in different document categories. TABLE III shows that crime category consists of large number of documents and environment category consists of smallest number of documents respectively. All data are stored in .txt format.

TABLE III. NUMBER OF CATEGORIES USED FOR CLASSIFICATION

Category Name	Number of training documents	Number of testing documents
Accident(A)	996	492
Crime(C)	2120	1089
Economics(EC)	850	345
Entertainment(EN)	1400	686
Environment(ENV)	355	40
International(I)	900	412
Politics(P)	659	274
Science_tech(ST)	1150	513
Sports(SP)	1570	800
Total	10000	4651

V. EVALUATION MEASURES

In order to evaluate the propose system, we used several evaluation metrics such as, precision, recall, F_1-measure, and confusion matrix.

- ***Precision:*** In the field of documents categorization, precision is the fraction of retrieved documents that are relevant to the query.

$$precision = \frac{\left|Rd \cap Re\right|}{\left|Re\right|} \tag{9}$$

- ***Recall:*** In the field of documents categorization, recall is the fraction of the relevant documents that are successfully retrieved.

$$recall = \frac{\left|Rd \cap Re\right|}{\left|Rd\right|} \tag{10}$$

Here Rd and Re are the relevant and retrieved documents.

- ***F_1 – measure***: A measure that combines precision and recall is the harmonic mean of precision and recall, the traditional F_1 measure.

$$F_1 = \frac{2 * precision * recall}{precision + recall} \tag{11}$$

- ***Confusion Matrix:*** In the field of machine learning and specifically the problem of statistical classification, a confusion matrix, also known as an error matrix. Each row of the matrix represents the instances in a predicted class while each column represents the instances in an actual class

VI. RESULTS

TABLE IV shows the precision, recall, F_1-measure and support values in different class.

TABLE IV. STATISTICALLY EVALUTION SUMMARY

Category Name	Precision	Recall	F_1-score	Support
Accident (A)	0.91	0.93	0.92	492
Crime (C)	0.91	0.95	0.93	1089
Economics (EC)	0.90	0.87	0.88	345
Entertainment (EN)	0.96	0.98	0.97	686
Environment (ENV)	0.96	0.65	0.78	40
International (I)	0.90	0.84	0.87	412
Politics (P)	0.96	0.89	0.93	274
Science_tech (ST)	0.91	0.95	0.93	513
Sports (Sp)	0.99	0.96	0.97	800
Avg/total	0.93	0.93	0.93	4651

Fig. 3 shows the precision vs. recall curve which revealed that the documents categorization system achieved the better AUC result (97.00%).

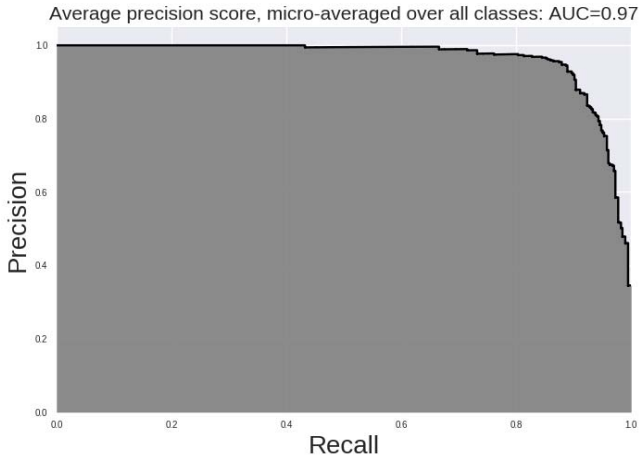

Fig. 3. The precision-recall curve for documents categorization.

TABLE V shows confusion matrix for documents categorization with error. Only A (accident) class overlap with C (crime) class. The sports class achieved the best accuracy and the environment class achieved the minimum accuracy. Due to number of training data and data pattern variations are the main reason of the accuracy discriminations.

TABLE V.CONFUSION MATRIX

	A	C	EC	EN	ENV	I	P	ST	SP
A	457	35	0	0	0	0	0	0	0
C	31	1037	8	1	0	9	1	1	1
EC	1	6	300	4	0	4	2	28	0
EN	1	6	0	673	0	1	3	1	1
ENV	1	5	1	0	26	7	0	0	0
I	2	19	6	15	1	346	3	16	4
P	3	16	6	0	0	3	346	3	16
ST	2	1	9	2	0	7	0	487	5
SP	2	9	3	5	0	9	1	2	769

Fig. 4 represents the ROC curve for different classes. In Fig. 4 the true positive rate and the false positive rate for different cut-off points of a parameter. Each point on the ROC curve represents a sensitivity or specificity pair corresponding to a particular decision threshold. The area under the ROC curve (AUC) is a measure of how well a parameter can distinguish amongh the classes.

We compared accuracy of the propose system with the existing Bangla text classification techniques [1, 11]. Where accuracy means the average accuracy .Table VI summarizes the comparison which shows the propose system is outperforms the existing techniques with higher accuracy 93.33%.

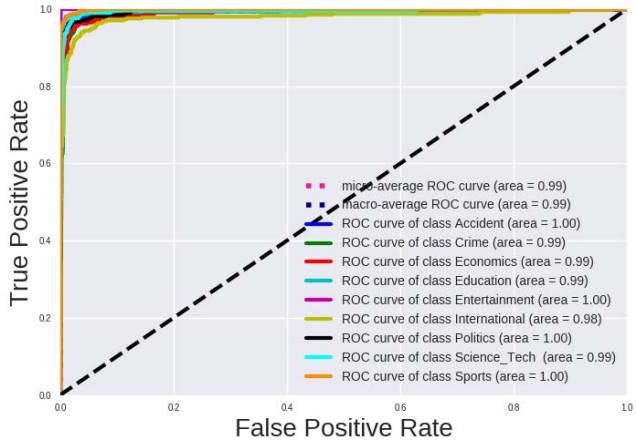

Fig. 4. ROC curve for multi-class categorization.

TABLE VI. PERFORMANCE COMPARISION

Method	No. of training documents	No. of testing documents	Total class	Accuracy (%)
TF-IDF+SVM [1]	1000	118	5	89.14
Word2Vec + K-NN + SVM [2]	19705	4713	7	91.02
Proposed	10000	4651	9	93.33

VII. CONCLUSION

Bangla text document classification is an important research issues in Bangla language processing. In recent year, due to the huge availability of Bangla texts in digital form we need to develop an automatic classification system to manage or organize these texts. In this paper, we propose Bangla text classification system using machine learning technique. Semantic feature of Bangla input texts is extracted using Word2Vec algorithm and developed a documents categorization system using multi class SVM with SGD. The proposed system is tested using author generated dataset and compare with the existing technique which shown the better performance. We will consider more classes with larger datasets in future that will improve the overall accuracy of the system.

References

[1] A. K. Mandal and Rikta Sen, "Supervised Learning Methods for Bangla Web Document Categorization," International Journal of Artificial Intelligence & Applications (IJAIA), vol. 5, no. 5, pp.93-105, 2014.

[2] K. Xu, Y. Feng, S. Huang and D. Zhao, " Semantic Relation Classification via Convolutional Neural Networks with Simple Negative Sampling, "Empirical Methods in Natural Language Processing, pp. 536–540, Lisbon, Portugal, 2015.

[3] X. Zhang, J. Zhao, and Y. LeCun, "Character-level Convolution Networks for Text Classification," Journal of CoRR, 2016.

[4] D. Tang, B. Qi, and Ting Liu, “Document Modeling with Gated Recurrent Neural Network for Sentiment Classification,” Empirical Methods in Natural Language Processing, pp.1422–1432, Lisbon, Portugal, 2015.
[5] J. Y. Lee and F. Dernoncourt, “Sequential Short-Text Classification with Recurrent and Convolutional Neural Networks,” Journal of CoRR, 2016.
[6] J. Pennington, R. Socher and C. D. Manning, “GloVe: Global Vectors for Word Representation,” Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1532-1543, 2014.
[7] T. Mikolov, K. Chen, G. Corrado and J. Dean, “Efficient Estimation of Word Representations in Vector Space,” Journal of CoRR, 2013.
[8] S. Ismail and M. S. Rahman “Bangla Word Clustering Based on N-gram Language Model,” International Conference on Electrical Engineering and Information &Communication Technology (ICEEICT), 2014.
[9] Z. Islam, A.Mehler, and R. Rahman, “Text Readability Classification of Textbooks of a Low-Resource Language,” 26th Pacific Asia Conference on Language, Information & Computation, pp 545–553, 2012.
[10] A. Ahmad and M. R. Amin, “Bengali Word Embeddings and its Application in Solving Document Classification Problem,” 19th International Conference on Computer and Information Technology, pp.425-430, 2016.
[11] M. Krendzelak and F. Jakab, “Text categorization with machine learning and hierarchical structures,” in Proc. of 13th Int. Con. on Emerging eLearning Technologies and Applications, pp.1-5, 2015.
[12] A. N. Chy, M.H. Seddiqui, and S. Das, “Bangla News Classification using Naive Bayes classifier,” in Proc. of 16th Int. Conf. Computer & Information Technology, pp. 366-371, 2014.
[13] C. Liebeskind, L. Kotlerman and I. Dagan, “Text Categorization from category name in an industry motivated scenario,” Journal of Language Resources and Evaluation, vol. 49, no. 2, pp. 227-261, 2015.
[14] M. Kaya, G. Fidan and I. H. Toroslu, "Sentiment Analysis of Turkish Political News," IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technology, vol. 01, pp. 174-180, 2012.
[15] S. Alsaleem, "Automated Arabic Text Categorization Using SVM and NB," International Arab Journal of e-Technology, vol. 2, no. 2, June 2011.
[16] The Daily Prothom Alo, Online, http://www.prothom-alo.com
[17] The Daily Jugantor, Online, https://www.jugantor.com
[18] The Daily Ittefaq, Online, http://www.ittefaq.com.bd
[19] The Daily Manobkantha, Online, http://www.manobkantha.com