Bangla News Recommendation Using doc2vec

Conference Paper · September 2018

DOI: 10.1109/ICBSLP.2018.8554679

5 authors, including:

Rabindra Nath Nandi
Khulna University of Engineering and Technology
14 PUBLICATIONS **32** CITATIONS
SEE PROFILE

M. M. Arefin Zaman
Socian Ltd
4 PUBLICATIONS **20** CITATIONS
SEE PROFILE

Tareq Al Muntasir
Socian Ltd
3 PUBLICATIONS **15** CITATIONS
SEE PROFILE

Sakhawat H Sumit
BJIT Ltd
5 PUBLICATIONS **26** CITATIONS
SEE PROFILE

Bangla news recommendation using doc2vec

Rabindra Nath Nandi*, M.M.Arefin Zaman*, Tareq Al Muntasir, Sakhawat Hosain Sumit, Tanvir Sourov and Md. Jamil-Ur Rahman
Socian Ltd
Dhaka, Bangladesh
{rabindra, arefin, tareq, sumit, tanvir, jamil}@socian.ai

***Abstract*—We present a content-based Bangla news recommendation system using paragraph vectors also known as doc2vec. doc2vec is a neural network driven approach that encapsulates the document representation in a low dimensional vector. doc2vec can capture semantic relationship effectively between documents from a large collection of texts. We perform both qualitative and quantitative experiments on a large Bangla news corpus and show that doc2vec performs better than two popular topic modeling techniques LDA and LSA. In the top-10 recommendation scenario, the suggestions from doc2vec are more contextually correct than both LDA and LSA. doc2vec also outperforms LDA and LSA on human-generated triplet dataset with 91% accuracy where LDA and LSA give 85%,84% accuracy respectively.**

Keywords—content-based, Bangla news recommendation, paragraph vectors, doc2vec, LDA, LSA.

I. INTRODUCTION

Digital information over the internet is growing exponentially day by day. With this tremendous amount of information, users face information overload problem. News recommendation systems solve the information overload problem by providing information quickly and efficiently to the massive users [1].

Recommendation systems are mainly categorized into three types: (1) Collaborative Filtering, (2) Content-based Filtering and Hybrid System [2]. Collaborative approach estimates an interest factor of an item for a user by analyzing the preference of other users who have already experienced the item. This approach doesn't focus on the content and features of the item. For this reason, when a system starts or bootstraps without any user's prior information, it faces cold-start problem. On the other hand, Content-based recommendation systems try to recommend an item to a user based on the description of the item and profile of the user's interests if user information available [3]. Content-based recommendation systems analyze item descriptions to identify items that are of particular interest to the user. There are some hybrid systems that integrate both collaborative and content-based idea altogether to build a more robust system by overcoming the drawbacks of these two approaches. Most of modern industry based recommendation systems are actually hybrid systems.

The quality of a content-based recommendation system mainly depends on the representation of the content of the items and user profiles. Most of the item's content is in the textual format such as document, news, movie reviews. So, working with textual information to extract latent features is a major concern in this type of recommendation systems. Topic modeling is an active research area in the document modeling domain to find out the topics of a set of documents and similar documents which are semantically and contextually similar. Several studies are available on topic or concept discovery techniques e.g. Latent Semantic Analysis (LSA), Latent Discriminate Analysis (LDA) and their variants for Document Recommendation [4, 5].

Bangla is one of the most widely spoken languages in the world as about 250-300 million people speak Bangla as their first language. Due to the lack of availability of sufficient research works and language modeling resources, developing an intelligent system like news recommender for Bangla is comparatively challenging than other rich languages e.g. English, Spanish and German etc. The authors in [13] proposed an ontology-based recommendation system for cross-lingual languages which includes Bangla and English news.

In this work, we use a recent idea from distributional semantics called document embedding using doc2vec [8] which is highly scalable and works without any preprocessing or feature extraction steps except for only tokenization. doc2vec learns the semantics and compositionality of the linguistic components by using a deep learning architecture. This neural architecture is simple and reduces human effort significantly. It compresses the whole contextual and structural information into a one-dimensional numeric vector. Although this way is theoretically interesting and straightforward, the main challenge is that it needs a lot amount of data to build a high dimensional semantic space where documents are placed perfectly with their latent version. We have collected about 0.3 million news from different Bangla news website to build a Bangla news corpus.

Document embedding models are trained on uncategorized articles and the trained model can generate an embedding to an input article. This embedding can be used as a feature vector and it can be used for document categorization, document ranking and information retrieval tasks.

We have evaluated the performance of the document vector on information retrieval tasks against LDA, LSA and n-gram based feature extraction methods. We have shown that the performance of the document vector is better and easy to deploy as an information retrieval system. We purely focus on the content-based recommendation since our crawled dataset does not contain any user information, the similar approach has been followed in [7].

II. METHODS

In this section, we discuss the doc2vec model that is used for the recommendation task. Furthermore, we

*The authors contributed equally

978-1-5386-8207-4/18/$31.00 ©2018 IEEE

briefly explain two popular topic modeling methods namely LSA and LDA which are also used for modeling unstructured texts.

A. doc2vec

doc2vec is an approach to learn a model that can generate an embedding to a given document [8]. Unlike some of the commonly used methods such as bag-of-words (BOW), n-gram models or averaging the word vectors, this method is very much generic and can be used to generate embeddings from texts of any length. It can be trained in a totally unsupervised fashion from large corpora of raw text without needing any task-specific labeled dataset. doc2vec performs really well in the case of representing longer documents [12].

doc2vec is an extension to the existing word embedding models. A very well-known technique for learning the word vectors is to predict a word given the other words in a context.

For a sequence of training words w1, w2, w3, ..., wT, the objective of the model is to maximize the average log probability,

$$\frac{1}{T}\sum_{t=k}^{t+k} \log p\,(w_t \mid w_{t-k}, \ldots\ldots\ldots, w_{t+k}) \qquad (1)$$

The probability is calculated using the softmax function which is defined as:

$$p(w_t \mid w_{t-k}, \ldots\ldots\ldots, w_{t+k}) = \frac{e^{y w_t}}{\sum_i e^{y_i}} \qquad (2)$$

Paragraph vectors are jointly trained with the word vectors. At first, the paragraph vector and the word vectors are initialized randomly. While training the language model, both of these vectors learn a semantic representation of the sequence of sentences. The paragraph vector also contributes to the prediction task along with the word vectors.

Two kinds of frameworks have been proposed by Le and Mikolov to learn the doc2vec [8]. These are:

(1) doc2vec with the distributed memory model

(2) doc2vec with the distributed bag-of-words.

In the distributed memory model, paragraphs and words are jointly trained using a stochastic gradient descent optimizer. Each paragraph is mapped to a fixed dimension unique vector represented by a column in matrix D and each word is represented by a column in matrix W. The paragraph vector and word vectors are averaged or concatenated to predict the next word in a context. The model is shown in Fig. 1.

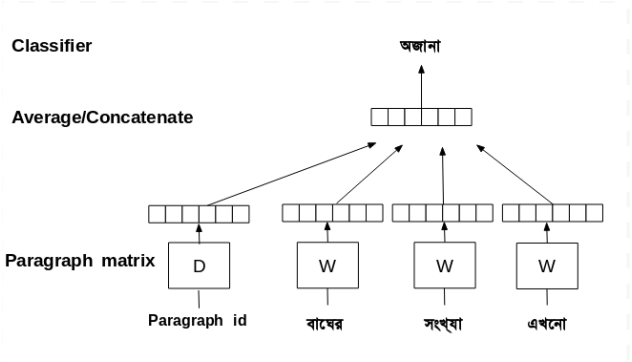

Fig. 1. The distributed memory model of doc2vec for a Bangla sentence

In this model, a paragraph vector is fixed for all samples generated by a sliding window from a document and the word vectors are shared across all documents. The total learning parameters excluding the softmax parameters are $N \times p + M \times q$, where p= the length of paragraph vector, q= the length of word vector, N=no of paragraphs, M= no of words in the vocabulary.

In the distributed bag of words model in Fig. 2, the model is strictly trained to predict words randomly sampled from a paragraph and no context word is used as a part of the input data. For each iteration, a text window is sampled and then a random word is chosen from the text window to form a classification task given the paragraph vector.

The basic difference with the distributed memory model is that this model needs less parameter and model size is relatively small but it cannot preserve word order. This model is similar to word2vec skip-gram model.

B. Latent Semantic Analysis

Latent Semantic Analysis (LSA) is a popular technique in distributional semantics to analyze the semantic relationship between a set of documents by using the term-document matrix and singular value decomposition technique [4]. LSA outputs a term-document matrix where similar documents and similar words are placed closer. The similarity between two documents is computed by the cosine similarity between their corresponding two column vectors and in a similar way, the correlation between two words are computed by their corresponding row vectors. LSA captures some basic linguistic properties such as synonymy and polysemy.

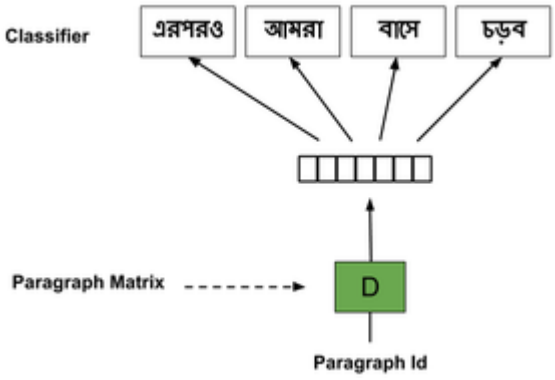

Fig. 2. The distributed bag-of-words model of doc2vec for a Bangla Sentence

C. Latent Dirichlet Allocation

Latent Dirichlet allocation (LDA) is a generative probabilistic model of a corpus to extract the hidden structure and topics [5]. The key concept is that documents are represented as random mixtures over latent topics where each topic is characterized by a distribution over words [9]. LDA model projects documents in a topical embedding space and it generates a topic vector from a document which can be used as the features of the document.

III. EXPERIMENTAL STUDIES

This section describes data acquisition techniques, validation parameters and performance evaluation of doc2vec and other document modeling techniques.

A. Data Acquisition

We have collected data from fifteen Bangla newspaper by running a news crawler for 1 week using Apache Nutch. The news corpus has more than 3,00,000 uncategorized articles which have been used for training the paragraph vector model. We collected 37,000 labeled news articles consisting of 12 different categories for the purpose of evaluating the recommendation performance.

In the preprocessing step, the HTML pages are cleaned using python-goose library and further distilled by using python beautifulsoup4 library. In the tokenization step we firstly remove special characters (e.g. '।', '!', '?') and non-Bangla characters and then a whitespace tokenizer is used for word segmentation. We don't use any stop-word removal technique as it causes to lose structural information and also makes context understanding difficult for the model.

B. Experimental Setup

Though distributed bag-of-words model loses the word order and should be inferior to the memory model according to the original paper [8], our experiments show that given enough training samples, this model outperforms the distributed memory model. Therefore, our choice of architecture for training doc2vec is dBoW. These models have been implemented using the gensim library [14], which is one of the most popular python libraries for text mining and statistical semantics. The model is trained only using CPU.

C. Validation Parameters

The accuracy of the model is evaluated by measuring the performance on triplet dataset using Eqn. 3 where M= total no. of triplets, sim(A_i, P_i) = Similarity between the anchor and the positive document and sim(A_i, N_i) = similarity between the anchor and the negative document of i^{th} element of the triplet dataset.

$$Accuracy = \frac{\sum_{i=1}^{M} SIM(A_i,P_i,N_i)}{M},$$

$$\text{where, } SIM(A_i,P_i,N_i) = \begin{Bmatrix} 1 & if\ sim(A_i,P_i) > sim(A_i,N_i) \\ 0 & otherwise \end{Bmatrix} \quad (3)$$

The accuracy is defined by the total no. correctly recommended triplets where the similarity is higher between the anchor and positive document rather than between anchor and negative document.

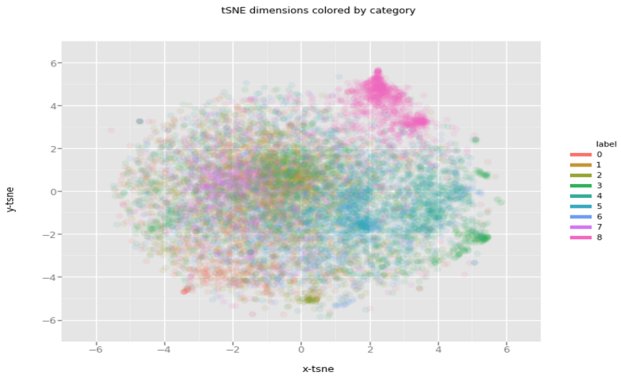

Fig. 3. t-SNE visualization of Bangla news embedding using doc2vec

D. Results and Discussion

We present both qualitative and quantitative experiments for a better understanding of model performance.

First, the model is trained on our uncategorized news articles and visualized the trained vectors using t-SNE [10]. t-SNE (t-Distributed Stochastic Neighbor Embedding) is a visualization technique for large-scale high dimensional data. It projects high dimensional data into low-dimensional space by capturing local and global structure perfectly resulting in a map containing clusters of similar data points placed nearby. From the visualization map in Fig. 3, it can be seen that the articles from the same category are clustered together.

Next, we qualitatively look at the nearest neighbors of news articles in the document embedding space. We generate top ten recommendations for different query document to test the model performances and we show that doc2vec gives more acceptable recommendation than LDA and LSA. For an example, we tabulated the result for a document titled 'প্রতিমা বিসর্জনে শেষ হলো দুর্গোৎসব' in Table I. The query article is related to 'Durga Puja', an annual festival of Hindu religion and the recommended articles are expected to be related to this festival. All of the top ten recommendations by doc2vec are closely related to the event of 'Durga Puja'.

On the other hand, the first recommendation from LDA is titled 'আকাশে উড়লো ফানুস, নদীতে ভাসলো নৌকা' and this news is related to 'Buddha Purnima', not expected as the first recommendation because there are lots of news related to the 'Durga Puja'. Another recommendation of LDA is 'ঐতিহ্য আর সম্প্রীতিতে পবিত্র আশুরা পালিত সৈয়দপুরে' that is related to one of the major religious days 'Ashura' of Islam. This recommendation is correct in a sense that it is religious news but not sufficient enough to be in the top-10 recommendations. The incorrect recommendation of LDA is 'সম্মেলনের প্রস্তুতি চলে রাতভর' which is about a totally different event.

LSA makes even more mistakes than LDA for this query. The articles recommend by LSA, titled 'হুমায়ূন আহমেদের জন্মদিন আজ', 'রুদ্র মুহম্মদ শহিদুল্লাহ'র তম জন্ম বার্ষিকী আজ' and 'হুমায়ূন আহমেদের জন্মদিন আজ' are totally different from the query article. These unexpected recommendations are

TABLE I. TITLES OF TOP 10 RECOMMENDATIONS FOR QUERY ARTICLE "প্রতিমা বিসর্জনে শেষ হলো দুর্গোৎসব" BY DOC2VEC, LDA AND LSA

doc2vec	LDA	LSA
কুমারী পূজায় জনজোয়ার	**আকাশে উড়লো ফানুস, নদীতে ভাসলো নৌকা**	মহাষ্টমী ও কুমারীপূজা আজ
মণ্ডপে মণ্ডপে হাহাকার, চলছে প্রতিমা বিসর্জনের প্রস্তুতি	মণ্ডপে মণ্ডপে হাহাকার, চলছে প্রতিমা বিসর্জনের প্রস্তুতি	ফরিদপুরে রামকৃষ্ণ মিশন আশ্রমে
সিঁদুর খেলায় মেতে উঠলো কলকাতা	দুর্গোৎসবের মহানবমী আজ	প্রতিমা বিসর্জনে শেষ হলো দুর্গোৎসব
মহাষ্টমী ও কুমারীপূজা আজ	সাভারে আনন্দ উল্লাসে মহাষ্টমী পালিত	আগরতলায় কুমারী পূজায় পূণ্যার্থীর ঢল
বৃন্দাবনের বিকল্প দুবলার চরের রাস মেলা	**ঐতিহ্য আর সম্প্রীতিতে পবিত্র আশুরা পালিত সৈয়দপুরে**	**কুষ্টিয়ার ছেঁউড়িয়ায় লালন স্মরণ উত্সব আজ থেকে**
সাম্প্রদায়িক সম্প্রীতির চিরন্তন	কুমারী পূজায় জনজোয়ার	প্রতিমা বিসর্জনের মাধ্যমে শেষ হলো শারদীয় উৎসব
দেবী দুর্গা বিসর্জনে মণ্ডপগুলোতে বিষাদের সুর	প্রতিমা বিসর্জনের মাধ্যমে শেষ হলো শারদীয় উৎসব	**হুমায়ূন আহমেদের জন্মদিন আজ**
দুর্গতিনাশিনী দেবী দুর্গা	**সম্মেলনের প্রস্তুতি চলে রাতভর**	**রুদ্র মুহম্মদ শহিদুল্লাহ'র তম জন্ম বার্ষিকী আজ**
বান্দরবানে দুর্গা প্রতিমা বিসর্জন	তারায় তারায় মিলেছে রং-বেরংয়ের ফানুস	ঝালকাঠিতে দুর্গোৎসব উপলক্ষে ঐতিহ্যবাহী 'দশহরার মেলা
নানা রঙ আর বৈচিত্র্য	বর্ষণ ও ধর্মীয় উৎসবমুখর পরিবেশে প্রতিমা বিসর্জন	**হুমায়ূন আহমেদের জন্মদিন আজ**

about the birthday of two popular writers of Bangladesh. Another recommendation by LSA 'কুষ্টিয়ার ছেঁউড়িয়ায় লালন

স্মরণ উত্সব আজ থেকে' which is related to a prominent Bengali philosopher and poet 'Lalon Shah'.

The recommendations by doc2vec are significantly better than both LDA and LSA model for this query as its top 10 recommendations are all about 'Durga Puja'.

We perform a quantitative evaluation to measure how well doc2vec learned semantic representation using a triplet dataset as described in [6].

The triplets were generated manually by human from a Bangla news corpus. The dataset consists of 330 triplets. Each triplet (anchor, positive, negative) is chosen by carefully analyzing the content of the three articles. The first two documents are not only from the same category of the newspaper (e.g. entertainment, sports) but also they are semantically equivalent. The negative document is from a different category and also contextually different from the anchor article. The objective of triplet evaluation is to explore the performance of a model to ensure high relevance between similar document pairs (anchor, positive) than (anchor, negative) pairs. The model accuracy is counted by using the Eqn. 3 which depends on the number of triplets for which the model outputs a high similarity score between the anchor and positive document.

The experimental result on triplet dataset is given in Table II. We extract 100 topics from both LDA and LSA to use feature size same for all models.

TABLE II: PERFORMANCES OF DIFFERENT METHODS ON HUMAN-GENERATED TRIPLETS

Model	Embedding Size/ Topics	Accuracy
Bag-of-words	N/A	67%
LSA	100	84%
LDA	100	85%
doc2vec	100	91.0%

We see that doc2vec gives 91.0% accuracy which is better than other methods. LDA and LSA give 85%, 84% accuracy respectively. We can conclude that doc2vec can capture the semantic similarity between Bangla news documents with long text more efficiently than other baseline methods.

IV. CONCLUSION

Content-based news recommendation system is strongly connected to the semantic relevance measure of news articles. We develop a Bangla news recommender system using doc2vec. The most attractive part of this model is its language-independent learning and adaptation capability to a large corpus. With a sufficient amount of data, it can learn a lot of intrinsic property, structural maps and semantic variations of a language.

Our experiments show that doc2vec surpasses two popular topic modeling techniques; LDA and LSA for building a content-based Bangla news recommendation system. This model can also be used for many different applications like Bangla news clustering, news summarization and question answering.

REFERENCES

[1] G. Adomavicius and A. Tuzhilin, "Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions", IEEE Transaction on Knowledge and Data Engineering, Volume 17, Issue 6, pp.734-749, 2005.

[2] M. Balabanović and Y. Shoham, "Fab: content-based, collaborative recommendation", Communications of the ACM, vol. 40, issue 3, 1997, pp. 66-72.

[3] M. Madhukar, Challenges & Limitation in Recommender Systems. International Journal of Latest Trends in Engineering and Technology (IJLTET), vol. 4, issue 3, pp. 138-142, September 2014.

[4] P. Velvizhi, S. Aishwarya and R. Bhuvaneswari, "Ranking of Document Recommendations from Conversations using Probabilistic Latent Semantic Analysis", International Conference on Innovations in Engineering and Technology (ICIET), 2016: 133-138.

[5] T. Chang and W. Hsiao, "LDA-based Personalized Document Recommendation", PACIS 2013 Proceedings, 2013.

[6] A. M. Dai, C. Olah, Q. V. Le, and G. S. Corrado, "Document embedding with doc2vec." NIPS Deep Learning Workshop, 2014.

[7] Md. N. M. Adnan, M. R. Chowdury, I. Taz, T. Ahmed and R. M Rahman, “Content Based News Recommendation System Based on Fuzzy Logic”, 3rd International Conference on Informatics, Electronics & Vision, 2014.

[8] Q. V. Le and T. Mikolov. “Distributed representations of sentences and documents”, International Conference on Machine Learning, 2014.

[9] D. M. Beli, A. Y. Ng and M. I. Jordan, “Latent dirichlet allocation”, Journal of machine Learning research, vol. 3, pp. 993-1022, 2003.

[10] L. J. P. van der Maaten and G. E. Hinton. “Visualizing high-dimensional data using t-SNE”. Journal of Machine Learning Research, 2008.

[11] S. T. Dumais. ‘’Latent Semantic Analysis’’. Annual Review of Information Science and Technology. vol. 38, pp. 188–230, 2005.

[12] J. H. Lau and T. Baldwin, “An Empirical Evaluation of doc2vec with Practical Insights into Document Embedding Generation”, Proceedings of the 1st Workshop on Representation Learning for NLP, pp. 78–86,, 2016.

[13] S. N. Ferdous and M. M. Ali, ‘A Semantic Content Based Recommendation System for Cross-Lingual News’, International Conference on Imaging, Vision & Pattern Recognition, 2017.

[14] R. Rehurek and P. Sojka, ‘Software Framework for Topic Modelling with Large Corpora’, Proceeding of the LREC 2010 Workshop on New Challenges for NLP Frameworks, 2010.