**Sentence Based Topic Modeling using Lexical Analysis**

Shahinur Rahman1, Shikh Abujar1, S. M. Mazharul Hoque Chowdhury1, Mohd Saifuzzaman1, Syed Akhter Hossain1

{shahinur3606, sheikh.cse, mazharul2213, saifuzzaman.cse}@diu.edu.bd, aktarhossain@daffodilvarsity.edu.bd

1Department of computer science and engineering, Daffodil International University, Dhaka, Bangladesh

**Abstract**. In every second in this world we are generating tons of data in the internet in different format. Most of them are in text format. Therefore, the demand of topic modeling is higher than ever right now. Data scientists are working day and night to make it more effective and accurate using different methods. Topic modeling focuses on the keywords that can express or identify the topic discussed in the document. Topic modeling can save a lot of time by releasing its user from page to page manual reviewing. In this paper a model has been proposed to find out topic of a document. This model works based on the relations between most frequent words and their relation with sentences in the document. This model can be used to increase the accuracy of the topic modeling.

**Keywords:** algorithm, frequent, method, relation, segmentation, sentence, summarization, topic modeling, word.

1. **Introduction:**

Now-a-days with development and increasing use of the internet, a reader not only read the document but also they contribute information publicly. From that large amount of information, it is quite difficult to sort a particular or desired word. So it has becomes more important to compress and summarize data. From that huge amount of data, extracting information using manual method is actually incompetents [1].

In text mining one of the most important problem is text summarization. In text summarization a lots of collection text are make smaller and neatly packed together text which is represent the meaning of main text. Text summarization helps to understand huge amount of text easily which saves a lot of time.

Text summarization are divided into two types. They are single document text summarization and multi document text summarization. In a single text summarization a large size of single text is summarized to another single document summary. In multi text summarization a set of documents are summarized to a single document summary. In both approach a large amount of data are summarized and stored in a single file.

In text summarization, extractive summarization is a common and mature technique which is extracted important sentence then recombine them and generate a summary base of this sentence.

In topic modeling, Latent Dirichlet Allocation (LDA) model is used to explore topic firstly. First, from a document, we extract sentence association with the most frequent word. Now it is possible to find out relations from high scored words and high length sentences.

1. **Literature Review**

Nowadays there are lots of numbers studies about topic modeling but rarely have been small number of studies in Bengali Sentence Based topic modeling. In thin paper we use LDA model with lexical analysis to extract Topic from large collect of information.

LDA is one of the most common method to extract topic modeling from different types of data examples that use auxiliary information are the author-topic model (Rosen-Zvi et al.,2004), tag -topic model(Tsai,2011) and Topic-Link LDA (Liu et al., 2009). All of this work has been done in English. Previously (Geetanjali and Pushpak) analysis Bengali Poet classification and Identification and Amitava and Sivaji (2010) have done to analysis a document and find out Opinion base summarization but there is no work related about Topic modeling in Bengali.

In this current year, some good research has been by different researchers all over the world. Jiang and Zhou (2017) worked on topic modeling based on the poison decomposition. In this work they tried to find out statistical results based on multidimensional characteristics of the topic [7]. Now at the same time, Truică and his team worked on this same topic using contextual cause. They applied automatic term recognition system using contextual cause for topic modeling. Another researcher named Ruohonen (2017) tried to classify web exploits using topic modeling [8]. Karami, Gangopadhyay, Zhou and Kharrazi (2017) worked on Fuzzy approach. Their target was health and media corpora topic modeling. Fuzzy approach was used to analyze medical document and extract information [9]. Work on probabilistic topic model has been done by Zhai (2017) for text data retrieval and analysis [10].

In this research work topic modeling has been done based on sentiment analysis. As sentiment is a very important part of a sentence and main focus depends on the sentiment, therefore by finding out the sentiment expressed in the sentence it is possible to find out main keywords. Those keywords was used for topic modeling.

1. **Proposed Method**

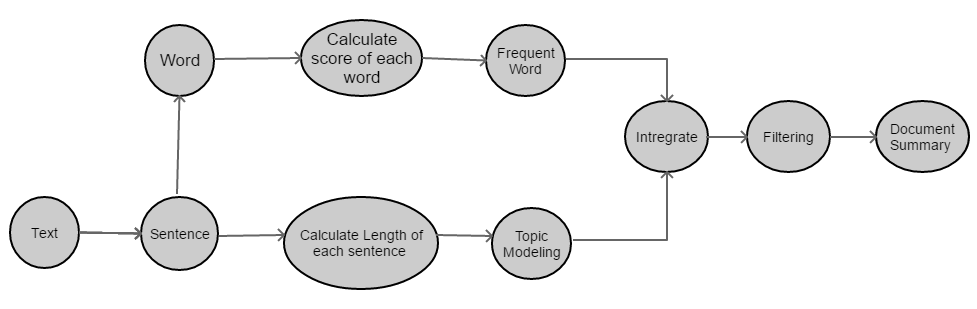


Figure 1: Data flow of proposed model

1. **Data Preprocess**

For text summarization, total collected data from online news portal and its more than 12000 thousand. To summarize each summary consists of three parts and its Beginning, body and ending part. Before summarization it is needed to preprocess the text document. Therefore, chapter segmentation, sentence segmentation, word segmentation occurs then it removes stop words and finally stemming.

* 1. **Chapter segmentation:** Basically, a text document consist of lots of chapters and each chapter is dependent on the other chapter. First, it is needed to separate text into many chapters that’s why chapter segmentation was used to divide the whole text. Chapter segmentation is a process of dividing the text into the meaningful unit. From each unit, different meaning can be extracted to summarize data.
  2. **Sentence Segmentation:** From a large collection of data, it's much difficult to make a meaningful summary. In a text document each sentence disclosure a meaning. Sentence segmentation is splitting the text into a sentence. After that whole text is exposed a set of sentences. We use NLTK toolkit in our work to separate sentence from given text. i.e.

S **=** মানুষ হিসেবে সবারই কিছু ইতিবাচক দিক রয়েছে।আবার বিপরীত মেরুতেই রয়েছে বিভিন্ন দোষত্রুটি।নিজেরাই অনেক সময় ভুলগুলো নিয়ে সচেতন থাকি না।

Table 1: Sentences separated from document

|  |  |
| --- | --- |
| Sentence Number | Sentence |
| S1 | মানুষ হিসেবে সবারই কিছু ইতিবাচক দিক রয়েছে |
| S2 | আবার বিপরীত মেরুতেই রয়েছে বিভিন্ন দোষত্রুটি |
| S3 | নিজেরাই অনেক সময় ভুলগুলো নিয়ে সচেতন থাকি না |

* 1. **Word segmentation:** Word segmentation is referring to extract sentence as a set of independent words. Generally, in Bengali and some other language using space is separator from one word to another word.
  2. **Remove stop words:** Stop words are a set of commonly used words in any language which has no actual meaning. It’s just use to make a sentence but doesn’t carry any tangible meaning. So, we need to remove the stop word from the text. To remove stop word in our paper we use NLTK tools.
  3. **Stemming:** In this step clean data are collected and stored in a document. Data are now ready to be analyzed and further processing.

1. **Calculate Sentence Score**

In sentence segmentation, just separation of sentence happened and set of sentences were created. But basically no one has clear idea about the value of the word inside the sentence. To calculate the length of each sentence we count how many words are in a sentence and scoring. Then sorting this sentence based on high score to low. i.e. we have set of sentences which is

S= {নদী আর নদী পাড়ের মানুষের জীবন জীবিকা নিয়ে এর আগে গৌতম ঘোষের পদ্মা নদীর মাঝি দেখেছিলাম।মানিকের উপন্যাসের মত করে নয়, গৌতম নিজের মত করে পদ্মা পাড়ের জেলেদের জীবনের বাস্তবতাকে ফুটিয়ে তুলেছিলেন}

Table 2: Sentence Scoring

|  |  |  |
| --- | --- | --- |
| **Sentence Number** | **Sentence** | **Length** |
| **S1** | নদী আর নদী পাড়ের মানুষের জীবন জীবিকা নিয়ে এর আগে গৌতম ঘোষের পদ্মা নদীর মাঝি দেখেছিলাম | **16** |
| **S2** | গৌতম নিজের মত করে পদ্মা পাড়ের জেলেদের জীবনের বাস্তবতাকে ফুটিয়ে তুলেছিলেন | **11** |

1. **Calculate Word Score**

To find out the most frequent word we have to calculate the score of each word is most important because after calculating the score of each word it's visible which group word is most used and we take which group of words has scored more than average. If a word used in many sentences for ten times that means this word score is 10.

1. **Topic modeling**

Latent Dirichlet Allocation (LDA) is one of most popular topic modeling which is followed by data preprocessing. It helps to search the topic words related to a document. LDA obtain feature words list for each topic. For example, if in a collection of words, there are numbers of words like as “sports”, “play” are used mostly then consider that it becomes a sports news. Consider the following equation

X = ……………………….. (1)

Count ………………………... (2)

Here, X is the total value of valid word and is the Total word, is Total Preposition, is Total Nouns, is Total Articles and is to be verbs. After calculating the total value of valid word, we find out probability of each word following the equation.

……………………………………. (3)

…………………………………….. (4)

…………………………………….. (5)

If probability of is bigger then then is respectively most probability to become the related topic. On the other hand, if the probability of are bigger then the topic would be the biggest probability related word.

1. **Filtering/Smoothing**

In this stage we are going to find out the score for the sentences and relations between the sentence and words. For that we are proposing a method. As before we explained that sentence length is the score for the sentence. This time sentence score will be the number of valid words were used in the sentence. To do that it is necessary to remove determinate, nouns, articles etc. Therefore, a set of clear data will be stored in the sentence. Now it is time to find out the expected topic of the document. We can write the process as,

S1 = [w1, w2, w3……...wn] ……………………. (6)

S2 = [w1, w2, w3……...wn] ……………………. (7)

For equation 6 and 7 if process starts, then

If S1[i] = S2[i]

Match found

Else S2[i++]

Go to Start

1. **Document Summary**

Suppose, the word ‘Bangladesh’ got the highest value and for this word we got 7 sentences. This time those sentences can be compared with each other using a two-dimensional array. This process will go on until next five top words to ensure maximum accuracy. After processing, most frequent matching will be taken as the best value and corresponding word will be the topic of the document.

1. **Final Outcome**

In this process it is possible to find out best possible output topic for a document. This can be used for any language as well as Bangla. Using Bangla language corpus data processing and other things can be done much more efficient way.

**Conclusion**

In this era of technology data bought us new opportunity as well as new complexity. Handling new data require new method, sometimes new technology. Reading files one by one to find out topic of the document is one of the toughest task now a day. If a simple system can solve this why should not we use this to proper use our brain and time in other important tasks. Topic modeling is a very important sector of data science and requires very large amount of research work. This system can be a simple but efficient method to use topic modeling for different languages.

**Acknowledgement**

We would like to thank DIU - NLP and Machine Learning Research LAB for all their support and help. Any error in this research paper is our own and should not tarnish the reputations of these esteemed persons.

**References**

1. [Mahak Gambhir, Vishal Gupta, Recent automatic text summarization techniques: a survey, Artificial Intelligence Review, v.47 n.1, p.1-66, January 2017](https://dl.acm.org/citation.cfm?id=3041126&CFID=837279816&CFTOKEN=87351615) [doi>[10.1007/s10462-016-9475-9](http://dx.doi.org/10.1007/s10462-016-9475-9)]
2. Rosen-Zvi, M., Griffiths, T., Steyvers, M., and Smyth, P. (2004). The author-topic model for authors and documents. In Proceedings of the 20th conference on Uncertainty in artificial intelligence, pages 487–494
3. Tsai, F. S. (2011). A tag-topic model for blog mining. Expert Systems with Applications, 38(5):5330–5335
4. Liu, Y., Niculescu-Mizil, A., and Gryc, W. (2009). Topic-link LDA: joint models of topic and author community. In Proceedings of the 26th Annual International Conference on Machine Learning, pages 665–672.
5. Geetanjali Rakshit, Anupam Ghosh, Pushpak Bhattacharyya, Gholamreza Haffri. Automated Analysis of Bangla Poetry for Classifiation and Poet Identifiation. IITB-Monash Research Academy, India, 2IIT Bombay, India 3Monash University, Australia
6. Amitava Das and Sivaji Bandyopadhyay. Topic-Based Bengali Opinion Summarization.
7. Jiang, H., Zhou, R., Zhang, L., Zhang, Y.: A Topic Model Based on Poisson Decomposition. In: Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, pp- 1489-1498, Singapore, November. (2017)
8. Ruohonen, J.: Classifying Web Exploits with Topic Modeling. In: 28th International Workshop on Database and Expert Systems Applications, DOI. 10.1109, DEXA. 2017.35, Lyon, France (2017)
9. Karami, A., Gangopadhyay, A., Zhou B., Kharrazi, H.: Fuzzy Approach Topic Modeling for Health and Medical Corpora, In: International Journal of Fuzzy Systems (2017)
10. Zhai, C.: Probabilistic Topic Models for Text Data Retrieval and Analysis. In: 40th International ACM SIGIR Conference, pp. 1399-1401, Shinjuku, Tokyo, Japan (2017)