Journal of Computer Science

Review

A Systematic Literature Review on English and Bangla Topic Modeling

Md. Basim Uddin Ahmed, Ananta Akash Podder,
Mahruba Sharmin Chowdhury and Mohammad Abdullah Al Mumin

Department of Computer Science and Engineering, Shahjalal University of Science and Technology, Sylhet, Bangladesh

Article history
Received: 21-09-2020
Revised: 15-11-2020
Accepted: 07-01-2021

Corresponding Author:
Md. Basim Uddin Ahmed
Department of Computer Science and Engineering, Shahjalal University of Science and Technology, Sylhet, Bangladesh
Email: basimuddintuhin@gmail.com

Abstract: Due to the enormous growth of information and technology, the digitized texts and data are being immensely generated. Therefore, identifying the main topics in a vast collection of documents by humans is merely impossible. Topic modeling is such a statistical framework that infers the latent and underlying topics from text documents, corpus, or electronic archives through a probabilistic approach. It is a promising field in Natural Language Processing (NLP). Though many researchers have researched this field, only a few significant research has been done for Bangla. In this literature review paper, we have followed a systematic approach for reviewing topic modeling studies published from 2003 to 2020. We have analyzed topic modeling methods from different aspects and identified the research gap between topic modeling in English and Bangla language. After analyzing these papers, we have identified several types of topic modeling techniques, such as Latent Dirichlet Allocation (LDA), Latent Semantic Analysis (LSA), Support Vector Machine (SVM), Bi-term Topic Modeling (BTM). Furthermore, this review paper also highlights the real-world applications of topic modeling. Several evaluation methods were used to evaluate these models' performances, which we have discussed in this study. We conclude by mentioning the huge future research scopes for topic modeling in Bangla.

Keywords: English Bangla Comparison, Latent Dirichlet Allocation (LDA), Systematic Literature Review (SLR), Topic Modeling Bangla, Topic Modeling Methods, Topic Extraction

Introduction

Because of the rapid development of Information Technology (e.g., Internet, Social Media, Online Databases, etc.), the amount of data generated has exponentially exacerbated in recent years. This vast accumulation of data provides essential support for training machine learning models and easy access to search engine queries. On the other hand, because of this massive flourish of information, extracting the knowledge of interest from these data has become a matter of general concern (Xu *et al*., 2019). According to the study of DOMO (a cloud-based business service system), roughly 2.5 Quintilian bytes of data are produced daily and 90% of that data in the world has been created in the last two years only (according to 2018 studies) (Al Helal and Mouhoub, 2018). So it is not feasible for any person to sieve useful information from these vast amounts of data manually. Moreover, the National Science Foundation (NSF) identified 'large-scale scientific data management and analysis' as one of the data-intensive challenges and as an area for future study (Karami *et al*., 2018). So it is crucial to precisely and efficiently estimate the numerical characteristics, to determine the appropriate statistical distributions for modeling text corpora (Jiang *et al*., 2017).

Topic modeling is a probabilistic approach that can be observed as an instrument of measurement for the hidden structures in a document (Shi *et al*., 2019). To infer these hidden structures, we have to pre-process the documents. At first, extraneous words and stop words are removed from the text. Punctuations are also usually removed, but some researchers have kept punctuations if they carry certain emotions or meanings. Then the words are stemmed (converted to its' root form). Some models consider Bi-grams (adjacent words that often appear together), tri-grams, etc. as one word. The resulting list is then transformed into a bag of words (words with

count) for that document. Weights are assigned to each word by analyzing them. To give importance to particular words, their weighting factors can be changed (e.g., using the term frequency-inverse document frequency or TF-IDF). The pre-processed data is fed to a Machine Learning algorithm (LDA, LSA, BTM, etc.). These algorithms iterate through the training data several times and try to accurately infer latent topics from those collections of documents as much as possible. Various parameters and hyperparameters are used in these algorithms. They are tuned during the training phase of the model. These models output documents as a distribution over topics and topics as a mixture of words. Again, sets of topically-related words are generated as 'topic', which can be associated with the documents of that corpus (Hasan *et al*., 2019).

By inferring topics from all these huge collections of documents and electronic archives, topic modeling is used in many real-world applications such as: Topic Extraction from social sites (e.g., Twitter, Facebook) (Alkhodair *et al*., 2018), classification and clustering (Ruohonen, 2017; Al Helal and Mouhoub, 2018), medical science (Shovkun *et al*., 2018), recommendation and feedback system (Uteuov, 2019; Rahman and Dey, 2018), spam detection (Li *et al*., 2013), text summarization (Chowdhury *et al*., 2017), etc.

Most of the topic modeling algorithms and research papers are specific to the English language. From frequently updated surveys, we can see that internet content is 25.9% in English[1] and this percentage may decrease over the coming years. Therefore, developing similar tools for other languages is essential. Bangla is such language and has become one of the most popular languages in the world after the announcement to annually observe February 21st as the International Mother Language Day by UNESCO on November 17th, 1999. With around about 228 million native speakers and another 37 million as second language speakers, Bangla is the 5th most spoken native language and the 7th by the total number of speakers in the world (Wikipedia, 2020).

Though Bangla is a very popular language in the world, there are barely any Topic Modeling techniques and studies out there to find. So in this SLR, we provide a comprehensive view of topic modeling according to the literature and how algorithms and techniques differ between English and Bangla language. We performed a systematic study to acquaint the methods, domains, datasets, etc. related to topic modeling and showed them in a tabular and diagram form. This process helped us to understand the research community's views and observations about the methods in different domains. We also learned how to evaluate each algorithm and technique through many evaluation matrices, which are also described below.

[1] www.statista.com, www.internetworldstats.com

The paper is outlined as follows: We briefly explained the basics of topic modeling in section 2, then we detailed the review process in section 3 and presented the results of the SLR in section 4. We talked about the challenges and future scopes in section 5. We finally concluded our work in section 7.

Background

This section provides a brief description of the topic modeling methods used in the selected papers. The evaluation methods for topic modeling are also introduced here. It will hopefully give the reader a basic idea about the models included in the reviewed papers.

Topic Modeling Methods

Topic Modeling is an emerging machine learning technology that is widely used in various fields of research (Yuan *et al*., 2015). The basic idea can be simply described as: Documents consist of various topics, which are modeled as distributions over a vocabulary (Arora *et al*., 2013). However, implementing an efficient working algorithm may not be so simple. Various topic modeling algorithms have been developed to work with many technical challenges and diverse text documents (Shi *et al*., 2019). From those, a few of the topic modeling methods used in our reviewed papers are described in brief here.

LDA

Latent Dirichlet Allocation (LDA) is one of the most widely used topic modeling techniques. It is a generative probabilistic model for collections of distinct data like text documents (Blei *et al*., 2003). LDA treats each document as a mixture of different "topics," and each topic is treated as a mixture of different "words" (Li *et al*., 2013). It is a matrix factorization technique and statistical model. The input for LDA is a fixed-length vectors (bag-of-words) (Hasan *et al*., 2019). LDA is very old and there have been many researchers who have modified the basic LDA structure published in (Blei *et al*., 2003) and used the modified versions to their uses (Yuan *et al*., 2015; Ramage *et al*., 2009b; Gao *et al*., 2018; Ramage *et al*., 2009a; Hasan *et al*., 2019).

BTM

Biterm Topic Modeling (BTM) is a useful topic modeling technique when it comes to extracting topics from short texts. As the growing social media platforms are generating a huge amount of short texts, BTM is becoming much more popular to work with short text topic models. The main theme of this technique is that it converts the short text in an unordered pair of words. If two words are frequently co-occurring, then the possibility increases that they are of the same topic (Cheng *et al*., 2014; Li *et al*., 2019a). Several researches have modified the basic BTM and developed more

specific models such as Sentiment Biterm Topic Modeling (SBTM), Multiterm Topic Modeling (MTM).

LSI

Latent Semantic Indexing (LSI) is an automatic retrieval and indexing model for topic modeling, used to identify higher-order structures and categories that associate terms with documents. It tries to find out the hidden semantic structures in documents using word co-occurrence. It uses a linear algebra technique called Singular Value Decomposition (SVD) matrix to identify statistical patterns between words and concepts in a text (Potha and Stamatatos, 2019). LSI tries to capture the many-to-many mapping between terms and concepts, outranking conventional vector-based models (Bertalan and Ruiz, 2019).

GPU-DMM

GPU-DMM model is built by associating Generalized Polya Urn (GPU) model with Dirichlet Multinomial Mixture (DMM). DMM is a probabilistic generative model that uses the assumption that a document is generated from a single topic (Li *et al*., 2016). This assumption enriches the word co-occurrences and makes the model better for short texts (Li *et al*., 2019c). GPU model makes a word pair distribution of semantically related words. For the currently sampled word w, the semantically related words w' are selected such that w' has strong ties with the sampled topic. Because of pairing the sampled word w to its semantically related words, sampling the word w in topic t will also increase the association between the topic and w's semantically related words, not only the association of w itself. In the GPU-DMM model, the generative process is the same as in DMM, but in the inference process GPU model is applied (Li *et al*., 2016).

HDP

Hierarchical Dirichlet Process (HDP), proposed by (Teh *et al*., 2006), is a non-parametric extension of LDA where texts are viewed as groups of observed words, topics are distributions over terms and each document exhibits its topics with different proportions (Bertalan and Ruiz, 2019). HDP infers the number of topics from the documents. This approach provides a prior distribution for the number of mixture components within each group.

SVM

Support Vector Machine (SVM) is a classifier model. It is a machine learning technique that solves the problem like matching patterns, acquiring symbolic theme that depends on syntax as well as semantic meaning (Das and Bandyopadhyay, 2010b). Given a set of training documents, each document marked with a particular category/topic, an SVM training algorithm can be used for topic modeling to categorize documents by assigning new documents into one of the predefined categories/topics (Ahmad and Amin, 2016).

Topic Modeling Evaluation Methods

As we have seen above, there are numerous methods to apply for topic modeling. To evaluate the performances of these models, many evaluation methods are used.

The most used evaluation method is Precision, Recall, F1-Measure (PRF). Confusion Matrix, Area Under Curve (AUC) were used in some papers as well. Another widely used evaluation method is Topic Coherence, which is briefed here.

Topic Coherence (PMI and UCI/U-Mass Scores)

One evaluation method taking off recently is topic coherence, which is calculated based on co-occurrences of words. It is considered a reliable evaluation system since it is highly consistent with human-produced results (Li *et al*., 2016). A popular metric to work with topic coherence is the Pointwise Mutual Information (PMI-Score).

Given the T most probable words of a topic k, (w_1,…, w_T), PMI-Score measures the pairwise association between them:

$$PMI - Score(k) = \frac{1}{T(T-1)} \sum_{1 \le i \le j \le T} PMI(w_i, w_j)$$

where, $PMI(w_i, w_j) = \log \frac{P(w_i, w_j)}{P(w_i)P(w_j)}$, $P(w_i, w_j)$ and $P(w_i)$ are the probabilities of co-occurring word pair (w_i, w_j) and word w_i estimated empirically from the external data sets, respectively (Cheng *et al*., 2014).

Besides PMI, UCI (Newman *et al*., 2010) and U-Mass (Mimno *et al*., 2011) scores are also used to measure topic coherence. The UCI-Coherence is calculated by the following formula (Jiang *et al*., 2017):

$$UCI - coherence(topic) = \sum_{i=2}^{T} \sum_{j=1}^{i-1} PMI(w_i, w_j)$$

Both *PMI* and *UCI* use external sources of large scales, which makes them model-independent. That is why both are fair for all topic models (Cheng *et al*., 2014).

Given the T most probable words of a topic k, (w_1,…,w_T), the U-Mass coherence is calculated by:

$$UMass - coherence(topic) = \sum_{i=2}^{T} \sum_{j=1}^{i-1} \log \frac{p(w_i, w_j) + \frac{1}{M}}{p(w_j)}$$

here, the definition for $p(w_i, w_j)$ and $p(w_j)$ are as described for *PMI*. $\frac{1}{M}$ is the smoothing factor added to avoid the possibility of calculating the logarithm of zero (Jiang *et al*., 2017).

Human Judgement

Human judgement is very reliable for matching extracted topics from a document. But it is not always feasible. Because it is prone to bias since no two human beings will produce the same summary and besides very much time-consuming (Sarkar, 2012b). However, it has been used in several studies (Das and Bandyopadhyay, 2010b; Akter and Aziz, 2016; Abujar *et al*., 2017; Sarkar, 2012a; Efat *et al*., 2013; Sarkar, 2014; Haque *et al*., 2015; Ahmad *et al*., 2018; Shi *et al*., 2019).

Research Methodology

Research methodology gives an overview of how this review process was conducted. This section covers what points we were looking for in the papers, how we searched and collected the papers, what sources they were gathered from, when they were published, what types of papers were collected and which criteria were chosen for paper selection, etc.

Research Questions

The purpose of this Systematic Literature Review (SLR) is to find a proper overview of Topic Modeling and comparison between Bangla and English topic modeling schemes. We asked the questions in Table 1 to extract data from the papers and conduct the review process. The answers to these questions found after reviewing the selected papers are listed and discussed in section 4.

Search Strategy

For searching, we developed some criteria (shown in section 3.4) and followed them. At first:

(a) We searched with some major/key terms related to topic modeling such as ‘topic modeling,’ ‘topic modeling in Bangla,’ ‘topic modeling in Bengali,’ ‘text summarization in Bangla,’ etc.
(b) For every key term, we searched once for English and then added necessary wording to search for Bangla paper in the same context
(c) We also tried out different synonyms and alternate names of these key terms to gather the collection
(d) After collecting the papers, we removed the papers that were duplicate
(e) We collected papers from the references of already included papers and removed duplicate papers again

The search process is illustrated in Fig. 1.

Table 1: Research questions - the questions asked for conducting the review

ID	Questions
RQ 1	What is the most used method for topic modeling?
RQ 2	What are the sources of the datasets used?
RQ 3	What evaluation methods are used to compare the models?
RQ 4	Which are the main fields of application for topic modeling?
RQ 5	What are the techniques that have been used in English topic modeling but not yet used in Bangla?

Table 2: Sources - number of papers from each source

Sources	Finally selected (Initially Selected)
IEEE	24 (28)
ACM Digital Library	15 (18)
Springer	5 (10)
Science Direct	4 (8)
Wiley Online Library	5 (9)
Others	18 (21)
Total	71 (94)

Table 3: Criteria for the paper selection process

Inclusion Criteria
Papers in Bangla and English
Journals and Conference papers are included
Papers describing topic modeling or related algorithms.
For Bangla topic modeling, text summarization and sentimental analysis were emphasized.
Exclusion Criteria
Languages other than Bangla and English
Books, thesis, editorials, prefaces, article summaries
Duplicate Papers
Papers that don’t describe any relevant algorithm.
Review papers

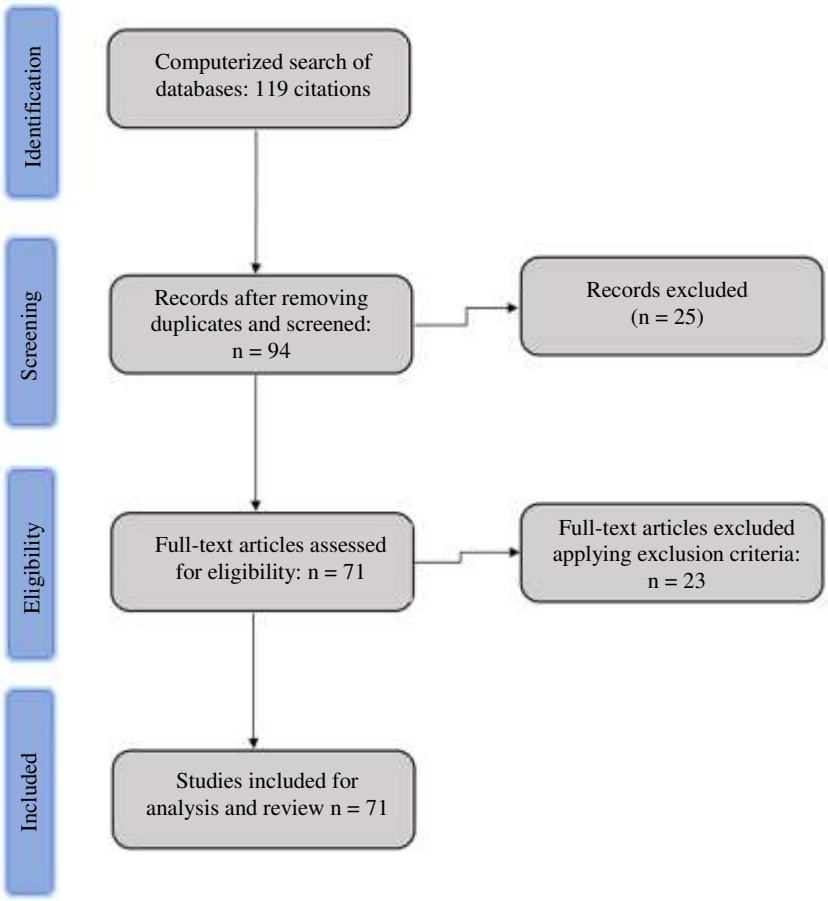

Fig. 1: Paper collection process

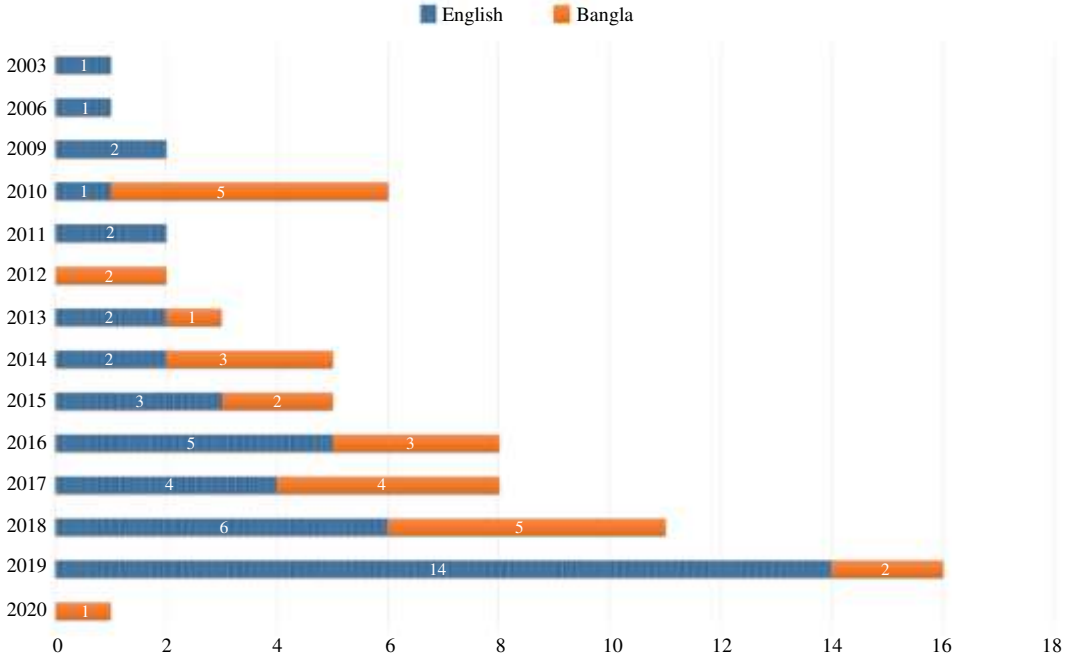

Fig. 2: Distribution of paper over publication year

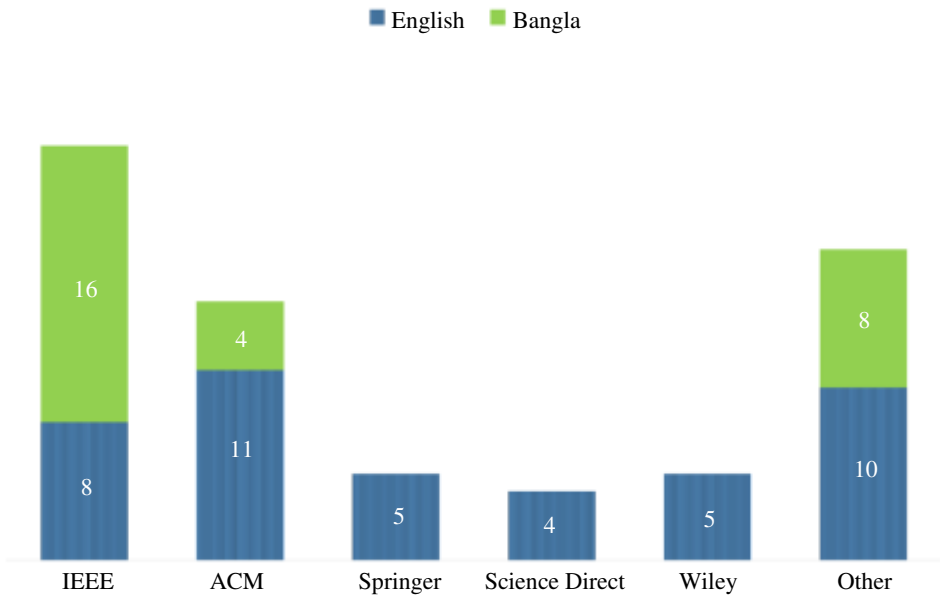

Fig. 3: Paper sources - for English ( ) and Bangla ( )

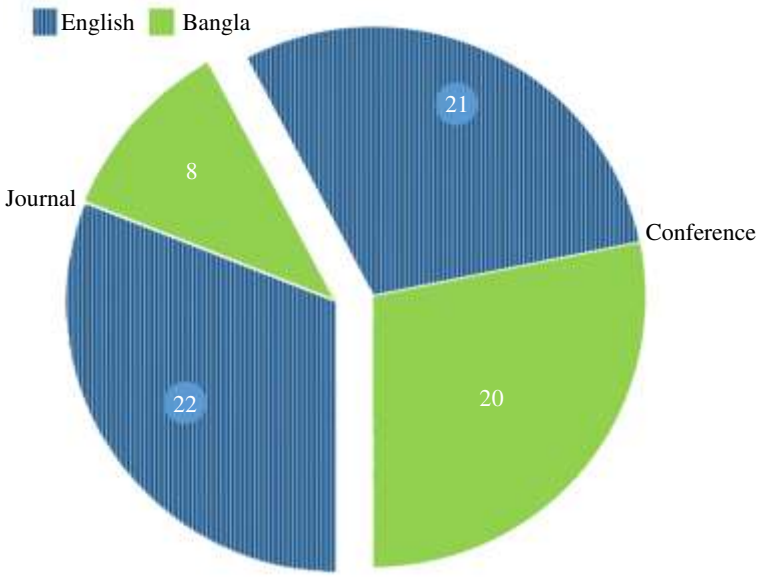

Fig. 4: Types of papers - number of papers from Journals and Conferences for English ( ) and Bangla ( )

Sources

From Table 2 we can see that the main sources of our collected papers were IEEE, ACM Digital Library, ScienceDirect, Springer, Wiley Online Library and Google Scholar. The selected papers were published from January 2003 to May 2020 and their distribution is given in Fig. 2. From Fig. 2, it is apparent that over time the researches on Topic Modeling have increased, especially later in the decade. Figure 3 shows paper sources (Bangla and English separately). We have collected papers published in Journals and Conferences, as shown in Fig. 4.

Selection Criteria

In this section, we arranged our paper selection criteria. Table 3 includes the criteria for including or excluding a paper. Topic Modeling techniques can be used to find the hidden structures of text documents (Al Helal and Mouhoub, 2018). Models used in topic modeling have also been used for text summarization (Chowdhury *et al.*, 2017) and sentiment analysis

(Akter and Aziz, 2016) in Bangla. So as mentioned in the inclusion criteria, apart from topic modeling, we have also included some text summarization and sentiment analysis related papers in Bangla. In Table 3, 'relevant algorithm' means algorithms that are related to topic modeling, text summarization, or sentiment analysis.

Results

In this section, we describe the outcome of this review process. We will go through the results by answering the questions asked in Table 1.

RQ1 - What is the most used method for topic modeling?

Latent Dirichlet Allocation (LDA) is the most used method for topic modeling. 22 of our reviewed papers used LDA for topic modeling, 7 of the other papers modified the basic LDA approach and used them. BTM model stands out in extracting topics from short texts and was used by 6 papers. Another conventional technique LSA (LSI before improved) was used by 5 papers. Many other models are being developed and used by researchers. Table 4 shows all the methods used in our reviewed papers. The most used methods are illustrated in Fig. 5 for easier comparison (separately for English and Bangla).

RQ2 - What are the sources of the datasets used?

There are two major sources of documents that are collected as datasets for the topic modeling techniques. One is the online newspapers and the other is the immensely growing social media sites. Most of the papers collected their data sets from either of these two sources and it is mostly true for both English and Bangla research. Researchers working on English languages collected their documents from sources such as Twitter, NY Times, BaiduQA (A Chinese Q&A website), BBC, Reuters, Yahoo, NIPS and many other online resources. Whereas for Bangla documents, researchers looked up into The Daily Prothom-Alo, The Daily Jugantar, Anandabazar Patrika, Twitter, Facebook, Comments from YouTube, etc. However, collecting a proper dataset in Bangla is challenging, since there are not many standard datasets available in Bangla (Alam *et al.*, 2017). English research has another advantage, that is, using research article Titles, Abstracts as datasets. The same is not possible for Bangla datasets since research articles are not written in Bangla. The datasets used in each of the papers are shown in Table 5 (for English) and Table 6 (Bangla).

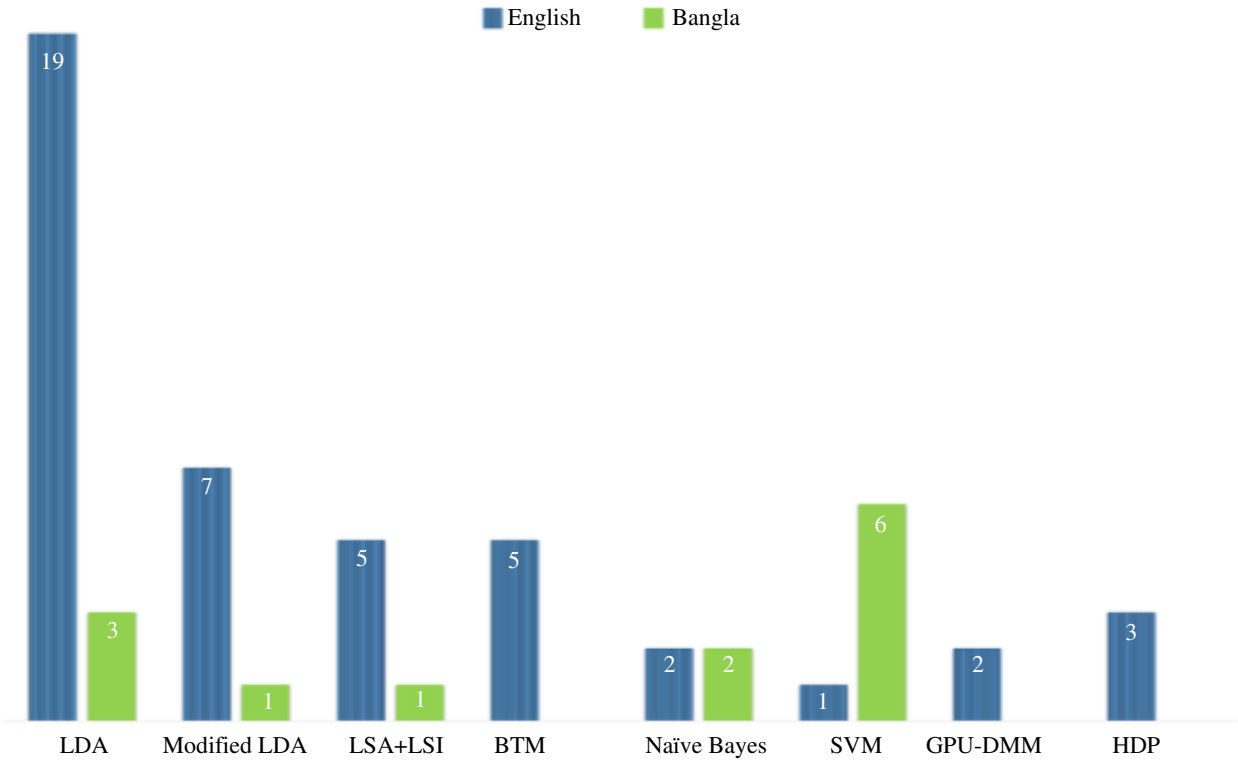

Fig. 5: Usage of models - shows the most used models separately for English ( ) and Bangla ( )

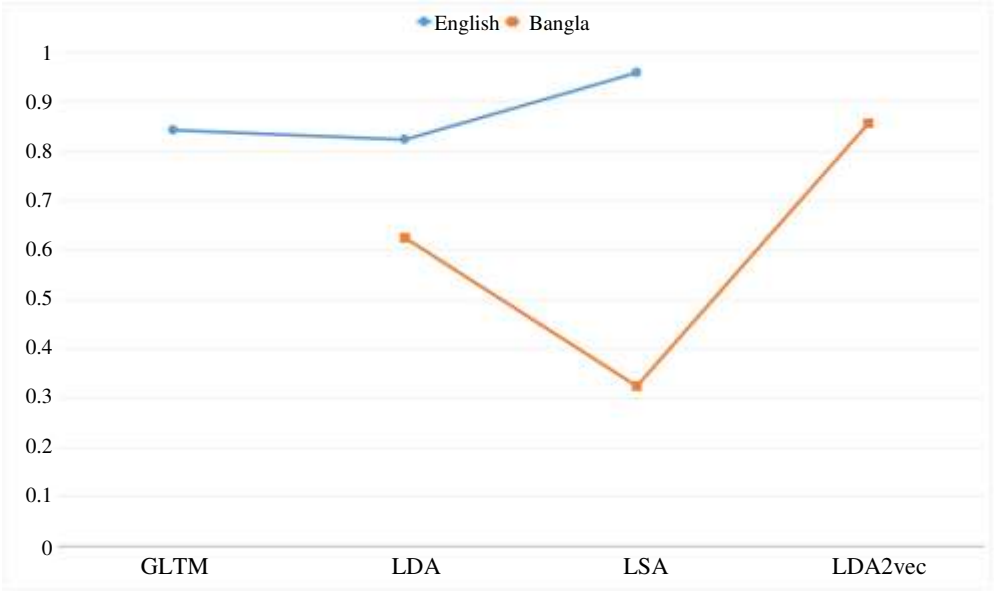

Fig. 6: F1 accuracy of the models

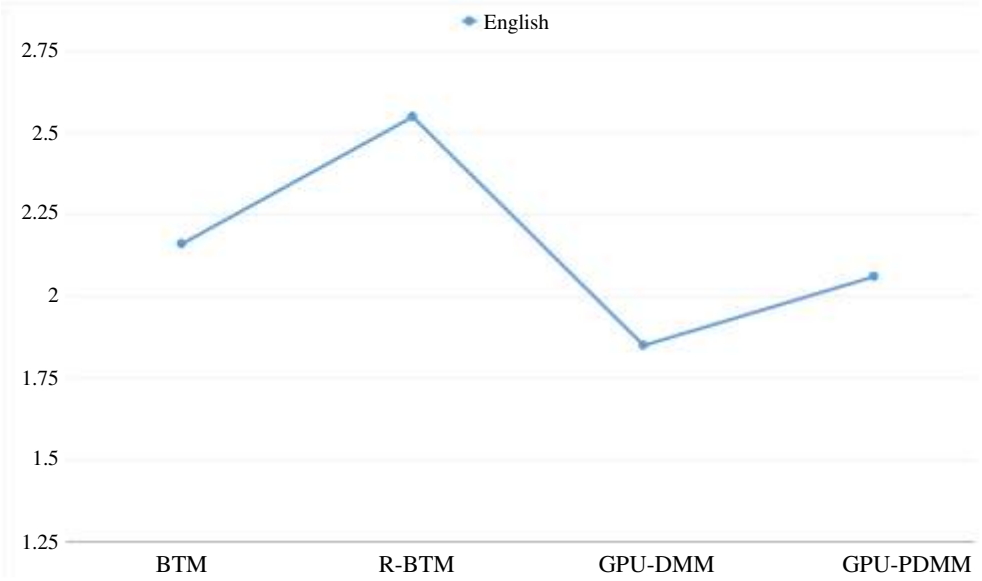

Fig. 7: Topic Coherence (PMI) accuracy of the models

Table 4: Models - lists the year wise usage of the models in the papers

Model	Year	English	Bangla
LDA	2003	Blei *et al.* (2003)	-
	2010	Hong and Davison (2010)	-
	2011	Wang and Blei (2011; Tsai, 2011)	-
	2013	Li *et al.* (2013)	-
	2016	Debortoli *et al.* (2016; Tong and Zhang, 2016)	-
	2017	Ruohonen (2017; Kho *et al.*, 2017)	-
	2018	Hidayatullah *et al.* (2018; Shovkun *et al.*, 2018)	Al Helal and Mouhoub (2018)
	2019	Sun and Platoš (2019; Xiong *et al.*, 2019; Uteuov, 2019; Potha and Stamatatos, 2019; Lesnikowski *et al.*, 2019; Song *et al.*, 2019; Xu *et al.*, 2019; Bertalan and Ruiz, 2019)	Hasan *et al.* (2019)
	2020	-	Sadeq *et al.* (2020)

Table 4: Contd.

Modified LDA	2009	Ramage *et al*. (2009a; 2009b)	-
	2014	Hu *et al*. (2014)	-
	2015	Yuan *et al*. (2015)	-
	2018	Gao *et al*. (2018; Alkhodair *et al*., 2018)	-
	2019	Shi *et al*. (2019)	Hasan *et al*. (2019)
LSA	2017	-	Chowdhury *et al*. (2017)
	2018	Karami *et al*. (2018)	-
	2019	Uteuov (2019)	-
BTM	2014	Cheng *et al*. (2014)	-
	2016	Pang *et al*. (2016)	-
	2019	Li *et al*. (2019a; 2019b)	-
MTM	2019	Wu and Li (2019)	-
Bigram topic model	2006	Wallach (2006)	-
Sentence scoring	2012	-	Sarkar (2012a; 2012b)
	2013	-	Efat *et al*. (2013)
	2014	-	Sarkar (2014)
	2015	-	Haque *et al*. (2015)
Naïve Bayes	2013	Arora *et al*. (2013)	-
	2016	-	Akter and Aziz (2016)
	2017	-	Phani *et al*. (2017)
	2019	Sun and Platoš (2019)	-
SVM	2010	-	Das and Bandyopadhyay (2010b)
	2014	-	Chowdhury and Chowdhury (2014)
	2015	-	Rakshit *et al*. (2015)
	2016	-	Ahmad and Amin (2016)
	2017	-	Phani *et al*. (2017)
	2019	Sun and Platoš (2019)	Bodini (2019)
GPU-DMM	2016	Li *et al*. (2016)	-
	2017	Li *et al*. (2017)	-
PDMM, GPUPDMM	2017	Li *et al*. (2017)	-
Topic Mapping	2019	Shi *et al*. (2019)	-
CNN	2017	-	Alam *et al*. (2017)
	2018	-	Rahman and Dey (2018) Tripto and Ali (2018)
	2019	-	Bodini (2019)
RNN	2016	-	Hassan *et al*. (2016)
	2020	-	Sadeq *et al*. (2020)
HDP	2018	Shovkun *et al*. (2018)	-
	2019	Shi *et al*. (2019; Bertalan and Ruiz, 2019)	-
Fuzzy approach	2018	Karami *et al*. (2018)	-
	2019	Rashid *et al*. (2019)	-
Theme relational graphical representation	2010	-	Das and Bandyopadhyay (2010d; 2010a)
LSTM	2016	-	Hassan *et al*. (2016)
	2018	-	Tripto and Ali (2018)
LASSO	2016	Debortoli *et al*. (2016)	-
SATM	2015	Quan *et al*. (2015)	-
PTM & SPTM	2016	Zuo *et al*. (2016)	-
PDM	2017	Jiang *et al*. (2017)	-
D ETM	2019	Dieng *et al*. (2019)	-
LSI	2019	Sun and Platoš (2019; Bertalan and Ruiz, 2019; Potha and Stamatatos, 2019)	-
GLTM	2018	Liang *et al*. (2018)	-
JST	2015	Lin *et al*. (2015)	-
VSM	2018	-	Roy *et al*. (2018)
Doc2Vec	2018	-	Al Helal and Mouhoub (2018)
Contextual valency analysis	2014	-	Hasan and Rahman (2014)
HCA	2018	-	Ahmad *et al*. (2018)
Extractive summarization	2017	-	Abujar *et al*. (2017)
Unnamed Model	2010	-	Das and Bandyopadhyay (2010c; 2010e)

Table 5: Dataset attributes - for English papers

Paper	Dataset source	Number of docs (train-test split)	Vocabulary	Number of Tokens	Average Doc length
Wang and Blei (2011)	CiteULike.org	16,980	8,000	1.6 M words	N/A
Wallach (2006)	Psychological review (Abstract only)	150 (100-50)	1,374	13,414	N/A
	20Newsgroup[a]	150 (100-50)	2,281	27,478	
Hong and Davison (2010)	Twitter	2M Tweets	N/A	3.7M	N/A
Hu *et al*. (2014)	20Newsgroup	18,770	Top 5,000	632,032	N/A
	NY Times	13,284	Top 5,000	2,714,634	
Cheng *et al*. (2014)	BaiduQA[b]	189,080 Questions	N/A	N/A	3.94
	Tweets2011[c]	4,230,578 Tweets			5.21
	Weibo	155,617,473 Blogpost			5.87
Li *et al*. (2016)	BaiduQA	179,042 Questions	26,560	N/A	4.11
	Web Searches	12,265 Web Searches	5,581		10.72
Shi *et al*. (2019)	Reuters	16,077	N/A	N/A	N/A
	Web of Science (Title and Abstract)	40,526			
	20Newsgroup	37,602			
Quan *et al*. (2015)	NIPS	1,740 (Title, Abstract, Body)	10,297	N/A	N/A
	Yahoo Answers	88,120 Questions	5,972		
Yuan *et al*. (2015)	NY Times	299,752	101,636	99,542,125	332
	Bing 'web chunk'	1.2 Billion Web Pages	1 M words	200 Billion	167
Li *et al*. (2019a)	Amazon Reviews	10,000 (Camera) (90-10%)	3,688	N/A	N/A
		10,000 (Cellphone) (90-10%)	2,797		
		10,000 (Computer) (90-10%)	4,018		
		10,000 (Watch) (90-10%)	2,731		
Zuo *et al*. (2016)	NY Times-Reuters	29,200 news	11,007	N/A	12.4
	Research Papers	55,290 Titles	7,525		6.4
	BaiduQA	142,690 Questions	26,470		4.6
	Twitter	182,671	21,480		
Li *et al*. (2017)	Web Searches	12,265 docs	5,581	N/A	10.72
	BaiduQA	179,042 Questions	26,560		4.11
Arora *et al*. (2013)	NY Times	295,000 docs (59,000)	15,000	N/A	298
	NIPS	1,100 Abstracts (230)	2,500		68
Gao *et al*. (2018)	ZTE STB	1,000 user-7,000 program pair	N/A	N/A	N/A
Li *et al*. (2019b)	Tweets2011	5.42M Tweets	112,000	27.3 M	5.05
	Kaggle.com (Stackoverflow)	3.37M (50,000 Questions)	30,000	18 M	5.34
Tsai (2011)	Nielsen BuzzMetrics	3,096 blog posts	N/A	4,111	N/A
Jiang *et al*. (2017)	NIPS	1,740 (1,557-183)	12,113	13,649	N/A
	Reuters	8,293 (5,946-2,347)		18,933	
	TDT2 Corpus	9,394 News		36,771	
	20 Newsgroup	18,774 News		61,188	
Ruohonen (2017)	Exploit Database (EDB)	36,184 PoC exploits	4,844	N/A	N/A
Karami *et al*. (2018)	Springer medical abstracts	1,527 Abstracts	14,411	245,931	96.3
	Nursing notes	1,607 Notes	11,059	299,449	124.8
	Medical research abstract	2,092 Abstracts	15,768	198,998	95.1
	Medical Tweets	58,927 Tweets	25,310	395,635	6.7
Kho *et al*. (2017)	Breast Cancer Genes	229 Breast Cancer Sample	N/A	23,424 Gene/sample	N/A
	Lung Cancer Genes	98 Lung Cancer Sample		23,996 Gene/sample	
Blei *et al*. (2003)	Reuters	8,000	15,818	N/A	N/A
Pang *et al*. (2016)	SemEval-2007 Dataset	1,246 News	N/A	N/A	N/A
Alkhodair *et al*. (2018)	Twitter (Off Acc dataset)	83,404 Tweets	29,155	N/A	N/A
	Twitter (FashionKW dataset)	38,038 Tweets	35,016		
Wu and Li (2019)	Tweets2011	2,472 Tweets	5,099	N/A	8.56
	Kaggle.com Stackoverflow	19,965 Questions	12,087		5.35
Dieng *et al*. (2019)	UN debates	207,853 (23,097)	12,466	N/A	N/A
	Science Magazine	14,713 (1,634)	25,987		
	ACL Abstracts	9,463 (1,051)	35,108		
Sun and Platoš (2019)	BBC News	2,225 news	N/A	N/A	N/A
Hidayatullah *et al*. (2018)	Football Tweet	120,639 Tweets	N/A	N/A	N/A
Lesnikowski *et al*. (2019)	Conference of Parties speeches	1,315 docs	3,069	N/A	N/A
	Local govt. docs	1,814 docs	21,243		
Liang *et al*. (2018)	Web Searches	12,340	5,432	N/A	14.6
	Amazon Review	19,980	14,331		17.5
	Yahoo Answers	6,310	15,776		117.4
	Tweets2011	30,946	8,536		7.5
Rashid *et al*. (2019)	Web Searches	12,340	30,452	N/A	N/A
	BaiduQA	179,022	26,565		
	Tweets2011	2 M	121,788		
Shovkun *et al*. (2018)	Paitent Question Answers	21,085	N/A	N/A	N/A

Table 5: Contd.

Uteuov (2019)	User Text Group Text User Profile Group Profile Labeled users text	250,000 220,000 400,000 1.6M 8,000	N/A	3.5 M 1.6 M 800 M 1.3 M 200,000	N/A
Xiong *et al*. (2019)	Brazilian Social Network	32,014	N/A	N/A	
Bertalan and Ruiz (2019)	Web of Science Journal (Abstract)	82,248 Abstracts	15,259	N/A	N/A

[a]http://people.csail.mit.edu/jrennie/20Newsgroups/
[b]https://zhidao.baidu.com/
[c]https://trec.nist.gov/data/tweets/

Table 6: Dataset attributes - for Bangla papers

Paper	Dataset source	Number of docs (Train-Test Split)	Vocabulary	Number of Tokens	Average doc length
Roy *et al*. (2018)	Newspaper sites	9,000 docs	N/A	1.8M	N/A
Hasan *et al*. (2019)	News docs (PIPILIKA)	22,675 docs	N/A	N/A	N/A
Rahman and Dey (2018)	Facebook (Cricket) Facebook (Restaurant)	2,985 comments 2,053 comments	N/A	N/A	N/A
Rakshit *et al*. (2015)	Bangla Poems	2,399 poems	N/A	N/A	N/A
Chowdhury and Chowdhury (2014)	Twitter	1,300 (1,000-300)	N/A	N/A	N/A
Hassan *et al*. (2016)	Facebook Twitter YouTube News Sites Review Comments	4,621 2,610 801 1,255 50	N/A	N/A	N/A
Das and Bandyopadhyay (2010b)	Newspaper Sites	20	3,455	5,761	288
Akter and Aziz (2016)	Facebook Posts	3,600	N/A	N/A	N/A
Alam *et al*. (2017)	Online Media Comments	120,000 comments	N/A	N/A	N/A
Sarkar (2012a)	Newspaper sites	38 news docs (28-10)	N/A	N/A	N/A
Tripto and Ali (2018)	YouTube comments	15,689 comments	N/A	N/A	N/A
Efat *et al*. (2013)	Newspaper sites	45 news docs	N/A	N/A	N/A
Das and Bandyopadhyay (2010a)	Newspaper sites	100 news docs	17,166	28,807	288
Al Helal and Mouhoub (2018)	Prothom Alo	7,143 news docs	N/A	N/A	N/A
Haque *et al*. (2015)	Prothom Alo, Jugantar	20 news docs (15,5)	N/A	N/A	N/A
Sarkar (2012b)	Newspaper sites	38 news docs (28, 10)	N/A	N/A	
Bodini (2019)	Facebook (Cricket) Facebook (Restaurant)	2,900 comments 2,600 comments	N/A	N/A	N/A
Phani *et al*. (2017)	Literature	3,000 (1,500-750(dev)-750)	N/A	N/A	N/A
Ahmad *et al*. (2018)	Newspaper sites	500 news docs	N/A	N/A	N/A
Das and Bandyopadhyay (2010e)	SentiWordNet	1,100 sentences	N/A	N/A	N/A
Ahmad and Amin (2016)	Newspaper sites	20,000 news docs	N/A	N/A	N/A
Sadeq *et al*. (2020)	Text Corpus (42 Bangla Websites) Speech Corpus (Google) Speech Corpus (developed)	10M sentences 217,902 utterances 220 hrs 28,973 utterances 50 hrs	N/A	N/A	N/A

Table 7: Evaluation Methods - lists all the evaluation methods and the papers they were used in

Evaluation method	English	Bangla
PRF	Wang and Blei (2011; Liang *et al*., 2018; Gao *et al*., 2018; Ramage *et al*., 2009a; Xu *et al*., 2019; Karami *et al*., 2018; Song *et al*., 2019; Rashid *et al*., 2019; Hong and Davison, 2010)	Chowdhury and Chowdhury (2014; Rahman and Dey, 2018; Sarkar, 2014; Sarkar, 2012a; 2012b; Haque *et al*., 2015; Ahmad and Amin, 2016; Bodini, 2019; Efat *et al*., 2013; Chowdhury *et al*., 2017; Das and Bandyopadhyay, 2010a-e; Ahmad *et al*., 2018; Phani *et al*., 2017)
Topic coherence (PMI)	Cheng *et al*. (2014; Zuo *et al*., 2016) Lesnikowski *et al*. (2019; Li *et al*., 2016; Wu and Li, 2019; Dieng *et al*., 2019; Jiang *et al*., 2017; Alkhodair *et al*., 2018; Bertalan and Ruiz, 2019)	Al Helal and Mouhoub (2018) -
Topic coherence (UCI, UMass)	Jiang *et al*. (2017; Zuo *et al*., 2016; Arora *et al*., 2013)	-
Confusion matrix	Uteuov (2019)	Roy *et al*. (2018; Rakshit *et al*., 2015; Hassan *et al*., 2016)
Classification performance	Ruohonen (2017; Wu and Li, 2019)	Roy *et al*. (2018)
Probability	Tsai (2011; Xiong *et al*., 2019; Tong and Zhang, 2016)	-
L1 Error	Arora *et al*. (2013)	-
Purity Metric	Quan *et al*. (2015)	-
AUC Scores	Potha and Stamatatos (2019)	-

RQ3 - What evaluation methods are used to compare the models?

There are several different evaluation matrices and methods used in the papers we reviewed. The most used methods are Precision, Recall, F1 measure (PRF), Topic Coherence (with PMI, UCI and UMass), Confusion Matrix, Probability, etc. From these, PRF is the most used evaluation metric for topic modeling. 26 of our reviewed papers used PRF. Some papers measured the F1 Score but not the recall and precision; some measured recall only and not the other two. Topic Coherence was used in 13 papers. Also, Purity Metric, L1 Error, Confusion Matrix, etc. evaluation matrices were used by a few of the papers. Table 7 shows which evaluation methods were used by which papers.

In Fig. 6, we have compared a few models which were evaluated using F1-score. Here, LSA acquired 0.9591 (Sun and Platoš, 2019) and LDA2vec acquired 0.8566 (Hasan *et al*., 2019), which are highest in English and Bangla language, respectively. Again, we compared a few models using Topic Coherence matrix in Fig. 7. Here, R-BTM scored highest with 2.55 (Li *et al*., 2019b). As different evaluation matrices were used to evaluate different models, the acquired model-accuracy data was inadequate to represent all the models in the figure. So, only the most used models are shown in Fig. 6 and. 7.

Some of the papers did not use any proper evaluation system. A few of them determined accuracy by comparison and other papers just provided the determined topics by their respective system without evaluating the quality.

RQ4 - Which are the main fields of application for topic modeling?

To find and extract information from vast collections of documents is a very hard, toiling and time-consuming process. So, to find individual documents from large document collections and to understand the general themes present in the collection, topic models are used as a statistical framework. How topic modeling can be used for researches in real-world circumstances (According to papers in our review) is given in Table 8.

Information from Social Sites

Nowadays, social media sites are significant sources of data. Many research works in topic modeling have focused on the use of these data. To get the required data from the websites, compatible APIs (Application Programming Interface) are used. These data may include web page titles, image captions, questions in Q&A websites, text advertisements and posts, messages, tweets in social media sites. LDA and BTM models were then used as standard tools for topic modeling on those data (Hong and Davison, 2010; Tong and Zhang, 2016; Li *et al*., 2019b; Cheng *et al*., 2014). There is another version of LDA, modified to better work on Twitter-dataset, called Twitter-LDA (Alkhodair *et al*., 2018).

Linguistic Science

Understanding the underlying meanings of texts and accordingly classifying the documents is an essential task in linguistic science. Similarly, to understand the emotions or sentiments of documents is also a part of it. Various topic modeling algorithms, e.g., LDA, SVM, Doc2Vec are used for document classifications and SVM, Long Short Term Memory (LSTM), Recurrent Neural Network (RNN) and Contextual Valency analysis (CVA) are used in sentiment analysis.

Author Verification

Author Verification system enables an author to check in which online or offline documents s/he has given the right to use his/her writing. Through this system, no particular organization or person can use another person's writing without authorization. LDA, Pachinko Allocation and Hierarchical LDA (Phani *et al*., 2017), LSI (Potha and Stamatatos, 2019) algorithms are used in this type of systems.

Medical/Biomedical Science

As most of the things are in digitized form, even all the medical and biomedical fields use all sorts of digital documents to conduct their work and research. To diagnose cancer and get data from gene expression or sequence, LDA was used (Kho *et al*., 2017).

Scientific Literature

Online archives are now the most common platform for research articles. While searching for necessary research articles, relevant search results are vital. So, in that case, topic modeling can be beneficial to get useful information. Topic modeling can be further extended to recommend similar articles and documents (Wang and Blei, 2011).

Recommendation System

A user's particular interest field can be predicted by using topic modeling. This process can help to recommend similar sorts of things to the user (Uteuov, 2019). Also, suggesting a product by extracting information from that particular product's review through topic modeling is very sophisticated. In this type of system, BTM is one of the most preferred algorithms (Li *et al*., 2019a).

Political Science

For politicians and political professionals, knowing the main debating topics of mass people is a valuable asset. People nowadays express their thoughts on social media or other online platforms and hence, people's political views can be extracted from those platforms. To extract information from various political websites LSI, LDA and Hierarchical Dirichlet Process (HDP) were used (Bertalan and Ruiz, 2019).

RQ5 - What are the techniques that have been used in English topic modeling but not yet used in Bangla?

There have not been many works in topic modeling in Bangla yet and the few models that have been used are basic models such as LDA, LSA and in some cases, classification models like SVM, Convolutional Neural Network (CNN), etc. There are so many models in English that are yet to be tried for Bangla. BTM, GPU-DMM, Generalized Polya Urn Poisson-based Dirichlet Multinomial Mixture (GPU-PDMM), Global and Local word embedding-based Topic Modeling (GLTM), Self Aggregation based Topic Model (SATM), Pseudo-document-based Topic Model (PTM), Word Embedding Approach, Joint Sentiment Topic Modeling (JST), Poisson Distribution Model (PDM) are some of the models that have been used in English but not yet in Bangla.

Table 8: Applications of Topic Modeling - in which areas the topic modeling techniques were applied in

Application field	Model	English	Bangla
Topic extraction (short text)	LDA	Hong and Davison (2010)	-
	Twitter-LDA	Alkhodair *et al*. (2018)	-
	SATM	Quan *et al*. (2015)	-
	MTM	Wu and Li (2019)	-
	GPU-DMM	Li *et al*. (2016; 2017)	-
	GPU-PDMM	Li *et al*. (2017)	-
	SBTM	Pang *et al*. (2016)	-
	BTM	Li *et al*. (2019b)	-
	PTM & SPTM	Zuo *et al*. (2016)	-
Topic extraction (large text)	LDA	Xu *et al*. (2019; Tsai, 2011; Hidayatullah *et al*., 2018; Blei *et al*., 2003)	Hasan *et al*. (2019)
	D-ETM	Dieng *et al*. (2019)	-
	FTM	Rashid *et al*. (2019)	-
	VSM	-	Roy *et al*. (2018)
Classification and clustering	LDA	Ruohonen (2017)	Al Helal and Mouhoub (2018)
	LSI	Sun and Platoš (2019)	-
	SVM	Sun and Platoš (2019)	Rakshit *et al*. (2015; Ahmad and Amin, 2016)
	WDM	-	Ahmad *et al*. (2018)
Medical science	FLSA	Karami *et al*. (2018)	-
	LDA	Kho *et al*. (2017; Song *et al*., 2019; Shovkun *et al*., 2018)	-
	LSA, HDP	Shovkun *et al*. (2018)	-
Political interest	LDA	Lesnikowski *et al*. (2019)	-
	LSA, HDP	Bertalan and Ruiz (2019)	-
Author verification	LDA	Potha and Stamatatos (2019)	Phani *et al*. (2017)
	LSI	Potha and Stamatatos (2019)	-
	HLDA	-	Phani *et al*. (2017)
Social sciences	LDA	Shovkun *et al*. (2018; Ramage *et al*., 2009b)	-
	LSA, HDP	Shovkun *et al*. (2018)	-
Sentiment analysis	JSTM	Lin *et al*. (2015)	-
	SBTM	Pang *et al*. (2016)	-
	SVM	-	Das and Bandyopadhyay (2010b)
		-	Chowdhury and Chowdhury (2014)
	LSTM	-	Tripto and Ali (2018; Hassan *et al*., 2016)
	RNN	-	Hassan *et al*. (2016)
	Naıve Bayes	-	Akter and Aziz (2016)
	CVA	-	Hasan and Rahman (2014)
	CNN	-	Alam *et al*. (2017; Tripto and Ali, 2018)
	Unnamed model	-	Das and Bandyopadhyay (2010c; 2010e)
Voice recognition	RNN	-	Sadeq *et al*. (2020)
	Labeled-LDA		Sadeq *et al*. (2020)
Spam detection	LDA	Li *et al*. (2013)	-
Recommendation and feedback system	LDA	Uteuov (2019)	-
		Wang and Blei (2011)	-
	cLDA	Gao *et al*. (2018)	-
	PLSA, ARTM	Uteuov (2019)	-
	Seeded-BTM	Li *et al*. (2019a)	-
	Vanilla-LDA	Hu *et al*. (2014)	-
	CNN	-	Rahman and Dey (2018; Bodini, 2019)
	SVM	-	Bodini (2019)
Scientific Literature	LDA	Xiong *et al*. (2019; Wang and Blei, 2011; Debortoli *et al*., 2016)	-
	LASSO	Debortoli *et al*. (2016)	-
Text Summarization	LSA	-	Chowdhury *et al*. (2017)
	TRGR*	-	Das and Bandyopadhyay (2010d; 2010a)
	Sentence scoring	-	Sarkar (2012b; 2012a; 2014; Efat *et al*., 2013; Haque *et al*., 2015)
	Extractive summarization	-	Abujar *et al*. (2017)

* Theme relational graphical representation

Table 9: Indic comparison - comparing the use of topic modeling in Bangla to other Indic Languages such as Hindi and Urdu.

Topic modeling in other Indic languages	Topic modeling in Bangla
New topic modeling algorithms such as Lexical LDA (Lex-LDA), Sliding Window-based Weighting LDA (LDASWSW), relative sentence weighting LDA(LDA-RSW), Integrated Sentence weighting LDA (LDA-ISW) (Rani and Lobiyal, 2020), Non-negative Matrix Factorization (NMF) (Ray *et al*., 2019) were used.	Only LDA and LDA2vec were used.
Application areas include Music Mood Classification (Chauhan and Chauhan, 2016), Statistical Machine Translation.	Application areas include topic extraction and news classification.
Since morphology, structure and syntax in Urdu is different, a special modified version of LDA is used named ULDA (Urdu-LDA) (Shakeel *et al*., 2018).	There is no specific version of LDA dedicated for Bangla languages' morphology, structure and syntax.
The number of research in this field is minimal.	The same is true for Bangla as well.

This paper mainly analyzed the studies of topic modeling in English and Bangla language. However, we think ideas from studies in languages closely related to Bangla can also help improve the understanding of the current state of topic modeling in Bangla. Hence, we collected papers in some other Indic languages (Hindi and Urdu) and compared them with existing Bangla works. The comparison results are shown in Table 9.

Discussion

In the results section, we have seen that LDA is the most used topic modeling technique for both Bangla and English language. One of the reasons behind the widespread use of LDA is that it is very flexible. LDA can be combined with many other models to perform tasks such as classifications, summarization, clustering, spam detection, etc. LDA has been used even out of the scope of NLP in Computer Vision to color naming (Benavente *et al*., 2012). LDA is time-efficient compared to many other topic modeling techniques. It is also unsupervised, which makes it a good choice while working with unlabelled data. The basic LDA is an old model and many task-specific models were innovated from the basic LDA later. LDA has been tested on many circumstances, domains and datasets and it has provided good results, which makes it reliable. However, there are also some drawbacks associated with the model. The most important of them is that LDA tends to work poorly if the input documents are very short in word length. With the rise of social media, text mining for short texts is becoming essential. Although modifications of the basic LDA like Twitter-LDA (Alkhodair *et al*., 2018) were created, yet the results need improvements. LDA also fails to provide satisfactory results if a document does not consistently discuss a single topic. LDA does not build any correlation between words. These are some areas where LDA still needs improvements. Although LDA is not the single best state-of-the-art model for topic modeling, it is still a good choice under most circumstances for topic modeling.

In the above discussion, we have talked about the advantages and disadvantages of only LDA model. But overall, the topic modeling techniques also have some challenges that need to be addressed. Topic modeling algorithms mainly focus on frequently co-occurring words. The semantic meaning of a word may change according to the context it is used in. But topic modeling algorithms treat a word the same in every context, which adds noise to the word distributions. Another issue with some of the models is that the number of topics needs to be specified before training. But it is not possible to know how many topics will work best beforehand. This leads to iterating over the dataset and trying out different numbers of topics, which is time-consuming. The evaluation methods for topic modeling may test if the model is working but cannot give an absolute measurement of the models' overall quality. So, many topic modeling applications need manual checking or other extrinsic evaluations (if labeled data is available). The studies of topic modeling in Bangla used only the LDA model and its modifications. No other models were experimented with for topic modeling in Bangla.

In light of the challenges discussed above and many other possibilities, topic modeling has an open field for future research. With the rise of social media, a large portion of the generated data is in the form of short texts. Conventional topic modeling methods are found to be performing poorly on short texts. Recently, some researchers have tried developing a few models to work on short texts. Topic modeling on short texts is an area for future research. Evaluation methods for topic modeling are not still well established. Topic coherence is an intrinsic evaluation method for topic modeling that can provide only a relative measure of performance between two models. For absolute measurements, metadata of extrinsic applications (text classification, sentiment analysis, etc.) are used (Shi *et al*., 2019). Developing evaluation methods for topic modeling can be a potential research area. Another exciting application of topic modeling techniques can be the medical domain. Already in (Kho *et al*., 2017), researchers have used topic modeling to understand the genetic expressions of cancer cells. Also, topic modeling can be useful to process medical big-data. In Bangla, the scope for future research is even broader. The scarcity of topic modeling research in Bangla leaves many potential areas untouched. Recommendation system, document classification, sentiment analysis, detecting and tracking

trending topics in social platforms, trigger word or voice command recognition and many other NLP tasks in Bangla can be performed using topic modeling.

Merits and Demerits of the Studies

During this review process, we have encountered papers that have both advantages and shortcomings. In most of the papers, the proposed models were well defined and had detailed explanations in them. The overall difference between English and Bangla Topic Modeling is very apparent because of the lack of research in the Bangla language. Thus, it was not difficult to draw comparisons between them. Also, the purpose and motivation of the authors were properly mentioned in all papers.

On the other hand, different evaluation matrices were used in different papers, which makes it difficult to compare the models together. Very few papers shared the same evaluation matrix for the same model. Moreover, attributes of the datasets (i.e., size, vocabulary and other trivial parameters) were not always properly described (Especially the Bangla papers). So, it was not easy to collect and organize that information from those papers. Authors of Bangla papers should give more attention to representing the datasets properly.

Conclusion

We analyzed the current state of topic modeling and the lack of study done in Bangla language topic modeling. After exhaustively searching for papers, we finally selected 71 papers from an initial collection of 94 papers for review. These papers were published between 2003 and 2020. We gathered data concerning several aspects such as method types, datasets, evaluation methods, application fields, etc. These extracted data were later used to answer the specified research questions and give proper insight into this field.

In our reviewed papers, the LDA method stands out as the most commonly used topic modeling technique. Furthermore, the BTM model performs best in extracting topics from short texts. A variety of evaluation methods were used to judge the performances of the models. Precision, recall rate, F1-score are the most used evaluating systems. Another explicit topic model evaluating method, Topic Coherence, was also used by many researchers. Besides the models and the evaluation methods, we also highlighted the field of topic modeling applications.

We believe this paper will help researchers to have a straightforward overview of topic modeling. There is a wide range of scopes available for topic modeling in the Bangla language compared to English as described in section 5. By reading this paper, researchers can easily identify the gaps between English and Bangla topic modeling and conduct further research in this emerging field.

Author's Contributions

Md. Basim Uddin Ahmed: Collected the papers for review, analyzed the papers, organized the figures and tables, drafted the manuscript.

Ananta Akash Podder: Analyzed, extracted and organized all required information from the paper for review.

Mahruba Sharmin Chowdhury: Made considerable contributions to this research by critically reviewing the manuscript for significant intellectual content.

Mohammad Abdullah Al Mumin: Verified the works, reviewed the manuscript and supervised the whole project.

Conflict of Interest

The authors declare that they have no Conflict of Interest.

References

Abujar, S., Hasan, M., Shahin, M. S. I., & Hossain, S. A. (2017, July). A heuristic approach of text summarization for Bengali documentation. In 2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT) (pp. 1-8). IEEE.

Ahmad, A., & Amin, M. R. (2016, December). Bengali word embeddings and it's application in solving document classification problem. In 2016 19th International Conference on Computer and Information Technology (ICCIT) (pp. 425-430). IEEE.

Ahmad, A., Amin, M. R., & Chowdhury, F. (2018, September). Bengali document clustering using word movers distance. In 2018 International Conference on Bangla Speech and Language Processing (ICBSLP) (pp. 1-6). IEEE.

Akter, S., & Aziz, M. T. (2016, September). Sentiment analysis on facebook group using lexicon based approach. In 2016 3rd International Conference on Electrical Engineering and Information Communication Technology (ICEEICT) (pp. 1-4). IEEE.

Al Helal, M., & Mouhoub, M. (2018). Topic modelling in bangla language: An lda approach to optimize topics and news classification. Computer and Information Science, 11(4).

Alam, M. H., Rahoman, M. M., & Azad, M. A. K. (2017, December). Sentiment analysis for Bangla sentences using convolutional neural network. In 2017 20th International Conference of Computer and Information Technology (ICCIT) (pp. 1-6). IEEE.

Alkhodair, S. A., Fung, B. C., Rahman, O., & Hung, P. C. (2018). Improving interpretations of topic modeling in microblogs. Journal of the Association for Information Science and Technology, 69(4), 528-540.

Arora, S., Ge, R., Halpern, Y., Mimno, D., Moitra, A., Sontag, D., ... & Zhu, M. (2013, February). A practical algorithm for topic modeling with provable guarantees. In International Conference on Machine Learning (pp. 280-288).

Benavente, R., Van de Weijer, J., Vanrell, M., Schmid, C., Baldrich, R., Verbeek, J., & Larlus, D. (2012). Color names.

Bertalan, V. G., & Ruiz, E. E. S. (2019, October). Using topic modeling to find main discussion topics in brazilian political websites. In Proceedings of the 25th Brazillian Symposium on Multimedia and the Web (pp. 245-248).

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. Journal of machine Learning research, 3(Jan), 993-1022.

Bodini, M. (2019). Aspect Extraction from Bangla Reviews Through Stacked Auto-Encoders. Data, 4(3), 121.

Chauhan, S., & Chauhan, P. (2016, October). Music mood classification based on lyrical analysis of Hindi songs using Latent Dirichlet Allocation. In 2016 International Conference on Information Technology (InCITe)-The Next Generation IT Summit on the Theme-Internet of Things: Connect your Worlds (pp. 72-76). IEEE.

Cheng, X., Yan, X., Lan, Y., & Guo, J. (2014). Btm: Topic modeling over short texts. IEEE Transactions on Knowledge and Data Engineering, 26(12), 2928-2941.

Chowdhury, S. R., Sarkar, K., & Dam, S. (2017, December). An Approach to Generic Bengali Text Summarization Using Latent Semantic Analysis. In 2017 International Conference on Information Technology (ICIT) (pp. 11-16). IEEE.

Chowdhury, S., & Chowdhury, W. (2014, May). Performing sentiment analysis in Bangla microblog posts. In 2014 International Conference on Informatics, Electronics & Vision (ICIEV) (pp. 1-6). IEEE.

Das, A., & Bandyopadhyay, S. (2010a, August). Opinion summarization in Bengali: a theme network model. In 2010 IEEE Second International Conference on Social Computing (pp. 675-682). IEEE.

Das, A., & Bandyopadhyay, S. (2010b). Phrase-level polarity identification for Bangla. Int. J. Comput. Linguist. Appl.(IJCLA), 1(1-2), 169-182.

Das, A., & Bandyopadhyay, S. (2010c). Sentiwordnet for bangla. Knowledge Sharing Event-4: Task, 2, 1-8.

Das, A., & Bandyopadhyay, S. (2010d, August). Topic-based Bengali opinion summarization. In Coling 2010: Posters (pp. 232-240).

Das, D., & Bandyopadhyay, S. (2010e). Identifying emotion holder and Topic from Bengali emotional sentences. ICON.

Debortoli, S., Müller, O., Junglas, I., & vom Brocke, J. (2016). Text mining for information systems researchers: An annotated topic modeling tutorial. Communications of the Association for Information Systems, 39(1), 7.

Dieng, A. B., Ruiz, F. J., & Blei, D. M. (2019). The dynamic embedded topic model. arXiv preprint arXiv:1907.05545.

Efat, M. I. A., Ibrahim, M., & Kayesh, H. (2013, May). Automated Bangla text summarization by sentence scoring and ranking. In 2013 International Conference on Informatics, Electronics and Vision (ICIEV) (pp. 1-5). IEEE.

Gao, Y., Wei, X., Zhang, X., & Zhuang, W. (2018). A combinational LDA-based topic model for user interest inference of energy efficient IPTV service in smart building. IEEE Access, 6, 48921-48933.

Haque, M. M., Pervin, S., & Begum, Z. (2015, December). Automatic Bengali news documents summarization by introducing sentence frequency and clustering. In 2015 18th International Conference on Computer and Information Technology (ICCIT) (pp. 156-160). IEEE.

Hasan, K. A., & Rahman, M. (2014, December). Sentiment detection from bangla text using contextual valency analysis. In 2014 17th International Conference on Computer and Information Technology (ICCIT) (pp. 292-295). IEEE.

Hasan, M., Hossain, M. M., Ahmed, A., & Rahman, M. S. (2019). Topic modelling: A comparison of the performance of latent dirichl*et al*location and lda2vec model on bangla newspaper. In 2019 International Conference on Bangla Speech and Language Processing (ICBSLP), pages 1–5. IEEE.

Hassan, A., Amin, M. R., Al Azad, A. K., & Mohammed, N. (2016, December). Sentiment analysis on bangla and romanized bangla text using deep recurrent models. In 2016 International Workshop on Computational Intelligence (IWCI) (pp. 51-56). IEEE.

Hidayatullah, A. F., Pembrani, E. C., Kurniawan, W., Akbar, G., & Pranata, R. (2018, April). Twitter topic modeling on football news. In 2018 3rd International Conference on Computer and Communication Systems (ICCCS) (pp. 467-471). IEEE.

Hong, L., & Davison, B. D. (2010, July). Empirical study of topic modeling in twitter. In Proceedings of the first workshop on social media analytics (pp. 80-88).

Hu, Y., Boyd-Graber, J., Satinoff, B., & Smith, A. (2014). Interactive topic modeling. Machine learning, 95(3), 423-469.

Jiang, H., Zhou, R., Zhang, L., Wang, H., & Zhang, Y. (2017, November). A topic model based on Poisson decomposition. In Proceedings of the 2017 ACM on Conference on Information and Knowledge Management (pp. 1489-1498).

Karami, A., Gangopadhyay, A., Zhou, B., & Kharrazi, H. (2018). Fuzzy approach topic discovery in health and medical corpora. International Journal of Fuzzy Systems, 20(4), 1334-1345.

Kho, S. J., Yalamanchili, H. B., Raymer, M. L., & Sheth, A. P. (2017, August). A novel approach for classifying gene expression data using topic modeling. In Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics (pp. 388-393).

Lesnikowski, A., Belfer, E., Rodman, E., Smith, J., Biesbroek, R., Wilkerson, J. D., ... & Berrang-Ford, L. (2019). Frontiers in data analytics for adaptation research: Topic modeling. Wiley Interdisciplinary Reviews: Climate Change, 10(3), e576.

Li, C., Duan, Y., Wang, H., Zhang, Z., Sun, A., & Ma, Z. (2017). Enhancing topic modeling for short texts with auxiliary word embeddings. ACM Transactions on Information Systems (TOIS), 36(2), 1-30.

Li, C., Wang, H., Zhang, Z., Sun, A., & Ma, Z. (2016). Topic modeling for short texts with auxiliary word embeddings. In Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval, pages 165–174. ACM.

Li, J., Cardie, C., & Li, S. (2013). Topicspam: a topic-model based approach for spam detection. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 217–221.

Li, N., Chow, C.-Y., & Zhang, J.-D. (2019a). Seededbtm: enabling biterm topic model with seeds for product aspect mining. In 2019 IEEE 21st International Conference on High Performance Computing and Communications; IEEE 17th International Conference on Smart City; IEEE 5th International Conference on Data Science and Systems (HPCC/SmartCity/DSS), pages 2751–2758. IEEE.

Li, X., Zhang, A., Li, C., Guo, L., Wang, W., & Ouyang, J. (2019b). Relational biterm topic model: Short-text topic modeling using word embeddings. The Computer Journal, 62(3), 359–372.

Li, X., Zhang, J., & Ouyang, J. (2019c). Dirichlet multinomial mixture with variational manifold regularization: Topic modeling over short texts. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 33, pages 7884–7891.

Liang, W., Feng, R., Liu, X., Li, Y., & Zhang, X. (2018). Gltm: A global and local word embedding-based topic model for short texts. IEEE Access, 6, 43612-43621.

Lin, C., Ibeke, E., Wyner, A., & Guerin, F. (2015). Sentiment–topic modeling in text mining. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 5(5), 246–254.

Mimno, D., Wallach, H., Talley, E., Leenders, M., & McCallum, A. (2011). Optimizing semantic coherence in topic models. In Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, pages 262–272.

Newman, D., Noh, Y., Talley, E., Karimi, S., & Baldwin, T. (2010). Evaluating topic models for digital libraries. In Proceedings of the 10th annual joint conference on Digital libraries, pages 215–224.

Pang, J., Li, X., Xie, H., & Rao, Y. (2016). Sbtm: Topic modeling over short texts. In International Conference on Database Systems for Advanced Applications, pages 43–56. Springer.

Phani, S., Lahiri, S., & Biswas, A. (2017). A supervised learning approach for authorship attribution of bengali literary texts. ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP), 16(4), 28.

Potha, N., & Stamatatos, E. (2019). Improving author verification based on topic modeling. Journal of the Association for Information Science and Technology, 70(10), 1074–1088.

Quan, X., Kit, C., Ge, Y., & Pan, S. J. (2015, June). Short and sparse text topic modeling via self-aggregation. In Twenty-fourth international joint conference on artificial intelligence.

Rahman, M. A., & Dey, E. K. (2018). Aspect extraction from bangla reviews using convolutional neural network. In 2018 Joint 7th International Conference on Informatics, Electronics & Vision (ICIEV) and 2018 2nd International Conference on Imaging, Vision & Pattern Recognition (icIVPR), pages 262–267. IEEE.

Rakshit, G., Ghosh, A., Bhattacharyya, P., & Haffari, G. (2015). Automated analysis of bangla poetry for classification and poet identification. In Proceedings of the 12th International Conference on Natural Language Processing, pages 247–253.

Ramage, D., Hall, D., Nallapati, R., & Manning, C. D. (2009a). Labeled lda: A supervised topic model for credit attribution in multi-labeled corpora. In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1, pages 248–256. Association for Computational Linguistics.

Ramage, D., Rosen, E., Chuang, J., Manning, C. D., & McFarland, D. A. (2009b). Topic modeling for the social sciences. In NIPS 2009 workshop on applications for topic models: text and beyond, volume 5, page 27.

Rani, R., & Lobiyal, D. (2020). An extractive text summarization approach using tagged-lda based topic modeling. Multimedia Tools and Applications, pages 1–31.

Rashid, J., Shah, S. M. A., & Irtaza, A. (2019). Fuzzy topic modeling approach for text mining over short text. Information Processing & Management, 56(6), 102060.

Ray, S. K., Ahmad, A., & Kumar, C. A. (2019). Review and implementation of topic modeling in hindi. Applied Artificial Intelligence, 33(11), 979–1007.

Roy, T. D., Khatun, S., & Begum, R. (2018, October). Vector space model based topic retrieval from bengali documents. In 2018 International Conference on Innovations in Science, Engineering and Technology (ICISET) (pp. 60-63). IEEE.

Ruohonen, J. (2017). Classifying web exploits with topic modeling. In 2017 28th International Workshop on Database and Expert Systems Applications (DEXA), pages 93–97. IEEE.

Sadeq, N., Ahmed, S., Shubha, S. S., Islam, M. N., & Adnan, M. A. (2020). Bangla voice command recognition in end-to-end system using topic modeling based contextual rescoring. In ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 7894–7898. IEEE.

Sarkar, K. (2012a). An approach to summarizing bengali news documents. In proceedings of the International Conference on Advances in Computing, Communications and Informatics, pages 857–862. ACM.

Sarkar, K. (2012b). Bengali text summarization by sentence extraction. arXiv preprint arXiv:1201.2240.

Sarkar, K. (2014). A keyphrase-based approach to text summarization for english and bengali documents. International Journal of Technology Diffusion (IJTD), 5(2), 28–38.

Shakeel, K., Tahir, G. R., Tehseen, I., & Ali, M. (2018, January). A framework of Urdu topic modeling using latent dirichlet allocation (LDA). In 2018 IEEE 8th Annual Computing and Communication Workshop and Conference (CCWC) (pp. 117-123). IEEE.

Shi, H., Gerlach, M., Diersen, I., Downey, D., & Amaral, L. A. (2019). A new evaluation framework for topic modeling algorithms based on synthetic corpora. arXiv preprint arXiv:1901.09848.

Shovkun, M., Fleischmann, K. R., & Xie, B. (2018). Computational social science using topic modeling: Analyzing patients' values using a large hospital survey. Proceedings of the Association for Information Science and Technology, 55(1), 892–893.

Song, C. W., Jung, H., & Chung, K. (2019). Development of a medical big-data mining process using topic modeling. Cluster Computing, 22(1), 1949–1958.

Sun, Y., & Platoš, J. (2019, November). Text Classification Based on Topic Modeling and Chi-square. In International Conference on Genetic and Evolutionary Computing (pp. 513-520). Springer, Singapore.

Teh, Y. W., Jordan, M. I., Beal, M. J., & Blei, D. M. (2006). Hierarchical dirichlet processes. Journal of the American Statistical Association, 101(476), 1566–1581.

Tong, Z., & Zhang, H. (2016, May). A text mining research based on LDA topic modelling. In International Conference on Computer Science, Engineering and Information Technology (pp. 201-210).

Tripto, N. I., & Ali, M. E. (2018, September). Detecting multilabel sentiment and emotions from bangla youtube comments. In 2018 International Conference on Bangla Speech and Language Processing (ICBSLP) (pp. 1-6). IEEE.

Tsai, F. S. (2011). A tag-topic model for blog mining. Expert Systems with Applications, 38(5), 5330–5335.

Uteuov, A. (2019). Topic model for online communities' interests prediction. Procedia Computer Science, 156, 204–213.

Wallach, H. M. (2006, June). Topic modeling: beyond bag-of-words. In Proceedings of the 23rd international conference on Machine learning (pp. 977-984).

Wang, C., & Blei, D. M. (2011, August). Collaborative topic modeling for recommending scientific articles. In Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 448-456).

Wikipedia. (2020). Bengali language. Wikipedia. https://en.wikipedia.org/wiki/Bengali_language.

Wu, X., & Li, C. (2019, July). Short Text Topic Modeling with Flexible Word Patterns. In 2019 International Joint Conference on Neural Networks (IJCNN) (pp. 1-7). IEEE.

Xiong, H., Cheng, Y., Zhao, W., & Liu, J. (2019). Analyzing scientific research topics in manufacturing field using a topic model. Computers & Industrial Engineering, 135, 333–347.

Xu, G., Meng, Y., Chen, Z., Qiu, X., Wang, C., & Yao, H. (2019). Research on topic detection and tracking for online news texts. IEEE Access, 7, 58407–58418.

Yuan, J., Gao, F., Ho, Q., Dai, W., Wei, J., Zheng, X., ... & Ma, W. Y. (2015, May). Lightlda: Big topic models on modest computer clusters. In Proceedings of the 24th International Conference on World Wide Web (pp. 1351-1361).

Zuo, Y., Wu, J., Zhang, H., Lin, H., Wang, F., Xu, K., & Xiong, H. (2016, August). Topic modeling of short texts: A pseudo-document view. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining (pp. 2105-2114).