Sādhanā (2024)49:198
https://doi.org/10.1007/s12046-024-02470-7

Likelihood corpus distribution: an efficient topic modelling scheme for Bengali document class identification

DEBAPRATIM DAS DAWN[1,*], ABHINANDAN KHAN[1,2], SOHARAB HOSSAIN SHAIKH[3] and RAJAT KUMAR PAL[1]

[1]Department of Computer Science and Engineering, University of Calcutta, Acharya Prafulla Chandra Roy Shiksha Prangan, JD-2, Sector-III, Saltlake, Kolkata 700106, India
[2]Product Development and Diversification, ARP Engineering, 147 Nilgunj Road, Kolkata 700056, India
[3]Department of Computer Science and Engineering, BML Munjal University, National Highway 8, 67KM Milestone, Gurugram, Haryana 122413, India
e-mail: debapratimdd@gmail.com; khan.abhinandan@gmail.com; soharab.hossain@bmu.edu.in; pal.rajatk@gmail.com

MS received 1 July 2022; revised 26 September 2023; accepted 16 January 2024

Abstract. The learning quality of humans depends on the sense of contemplation. Textual documents are a huge part of the literature on contemplation which effortlessly creates perception. Automatic document class identification or organisation is a machine learning function to understand the psychological and emotional content of the text in a concise way. The problem of identification of documents falls in the field of library science, information science and artificial intelligence. The research progress of class identification of documents has been made in various most spoken languages. Numerous research works have been published in European and Asian languages. However, there is a gap in the literature when it comes to any less resource language, especially Bengali. Consequently, this work portrays an efficient topic modelling approach for Bengali document class identification. It proposes a Dirichlet-polynomial clustering model *likelihood corpus distribution* (LCD), which is based on a Bayesian numerical prototype. Experiments are done to prove the efficiency of LCD over various topic modelling algorithms, such as *latent Dirichlet allocation* (LDA), LDA with *bag-of-words* (LDA-BOW), *latent semantic indexing* (LSI), and *hierarchical Dirichlet process* (HDP). For performance evaluation, we considered five real-world datasets of Bengali corpora, such as `science`, `sports`, `computer`, `season`, and `epic` in this work. The **coherence score** of different modelling algorithms is compared to find the best model for each dataset separately.

Keywords. Automated document organisation of low-resource languages; corpus building; document class identification; likelihood corpus distribution; text labelling; topic modelling.

1. Introduction

A statistical probability of the perception in the scripted statement depends on the degree of specificity of the area under discussion with the mixing of keywords. The problem of document classification is a component of computerised content analysis. It is a process of automatic allocation of an undefined document to one or more predefined classes or categories. It is an interdisciplinary research work of library science, information science, and computer science (especially, artificial intelligence). It is also an indispensable part of text mining. The probe of the words in the textual presence provides some hints to determine its content. The practical importance of document classification has been increased day by day due to the growth of availability of a vast number of textual documents [1].

Consequently, it is a very challenging task for humans to analyse and manage a huge amount of these texts manually. Conventionally, this assignment was elucidated by physically. Nevertheless, the manual document classification was expensive to mount and intensive to manual workers. Due to the boost of technology, the volume of digital text is increasing rapidly. Over the last seven decades, a lot of papers have been published in the field of document classification [2]. The natural language research community performed various experimental works to break the ambiguity of decision-making of document classification. They proposed various state-of-the-art techniques to speed-up this decision-making procedure.

*For correspondence
Published online: 08 June 2024

Computational linguistic researchers have carried out research work on document classification since 1963 [2]. The exploration into text categorization strives for the partition of unstructured sets of documents into classes that express the subject matter of the documents. Each document may lay off in multiple classes or single class or no/undefined class. According to the nature of document classification, the formulation of classification is split into following categories:

(i) Hard Classification: *Hard classification* (HC) based approaches classify text into only one class, and a specific label is unambiguously allocated to the instance. Generally, machine learning based approaches are used for this type of classification. E.g.:

- **Snapshot**: খেলাধুলা মানসিক আনন্দ প্রদান ছাড়া দৈহিক বৃদ্ধি এবং শারীরিক সুস্থতার অন্যতম উৎস। যান্ত্রিক সভ্যতার পূর্বে মানুষ দু-পায়ের শক্তিতে ভর করে মাইলের পর মাইল হেঁটে একস্থান থেকে অন্যস্থানে যাতায়াতে অভ্যস্থ ছিল। গ্রীক নগর-রাষ্ট্রে শরীরচর্চা ও খেলাধুলা খুব জনপ্রিয় ছিল। তারা নিয়মিত খেলাধুলার আয়োজন করত।
- **Transliteration**: *Khēlādhulā mānasika ānanda pradāna chāṛā daihika brd'dhi ēbaṁśārīrika susthatāra an'yatama uṯsa. Yāntrika sabhyatāra pūrbē mānuṣa du-pāẏēra śaktitē bhara karē mā'ilēra para mā'ila hēmtē ēkasthāna thēkē an'yasthānē yātāẏātē abhyastha chila. Grīka nagara-rāṣṭrē śarīracarcā ō khēlādhulā khuba janapriẏa chila. Tārā niẏamita khēlādhulārā āẏōjana karata.*
- **Translation**: *Games and sports give us not only mental recreation but also physical fitness. In the pre-mechanical era, people were habituated to cover many miles on foot. Physical training through games and sports were very popular with the Greeks. They used to regularly organise various games and sports.*

This document belongs to the "Sports" category. The formulation of this class is described in Equation (1), where d indicates specific testing document.

$$HC : |Class(d)| = 1 \tag{1}$$

(ii) Soft Classification: The *soft classification* (SC) based approach is the classified text in one or more than one category. Likelihood importance is assigned to the test data. This type of approach has maintained some kind of mathematical trade-off between various classes. Generally, fuzzy-based techniques are followed by soft classification rule. For example:

- **Snapshot**: "ছাত্রানাং অধ্যয়নং তপঃ।" ছাত্রছাত্রীদের প্রধান কাজ হল পড়াশুনা করা। পড়াশুনাকে তপস্যার মতো করেই করা দরকার। তবে পড়াশুনার সাথে সাথে শরীরচর্চা ও খেলাধূলার বিশেষ প্রয়োজন আছে। কারণ অসমর্থ ও দুর্বল শরীরে কখনো ভালো পড়াশুনা হয় না। এই প্রসঙ্গে তরুণদের উদ্দেশ্যে স্বামীজী বলেছেন "গীতাপাঠ অপেক্ষা ফুটবল খেলিলে তোমরা স্বর্গের আরও নিকটবর্তী হইবে।" ফুটবল খেলা অর্থাৎ নিয়মিত খেলাধূলা মন ও শরীরকে সতেজ ও সুস্থ রাখে একথা সবাই স্বীকার করে।
- **Transliteration**: *"Chātrānāṁ adhyaẏanaṁ tapaḥ." Chātrachātrīdēra pradhāna kāja hala paṛāśunā karā. Paṛāśunākē tapasyāra matō karē'i karā darakāra. Tabē paṛāśunāra sāthē sāthē śarīracarcā ō khēlādhūlāra biśēṣa praẏōjana āchē. Kāraṇa asamartha ō durbala śarīrē kakhanō bhālō paṛāśunā haẏa nā. Ē'i prasaṅgē taruṇadēra uddēśyē sbāmījī balēchēna "gītāpāṭha apēkṣā phuṭabala khēlilē tōmarā sbargēra āra'ō nikaṭabartī ha'ibē." Phuṭabala khēlā arthāṯ niẏamita khēlādhūlā mana ō śarīrakē satēja ō sustha rākhē ēkathā sabā'i sbīkāra karē.*
- **Translation**: *"To study should be the sole austerity for the students." The first and foremost duty of a student is to study hard with full concentration of mind. However, it should be remembered that formal education is not enough along with such education. Physical exercise and participation in games and sports are essential only because a student of the sickly body and ill health cannot educate himself properly. Swamiji has said, "You will be nearer to God if you devote yourself to playing football instead of reading the Gita." It is universally acknowledged that regular participation in games and sports make one physically fit and mentally alert.*

This document belongs to both "Education" and "Sports" category. The formulation of this class is described in Equation (2).

$$SC : |Class(d)| \geq 1 \tag{2}$$

Conceivably, document classification is a very popular problem for linguistic researchers. A number of algorithms have been proposed to this domain in various most spoken languages, like English, Chinese, Arabic, Malay, Urdu, Portuguese, Indonesian, Polish, Dutch, Japanese, German, etc. The computational linguistic researchers proposed a wide variety of document classification algorithms in these most spoken languages [3]. In the case of Asian languages, various papers have been published such as Nepali, Tibetan, Persian, and so on. Along with, the research work in document classification has been performed in various

Indian languages such as Hindi, Punjabi, Telugu, Tamil, Marathi, Assamese, Gujarati, and so on [4]. However, due to the unavailability of resources, the research progress is not up to the mark in most of the low-resource languages, especially Bengali. This language is the associate of the Indo-Aryan group, sub-class of Indo-Iranian, and member of Indo-European language. It has great importance in world literature. Thus, this work presents an efficient method to disambiguate the decision-making process of classification of the Bengali text document.

Over the last two decades, a few of Bengali document classification algorithms have been proposed [5]. A generic Bengali document classification system has been followed by five basic steps. The pre-processing task of the text includes text normalization, removal of unwanted symbols, and so on. In some cases, *part-of-speech* (POS) tagging is also included in the pre-processing task. However, text lemmatization process is not effective in Bengali as compared to English. Since, Bengali language has a large number of inflected words with a huge variety, and deriving a root word from the inflected form is a very challenging task. The key-attribute identification is a very important step to identify keywords throughout the text. It may be depending on single or multiple text documents, according to the algorithm of feature extraction. If features are assembled from a single document, then it is called a local feature. On the other hand, if features are accumulated from more than one text document, then it is called a global feature. The feature collection step is responsible to do the formulation of these feature sets. Finally, a suitable classification technique is used to classify or label the unknown texts. Figure 1 represents a step by step procedure of a generic Bengali document class identification system in a simple way.

Bengali document classification has important applications in various linguistics tasks, some of which are question classification [6], news classification [7], web document cataloguing [8], word clustering [9], word embedding [10], etc.

The remainder of this paper is prearranged as follows: Section 2 illustrates related approaches of Bengali document classification. Section 3 describes an efficient method of document classification. Section 4 describes the dataset of document classification. Section 5 presents the results of experimental works of the proposed algorithm. Section 6 analyses the result in a variety of circumstances. The last section wraps up this work.

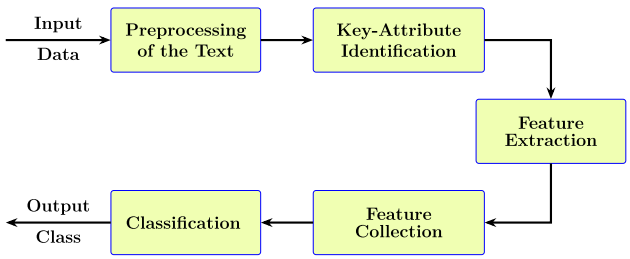


Figure 1. Flow diagram of a generic Bengali document class identification system.

2. Related works

A lot of algorithms has been proposed by linguistic researchers for Bengali text classification [5, 11]. These algorithms are categorically divided in two approaches, such as hard classification and soft classification. The hard classification approach includes supervised, unsupervised, and semi-supervised techniques. On the other hand, soft classification approach includes fuzzy-based technique, and topic modelling-based approach. In this section, this work summarises most of the relevant works of Bengali text classification

2.1 *Hard classification based approaches*

The literal name of hard classification is "winner-takes-all", i.e., a document is labelled only for a single class. Another name of this scheme is "bloc voting". A brief picture of hard classification approaches is given below.

2.1.1 *Unsupervised learning based approaches:* In 2005, Mansur *et al* [5] have in mind of an *n*-gram based method for automatic text categorization by using *n*-character. *n*-gram refers sequel of *n* items in a specified sequence. It shows a discrepancy in two levels, such as character-level and word-level. The character-level *n*-gram is used to predict the character and word-level *n*-gram is used to predict the word in a given sequence. Subsequently, this algorithm has performed character-level *n*-gram such that every *n*-gram generates a unique number. This unique number is used as a key in the hash map table, which has mapped frequency of each *n*-gram in the hash map table. A profile distance measurement matrix is created to find the distance between category samples to a new document sample. The minimum profile distance has been used to recognize the class of a new sample. However, this technique is unable to put any idea regarding word-level *n*-gram. It is unable to capture all the important words in sliding window basis, because this algorithm is mulled over two-, three-, and four-gram only.

In 2017, Dhar *et al* [11] performed a multi-domain document classification task using distance measurement matrix. This paper has collected data from online news corpus in five domains, such as Business, Medical, Science and Technology, Sports, and State. This dataset consists of one thousand documents, and two hundred text documents in each domain. This paper has performed a lot of pre-processing tasks, such as elimination of stop words, post-positions, English equivalent words, conjunction and numeric values. It has created a vector space model, such as

V_{ij}, *term frequency* (TF) of an *i*th term in document *j*. It has calculated accuracy with the discarding of 90th percentile formula by measuring various distance matrices, such as cosine similarity, Euclidean distance, *naïve Bayes* (NB) multinomial, *random forest* (RF), naïve Bayes, J48, and simple logistics. It has received a good accuracy level in "Business" and "Sports". However, the mean rank of simple logistic and Naïve Bayes classification approach is not good. These two approaches have been shown very poor result in the Friedman test [11].

The research work about the utility to discover the hidden semantic structure in the text has been increased day by day. A topic comprehends a bunch of words that commonly arise together. A topic modelling approach is able to associate words with analogous meanings and discriminate between utilisations of words with numerous meanings. On this occasion, Helal *et al* recommended a topic modelling approach to enhance topics and news classification in 2018 [12]. In the preprocessing phase, this paper performed tokenisation, exclusion of stop words, and formation of bigram. A bigram is an order of two contiguous tokens in the corpus appearing repeatedly. This algorithm has performed trial run on various coherence scores to identify an optimum number of topics. It performed a cosine similarity measurement of pairs of documents in both LDA and Doc2Vec model and proved LDA is a better model compared to Doc2Vec model. However, similarity percentage is very insignificant in this paper. One of the major reasons for this result is that the algorithm is over-fitted by selecting too many topics.

2.1.2 *Supervised learning based approaches:* In 2014, Chy *et al* [7] have attempted to find out user-specific news article for the newsreader by performing news classification task in a supervised way. In the preprocessing phase, this algorithm has followed text tokenisation, stop word removal, and word stemming. It has used a light Bengali lexicon to identify whether a word is a root word (or not) and replace a root word in place of an inflected word. *Inverse document frequency* (IDF) technique has been used to collect the feature set and to sort the feature set according to the IDF score. This score has further used in Naïve Bayes classifier, which is based on a probabilistic approach. However, this algorithm is unable to achieve a good precision value. It is incapable to sort out word-level overlapping characteristics.

In the postponement of work [5], Islam *et al* utilised character-level *n*-gram model for text classification in a supervised way [13]. This paper has cleaned up the dataset by text tokenisation, word stemming, and removal of special symbol, pronouns, conjunction, etc. It has designed a normalised vector space model by calculating the *term frequency inverse document frequency* (TFIDF) of each feature to prepare multi-class *support vector machine* (SVM). The experimental set up has diverged with the values of n in n-gram, i.e., $n = 1, 2$, and 3. On the contrary to the comparison of feature selection technique, this algorithm has exhibited best results on the unigram model, where the value of $n = 1$. However, 1-gram character-level is unable to capture in any lexical and semantic information in a text. Due to the lack of deliberation of domain-related features, the *F*-measure score of "Art" and "Opinion" category is very low.

In addition, word embedding is very popular in automatic document classification in the last couple of years. A lot of algorithms present in English-based document classification. There is a various recognised set of rules for word embedding but Word2Vec and GloVe are ready for action algorithm. Word embedding is a course of action of feature mining liable on syntactic and semantic associations. The combination of Word2Vec or GloVe with SVM is widely held by NLP researchers [14]. On that occasion, Hossain *et al* reported an algorithm of statistical machine learning by following word embedding-based feature extraction [15]. It followed Word2Vec in feature extraction and SVM in classification. However, this algorithm extracts only semantic feature without consideration of syntactic feature. The sub-linear interactions are not clearly stated. Due to the training difficulty in SoftMax function, it is incompetent to supervise a large number of document classes.

2.2 *Soft classification based approaches:*

Soft classification based approaches are proficient in classifying multi-domain content. The task of fuzzification and predictability is grounded on the level of truthiness. A fuzzy set interprets the vagueness of ideas in textual documents. Generally, fuzzy logic is applied in the classification phase to reinforce the decision-making process in the cases of uncertainty and unpredictability in the text. The fuzzy set concept apportions with the depiction of classes whose margins are not well demarcated. It is defined as [0, 1], where membership values, 0 and 1 indicate the marginal values, i.e., no membership and full membership, respectively [16]. A lot of research papers have been published on fuzzy based soft classification techniques in various languages. However, due to the undecidability and uncertainty of human perception research work in document classification is not as popular as supervised, unsupervised, or semi-supervised learning based techniques.

2.2.1 *Fuzzy based approaches:* Dhar et al [17] proposed a fuzzy based soft classification technique for Bengali web document. The authors collected data from various Bengali news websites, online magazines, web pages, and created the dataset in eight predefined domains, namely *Business*, *Entertainment*, *Science and Technology*, *Food and Recipe*, *Medical*, *State*, *Sports*, and *Travel*. The dataset was preprocessed by removing stop words and

tokenising the text. Preprocessing is very important in fuzzy based classification because it removes the probability of strong confusion in the text document. The proposed technique utilised the concept of multiplication of TFIDF and *inverse class frequency* (ICF) score of each term to extract features. ICF measures the class information of term, i.e., TF of a word on a category or class basis. The authors [17] also modified the *fuzzy C-means* (FCM) algorithm for obtaining the best outcome to overlay dataset and *k*-means algorithm. They obtained a good accuracy value in the case of the *Food and Recipe*, *State*, and *Science and Technology* datasets. However, this modified FCM needs to know the number of clusters beforehand, which is a time-consuming task and very sensitive to noise and preliminary predictions. It can also give rise to cluster errors for wrong seeds.

3. Proposed methodology

Recent advances in computational linguistics and the modern age of information retrieval technology have led to the creation of many textual data in digital format in Bengali. Due to the large number of people speaking Bengali, the digital data used in Bengali texts is growing rapidly. These data exist in the following format: unstructured, semi-structured, and structured [17]. In most cases, these documents convey interconnected and identical concepts. For that reason, it is very important to classify these documents in a more efficient way.

This work presents a topic modelling approach for document class identification of Bengali text. Topic modelling is a formidable tool of document classification in an unsupervised way. It follows two approaches of text mining, i.e.,

(i) *Abstractive Mining*: It conveys the ideas of a source document by employing different words.

(ii) *Extractive Mining*: It conveys the summary by utilising sentences or phrases in the text document.

Subsequently, in this work, we have implemented an *extractive* mining based methodology.

Topic modelling is a statistical tool for determining the abstractness of topics in an assemblage of documents. It is a probabilistic reproductive prototype with exactness in information retrieval and text mining. It follows the unsupervised machine learning-based approach. It is not prerequisite to declare labelled data, grouped by a human. Thus, it is an effective tool for data assessment and evaluation, where documents are unknown or unseen. Topics coexist in a document in terms of association of variable degrees.

On the perspective of language processing, topic modelling illustrates a way of revealing the hidden meaning of the text. It enables computerised extraction of topics from textual documents. It creates a term-document matrix based on the probability or matrix decomposition operation and measures the semantic coherence of topics. A topic is an allocation over a static dictionary. A key advantage of topic modelling is to assign more than one topic to a single document. Topic modelling can be an alternative tool for soft classification apart from the fuzzy-based techniques. Several algorithms for topic modelling exist in literature, such as (*i*) *latent semantic indexing* (LSI), (*ii*) *Latent Semantic Analysis* (LSA), and (*iii*) *latent Dirichlet allocation* (LDA), (*iv*) *hierarchical Dirichlet process* (HDP), etc.

This work identifies an efficient topic modelling algorithm for Bengali document classification. A flow diagram of the proposed methodology is mentioned in figure 2. This section of this work illustrates an efficient approach of topic modelling in step by step, as has been elucidated in the following subsections.

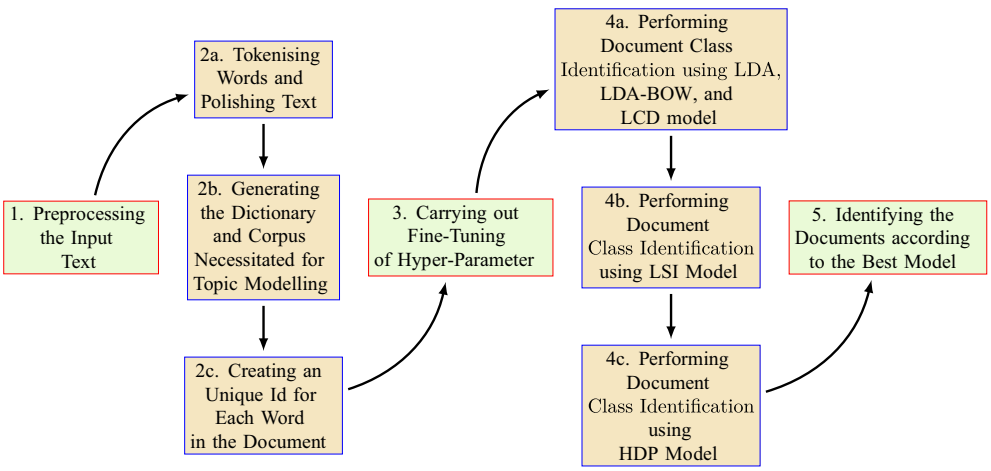


Figure 2. Flow diagram of the proposed methodology.

3.1 *Text preprocessing*

The fast-spreading out of the Internet, electronic message, and consumer-angled media such as communal interacting sites, blog, micro-blog, and news data increase the resources of text classification. These resources are available in the following two formats:

(i) *Romanised Bengali*: It is the depiction of written Bengali in the Latin script. For example, *I am fine* is written as *Āmi bhālō āchi*.
(ii) *Bengali*: It is written in Bengali script. For example, *I am fine* is written as আমি ভালো আছি.

This work deals with the Bengali script, which inherits all the linguistic features of Bengali. The dataset comprises a collection of documents from numerous sources. A lot of miscellany presence in the dataset because of inconsistency, noise, and superfluous entity. These may perhaps mislead the result. Of the essence, the task of preprocessing is very essential of any raw text before processing. The main goals of the preprocessing steps are as follows:

- enhancing the quality of the text, and
- reducing the computational complexity in the classification stage.

As a consequence, many steps are required for preprocessing Bengali text, such as removing emails, social networking id, web information, unwanted symbols, stopping words, etc.

3.2 *Tokenising words and polishing text*

Tokenisation is a course of action by which a large amount of text is split into smaller chunks called tokens. It has two classes:

- sentence tokenisation, and
- word tokenisation.

The sentence tokenisation includes splitting up of a paragraph into a number of sentences. Whereas, word tokenisation involves partition of a paragraph into a number of words. These words are very useful to comprehend better in text classification. It plays a pivotal role in the conversion of string to numerical data. For this reason, this work follows the word tokenisation step.

3.3 *Generating the word dictionary*

Of the essence, a numeric feature space is very essential in any machine learning-based linguistic processing. This feature vector consists of features and word instances. Generally, a topic modelling approach has required a dictionary and corpus for text classification. A dictionary contains all the unique words in the corpus. A text condenses a mapping between unique word and the number of occurrences in the corpus.

3.4 *Creating a unique identification number for each word in the document*

In a language-conscious investigation, the step of feature extraction is very essential for vectorization. A vector may be sparse or dense. A sparse vector is generated when a dictionary or vocabulary of the corpus is too large. However, time and space complexity of sparse vector is quite high. The hashing technique is very useful on that occasion. Subsequently, this step is responsible for the creation of a unique id of each token, extracted from the previous step.

3.5 *Finding the optimum number of topics*

Distributional hypothesis on linguistics utters those terms with analogous sense have a tendency to co-occur within a parallel corpus. The dissimilarities among topics are evaluated with a score of topic coherence. The topics are taken into consideration of coherent score, if these are entirely related or partially (for example, top *m* words) [18].

3.5.1 *Computing coherent scores for various numbers of topics*

The coherence property tells that an array of topics is coherent if there exists a correlation between the topics. As example set coherent statements are:

- **Statement 1**: বাংলায় কলিকাতা বা কলকাতা নামটি প্রচলিত হলেও ইংরেজি ভাষায় এই শহর আগে ক্যালকাটা নামে পরিচিত ছিল।

  (*Bānlāẏa kalikātā bā kalakātā nāmaṭi pracalita halē'ō inrēji bhāsāẏa ē'i śahara āgā kyālakātā nāmē paricita chila.*: *Although the name Kolkata or Kalikata is common in Bengal, the city was formerly known as Calcutta in English.*)

- **Statement 2**: কলকাতা শহরটি হুগলি নদীর পূর্ব পাড়ে অবস্থিত।

  (*Kalakātā śaharaṭi hugali nadīra pūrba pāṛē abasthita.*: *The city of Calcutta is located on the east bank of the Hooghly river.*)

- **Statement 3**: কলকাতার জলবায়ু ক্রান্তীয় সাভানা প্রকৃতির।

  (*Kalakātāra jalabāẏa krāntīẏa sābhānā prakrtira.*: *The climate of Calcutta is tropical savanna in nature.*)

A topic coherence quantifies the score of a single topic by determining the level of semantic resemblance among the top scoring terms in the topic [19]. A pair of documents is comprehensible or coherent if the likeness between these two documents goes above a given threshold. Let, whole set D of documents consists of multiple documents d_i, where i varies from 1 to N. N is the total number of documents. The similarity Δ between a pair of documents d_x and d_y is

measured according to Equation (3). The value of cosine similarity, ζ of two document vectors, ∇_x and ∇_y is calculated according to Equation (4). The threshold value, ϑ is varied with the database. This work considers the average value of Δ as a threshold value. The set coherence value of all documents, $\Upsilon(D)$ is calculated according to Equation (5).

$$\Delta(d_x, d_y) = \begin{cases} 1 & \text{if } \zeta(d_x, d_y) \geq \vartheta \\ 0 & \text{otherwise} \end{cases}, \tag{3}$$

$$\text{where } x \cap y = \emptyset \text{ and } x, y \in N$$

$$\zeta(d_x, d_y) = \frac{\nabla_x \times \nabla_y}{\sqrt{\nabla_x^2 \times \nabla_y^2}} \tag{4}$$

$$\Upsilon(D) = \frac{\sum_{i=1}^{N} \Delta_i}{N(N-1)} \tag{5}$$

A topic coherence pipeline consists of four stages, such as:

(i) Segmenting the Corpus: It divides the whole corpus into a number of subsets, such that a good quantity of dissimilarity is maintained between the corpora.
(ii) Estimating Probability: The probability distribution of each subsample, segmented by the previous step, is measured.
(iii) Validating the Estimation: It validates each subsample with the measurement of metric, and a unique number is allocated to each subsample concerning its quantity.
(iv) Aggregation: This step is responsible to combine the subsamples according to the arithmetic mean and assigns a unique number for a group of each subsample. It generates the coherence value.

Figure 3 is shown in the form of a flow diagram of topic coherence pipeline. A lot of coherence measures is present in the literature [19], such as *c_v*, *c_p*, *c_uci*, *c_umass*, *c_npmi*, *c_a*, etc. In this work, we have performed a comparative analysis amongst all of them and selected *c_v* as the text classification coherence measure. The reasons are mentioned below:

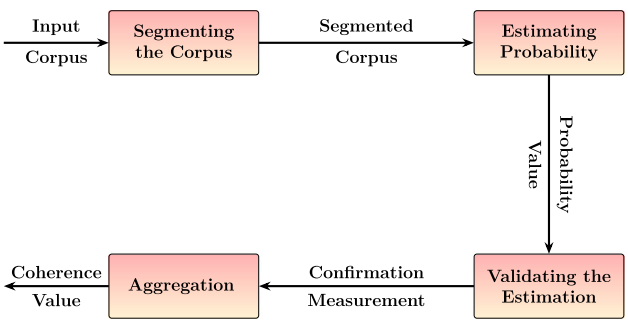


Figure 3. A flow diagram of topic coherence pipeline.

(i) Based on a sliding window, the *c_v* measure is performed in separation of the corpus into a set of words.
(ii) Individual segmentation set of top terms is used to calculate the word and word set probability.
(iii) This measure handlings *normalised pointwise mutual information* (NPMI) for the estimation of validation of a measure. The cosine similarity is used to find the degree of cohesion that how to quantify is related to one-word set.

3.5.2 *Acquiring coherence score corresponding to the number of topics* The selection of a number of topics is very crucial in any topic modelling algorithm. For examples:

- If it selects too many topics, then the topic modelling algorithms will be over-fitted.
- If it selects a small number of topics, then topic generalisation is performed, and it comes up with no sense.

Hence, this step is responsible to find out the optimal number of topics.

3.5.3 *Choosing the best value as the number of topics* Coherence deals with the comparative distance between terms within a topic. The range of *c_v* measure score is mentioned in Equation (6).

$$0 < c\_v < 1 \tag{6}$$

However, it is very rare to see $c\_v = 1$. It happens only for identical words. The average coherence score of a topic is the mean of the distances between terms.

3.6 *Performing document class identification using LDA model*

LDA [20] is a continuous unsupervised learning algorithm. It is a flexible, propagative, probabilistic model for a compilation of distinct data from text corpora. It is a hierarchical three-tier Bayesian prototype. Topic probabilities deliver an unambiguous exemplification of a document. A document is expressed as a distribution of topics. A topic is illustrated as a distribution of words.

Before execution of LDA, three essential hyper-parameters tuning is required, such as a number of topics, alpha, and beta. The value of a number of topics is acquired by tuning the coherence measure. Coherence value measures the relative distance between words within a topic. It indicates how much a topic fond of a word. The value of

alpha indicates document-topic firmness, whereas beta indicates topic-term compactness.

An upper value of alpha indicates a greater number of topics present in the document. Initially, alpha is set to 0.1. A higher value of beta indicates a greater number of words is compiled in a single document. Initially beta is set to 0.5. This work adjusts these values in an experimental way. Additionally, another hyper-parameter number of terms under each topic is initially set to ten. In the experimental setup, we have varied the value of number of terms under each topic from 5, 10, 15, 20, 25, etc. and got the best result for 10.

The output explains that the top ten words are the most prominent words in each topic. The number associated with each term/word indicates the likelihood/probability of being related to the topic. Theoretically, these top ten words of each topic equip into a coherently meaningful domain. This model is not labelled with any topic. It only enables you to create a topic. The labelling of the topic can be done just like managing a library.

3.7 *Performing document class identification using LDA-BOW*

BOW model is an extensive tool for feature extraction from text. The insight of BOW is that the documents are analogous if the contents are very close or similar. The algorithmic steps are mentioned below.

- **Step 1**: Preprocessing the text.
- **Step 2**: Initialise a dictionary to keep up a bag of terms/words.
- **Step 3**: Tokenise each term in the sentence.
- **Step 4**: If a word present in the dictionary, then increment the frequency of the word by one. Otherwise, add the new word into the dictionary and update its count accordingly.
- **Step 5**: Initialise a vector to check a word in each sentence is frequent or not. If the word is frequent, then set one in the vector, otherwise zero.

As an outcome of the BOW, each document is epitomised as a vector, and the size of the vector is equal to the length of the vocabulary size. In the text classification, LDA is taken input BOW vector as a corpus and dictionary. The strengths of this model are the followings:

- It is a practical model for classification of multiple documents.
- It is capable to generalise the unknown documents.

However, LDA-BOW model suffers a lot of weaknesses. The cons of this model are the followings:

- The BOW representation is very impractical.
- Weigh-up all words equally.
- Unable to find rareness of terms.

3.8 *Performing document class identification using LCD*

Due to the high morphological complexity of the Bengali, the task of document set identification is a tedious task for any computational linguistics researcher [21, 22]. Bengali belongs to the Indo-Aryan class, and is highly evolutionary or inflectional in nature [21]. It is less computerised than any other most spoken European and Asian language [23].

Subsequently, this work presents a comprehensive topic modelling approach for clustering Bengali documents. LCD is a Bayesian model of the third level, capable of clustering in three levels [24], such as:

- word level,
- document level, and
- corpus or database level.

LCD solves the problem of word separation or morphological parsing and it is able to create meaningful words like morphemes from the Bengali text. To elaborate LCD, this work considers a `corpus` or `database` D, which is collection of M `documents`. A `document` is an array of N words or tokens.

The probability density of k-dimensional `dirichlet` random variable θ with parameter k-vector α is illustrated in Equation (7).

$$\rho(\theta|\alpha) = \frac{\Gamma\left(\sum_{i=1}^{k}\alpha_i\right)}{\prod_{i=1}^{k}\Gamma(\alpha_i)}\left[\sum_{i=1}^{k}\left(\theta_i^{\alpha_i-1}\right)^2\right], \tag{7}$$

where $\Gamma(\alpha)$ is the `Gamma` function, $0<\alpha<+\infty$, and θ is `Poisson` distribution of α. α is the parametric quantity of `dirichlet` before distributing the contents of each document. θ_i is the topic dispersion for document i. The joint distribution of topic aggregation θ is expressed in Equation (8).

$$\rho(\theta, z, w|\alpha, \beta) = \frac{\rho(\theta|\alpha)}{1+[\rho(\theta|\alpha)]^2}\prod_{n=1}^{N}\rho(z_n|\theta)\rho(w_n|z_n, \beta), \tag{8}$$

where z_{ij} is the topic for the jth token in ith document, w_{ij} is the specific word, and β is the parametric quantity of the dirichlet before on the per-topic word distribution.

$$\rho(w|\alpha, \beta) = \int \frac{\rho(\theta|\alpha)}{1+[\rho(\theta|\alpha)]^2} \left(\prod_{n=1}^{N}\sum_{z_n}\rho(z_n|\theta)\rho(w_n|z_n, \beta)\right)d\theta \tag{9}$$

$$\rho(D|\alpha,\beta)=\prod_{d=1}^{M}\int\frac{\rho(\theta_d|\alpha)}{1+[\rho(\theta_d|\alpha)]^2}\left(\prod_{n=1}^{N_d}\sum_{z_{dn}}\rho(z_{dn}|\theta_d)\rho(w_{dn}|z_{dn},\beta)\right)d\theta_d \tag{10}$$

The marginal distribution of a document is expressed in Equation (9). Equation (10) depicts the probability of a corpus or database through the product of the marginal probabilities of a document.

3.9 *Performing document class identification using LSI model*

LSI model is a statistical means to express the primary composition of the text. It enables to retrieve exact information from large corpora. It takes advantage of a term-document matrix to illustrate the presence of terms in text corpora, where rows relate to terms and columns match up to documents. The document similarity measure ensures whether certain words exist or not in some context [25]. The algorithmic steps of LSI are listed below.

- **Step 1**: Tokenising the text corpora into sentences.
- **Step 2**: Initialise a sparse matrix of size $r \times c$ for the identification of frequency of r unique words within c number of documents. It contains a word and corresponding counting variable. The corresponding count of a word is set to one, if the word be present in a sentence.
- **Step 3**: Normalise each term or word with the TFIDF score.
- **Step 4**: Reduce the sparse matrix or sentence vector into a multidimensional abstract space. The normalised word-sentence matrix is a multiplicative matrix of transform matrix, scaling matrix, and concept matrix.
- **Step 5**: Map similar terms to the analogous location on low-dimensional space. It maintains singular value decomposition to identify patterns in the mapping.
- **Step 6**: Selecting the sentences based on the absolute value in the sentence vector.

LSI model performs a good quality of text classification. It has the capability to cut down noise by lowering the dimensions. However, the negative sign indicates semantic dissimilarity in the topics. It indicates contribution of a word in more than one topic.

3.10 *Performing document class identification using HDP model*

HDP model is an impressive tool to enable summarization and organisation of a huge volume of text efficiently. A lot of applications of HDP is included in the domain of statistics, machine learning, and information retrieval. It is tagged of the non-parametric Bayesian methodology, which is a potent diverse membership model [26]. It is responsible to cluster categorised data in an unsupervised way. The algorithmic steps of HDP are given below.

- **Step 1**: Initialise topic boundary and corpus level limitation arbitrarily. Fix the current time to one.
- **Step 2**: Get hold of a document from the textual corpus in hit and miss way.
- **Step 3**: Bring up-to-date document level and calculate all parameters, such as alpha, gamma, kappa, tau, max_chunks, max_time, chunksize, and so on.
- **Step 4**: Update corpus level parameters, such as var_converge, doc_word_counts, doc_word_ids, unique_words, and so on.
- **Step 5**: Set the learning frequency by the utility of current time, kappa, and tau.
- **Step 6**: Augment the current time by one after the completion of Step 5.
- **Step 7**: Revise all the factors in the corpus level.

The major improvements of HDP over LDA and LSI are mentioned below.

- It deduces the number of topics from textual corpora automatically. There is no need to pre-declare the number of topics.
- It is a non-parametric generalisation model of LDA.
- The number of topics is unrestrained.
- It maintains a hierarchical architecture from bottom to top, such as a bag of words, topic, and document. A group consists of a bag of words, a topic is defined in each cluster, and a document is a collection of topics.
- Due to the incorporation of the infinite hidden Markov model, it holds limitless internal states for learning the data.
- It captivates the three following halting conditions:
  - completion of processing,
  - achieving chunk limit, and
  - termination of time quantum.

3.11 *Identifying documents according to the best model*

This step is responsible to acquire data from five topic modelling approaches, such as LDA, LDA-BOW model, LCD, LSI, and HDP model. It creates a matrix of dimension 5×1 to maintain the value of coherence score of each model. This matrix demonstrates a piece of pair-wise information between model and coherence score.

From the literature study of coherence score of topic modelling approaches [27–29], it is well known that the optimum score of a topic modelling approach should lie between 0.4 and 0.7. Explicitly, the facts are as follows:

- If the coherence score is less than 0.4, then the topics are classified in a more generalised way.
- If the coherence score is greater than 0.7, then the topics are classified in a more specialised way.

Hence, this work applies a threshold cutter to find the most relevant coherence model. Here, the best model is selected among the optimum models. This work selects the maximum coherence measure as the best measure and chooses the corresponding topic model as the best model.

All the coherence measure-based calculations depend on the dataset. The result may differ from dataset to dataset. In the above example, the best text classification approach is LCD model. So, the final result after classifying the LCD model is given below. The topic labelling is performed according to the maximum number of topics present in the cluster. In case of a tie, a cluster consists of more than one topic. Algorithm 1 illustrates the detailed steps of the proposed method.

Algorithm 1. Efficient Topic Modelling Algorithm

Input: database D and number of documents M
Output: appropriate topic
Preprocess all M documents in D
Generate N array of words or tokens for all M documents in D
Generate word dictionary w of unique elements of set N
Create unique identification number of each element of w
Measure similarity Δ between a pair off documents d_x and d_y using Equation (3)
Calculate the value of cosine similarity, ζ, of two document vectors ∇_x and ∇_y, according to Equation (4)
Set threshold value $avg\,(\Delta)$
Calculate the set coherence values of all documents, $\Upsilon\,(D)$, according to Equation (5)
Select c_v as text classification coherence measure using Figure 3
for all M documents in D **do**
    Choose best value of c_v using Equation (6)
end for
for all M documents in D **do**
    Perform document class identification of LDA model according to Equations (3) through (6)
    Perform document class identification of LDA-BOW model according to Section 3.7 Step 1 through Step 5
    while $N \neq \emptyset$ **do** ▷ LCD model
        Compute probability density of k-dimensional `dirichlet` random variable θ with parameter k-vector α using Equation (7)
        Calculate joint distribution of topic aggregation θ using Equation (8)
        Calculate marginal distribution using Equation (9)
        Compute product of the marginal probabilities using Equation (10)
    end while
    Perform document class identification of LSI model according to Section 3.9 Step 1 through Step 6
    Perform document class identification of LSI model according to Section 3.10 Step 1 through Step 7
end for
Create a matrix MAT of dimension 5×1 for acquiring the coherence score, cs, of the LDA, LDA-BOW, LCD, LSI, and HDP models
for $i = 1$ to 5 **do**
    if $0.4 \leq cs \leq 0.7$ **then**
        Assign $OPTIMUM = MAT[i]$
    end if
end for
MODEL: max($OPTIMUM$)
Predict more suitable label according to clustering of D using MODEL
Return the topic found

4. Dataset

Due to non-availability of Bengali dataset in document class identification, a dataset of 1K documents was presented in [30]. This dataset consists of five classes. It was collected from newspaper archives, books, blogs, online resources, magazines etc. A description of the dataset is discussed in this section.

- **Dataset 1**: `Science` Dataset
  - **Description**: Dataset 1 consists of 20% of the science data of the entire dataset. It consists of five branches of science, namely Biology, Chemistry, Mathematics, Medicine, and Physics. The sub-classes have a lot of similarities, e.g.: Biology and Medicine are very close to each other. Each subclass contains 20% of the total number of documents in the science dataset, which is 4% of the total number of documents in the entire dataset. Biology, Chemistry, Mathematics, Medicine, and Physics are comprised of 20.67%, 14.82%, 16.85%, 23.38%, and 24.26% of the number of single-occurrence terms of the science dataset; and 19.64%, 15.65%, 17.23%, 22.43%, and 25.03% of the number of distinct terms of the science dataset, respectively.
  - **Snapshot**: পদার্থবিজ্ঞানের ভাষায় শক্তি বলতে কাজ করার সামর্থ্যকে বুঝায়। কাজ বা কার্য হচ্ছে বল ও বলাভিমুখী সরণের গুণফল। কৃতকাজের পরিমাণ দিয়েই শক্তির পরিমাপ করা হয়।
  - **Transliteration**: *Padārthabijñānēra bhāṣāẏa śakti balatē kāja karāra sāmarthyakē bujhāẏa. Kāja bā kārya hacchē bala ō balābhimukhī saraṇēra guṇaphala. Krtakājēra parimāṇa diẏē'i śaktira parimāpa karā haẏa.*
  - **Translation**: *In the language of physics, energy refers to the ability to work. Work is the product of force and verbal movement. Energy is measured by the amount of work done.*
- **Dataset 2**: `Sports` Dataset
  - **Description**: Dataset 2 consists of 20% of the sports data of the entire dataset. It consists of five branches of sports, namely Badminton, Cricket, Football, Hockey, and Table tennis. The sub-classes have a lot of similarities, e.g.: Cricket and Football; Badminton and Table tennis are very close to each other. Each subclass contains 20% of the total number of documents in the sports dataset, which is 4% of the total number of documents in the entire dataset. Badminton, Cricket, Football, Hockey, and Table tennis are comprised of 22.19%, 17.55%, 22.74%, 20.63%, and 16.87% of the number of single-occurrence terms of the sports dataset; and 22.36%, 18.13%, 22.92%, 18.95%, and 17.60% of the number of distinct terms of the sports dataset, respectively.
  - **Snapshot**: টেবিল টেনিসের আরেক নাম পিং পং। বিশ্বের অন্যতম জনপ্রিয় খেলা হিসেবে বিভিন্ন দেশের ক্রীড়াবিদদের কাছে অত্যন্ত পরিচিত। ব্যক্তিগত কিংবা দলগত বিষয় হিসেবে এ খেলা টেবিলের উপরের অংশে খেলতে হয়। খেলার উপকরণ: ব্যাট, ছোট বল, জাল, এবং টেবিল।
  - **Transliteration**: *Ṭēbila ṭēnisēra ārēkanānāma piṁ paṁ. Biśbēra an'yatama janapriẏa khēlā hisēbē bibhinna dēśēra krīṛābidadēra kāchē atyanta paricita. Byaktigata kimbā dalagata biṣaẏa hisēbē ē khēlā ṭēbilēra uparēra anśā khēlatē haẏa. Khēlāra upakaraṇa: Byāṭa, chāṭa bala, jāla, ēbaṁ ṭēbila.*
  - **Translation**: *Another name of table tennis is ping pong. It is one of the most popular sports in the world. The game is played on the top of table as a personal or as a group game. Playing equipment's are bats, a small ball, a net, and a table.*
- **Dataset 3**: `Computer` Dataset
  - **Description**: Dataset 3 consists of 20% of the computer data of the entire dataset. It consists of five branches of computer, namely input devices, memory, operating system, output devices, and processing unit. The sub-classes have a lot of similarities, e.g.: memory and operating system are very close to each other. Each subclass contains 20% of the total number of documents in the computer dataset, which is 4% of the total number of documents in the entire dataset. Input devices, memory, operating system, output devices, and processing unit are comprised of 16.36%, 16.82%, 24.98%, 19.79%, and 22.03% of the number of single-occurrence terms of the computer dataset; and 17.92%, 16.34%, 23.15%, 20.89%, and 21.69% of the number of distinct terms of the computer dataset, respectively.
  - **Snapshot**: কোনও কম্পিউটার ব্যবস্থায় এক জায়গা থেকে আরেক জায়গায় তথ্যের জোগানকে বজায় রাখে কন্ট্রোল ইউনিট। বস্তুত, কন্ট্রোল ইউনিট একটি কম্পিউটারের বিভিন্ন যন্ত্রাংশের ক্ষেত্রে কেন্দ্রীয় স্নায়ুতন্ত্র বা সেন্ট্রাল নার্ভাস সিস্টেমের ভূমিকা পালন করে। একটি কম্পিউটারের বিভিন্ন অংশের মধ্যে তথ্যের আদান-প্রদানকে পরিচালনা করে কন্ট্রোল ইউনিট।
  - **Transliteration**: *Kōna'ō kampi'uṭra byabasthāẏa ēka jāẏagā thēkē ārēka jāẏagāẏa tathyēra jōgānakē bajāẏa rākhē kaṇṭrōla i'uniṭa. Bastuta, kaṇṭrōla i'uniṭa ēkaṭi kampi'uṭārēra bibhinna yantrāṅśēra kṣētrē kēndrīẏa snāẏutantra bā sēṇṭrāla nārbhāsa sisṭēmēra bhūmikā pālana karē. ēkaṭi kampi'uṭār-ēra bibhinna anśēra madhyē tathyēra ādāna-pradānakē paricālanā karē kaṇṭrōla i'uniṭa.*
  - **Translation**: *The control unit maintains the flow of information from one place to another in a computer system. In fact, the control unit plays the role of the central nervous system in the various parts of a computer. The control unit manages the*

exchange of information between different parts of a computer.

- **Dataset 4**: `Season` Dataset
  - **Description**: Dataset 4 consists of 20% of the season data of the entire dataset. It consists of five types of seasons, namely autumn, monsoon, spring, summer, and winter. The characteristics of a season are very similar to those preceding and following it. Each subclass contains 20% of the total number of documents in the season dataset, which is 4% of the total number of documents in the entire dataset. Autumn, monsoon, spring, summer, and winter are comprised of 22.78%, 16.18%, 21.57%, 16.29%, and 23.16% of the number of single-occurrence terms of the season dataset; and 24.06%, 17.53%, 20.95%, 15.51%, and 21.92% of the number of distinct terms of the season dataset, respectively.
  - **Snapshot**: বাংলা ছয় ঋতুর দেশ। এর মধ্যে গ্রীষ্মকাল হচ্ছে ঋতু গণনার প্রথম মাস, যা বৈশাখ ও জ্যৈষ্ঠ মাসকে ধারণ করে। আমরা সহজভাবে বলে থাকি বৈশাখ ও জ্যৈষ্ঠ -- এই দুই মাস নিয়েই গ্রীষ্মকাল। বৈশাখ মাস হচ্ছে বাংলা সালের প্রথম মাস।
  - **Transliteration**: *Bānlā chaẏa rtura dēśa. Ēra madhyē grīṣmakāla hacchē rtu gaṇanāra prathama māsa, yā baiśākha ō jyaiṣṭha māsakē dhāraṇa karē. Āmarā sahajabhābē balē thāki baiśākha ō jyaiṣṭha – ē'i du'i māsa niẏē'i grīṣmakāla. Baiśākha māsa hacchē bānlā sālēra prathama māsa.*
  - **Translation**: *Bengal is a country of six seasons. Summer is the first month of the season, which includes the months of Boishakh and Jyastha. We simply say Boishakh and Jyastha - these two months are summer. Boishakh is the first month of the Bengali year.*
- **Dataset 5**: `Epic` Dataset
  - **Description**: Dataset 5 consists of 20% of the epic data of the entire dataset. It consists of five types of epics, namely Iliad, Mahabharata, Odyssey, Paradise Lost, and Ramayana. These deal with stories of heroic men of rare bravery. Each subclass contains 20% of the total number of documents in the epic dataset, which is 4% of the total number of documents in the entire dataset. Iliad, Mahabharata, Odyssey, Paradise Lost, and Ramayana are comprised of 10.06%, 15.68%, 27.91%, 32.24%, and 14.08% of the number of single-occurrence terms of the epic dataset; and 8.95%, 13.98%, 29.56%, 32.35%, and 15.13% of the number of distinct terms of the epic dataset, respectively.
  - **Snapshot**: ইলিয়াড গ্রিক মহাকাব্য। প্রাচীন গ্রিসের ইলিওন শহরের নামানুসারে এই মহাকাব্যের নামকরণ করা হয়। মহাকবি হোমার এই মহাকাব্যের রচয়িতা। এটি গ্রিক ভাষায় রচিত ও ২৪ টি সর্গে বিভক্ত। এর বিষয় ট্রয়ের যুদ্ধ।
  - **Transliteration**: *Iliẏāḍa grika mahākābya. Prācīna grisēra ili'ōna śaharēra nāmānusārē ē'i mahākābyēra nāmakaraṇa karā haẏa. Mahākabi hōmāra ē'i mahākābyēra racaẏitā. Ēṭi grika bhāṣāẏa racita ō 24 ṭi sargē bibhakta. Ēra biṣaẏa ṭraẏēra yud'dha.*
  - **Translation**: *The Iliad is a Greek epic. The epic is named after the ancient Greek city of Ilion. The great poet Homer is the author of this epic. It is composed in Greek and is divided into 24 surges. The subject is the Battle of Troy.*

5. Experimental results

This section has demonstrated how the topics modelling approaches are suitable for Bengali document classification. The experimental study has incorporated a dataset, consists of five sub-datasets. This section shows all the experimental results of these five datasets.

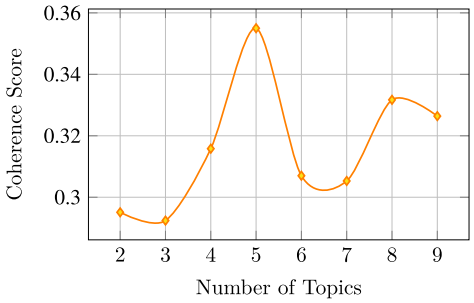
5.1 *Experimental results for dataset 1 (science dataset)*

The hyper-parameter tuning is a very challenging task in Dataset 1. This is because a high level of inter-domain similarity exists in this dataset. Medicine and Biology are very close subjects. On the other hand, Physics and Mathematics are very close subjects. Figure 4(a) shows a coherence measure versus the number of topics in a graph. This graph shows that five is the optimum number of topics, in which this algorithm receives the highest score.

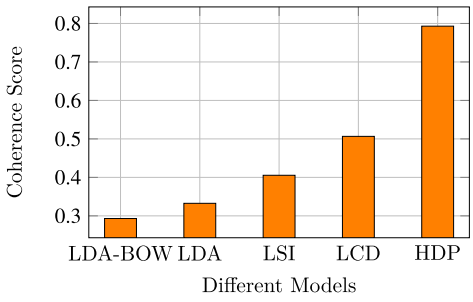
The number of topics is very crucial for further processing. After executing all the five topic modelling algorithms, this subsection plots graph figure 4(b) for five different models versus coherence score. This graph shows that LCD model is the best topic modelling approach in Dataset 1. It shows a coherence measure of five different topic modelling approaches, in the form of bar graphs.

5.2 *Experimental results for dataset 2 (sports dataset)*

The hyper-parameter tuning is performed in Dataset 2. Figure 5(a) shows a graph of coherence score versus the number of topics. This graph depicts that the coherence measures of the number of topics four, five, and six are very close. However, the highest measure is obtained when the number of topics is seven. This seven is considered as the number of topics for Dataset 2.

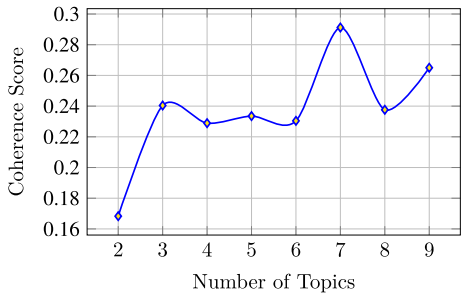


(a) Graphical plot of coherence score vs number of topics.

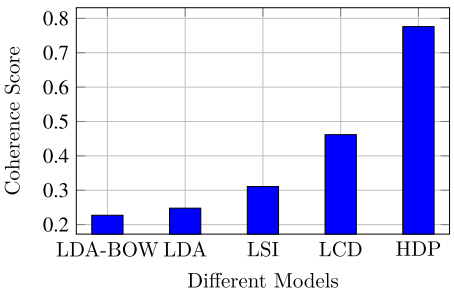


(b) Coherence score comparison graph of each model.

Figure 4. Experimental Results for Dataset 1.

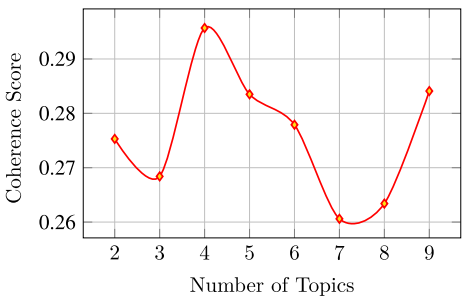


(a) Graphical plot of coherence score vs number of topics for Dataset 2.

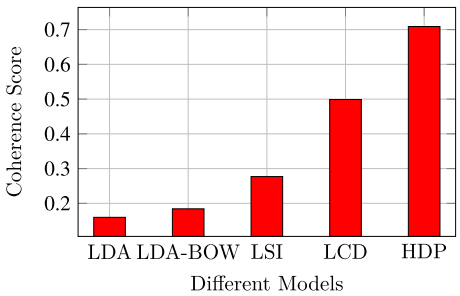


(b) Coherence score comparison graph of each model for Dataset 2.

Figure 5. Experimental Results for Dataset 2.



(a) Graphical plot of coherence score vs number of topics for Dataset 3.



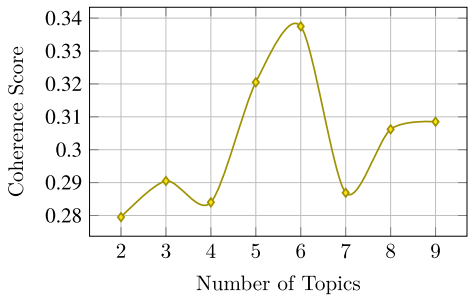
(b) Coherence score comparison graph of each model for Dataset 3.

Figure 6. Experimental Results for Dataset 3.

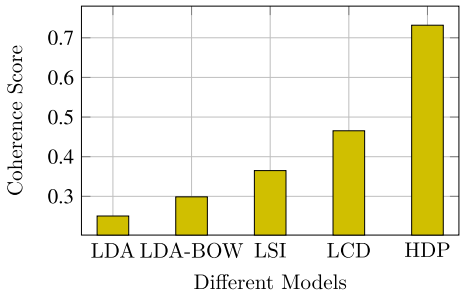
5.3 *Experimental results for dataset 3 (computer dataset)*

The hyper-parameter tuning of Dataset 3 is performed before the text classification. Figure 6(a) shows a graph of hyper-parameter tuning of this dataset. This graph depicts that it receives the highest coherence score for the number of topics of four. Along with, the coherence score of the number of topics, five and six are very close. However, this graph has the lowest coherence value at seven.

The algorithms of topic modelling are executed with the consideration of the number of topics same as four. Figure 6(b) shows coherence scores of all topic modelling algorithms of Dataset 3. This graph depicts that LCD model is the best model. However, the performances of LDA and LDA-BOW model are very poor. The performance of the LSI model is moderate.

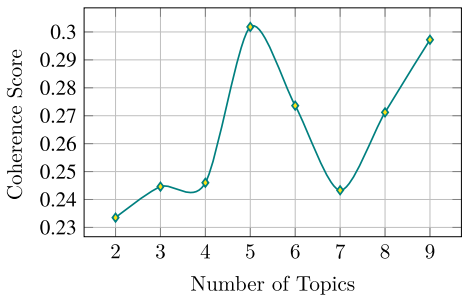


(a) Graphical plot of coherence score vs number of topics for Dataset 4.

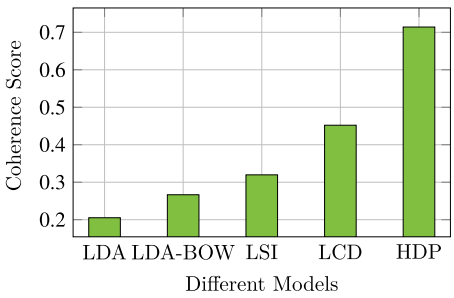


(b) Coherence score comparison graph of each model for Dataset 4.

Figure 7. Experimental results for dataset 4.



(a) A graphical plot of coherence score vs number of topics for Dataset 5.



(b) Coherence score comparison graph of each model for Dataset 5.

Figure 8. Experimental results for dataset 5.

5.4 *Experimental results for dataset 4 (season dataset)*

After preprocessing Dataset 4, hyper-parameter tuning is very imperative. In hyper-parameter tuning, this dataset has lower-level coherence value at the number of topics two, three, and four. However, the coherence score is improved drastically when the number of topics is five. The best value is obtained when the number of topics is six. Figure 7(a) shows coherence score versus number of topics for Dataset 4.

The potency of topic modelling algorithms has been measured for Dataset 4. The performance of LDA is very poor. The performances of LSI and LDA-BOW model is moderate. However, the best value is obtained in case of LCD model. It classifies Dataset 4 in six topics. Figure 7(b) shows a graph of coherence score to topic modelling approaches.

5.5 *Experimental results for dataset 5 (epic dataset)*

In the experimental result, Dataset 5 is tuned accurately for finding the number of topics. Figure 8(a) shows a graphical overview of coherence score vs the number of topics of Dataset 5. This graph depicts that the value of coherence score is lowest when the number of topics is two. The coherence scores are very poor when the numbers of topics are three and four. However, the best value is obtained when the number of topics is five. The coherence score is very close to five, when the number of topics is nine again.

In the process of execution, the value of number of topics same as five is considered for further processing. Five topic modelling approaches are applied for classifying Dataset 5. A coherence score of each model is shown in the figure 8(b). This figure depicts that the coherence score is lowest for the LDA model. The result status of LDA-BOW model and LSI model are moderate. Due to the high level of specialisation, the HDP is discarded. However, the highest coherence value is obtained at LCD model.

6. Result analysis

Automatic text classification is a computerised process of arrangement and categorisation of structured, semi-structured, and unstructured text documents. Acquiring efficient

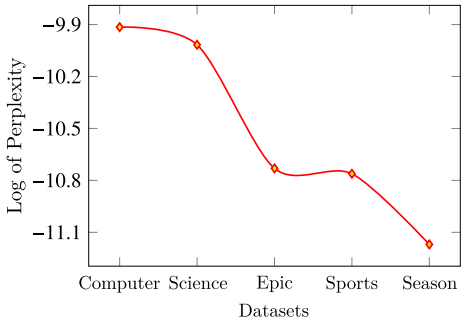


Figure 9. A graphical plot of perplexity for five datasets.

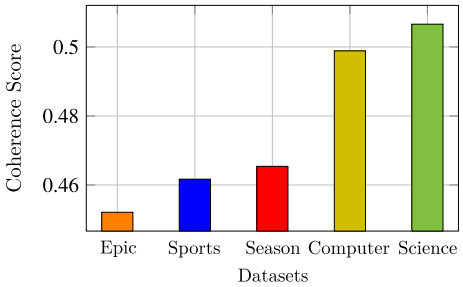


Fig. 10. Coherence score comparison graph of each dataset.

Bengali text classification algorithm is an essential issue, as Bengali is one of the most spoken languages around the world. This part of the paper describes the result analysis of five topic modelling approaches.

6.1 *Result analysis according to the perplexity*

On the analysis of predictive command of topic modelling approaches, this part of the paper presents the hold-out likelihood. The hidden deep space of a topic model comprises of topics. A topic consists of an allocation of words for each textual document. Interpretability of model is employed with perplexity or confusion metric. The perplexity of a model depends on two measures:

- **Word Intrusion**: It determines semantically cohesiveness of the topics suggested by a model.
- **Topic Intrusion**: It quantifies disintegration of a document as an assortment of topics.

In essence, perplexity designates the multiplicative probability of mass of the sample. Figure 9 shows the logarithm perplexity of five datasets. This work considers the value perplexity, according to the best model of each dataset. The negative sign indicates the logarithm of perplexity value. In this figure, computer dataset has the highest logarithm of perplexity value and season dataset has the lowest logarithm of perplexity value. According to the concept of perplexity [31], computer dataset is the best in terms of word and topic level intrusion. The logarithm of the perplexity of epic and sports dataset are very close. The science dataset has also revealed very close value with computer dataset. On the contrary, the perplexity is unable to discourse exploratory aspirations of topic modelling. Hence, analysis of coherence score is discussed next.

6.2 *Result analysis according to the coherence score*

In general, the coherence measure evaluates the excellence of topics. It quantifies the quality of extracted topics in low-dimensional feature space. Topic coherence is the summation of pairwise-probability values of the top two words in the topic. This work uses c_v measure to obtain a coherence score.

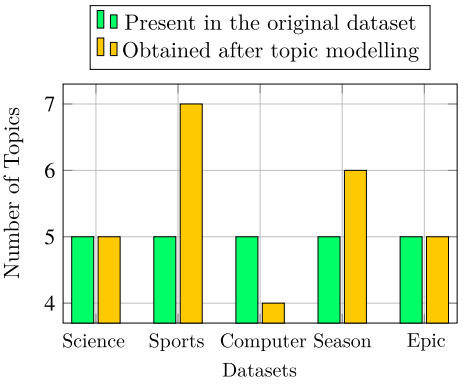


Figure 11. Graphical plot of number of topics of each dataset.

Table 1. Types of classification according to the number of topics.

Types of classification	Deviation rate	Example
Equivalence classification	$\delta = 0$	Science and epic dataset
Specialised classification	$\delta > 0$	Sports and season dataset
Generalised classification	$\delta < 0$	Computer dataset

The estimation of topic modelling approaches has been performed by calculating the probability of unknown held-out textual documents. A superior model always generates a higher probability value on held-out articles. This work uses five real-world datasets to compare the coherence score. The coherence score of each dataset, generated by the best topic modelling approach in shown in figure 10.

This figure depicts that the coherence measure in the epic dataset is lowest. In point of the fact that the flow of

Table 2. Content based classification: hard and soft classification.

Types of classification	Classification rule	Example
Hard classification	$T_u \geq T_p$	Science, computer, and epic dataset
Soft classification	$T_u < T_p$	Sports and season dataset

information from one class to another class is minimum in this dataset. This dataset consists of corpora of Iliad, Mahabharata, Odyssey, Paradise Lost, and Ramayana. These epics illustrate the social and cultural outline of contemporary life with different types of themes. In consequence, this result shows that the social and cultural patterns of human beings are varied according to the time and place.

The coherence score of science dataset is highest. This result depicts a lot of linking words present in the dataset to connect the ideas. A lot of inter-domain similarities present in between Physics with Mathematics corpora and Biology with Medicine corpora. The score of computer dataset is very closed to science dataset. The coherence score of sports and season dataset is moderate.

6.3 *Result analysis according to the number of topics*

In consequence, the result analysis according to the number of topics is illustrated in this section. The value of the number of topics is deducted from hyper-parameter tuning. All the topic modelling algorithms (except HDP) have taken this value as input. However, this work considers the number of topics according to the best algorithm of each dataset. Figure 11 shows a graphical overview of the number of topics present in the original dataset and the number of topics is obtained after text classification.

As on additional analysis of figure 11, a deviation rate is calculated according to Equation (11).

$$\delta = T_p - T_u, \tag{11}$$

where T_p indicates the number of topics obtained after document class identification and T_u indicates the number of topics present in the original dataset. This rate is responsible to classify the dataset in three classes, such as equivalence classification, specialised classification, and generalised classification. Equivalence classification indicates that the number of topics before and after classification must be the same. Generalised classification indicates the decreasing number of topics after text classification. Specialised classification specifies that the number of topics is increased after applying the topic modelling approach. Table 1 shows the types of classification according to the number of topics.

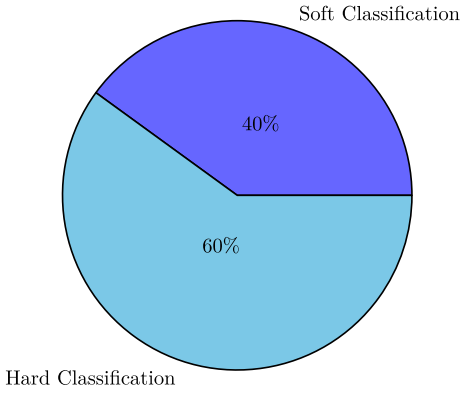


Figure 12. Hard and soft classification distribution over the dataset.

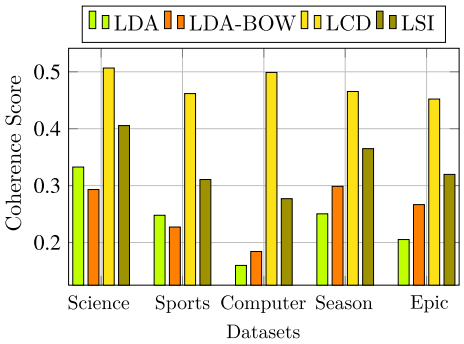


Figure 13. A graphical plot of coherence scores of each dataset for different topic modelling algorithms.

Table 3. Performance comparison of the best topic modelling approach of each dataset.

Dataset	First-rate	Improvement of LCD over		
		LDA (%)	LDA-BOW (%)	LSI (%)
Science	LCD model	52.30	72.84	24.90
Sports	LCD model	86.25	103.16	48.59
Computer	LCD model	212.26	170.84	80.08
Season	LCD model	85.86	55.79	27.53
Epic	LCD model	120.28	69.59	41.41

6.4 *Result analysis according to hard and soft classification*

In general, a classification algorithm is followed soft and hard-based classification rule. A soft-based classification rule enables to map multi-domain content, whereas a hard-based classification rule maps single-domain content. In this context, this work defines these two rules in the following way.

- **Hard Classification**: It specifies that the number of topics present in the original dataset is greater than or

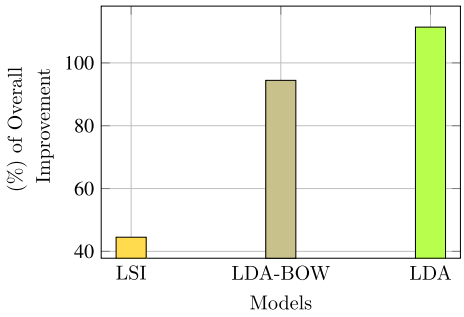


Figure 14. Average percentage of performance improvement of LCD over other models.

equal to the number of topics obtained after document class identification, i.e., $T_u \geq T_p$.

- **Soft Classification**: It specifies that the number of topics present in the original dataset is less than the number of topics obtained after document class identification, i.e., $T_u < T_p$.

Topic modelling approaches are followed by both hard and soft classification rule. Table 2 specifies hard and soft classification rules with example. All the data are populated according to the best approach to topic modelling. Figure 12 illustrates the hard and soft classification distribution over the entire dataset using pi-representation. This result specifies that the topic modelling algorithms follow both hard and soft-based classification rules.

6.5 *Result analysis for finding out an efficient topic modelling approach for Bengali text classification*

An efficient topic modelling approach enables to classify Bengali text in a more effective way. This work considers five topic modelling algorithms for experimental purposes, such as LDA, LDA-BOW, LCD, LSI, and HDP. The potency of these algorithms has been tested on the five datasets. However, HDP is discarded in this regard due to the idealistic classification of the number of topics. It is a non-parametric Bayesian model. It suffers in uncertainty regarding the selection of the number of topics. The maximum number of topics is unbounded. Due to the high morphological complexity, the document classification of HDP is not advantageous in the case of Bengali. Figure 13 shows a graphical plot of coherence scores of five datasets of four models, such as LDA, LDA-BOW, LCD, and LSI.

Subsequently, the performance of LCD model is obtained as the best. The percentage of performance improvement $\Psi(D)$ of dataset D is calculated according to Equation (12).

$$\Psi(D) = \frac{\Upsilon_B(D) - \Upsilon_O(D)}{\Upsilon_O(D)} \times 100\%, \tag{12}$$

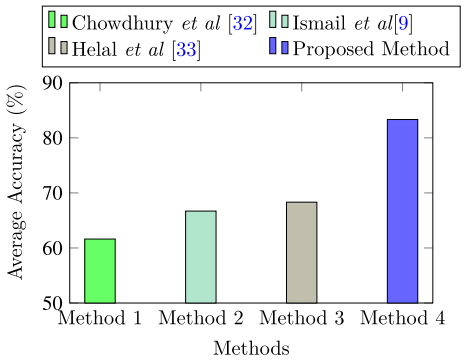


Figure 15. Performance comparison of our method to other state-of-the-art technologies.

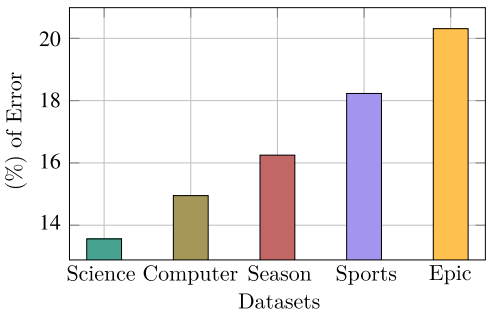


Figure 16. Error analysis of each dataset.

where $\Upsilon_B(D)$ indicates the best set coherence value and $\Upsilon_O(D)$ indicates the coherence value of other model of dataset D. In this regard, $\Upsilon_B(D)$ indicates the coherence score of LCD model. Table 3 illustrates the percentage of performance improvement of LCD model over other models of each dataset. Figure 14 shows the average percentage of performance improvement of LCD over LSI, LDA, and LDA-BOW model. This graph depicts that the performance of LSI is very close to LCD model. However, LCD model is the more improved version of LDA and LDA-BOW model.

6.6 *Comparative comparison*

Due to the lesser number of research papers in the context of Bengali, this work compares the performance of the proposed method with three state-of-art technologies. Figure 15 compares average accuracy of the proposed method to Chowdhury *et al* [32], Ismail *et al* [9], and Helal *et al* [33]. This figure illustrates the superiority of the proposed method. The performance of LCD is in top due to the consideration of likelihood `dirichlet` parameters, topic assignment, and document-specific topic distribution.

6.7 *Error analysis*

Since LCD is a more advanced version of the topic modelling algorithm for Bengali document class identification, we have considered the result set of this model for further error analysis. The error indicates incomplete clustering of the LCD algorithm. Figure 16 depicts the error (%) of each dataset. This figure shows that the science dataset got the lowest percentage of errors. Due to the presence of precise definitions of science datasets, the error rate in this case is minimal. On the other hand, epic dataset obtained the highest percentage of errors. Due to deficiency of sensitivity, the error rate is the highest in this case. All epics discuss stories of heroic men of rare bravery. In addition, the computer, season and sports datasets have moderate errors. This is due to the presence of redundant or incorrect text in the dataset, which follows the same meaning.

6.8 *Performance appraisal on comparative datasets*

The performance of the LCD model is evaluated with various comparative datasets. Bengali news dataset https://www.kaggle.com/datasets/csoham/classification-bengali-news-articles-indicnlp is a popular news repository, which was collected from three popular newspapers, viz. Anandabazar Patrika https://www.anandabazar.com/, Ebela https://banglahunt.com/ebela/, and Zee 24 Ghanta https://zeenews.india.com/bengali/. This dataset comprises following ten categories: sport, entertainment, travel, Kolkata, sports, state, nation, national, world, and international. The performance of LCD is around 70% on this dataset.

Afterwards, the performance of LCD is verified on the Wikipedia corpus https://www.kaggle.com/datasets/shazol/bangla-wikipedia-corpus?select=wiki_bangla.csv, which contains 56794 articles. This dataset has the following labels: art-and-literature, Bangladesh, durporobash, economy, education, entertainment, international, life-style, north America, sports, opinion, and technology. The achievement of LCD is about 66% on this dataset.

In addition to this, the performance of LCD is tested on the *Indian Languages Corpora Initiative* (ILCI) corpus, developed by the *Technology Development for Indian Languages* (TDIL) program of *Govt. of India* (GoI) https://www.isical.ac.in/~lru/downloadCorpus.html. This model performs well on the ILCI corpus, producing 60% average accuracy on this corpus.

7. Conclusion

This work presents an efficient topic modelling approach for Bengali document class identification. It introduces an unsupervised document clustering method LCD for Bengali document set identification. Five topic modelling approaches are considered, namely LDA, LDA-BOW, LCD, LSI, and HDP, for document class identification. The performance of these approaches are validated using five datasets: science, sports, computer, season, and epic. Hyper-parameter tuning is done for the five datasets separately for finding a good number of topics in the textual documents.

In this work, we propose an efficient approach for Bengali document clustering. A lot of algorithmic steps are discussed to find out an efficient topic modelling approach. In experimental results, the performance of LCD model is found to be the best model among all datasets. The efficiency of this approach is measured over other topic modelling approaches. This model produces a good coherence score on all datasets.

For evaluation-based findings, this work covers result analysis according to the perplexity, coherence score, number of topics, hard and soft classification, and post-analysis of LCD model. It classifies the dataset on the behalf of equivalence, specialised, and generalised class. It calculates the percentage of performance improvement of the LCD model over the other said models.

Data availability

The datasets generated and/or analysed during the current study are available in the "Kaggle" repository, https://www.kaggle.com/dsv/4761177 with DOI: 10.34740/KAGGLE/DSV/4761177.

References

[1] Luhn H P 1957 A statistical approach to mechanized encoding and searching of literary information. *IBM Journal of Research and Development* 1(4): 309–317

[2] Borko H and Bernick M 1963 Automatic document classification. *Journal of the ACM (JACM)* 10(2): 151–162

[3] Dhar A, Mukherjee H, Dash N S and Roy K 2021 Text categorization: past and present. *Artificial Intelligence Review* 54(4): 3007–3054

[4] Das Dawn D, Khan A, Shaikh S H, and Pal R K 2022 A dictionary based model for Bengali document classification. *Applied Intelligence*, pages 1–20

[5] Mansur M 2006 *Analysis of n-gram based text categorization for Bangla in a newspaper corpus*. PhD thesis, BRAC University

[6] Banerjee S and Bandyopadhyay S 2012 Bengali question classification: Towards developing QA system. In *Proceedings of the 3rd Workshop on South and Southeast Asian Natural Language Processing*, pages 25–40

[7] Chy A N, Seddiqui Md H, and Das S 2014 Bangla news classification using naive Bayes classifier. In *Proceedings of the 16th International Conference Computer and Information Technology*, pages 366–371

[8] Mandal A K and Sen R 2014 Supervised learning methods for Bangla web document categorization. *International Journal of Artificial Intelligence & Applications (IJAIA)*

[9] Ismail S and Rahman M S 2014 Bangla word clustering based on n-gram language model. In *Proceedings of the International Conference on Electrical Engineering and Information & Communication Technology*, IEEE, pages 1–5

[10] Ahmad A and Amin M R 2016 Bengali word embeddings and it's application in solving document classification problem. In *Proceedings of the 19th International Conference on Computer and Information Technology (ICCIT)*, pages 425–430

[11] Dhar A, Dash N S, and Roy K 2017 Classification of text documents through distance measurement: An experiment with multi-domain Bangla text documents. In *Proceedings of the 3rd International Conference on Advances in Computing, Communication & Automation (ICACCA)*, pages 1–6

[12] Al Helal M and Mouhoub M 2018 Topic modelling in Bangla language: An LDA approach to optimize topics and news classification. *Computer and Information Science*, 11(4)

[13] Islam Md S, Jubayer Md F E, and Ahmed S I 2017 A support vector machine mixed with TF-IDF algorithm to categorize Bengali document. In *Proceedings of the International Conference on Electrical, Computer and Communication Engineering (ECCE)*, pages 191–196

[14] Lilleberg J, Zhu Y, and Zhang Y 2015 Support vector machines and word2vec for text classification with semantic features. In *Proceedings of the 14th International Conference on Cognitive Informatics & Cognitive Computing (ICCI* CC)*, pages 136–140

[15] Hossain Md R and Hoque Md Moshiul 2018 Automatic Bengali document categorization based on word embedding and statistical learning approaches. In *Proceedings of the International Conference on Computer, Communication, Chemical, Material and Electronic Engineering (IC4ME2)*, pages 1–6

[16] Haruechaiyasak C, Shyu M-L, and Chen S-C 2002 Web document classification based on fuzzy association. In *Proceedings of the 26th Annual International Computer Software and Applications*, pages 487–492

[17] Dhar A, Dash N S, and Roy K 2018 A fuzzy logic-based Bangla text classification for web text documents. *Journal of Advanced Linguistics Studies*, 7(1-2)

[18] Syed S and Spruit M 2017 Full-text or abstract? Examining topic coherence scores using latent dirichlet allocation. In *Proceedings of the IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pages 165–174

[19] Röder M, Both A, and Hinneburg A 2015 Exploring the space of topic coherence measures. In: *Proceedings of the 8th ACM International Conference on Web Search and Data Mining*, pages 399–408

[20] Blei D M, Ng A Y and Jordan M I 2003 Latent Dirichlet allocation. *Journal of Machine Learning Research* 3: 993–1022

[21] Dasgupta S and Ng V 2006 Unsupervised morphological parsing of Bengali. *Language Resources and Evaluation* 40 (3): 311–330

[22] Lahiri A 2013 *Hierarchical restructuring in the creation of verbal morphology in Bengali and Germanic: Evidence from phonology*. De Gruyter Mouton

[23] Dolamic L and Savoy J 2010 Comparative study of indexing and search strategies for the Hindi, Marathi, and Bengali languages. *ACM Transactions on Asian Language Information Processing (TALIP)* 9(3): 1–24

[24] Lau J W and Green P J 2007 Bayesian model-based clustering procedures. *Journal of Computational and Graphical Statistics* 16(3): 526–558

[25] Hofmann T 1999 Probabilistic latent semantic indexing. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 50–57

[26] Teh Y W, Jordan M I, Beal M J and Blei D M 2005 Sharing clusters among related groups: Hierarchical Dirichlet processes. In: *Proceedings of the Advances in Neural Information Processing Systems*, pages 1385–1392

[27] He J, Weerkamp W, Larson M and De Rijke M 2009 An effective coherence measure to determine topical consistency in user-generated content. *International Journal on Document Analysis and Recognition (IJDAR)* 12(3): 185–203

[28] He J, Larson M and De Rijke M 2008 Using coherence-based measures to predict query difficulty. In *Proceedings of the European Conference on Information Retrieval*, pages 689–694

[29] Newman D, Bonilla E V, and Buntine W 2011 Improving topic coherence with regularized topic models. In: *Proceedings of the Advances in Neural Information Processing Systems*, pages 496–504

[30] Das Dawn D, Khan A, Shaikh S H, and Pal R K 2023 A 2-tier Bengali dataset for evaluation of hard and soft classification approaches. *IETE Journal of Research*, pages 1–23

[31] Chang J, Gerrish S, Wang C, Boyd-Graber J L, and Blei D M 2009 Reading tea leaves: How humans interpret topic models. In: *Proceedings of the Advances in Neural Information Processing Systems*, pages 288–296

[32] Chowdhury R R, Nayeem M T, Mim T T, Chowdhury Md, Rahman S, and Jannat T 2021 Unsupervised abstractive summarization of Bengali text documents. *arXiv preprint* arXiv:2102.04490

[33] Helal M A and Mouhoub M 2018 Topic modelling in Bangla language: An LDA approach to optimize topics and news classification. *Computer and Information Science* 11(4): 77–83