A Literature Review of Bangla Document Clustering

Arefin Niam
Department of Computer Science and Engineering
Shahjalal University of Science and Technology

Avijit Das
Department of Computer Science and Engineering
Shahjalal University of Science and Technology

Mahruba Sharmin Chowdhury
Department of Computer Science and Engineering
Shahjalal University of Science and Technology

Mohammad Abdullah Al Mumin
Department of Computer Science and Engineering
Shahjalal University of Science and Technology

ABSTRACT
Document clustering is a machine learning approach to categorize documents into related groups without any definition to the documents prior to the process. It helps to categorize very large chunks of documents into similar categories for making the process of finding a particular document easier. It also helps in retrieval of the data. There has been numerous works in document clustering in other languages but the amount of work in Bangla is still not sufficient. In this paper it has been aimed to evaluate the techniques that have been adopted in clustering Bangla documents. These techniques and their effectiveness has also been compared in contrast to the contemporary methods adopted by researchers around the world on other languages and a vision is proposed on current state of development in Bangla Document Clustering.

General Terms
Natural Language Processing, Artificial Intelligence.

Keywords
Data Mining, Document Clustering, Information Retrieval, Text Mining.

1. INTRODUCTION
Over the recent years there have been revolutionary innovation of computer technologies. With rapid improvement in technology we have entered an age where information can travel anywhere in large amounts within seconds. Every day huge amount of data is generated by interacting with various devices. This information is an invaluable assent in advancement of artificial intelligence and data science. But this large chunk of data cannot be processed manually. Therefore, various computer assisted techniques are used for the handling of this large data. Document clustering in one such machine assisted information retrieval technique that enables us to group documents into various groups based on their similarity for easier retrieval of information.

In data mining implicit, previously unknown and potentially valuable information is extracted from data. Document clustering can be categorized as a subset of data clustering, which includes concepts from fields of information retrieval, natural language processing and machine learning. There have been numerous works on Document Clustering for a long time. Over many years of extensive research by prominent Scientist and researcher the algorithms used for Document Clustering are now more efficient and robust. But most of the research works on Document Clustering have been on English Language.

Bangla is spoken by approximately more than 210 million people around the world as a first or second language [9]. But the amount of research work on Document Clustering done in Bangla is still in the beginning phase. There is no specific information on how much work has been done in Bangla on Document Clustering and how it compares to the amount of research works that has been done in other language, there is also no information describing if the works are in compliance with the recent trends in Machine Learning.

A search in google scholar generated 2,200,000 of articles in the result, when searched for the term "Document Clustering". This portrays an idea about the amount of research interests in the field of Document Clustering. It is a very important text mining technique. Document Clustering is a vital aspect of various machine learning approaches like Topic Extraction, Organization of Documents, Summarization and Information Retrieval. Every day millions of text-based data is being generated due to the expansion of internet all over the world. To efficiently extract information from this huge data Document Clustering is crucial.

There have been numerous works and development various methods for clustering of documents for efficient information retrieval and grouping of documents. These computer-assisted methodologies have been also adopted in Bangla language with time. But there has been no particular study that represents the state of the research works on Bangla Document Clustering. In the article [20], the authors have successfully been able to demonstrate the performance analysis of three different word embedding methods on Bangla corpus, which is one of key factor in document clustering. In the article [3], it was attempted to implement the Word Mover's Distance technique (based on the Earth Mover's Distance) to per-form the task of Document Clustering. In the article [7], the authors demonstrated an attempt to identify Multi word Expressions in Bangla. Attempts were also made in this research to review con-temporary works on Document Clustering. The researches done by [1] provides with an insight into what new Document Clustering algorithms are being developed in recent times and how efficiently they perform against other document clustering techniques under various and unique data-sets. Besides the method pro-vided by [15], review works done on document clustering were also taken into account in this research to gather detailed information on methodologies, ideologies, motivation and domain of reviewing the document clustering techniques on a particular language. An article by

[22] a detailed review provided with insight on how document clustering works and main discussion points in document clustering. Whereas, another research by [5] provided with information on recent trends and developments on document clustering.

This is a literature review of Bangla Document Clustering that aims to study the approaches and influences of Document Clustering methods, and how they compare to Document Clustering techniques in other languages.

The review process has been conducted in the process demonstrated by [15] for Systematic Literature Review. This research paper has been organized into the following parts:

- Preliminaries
  - What is Clustering
  - Document Clustering
  - Document Representation
  - Similarity Measure
- Methodology
  - Questions
  - Search Strategy
  - Study Selection
  - Assessment
  - Comparison
- Result and Discussions
- Conclusion
- References

Attempts were also made to formulate some question, by answering which it is aimed to conduct the research on reviewing Bangla Document Clustering. The following questions were asked in the process:

1) Which ML technique have been used in Bangla Document Clustering?
2) What is the overall accuracy of the implemented models?
3) What datasets are used in the researches?
4) Is there any difference in accuracy due implementation of the methods in Bangla?

2. PRELIMINARIES

This section highlights on understanding of document clustering and its accompanying concepts. Discussion on clustering, document clustering, document representation, and similarity measure helps understand the later discussion of the paper.

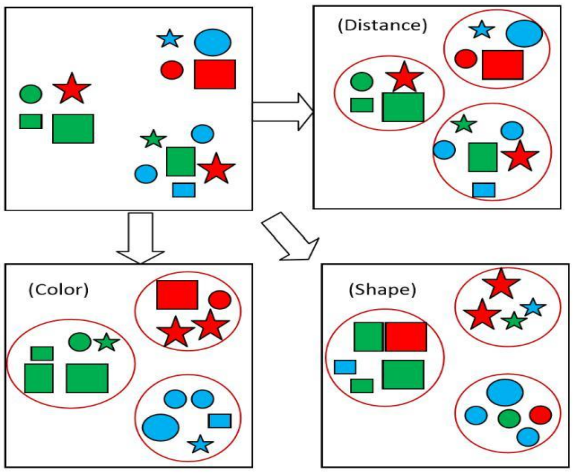

Fig 1. The same group of objects (1) clustered based on their relative distance and (2) feature they exhibit.

2.1 What is Clustering

The algorithms in clustering performs the action of grouping a set of documents into a number of subsets of the original selection or clusters. The goal of the algorithm is to generate clusters that are similar to each other internally, but externally very different from each other. That means, the documents that are grouped into one cluster should be very similar to each other and documents of each cluster should be as different as possible from documents in other clusters [17]. According to JSTOR, a digital library of academic journals, books, and primary sources, the term ”Data Clustering” appeared for the first time in an anthropological article from 1954. Over the time many researches from many academic disciplines have promulgated a large number algorithms on clustering. Their researches have built a strong structure upon which the clustering methods are built upon today. These clustering algorithms vary significantly from each other. Many researchers have improved upon the algorithms or have proposed new algorithms thus increasing the dimensionality of practical implementation of these clustering methods. It is an unsupervised learning method. Hence there are no predefined classes and so the classification is based on inherent statistical structure of the overall collection of input dataset. Whereas classification is a form of supervised learning.

An intuitive demonstration of this difference between classification and clustering is shown in Fig.1 The same objects are grouped de-pending on their relative distance, or feature — shape and color.

Classification of various clustering algorithms from a general perspective results into three different categories [11]:

- Partitional Clustering
- Density based Clustering
- Hierarchical Clustering

2.2 Document Clustering

Data mining is a machine learning process that can be described as extraction of previously unknown and potentially useful information from data. Whereas, document clustering is a subset of data clustering. It is a method of data mining which incorporates concepts and ideas from the fields of natural language processing and information retrieval. Document clustering algorithms perform the task of organizing documents into different classes or groups called

clusters, where the documents in each cluster share similar proper-ties according to defined similarity measure [22].

From a generalized view it can said that document classification is the task of assigning documents to one or more classes manually or algorithmically. In document classification document may be classified based on their various attributes. For example, a large number of document documents can be classified into a group of documents based on their expression of distinctive predefined sentiment. Therefore, document classification is a supervised learning method that needs manual interpretation. Whereas, document clustering is the most recognizable form of unsupervised learning. There is no need of manual interpretation for the assigning of documents to particular categories.

Clustering algorithms can produce either disjoint or overlapping partitions. In overlapping partitions, a particular document can be a member of multiple clusters [5] whereas in disjoint clustering, each document is a member of exactly one cluster.

At a high-level, the problem of document clustering is defined as follows. Given a set S of n documents, we would like to partition them into a predetermined number of k subsets S1, S2, . . . , Sk, such that the documents assigned to each subset are more similar to each other than the documents assigned to different subsets. Document clustering is an essential part of text mining and has many applications in information retrieval and knowledge management. Document clustering faces two big challenges: the dimensionality of the feature space tends to be high (i.e., a document collection often consists of thousands or tens of thousands unique words) and the size of a document collection tends to be large [21].

2.3 Document Representation

In various machine learning tasks, e.g. information retrieval, text mining, document clustering, the representation of textual document is done through various models of document representation. It is important to represent the unstructured text documents using an appropriate structured representation. The choice of document representation method has a profound impact on the overall quality of clustering.

Text clustering is defined as: Given D set of documents D = {d1,d2,….,dN} where, N is number of the whole documents in the data-set, d1 is the document number one, F(x) is an objective function to maximize the Cosine similarity measure or minimize the Euclidean distance measure. Its common measures are used to evaluate the performance of the clustering methods [13].

Each document is represented as a vector of terms weight, $d_i = w_{i1}, w_{i2},\ldots w_{ij},\ldots,w_{it}$ where w_{ij} is the weight of the document I, and j is the term number. The following equation is used to calculate the terms weighting:

$$W_{ij} = TFIDF(i,j) = tf(i,j) * \left(log \frac{N}{df(j)} \right) \qquad (1)$$

Where, **tf(i,j)** is the frequency of term j in document i, N is the number of all documents in the data set, df(j) is the number of documents which contains the term j.

2.4 Similarity Measure

The accuracy in the implementation of clustering algorithms de-pends on the definition of similarity, as it is difficult to define. While, similarity is an amount that reflects the strength of relation-ship between two data items, dissimilarity deals with the measurement of divergence between two data items [19]

A clustering algorithm must use a similarity measure for the comparison of different documents. It reflects the strength of the relationship between two documents, representing how similar the data components of the documents are. The clusters are formed in such a way that any two data objects within a cluster have a minimum distance value and any two data objects across different clusters have a maximum distance value [11].

According to [15] some widely used similarity measures are:

- **Euclidean Distance:** It is a standard metric for geometrical prob-lems. It is the ordinary distance between two points and can be easily measured with a ruler. It is also the default distance mea-sure used in K-means algorithm
- **Cosine Similarity:** The similarity of two documents corresponds to the correlation between the vectors. This is quantified as the cosine of the angle between vectors, that is, cosine similarity
- **Jaccard Coefficient:** The Jaccard coefficient compares the sum weight of shared terms to the sum weight of terms that are present in either of the two documents but are not the shared terms.
- **Pearson Correlation Coefficient:** It is another measure of the extent to which two vectors are related.
- **Averaged Kullback-Leibler Divergence:** The KullbackLeibler divergence (KL divergence), also called the relative entropy, is a widely applied measure for evaluating the differences between two probability distributions.

3. METHODOLOGY

The research process was planned and conducted by the Systematic Literature Review process suggested by Kitchenham and Charters [15]. Accordingly, the review process comprises of the following steps:

- Planning the review
- Conducting the review
- Reporting the review

We adopted the methodology and by following accordingly we have followed a number of steps to conduct the process of Systematic Literature Review. In the process we have developed a number of stages through which we have planned our review, conducted it accordingly and have reached a conclusion. The stages comprise of: Questions, Search Strategy, Study Selection, Result Assessment as shown in the Fig. 2

Fig. 2. The research methodology and the flow of activities

In the first stage, we asked some questions based on our objective in this literature review. And then in the next stage, the Search Strategy stage, we made decisions on how we are going to search for the topic related research articles and our methods of interpreting the search results.

In the third stage, we selected a number of papers, and then we applied some demonstrated methods of collecting a storing of re-search articles and writings that deemed to be valuable to our re-search work.

In the fourth stage, we studied the articles and assessed the result that have been derived by the researchers in the selected articles.

In the final stage, the result derived from the studied articles were compared to contemporary and more recent research works and were evaluated based on their qualitative importance and significance in the scope under study.

This review protocol is very critical to the study for preparing Literature Review. The following subsections elaborately present the details of the review protocol that has been adopted in this study.

3.1 Questions

This literature review aims to summarize the works that have been done till date on Bangla Document Clustering.

1. **Which ML techniques have been used in Bangla Document Clustering?**
   This question aims to identify the Machine Learning techniques that have been used to implement Document Clustering in Bangla Language. This shall help us estimate the advancement of Bangla Document Clustering from a technological perspective. In near future, researcher keen on working on this field shall have a broader perspective from this on the works that have already been done and those that still needs research efforts.
2. **What is the overall accuracy of the implemented models?**
   This help us look into the results of the researches under study and represent the accuracy attained in document clustering of Bangla language by the implemented methods. Estimation accuracy is the primary metric for evaluation.
3. **What datasets are used in the researches?**
   With this question we take a look into the datasets that have been used to conduct the researches. We compare the datasets and their favorability to the implemented algorithm. Thus, providing us with an insight on how robust the algorithms are on Bangla datasets and how they are different from internationally used datasets.
4. **Is there any difference in accuracy due implementation of the methods in Bangla?**
   Through this question we compare the conducted researches in Bangla to their prominent English counterparts. This provides us with a view of how the difference in language impacts the accuracy of a clustering model.

3.2 Search Strategy

In search strategy we devised how we would carry out the search to collect information on our field of research and gather the necessary articles. In the process we have followed some steps that are discussed below:

3.2.1 Searching the Terms

We started our work by looking into the earlier researches on Bangla document clustering. And there have been some remarkable works by very proficient researchers recently. The key terms related to Bangla Document Clustering were derived from these papers and were looked up on the internet for their definitions and background. We learned about the machine learning algorithms that are used for document clustering. The following information tasks were carried out during this phase:

- Machine Learning and the scope of machine learning in details.
- Definition of clustering and how clustering is applied on information and documents.
- Word embedding and placement of words from various document into a vector space.
- Finding relation between words by measuring their Euclidean distance in the vector space.
- Learning clustering algorithms like K- Means, Hierarchical Clustering, Hierarchical Density-Based Spatial Clustering of Application with Noise.

From studying the papers, we learned techniques of using the algorithms for clustering the documents and various methods of analyzing the accuracy of the algorithms.

3.2.2 Finding the Appropriate Resources

Various online and offline resources were considered as valuable source of information regarding the study at the initial phase of the study. Among the offline resources the former students of Shahjalal University of Science of Technology stood out as the most valuable resource for papers particularly on Bangla language. On the other hand, we also tried to make contact with a number of people who are currently working on Bangla language and document clustering for further resources.

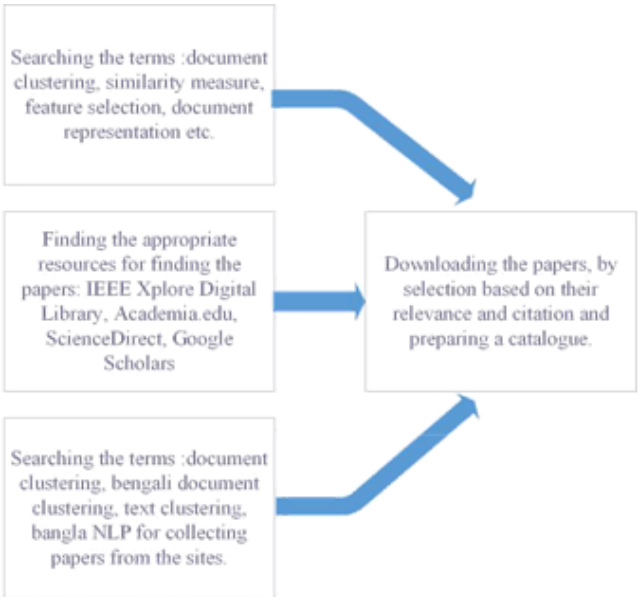

Fig. 3. Search strategy for gathering knowledge and research papers from various sources

On the other hand, academic websites such as, IEEE Xplore Digital Library (ieeexplore.ieee.org), Academia.edu (www.academia.edu), ScienceDirect.com (www.sciencedirect.com), Google Scholars (scholars.google.com) were used as resources for collecting papers on Bangla language and document clustering.

3.2.3 Searching Process

According to SLR we need to search all the relevant resources comprehensively. Therefore, the searching process was defined and subdivided into the following phases:

- **Searching by Specific Term:** In this phase we entered the following terms, ”Bangla Document Clustering”, ”Bengali Document Clustering”, ”Document Clustering”, ”Text Clustering”, ”Bangla NLP”, ”Bengali Clustering” as search terms in the elected resources. The result as found are demonstrated in Table 1.
- **Searching by Citations**: When we searched for the specific pa-pers on various resources there were citations provided along with the respective papers. We followed the citations to find other papers relative to the respective field of study. Thus, obtaining more related and recent papers from the resources.

3.3 Study Selection

After completing the searching phase, we focused onto the study selection phase of the research. In this phase we selected specific papers to catalogue under study from the papers that were found during the implementation of the Search Strategy.

The usefulness and importance of the candidate papers were used as metrics to catalogue the papers under study. The following characteristics of the candidate papers were evaluated during the generation of the study selection catalogue:

(1) Relevance to the topic under study.

(2) Relevance to the language of the topic under study

(3) Number of citations

(4) Contemporary and recent topics and subjects.

The candidate papers were selected on base of the aforementioned criteria. Then they were enlisted with detailed information as a catalogue with the help of Microsoft Excel. Three different excel files were created as Paper Index, Index Selection and Review Selection respectively. Paper Index contained a catalogue of the obtained papers along with the bibliography and link to the paper. Index Se-lection contained the key terms discussed in the paper along with the main objective and methods discussed in the paper, and lastly, Review Selection contained a catalogue containing the figures and tables in the papers along with their reference studies.

A total of 50 papers were obtained among which 19 were in Bangla and the rest of them were on English Language.

3.4 Assessment

From carrying out of the search an overview of the number of articles in Bangla and English Language can be seen as it is shown in 1. When searched by the terms it can be seen that the number of articles in Bangla are still very less when compared to the number if works that have been conducted on other Languages.

From the huge number of collected papers research papers were stored on local machine for further reading and analysis. The papers were selected on the basis of their relevance mostly by their tittle and citations and number of citations. The date of publication of the research work and the conference or journal they were published in were also taken into account in the process. As a catalogue of the papers was created, the duplicate papers, if there any, were removed in the process. The papers were studied individually. The results and methodology of the papers were discussed among the researchers. And finally the studies relevant to the findings of this review were considered to ensure the reliability of the research. From the catalogued papers in the index selection it can be seen that, the earliest research on Bangla text categorization was in 2006 [18]. And yet there are a lot of work left to be done in this field.

3.5 Comparison

For comparison the assessed researches were evaluated on the basis of how recent the implemented methods are and the quantity of work done in the particular field and their applications. Given the results and discussion at the end of the collected papers, distinct in-formation on how the implemented method in a particular research has performed in terms of their efficiency and accuracy have been acquired.

The papers were compared in terms of their relevance, likeness of methodologies and accuracy.

4. RESULT AND DISCUSSION

The number of results on Google Scholar when search for the terms ”Bengali Document Clustering” and ”Document Clustering” in a particular year from 2015 to 2019 have been demonstrated in 2 From the table it can be seen that the number of results for the term ”Document Clustering” has been on a constant decline from 2015 to 2019. Whereas, the number of result for the term “Bengali Document Clustering” has been slightly increasing at the same time.

From this table it is evident that, the diversity of the research works on document clustering are gradually converging, whereas the re-search works on Bangla document clustering is increasing over time by a very marginal number.

In 1, the search results on various research articles has been shown and papers hosting websites against the terms has been specified on the left side of the table. This table provides with a brief look at the expansion of researches on document clustering on various individual terms. From this table it is evident that the number of researches works on Bangla related to the field of document clustering is very little. Whereas the amount of researches on the other languages is ever-growing,

Search for the terms "Bangla Document Clustering" and "Bengali Document Clustering" combined brought up only one research paper that specifically aimed to perform the implementation of document clustering algorithm [3].

The research conducted in [20] has successfully been able to demonstrate the performance analysis of three different word embedding methods on Bangla corpus, which is one of key factor in document clustering. Word embed dings allow to exploit ordering of the words and semantics information from the text corpus.

Table 1. Search Results from the selected resources on key terms

	Google Scholar	Science Direct.com	Acade-mia.edu	IEEE-Xplore
Bangla Document Clustering	4,160	132	1,128	16
Bengali Document Clustering	16,000	257	1,754	10
Document Clustering	2,200,000	266,766	133,548	3,430
Text Clustering	3,220,000	213,860	218,344	4,162
Bangla NLP	2,930	14	634	37
Bengali Clustering	5,610	598	2,567	23

The research conducted in [20] has successfully been able to demonstrate the performance analysis of three different word embedding methods on Bangla corpus, which is one of key factor in document clustering. Word embed dings allow to exploit ordering of the words and semantics information from the text corpus.

According to [20] both similarities and dissimilarities were found among the clusters derived from the implemented approaches. Even though due to polysomic nature of the vocabulary of Bangla language, Fast Text Skip gram yielded the best results. On the other hand in a research, Word Mover's Distance was used to measure the distances between documents [3] to propose a pipeline architecture for Bangla Document Clustering using different clustering Algorithms. It used an already computed word embedding model in Word2Vec. But from [20] it was evident that among the various word embedding models (including Word2vec and Fast Text Skip gram) Fast Text skip proved to be more efficient than Word2Vec in Bangla language.

There have been some works in the field of Bangla document categorization. It is similar to document clustering in its goal but different in methodology as it uses supervised learning from accomplishing the task. A study on different types of approaches on Bangla Document categorization has compared three different supervised learning techniques for Bangla document categorization

[12]. In a research, graph-based edge-weighting approach was de-vised to measure semantic similarity between Bangla words [24]. It has been verified using user studies also. This particular approach may prove to be useful in later applications of similarity measure in Bangla document clustering.

In [14], the researchers have demonstrated comparison between various clustering techniques, particularly the two widely used document clustering techniques: agglomerative hierarchical clustering and K-means. It has been demonstrated in this research that agglomerative

Table 2. Search Results for the terms "Bengali Document Clustering" and "Document Clustering" Google Scholar

Year	Number of Search Results on Bangla	Number of Search Results on Others
2015	2,160	55,700
2016	2,390	53,300
2017	2,560	50,600
2018	2,610	45,000
2019	2,680	32,100

Hierarchical clustering performs poorly in contrast to the K-means algorithm and its variants. And bisecting K-means algorithm performs better than K-means algorithm. Due its superiority in performance K-means algorithm has been used in various application of document clustering in the subsequent years.

In [6], researchers have proposed Context Semantic Analysis which is novel knowledge-based method that is aimed to make estimation of inter-document similarity more efficient. It uses a knowledge base to compute inter-document similarity. Published in 2019 this is one of the most recent works in measurement of inter-document similarity, which is one of the vital aspects of document clustering methodologies.

Again for text similarity in vector space models it has been found in [23] that when compared the TFDIF (short for term frequency–inverse document frequency) is better than more complicated methods like extended TFIDF model, LSI topic model and D2V neural model. While the more complicated methods can yield slight better results, they require extensive works for tuning. There-fore the use of the more complicated methods is appropriate only for dense text with rough similarity detection.

In one of their papers, the researchers have used the very recent Krill Herd algorithm for solution of several complex global optimization problems [1]. They improved upon the hybrid krill herd algorithm by combining it with objective functions that yielded very respectable results based on the evaluation measures in terms of precision, F-measure, recall, purity, entropy and accuracy. In the same year they proposed a new feature selection method using particle swarm optimization algorithm to improve the document clustering process [2]. The feature selection improves the effectiveness of document clustering algorithm by introducing a new subset containing informative features as an input to the k-means algorithm. The application of Particle Swarm Optimization improves upon the clustering results on almost all of the experimental data sets.

Most learning methods treat clustering and dimensionality reduction separately. But in a very recent work it has been attempted to perform a joint dimensionality reduction and k-means algorithm for clustering [25], where the dimensionality reduction is achieved using a deep neural network. The approach proved to be effective on a variety of data-sets using real and synthetic data experiments. There have also been researches on clustering and learning representation jointly. When learning representation is favorable towards the clustering data and well adapted to the clustering algorithm, better results are yielded when the task is performed in jointly. In this paper [10] a new approach is proposed to cluster k-Means and learning representations jointly by considering the k-Means clustering loss as the limit of a differentiable function. This is the first approach that truly jointly optimizes, through simple stochastic gradient descent updates, representation and k-Means clustering losses.

Through comparisons the eligibility of the proposed methods were also tested.

A universal taxonomy has also been proposed in [4] that utilizes deep neural network. This shows that taxonomy facilitates the development of more sophisticated methods in a structured and analytical way. This makes development of new methods easier by enabling replacement and recombination of distinct aspects of existing methods without having to discover and implement every-thing on the related field. Very recently the works done in [8], has opened a new door for researches in NLP. Due its bidirectional contextualization it will help create superior data-sets for document clustering. In an implementation of document clustering based on weighted BERT model [16], pre-trained language representation model is used to generate contextualized sentence embed dings. Their experiment on four data sets showed higher accuracy in the results than unweighted average methods. The pre-trained Bangla data-set can be further fine-tuned for preparing state-of-the-art data-sets for document clustering task thus taking Bangla Document Clustering to new heights.

5. CONCLUSION

There have been introduction of numerous algorithms in the time-line of Document Clustering over time. New clustering algorithms continue to appear till today, which later, are subsequently applied to other fields like Document Clustering. But from the research it can be seen that the number of researches in Bangla is still very little compared to the works done in other Language. And the works that have been done are all very recent, thus not allowing us have a speculation of older works in the field the trends of the researches.

New methods and techniques are being proposed for NLP regularly, thus providing new methods to be adopted for document clustering. While new clustering methods are being implemented every once in a while, the methods that have been in the use for a long are also being improved upon by researchers. And over the time, the convergence of the techniques for standard performance over every variation of data set is promising. Whereas, looking at the results, the implementation of these methods on Bangla language is still at a very early stage and shall require more efforts by researchers to provide this field with a consistent base upon which to improve and implement.

6. REFERENCES

[1] Laith Mohammad Abualigah, Ahamad Tajudin Khader, and Essam Said Hanandeh. A combination of objective functions and hybrid krill herd algorithm for text document clustering analysis. Engineering Applications of Artificial Intelligence, 73:111–125, 2018.

[2] Laith Mohammad Abualigah, Ahamad Tajudin Khader, and Essam Said Hanandeh. A new feature selection method to im-prove the document clustering using particle swarm optimiza-tion algorithm. Journal of Computational Science, 25:456– 466, 2018.

[3] Adnan Ahmad, Md Ruhul Amin, and Farida Chowdhury. Bengali document clustering using word movers distance. In 2018 International Conference on Bangla Speech and Lan-guage Processing (ICBSLP), pages 1–6. IEEE, 2018.

[4] Elie Aljalbout, Vladimir Golkov, Yawar Siddiqui, Maxim-ilian Strobel, and Daniel Cremers. Clustering with deep learning: Taxonomy and new methods. arXiv preprint arXiv:1801.07648, 2018.

[5] Nicholas O Andrews and Edward A Fox. Recent devel-opments in document clustering. Technical report, Depart-ment of Computer Science, Virginia Polytechnic Institute & State . . . , 2007.

[6] Fabio Benedetti, Domenico Beneventano, Sonia Bergam-aschi, and Giovanni Simonini. Computing inter-document similarity with context semantic analysis. Information Sys-tems, 80:136–147, 2019.

[7] Tanmoy Chakraborty, Dipankar Das, and Sivaji Bandyopad-hyay. Semantic clustering: an attempt to identify multiword expressions in bengali. In Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World, pages 8–13, 2011.

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, 2018.

[9] Encyclopedia Britannica, 2019.

[10] Maziar Moradi Fard, Thibaut Thonet, and Eric Gaussier. Deep k-means: Jointly clustering with k-means and learning repre-sentations. Pattern Recognition Letters, 2020.

[11] Jasmine Irani, Nitin Pise, and Madhura Phatak. Clustering techniques and the similarity measures used in clustering: A survey. International journal of computer applications, 134(7):9–14, 2016.

[12] Md Islam, Fazla Elahi Md Jubayer, Syed Ikhtiar Ahmed, et al. A comparative study on different types of ap-

proaches to bengali document categorization. arXiv preprint arXiv:1701.08694, 2017.

[13] P Jaganathan and S Jaiganesh. An improved k-means algo-rithm combined with particle swarm optimization approach for efficient web document clustering. In 2013 International Conference on Green Computing, Communication and Con-servation of Energy (ICGCE), pages 772–776. IEEE, 2013.

[14] Michael Steinbach George Karypis, Vipin Kumar, and Michael Steinbach. A comparison of document clustering techniques. In TextMining Workshop at KDD2000 (May 2000), 2000.

[15] Barbara Kitchenham and Stuart Charters. Guidelines for per-forming systematic literature reviews in software engineering. 2007.

[16] Yutong Li, Juanjuan Cai, and Jingling Wang. A text docu-ment clustering method based on weighted bert model. In 2020 IEEE 4th Information Technology, Networking, Elec-tronic and Automation Control Conference (ITNEC), vol-ume 1, pages 1426–1430. IEEE, 2020.

[17] Christopher Manning, Prabhakar Raghavan, and Hinrich Schutze¨. Introduction to information retrieval. Natural Lan-guage Engineering, 16(1):100–103, 2010.

[18] Munirul Mansur. Analysis of n-gram based text categoriza-tion for bangla in a newspaper corpus. PhD thesis, BRAC University, 2006.

[19] Anil Kumar Patidar, Jitendra Agrawal, and Nishchol Mishra. Analysis of different similarity measure functions and their impacts on shared nearest neighbor clustering approach. In-ternational Journal of Computer Applications, 40(16):1–5, 2012.

[20] Zakia Sultana Ritu, Nafisa Nowshin, Md Mahadi Hasan Nahid, and Sabir Ismail. Performance analysis of different word embedding models on bangla language.I In 2018 Inter-national Conference on Bangla Speech and Language Pro-cessing (ICBSLP), pages 1–5. IEEE, 2018.

[21] Claude Sammut and Geoffrey I Webb. Encyclopedia of ma-chine learning. Springer Science & Business Media, 2011.

[22] Neepa Shah and Sunita Mahajan. Document clustering: a de-tailed review. International Journal of Applied Information Systems, 4(5):30–38, 2012.

[23] Omid Shahmirzadi, Adam Lugowski, and Kenneth Younge. Text similarity in vector space models: a comparative study. In 2019 18th IEEE International Conference on Machine Learning and Applications (ICMLA), pages 659–666. IEEE, 2019.

[24] Manjira Sinha, Tirthankar Dasgupta, Abhik Jana, and Anu-pam Basu. Design and development of a bangla semantic lex-icon and semantic similarity measure. International Journal of Computer Applications, 975:8887, 2014.

[25] Bo Yang, Xiao Fu, Nicholas D Sidiropoulos, and Mingyi Hong. Towards k-means-friendly spaces: Simultaneous deep learning and clustering. In international conference on ma-chine learning, pages 3861–3870, 2017.