Volume 9, Issue 10, October – 2024 *International Journal of Innovative Science and Research Technology (IJISRT)*
ISSN No:-2456-2165 *https://doi.org/10.5281/zenodo.14730649*
IJISRT25OCT080 *www.ijisrt.com*

A Systematic Literature Review of Similarity Analysis Techniques for Bangla Text

Hasan Mahmud[1]; Mahmud Hasan[2];
Farhana Ryhan Kabir[3]; Md. Zahiruddin Aqib[4]

[1]United International University, Bangladesh
[2]United International University, Bangladesh
[3]United International University, Bangladesh
[4]United International University, Bangladesh

Publication Date: 2025/01/24

Abstract

Natural language processing (NLP) includes similarity analysis of words, phrases, or texts in the context of lexical analysis and semantic analysis. Because Bangla is a language with few resources, this process is more difficult for this language. Different types of methods are used to extract the similarity based on meaning. Compared to lexical similarity analysis, semantic similarity analysis is more difficult. We primarily addressed the theoretical aspect of the semantic similarity analysis in this study. A small number of approaches are investigated and found to be effective in identifying the similarities in the context of Bangla NLP study. The corpus in Bangla WordNet is not generally available to work with. Bangla similarity is a concern based on research conducted thus far with WordNet, LDA, LSA, Word2Vec, Doc2Vec, and WMD. We have reviewed all these techniques and prepared a comparative study among them in this paper.

Keywords: *AI, WordNet, Jaccard Similarity; Semantic Textual Similarity; Statistical Similarity; Cosine Similarity; N-gram; Natural Language Processing; Character-based Similarity; Term-based Similarity.*

I. INTRODUCTION

We are currently witnessing a technological paradigm shift led by artificial intelligence (AI). AI's Natural Language Processing (NLP) domain is diverse and difficult. Similarity analysis is one of the many difficulties in NLP. Every language uses a similarity analysis to show how similar two words, phrases, paragraphs, or documents are to one another. For artificial intelligence, cognitive science, semantic similarity, and numerous other related applications, Natural Language Processing (NLP) is crucial [16]. Similarity analysis is used in these applications for sentiment analysis, machine translation, text summarization, and plagiarism detection, among other things. To differentiate evaluations and measurements between words, sentences and paragraphs in a language, Similarity Analysis is un- doubted required [13]. In language analysis, similarity analysis describes the cohesiveness of features shared by any two words. This is also used for the measurement of abstract similarity between two words, phrases, paragraphs, papers, or even two pieces of texts.

Comparing Bangla entries to English entries, determining resemblance is a more difficult task. English provides enough resources to compare two words, phrases, paragraphs, documents, or even text fragments. The number of terms that are present in both text segments is used to calculate the similarity score. However, these metrics are unable to determine the resemblance above a purely coincidental level. Moreover, this matching can only estimate the textual similarity but not semantic[16]. Let we think of two texts, "তিনি একজন শিক্ষক" (Meaning, He is a Teacher) and "তিনি একজন শিক্ষকের পিতা" (meaning, He is a father of a Teacher). According to the lexical matching, there are three lemmatized terms ("তিনি", "একজন", "শিক্ষক")exist in both sentences. That means the similarity between two sentences is nearly 0.75 on a scale of 1.0. In between two texts, there is no strong semantic connection between these two sentences. Lets we consider another example of two sentences, "তার একটি কলম আছে" (meaning, He has a Pen) and " সে লিখছে" (meaning, He is writing). There is no single word that exists between two sentences but they have semantic similarity [16].

Hasan Mahmud; Mahmud Hasan; Farhana Ryhan Kabir; Md. Zahiruddin Aqib, (2025), A Systematic Literature Review of Similarity Analysis Techniques for Bangla Text. *International Journal of Innovative Science and Research Technology*, 9(10), 3051-3058. https://doi.org/10.5281/zenodo.14730649

Table 1. Lexical matching between two sentences

Sentence 1	Sentence 2	Similarity Between Sentences
তিনি একজন শিক্ষক	তিনি একজন শিক্ষকের পিতা	Lexically Similar but not Semantically
তার একটি কলম আছে	সে লিখছে	Semantically Similar but not Lexically

A lexicon is like a vocabulary of a language consisting of the entries of words and expressions. The lexical entry contains two types of information: form and meaning [17]. The meaning of a word and gaining the information through the expression is known as lexicon semantics. Lexicon semantics is the core challenge in natural language processing tasks like web mining, text mining, information retrieval, information extraction, and machine translation. Different types of schemes are applied to represent the lexicon in different language database and resources like WordNet, FrameNet, ConceptNet, etc.

Since Bangla NLP is yet to be matured, few literature is available covering the similarity analysis technologies.

Among them some important research papers are reviewed and a summary is presented in this article.

II. MOTIVATION FOR CONDUCTING THE REVIEW

Bangla is an Indo-Aryan language and spoken by the Bangla in Bangladesh and India. The Language is the most widely spoken language of Bangladesh and second most widely spoken in India. In the world, approximately 228 million native speakers and another 37 million second-language speakers speak in Bangla[6]. Among 6500 languages Bangla is the seventh most-spoken language and the fifth most-spoken native language by total number of speakers in the world [9][3]. Although a large number of people speak this language there has been relatively little development in Bangla NLP. As a consequence, there is no well-established WordNet in Bangla Language. Researchers are trying to build a rich WordNet to ease similarity analysis techniques between Bangla entries. There are numbers of techniques to measure this similarity. The motivation for conducting the review is exploring modern techniques and finding popular techniques suitable for similarity analysis in Bangla language. This paper is organized as follows. Section 3 focuses on pre-processing techniques used to find similarities. In section 4, the classification of various techniques of similarity which are highlighted in Bangla language, and previous work on these techniques to find similarity are discussed. In section 5, we highlighted the pros and cons of different types of similarity analysis techniques. In section 6, we discussed the progress of the similarity analysis in Bangla language and the future of the similarity analysis in Bangla language and finally concluded the paper.

III. PRE-PROCESSING TECHNIQUES OF FINDING THE SIMILARITY

Similarity is computed between two texts by using a predefined word hierarchy which has words, meaning, and relationship with other words [7]. These similarities can be measured by different methods and there are various preprocessed techniques. Lemmatization, stemming, parts of speech tagging, word segmentation, chunk parsing, clustering, and translation from Bangla to English are the traditional first techniques in NLP to process the word for finding similarity [1]. The next step is to search for the similarity analysis techniques which can help to find the accurate similarity from between two corpora. Structure-based measures, Information Content-based (IC) measures, feature-based measures, and hybrid-based measures are the four strategic methods to measure similarity. Machine learning models and word embedding are also used as a part of the similarity analysis process. Gensim is one of the popular word embedding models. The process of similarity for finding similarity are shortly described in Figure-1.

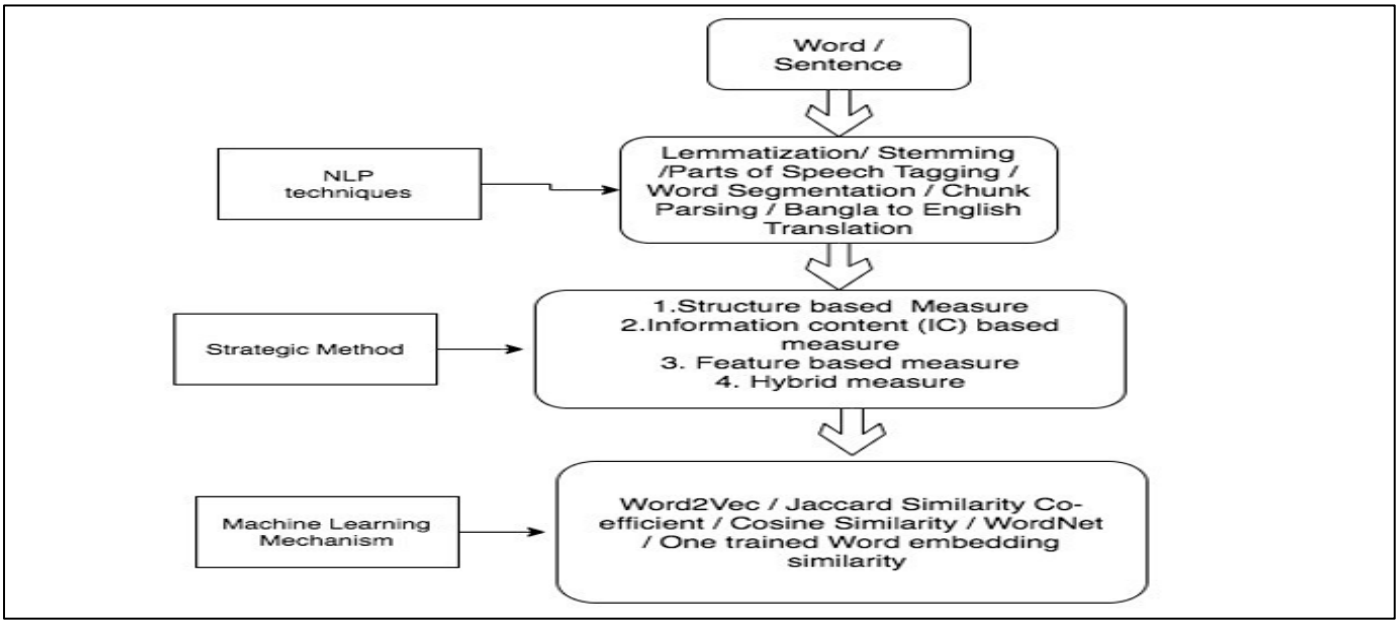

Fig. 1. Process of Similarity Analysis

IV. CLASSIFICATION OF SIMILARITY ANALYSIS TECHNIQUES

There are plethora of methods to calculate the similarity as long as NLP is concern. Different types of similarity analysis techniques are shown visually in the following Figure-2 [8] :

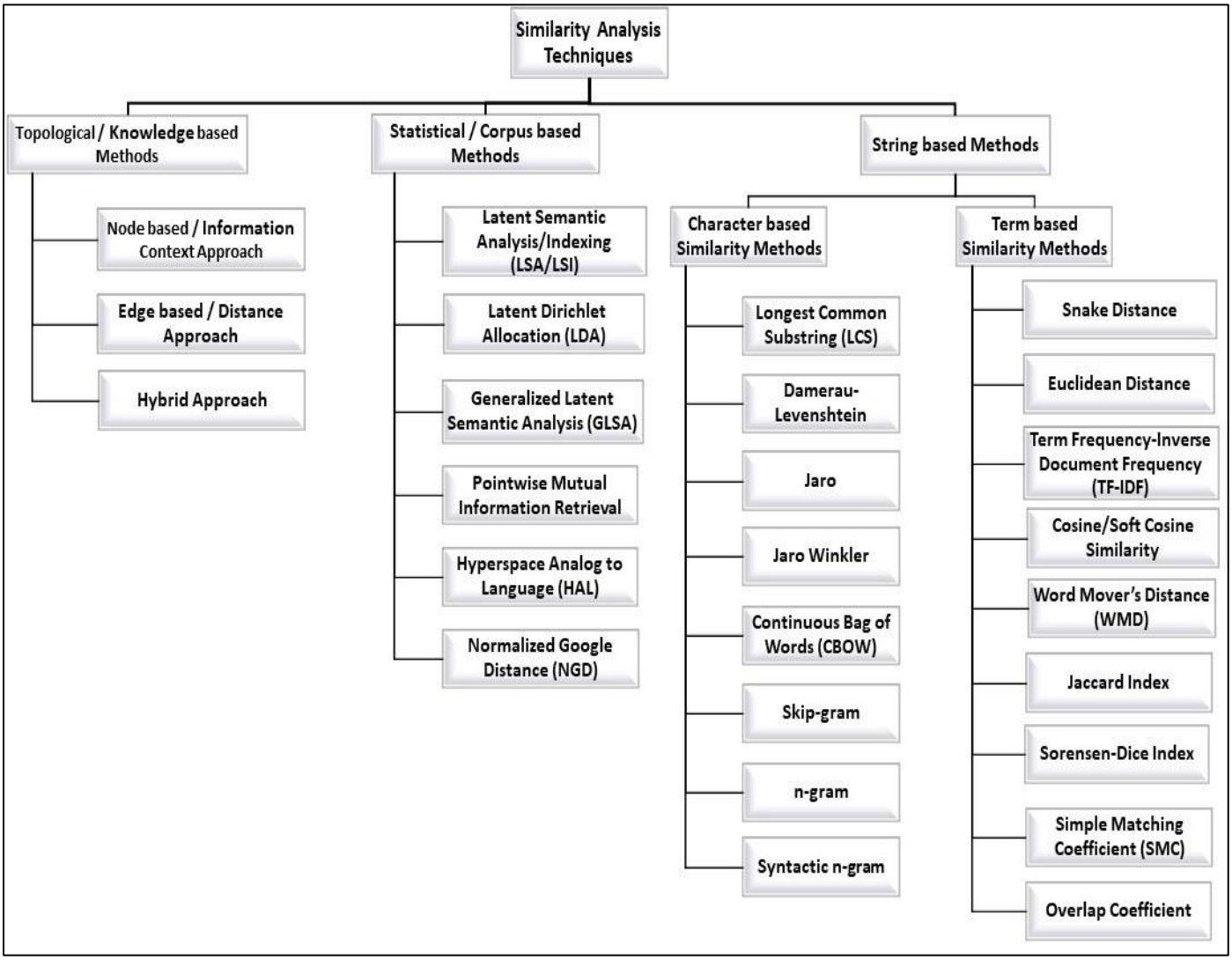

Fig 2 Classification of Similarity Techniques

Among all the methods there are very few methods that are explored through research works for Bangla similarity analysis. Some of the important papers are reviewed and summarized in this work.

A. Topological / Knowledge-based Method

- *Path Based.*

Path based similarity measures the distance using shortest paths between two concepts. To find the similarity be- tween two Bangla words or sentences researchers use path based similarity along with other similarity. Graph based edge weighting approach is used to measure the semantic similarity between two Bangla words. It also finds the short- est path between two words. Manjira Sinha et al.(2012) used Samsad Samarthasabdokosh by Ashok Mukhopadhyay for lexical representation to find the similarity[17]. Pandit et al.[13] translate the word from Bangla To English then calculate path based similarity using English WordNet which is a hierarchically organized lexical database. Path based similarity is also used to measure the similarity between two words from the WordNet as mentioned in[1].

- *Levenshtein Distance (Lev.)*

Levenshtein distance is used to measure the similarity between two strings. It is also called edit distance because it edits the string for one to another. Similarity score will be less if Levenshtein distance is greater. The paper[2] utilized the model to measure the spelling similarity between two strings and also in[1], the similarity metrics were used to measure the distance between two words after splitting the sentences.

- *Wu and Palmer (WUP similarity).*

Wu and Palmer (WUP) measures the similarity of two concepts from WordNet taxonomy. WUP measures the edge of two paths and calculates the depth of the Least Common Subsummer (LCS). LCS works to find the closest relation to both concepts. Sheikh Abjur et al.(2019) used WUP to find the similarity between two words[1]. They have used Bangla WordNet taxonomy and also have used WUP along with other metrics namely LevSim and Lin.

- *Lin Similarity.*

Lin similarity returns a score for calculating the similarity and it also shows how similar word senses are

based on Information Content based similarity strategy. Lin similarity finds the similarity of arbitrary objects. It relies on the structure of thesaurus. Thesaurus is a dictionary based model to find similarity. From the following equation we can find the combined similarity model derived from LevSim, WUP and Lin techniques.

$$total\_Similarity(t1,t2) = (Lev\_Sim(t1,t2) + WP\_Score(t1,t2) + Lin\_Score(t1,t2))/3$$

In the equation, total similarity of t1 and t2 tokens are calculated by the summation of Levenshtein distance, WUP distance and score of the Lin similarity.

➢ *WordNet.*

In paper [14], WordNet technique and cosine similarity were used to measure the similarity. WordNet is a semantic network where synonyms and word-senses are used as the nodes of the network and relation of them are the edges of the network. Synset term used in WordNet as a unique set of word-senses and synonyms. Bangla tweets were preprocessed to remove the noise and the remaining useful portions of the texts were considered. Two types of similarity measurements were considered in that paper: word level similarity and later used the word level similarity to measure the sentence level similarity. A same word can have different meanings that can be used in the sentences. According to that meaning a Synset is considered. If the scalar distance of the word fall into the same Synset or if the word has the same meaning then the distance will be 0. Otherwise the distance will be 1.

$$Sim(w_1, w_2) = \begin{cases} 1 & \text{(if d=0)} \\ 0 & \text{(if d=1)} \end{cases}$$

In sentence level similarity, let two sentences A and B creates a matrix according to their length of tweet. If the two matrix completely match then the score will be 1 otherwise 0. The matrix that is created will be represented as tweet A in a row or in the X-axis and tweet B in Y-axis. The equation of similarity can be determined as:

$$Sim(A,B) = \frac{\sum totalmatchingweight(A,B)}{\sum length(A,B)}$$

B. Statistical / Corpus based Method

➢ *Latent Semantic Analysis(LSA)-Distributed Semantic Model.*

LSA is used for text summarization. It gives the result as a matrix to present the corpus in the paragraph by using a bag of words(BOW). Nandi et al.[12] used this technique to measure the similarity and compared it with other techniques.

➢ *LDA and RNN.*

Mustakim Al Helal et al.[4] proposed a model for Bangla Topic Modelling and Sentiment Analysis in 2018. For topic modelling they proposed an unsupervised learning model using Latent Dirichlet Allocation(LDA) with bigram (a version of n-gram with n=2). The corpus used for the topic modelling research consists of 7,134 news articles from the online version of daily "Prothom Alo", a renowned newspaper of Bangladesh. They examined the perplexity and coherence of different topics. Finally they compared the performance of different models (LSI, HDP, LDA, Doc2Vec) and concluded that LDA performs better than other methods for Bangla Topic modelling. The proposed model is depicted in the following Figure-3:

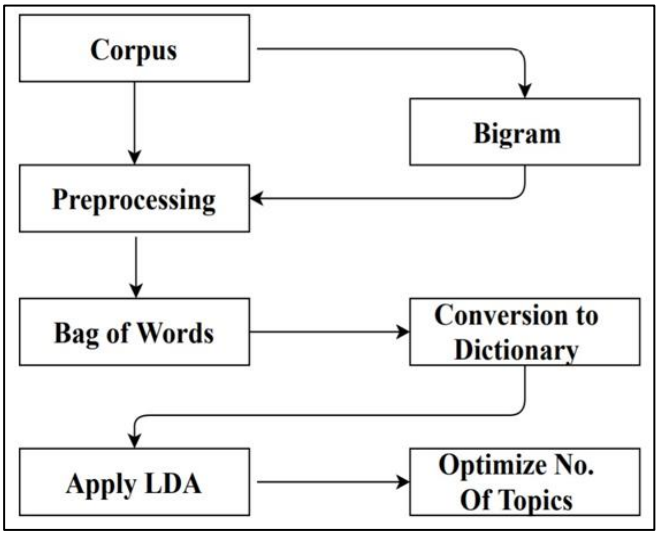

Fig 3 LDA Model Proposed by Mustakim Al Helal et al. [4]

To find out the optimal number of topics a coherence based method is also proposed. It prevents the model from being under-fitted or over-fitted. Moreover, for sentiment analysis, they proposed a Recurrent Neural Network (RNN) based model with Long-Short-Term Memory (LSTM) and Gradient Recurrent Unit (GRU). They have collected data from Facebook through Facebook graph API. The model has two versions: character level (Figure-4) and word level(Figure-5).

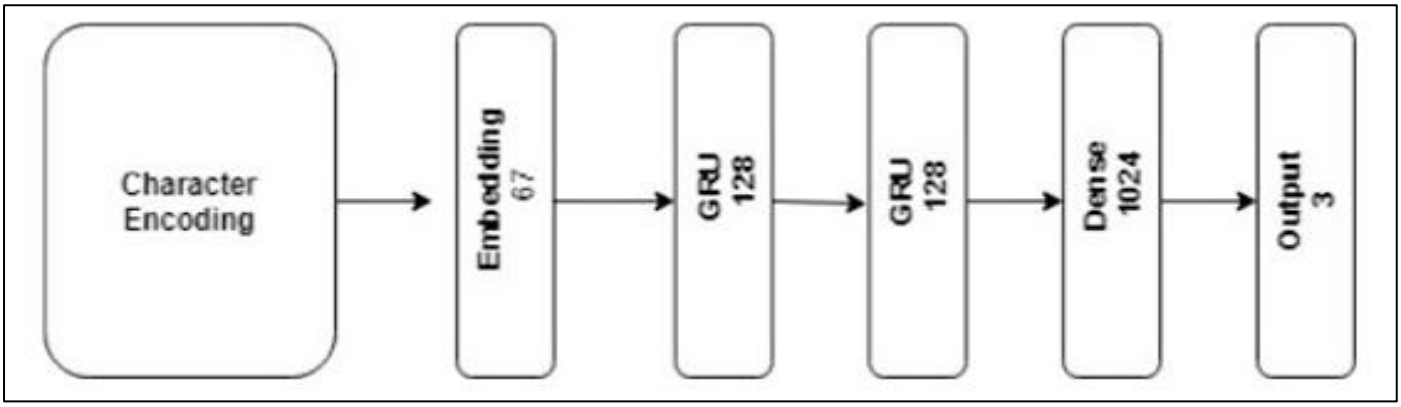

Fig 4 Character based Model Proposed by Mustakim Al Helal et al. [4]

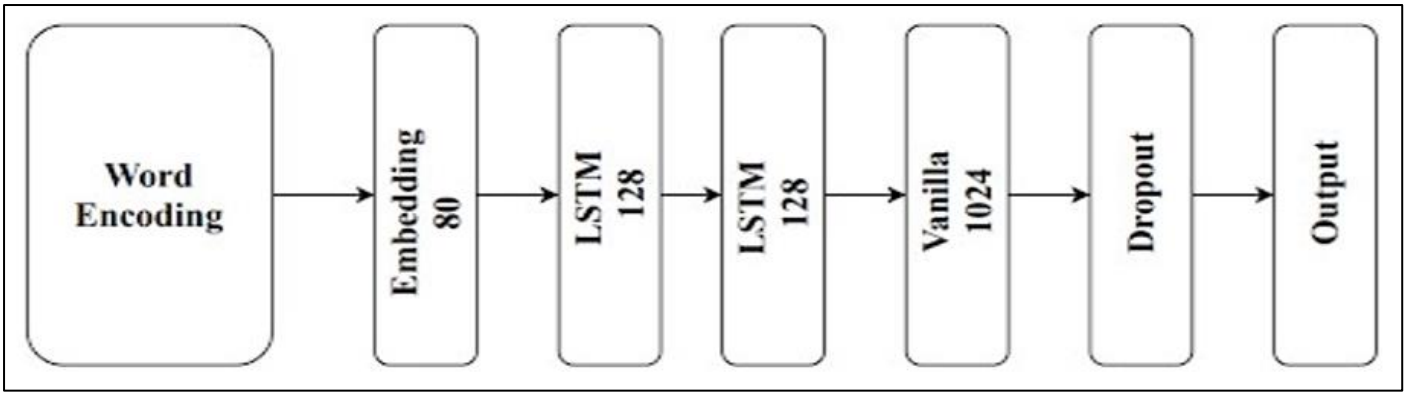

Fig 5 Word based Model Proposed by Mustakim Al Helal et al. [4]

Finally, from the research result, they concluded that the character level model performs better having 80% accuracy and the word level model has 77% accuracy from their baseline model.

C. String Based Similarity Method

➢ *Word2Vec-Word Embedding.*

Word2Vec, a word-embedding method is used to measure the similarity between two words. It is a distributional or corpus based semantic similarity model which consists of two models: Continuous Bag of Words(CBOW) and skip-gram. In [13][16], the CBOW model was used for measuring the distance.

➢ *Doc2Vec.*

In 2018, a content-based Bangla news recommendation system was proposed by Rabindra et al.[12]. They conducted a comparative study among doc2vec, LSA LDA techniques and found that doc2vec outperforms the other two techniques on a human-generated triplet dataset. doc2vec is a document embedding system that is based on a Neural Network(NN) driven architecture. It is an extension of a very renowned machine learning technique to gener- ate word vector namely word2vev. It is different from other common methods such as bag-of-words (BoW), n-gram models or averaging the word vectors. It can be trained in a fully unsupervised ML model from raw corpora without using any domain specific labeled dataset. It is very much scalable and no preprocessing or feature engineering is needed except tokenization. For training the doc2vec model they have chosen the architecture of distributed BoW (dBoW) and the popular gensim library for statistical semantics and text mining.

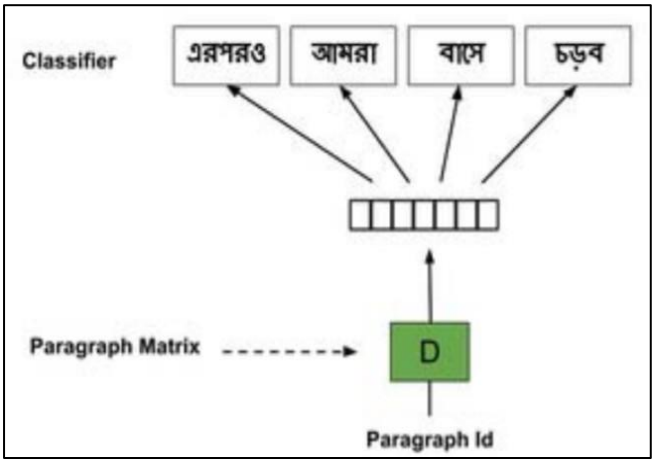

Fig 6 Doc2Vec Model from Rabindra et al. [12]

For the experiment, they have collected one week's data from 15 newspapers using a news crawler (Apache Nutch). The training corpus contained 300000+ uncategorized articles. For testing the performance of the news recommen- dation system, they have also collected 37,000 labeled news articles from 12 separate categories. The model exhibits language-independent learning and adaptation capability of a large corpus. According to the paper, they got 91% accuracy in doc2vec model while for LSA and LDA the accuracy level was 84% and 85% respectively. Figure-6 depicts the doc2vec model proposed by Rabindra et al. for Bangla language.

➢ *Jaccard Coefficient-Term Based.*

Jaccard similarity measures the similarity between two nominal attributes by insertion over union of two sets. As mentioned in [11], the results of Jaccard similarity are poor compared to Cosine similarity based measures.

➢ *Cosine Similarity-Term Based.*

Cosine similarity is the most popular effective similarity to measure similarity. This technique measures the similarity by using the cosine of angle between two vectors. Euclidean similarity gives less better results than Cosine similarity[11].

➢ *Word Mover's Distance (WMD).*

Word mover distance(WMD) is a new and effective technique in Ma- chine Learning(ML) for sentence or word similarity analysis. It can measures accurately the distance between two documents or sentences. It is designed to overcome the synonym problem but it is useful in large grammatical change. It gives better results than other techniques[11] to measure the similarity of two non-zero vectors and the similarity utilizing the cosine angle between sentences. It ascertains the closeness between sentences sets. The cosine similarity equation for is provided below:

$$Sim(A,B) = cos\theta = \frac{A.B}{||A||||B||}$$

Jaccard similarity is used to calculates the closeness between the finite set of text. Its values also are always less than or equal one.[10]

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}$$

Cosine similarity provides high score on the other hand Jaccard similarity shows a low score value. The Bangla text

similarity structure is very different from other languages. Those approaches work well for other languages since there are pre-trained word2vector file available. For the Bangla language, WMD is given an accurate similarity if compared with other approaches. Because its calculation is very simple for WMD methods to cause its use vector weight value from embedding file while others can't use.[10]

➢ *CBOW and Skip-Gram Word embedding.*

Sadman et al. presented a study regarding intrinsic evaluation of Bangla word embeddings in 2019 [15]. They trained CBoW and Skip-gram models using self-built corpus of more than 1.3 million unique words from 700,000 articles. They got average accuracy more than 90% for concept categorization and found that Skip-gram model performs better than the CBoW model.

➢ *n-gram.*

In report [5], the author proposes n-gram language model to find the Bangla similarity. Lets consider the following Table-2 to understand the technique.

Table 2 Pair of Similar Two Words

Similar Previous Two Words	Similar Next Two Words
সরকারের কাছে অনুরোধ সরকারের কাছে আবেদন	অনুরোধ করেন তিনি আবেদন করেন তিনি
ভিসার অনুরোধ ভিসার আবেদন	অনুরোধ করা হয়েছে আবেদন করা হয়েছে
রাজার কাছে অনুরোধ রাজার কাছে আবেদন	অনুরোধ করা হয় আবেদন করা হয়
কর্তৃপক্ষের কাছে অনুরোধ কর্তৃপক্ষের কাছে আবেদন	অনুরোধ জানানো হয় আবেদন জানানো হয়

Here, the sentences of অনুরোধ and আবেদন are paired between their similarities respectively. The list of three consecutive words w_i, w_{i+1}, w_{i+2} are created. For every word w_i two lists were proposed. One of them contain previous two word w_{i-2}, w_{i-1} and other list contain w_{i+1}, w_{i+2}. The matched words are calculated through the $list(w_{i-2}, w_{i-1})$ and $list(w_{i+2}, w_{i+1})$. Tri-gram model equation is used to measure the similarity between sentences. The pair of words are included into the same cluster. The similarity determining equations for n-gram technique is given below:

$$P(w_n|w_{n-1}, w_{n-2}) = count(w_{n-2}w_{n-1}w_n)/count(w_{n-2}w_{n-1})$$

or

$$P(w_n|w_{n+1}, w_{n+2}) = count(w_{n+2}w_{n+1}w_n)/count(w_{n+2}w_{n+1})$$

For finding two previous words, similarity of first word with the second word in the paired list will be:

$$P(w_i|w_j) = \frac{count(match(list(w_{i-2}, w_{i-1}), list(w_{j-2}, w_{j-1})))}{count(list(w_{i-2}, w_{i-1}))}$$

For finding two next words, similarity of first word with the second word in the paired list will be:

$$P(w_i|w_j) = \frac{count(match(list(w_{i+2}, w_{i+1}), list(w_{j+2}, w_{j+1})))}{count(list(w_{i+2}, w_{i+1}))}$$

V. SIMILARITY OF BANGLA ENTRIES BASED ON METRICS/TECHNIQUES

Different kinds of Bangla entries can be observed in the form of words or text, tweets, strings or documents. There are several metrics and techniques for identifying the different kinds of similarities. The findings of our study for detecting similarity among various forms of Bangla texts are summarized in the following Table-3.

Table 3 Similarity of Bangla Entries Based on Metrics/Techniques:

Sl no.	Reference	Similarity Type	Bangla Entry	Metrics/Techniques	Pros and Cons	Result
1	Sinha et al(2014) [17]	Semantic Similarity	Words	Graph Based Edge Weighting approach	Pros- Simple method is used. Identify the verbs, mythological words in this method. Cons- Words are divided into only six clusters.	Accuracy is good .
2	Pandit et al(2019) [13]			Path based similarity (WordNet) & Distributional (Word2Vec) semantic similarity	Pros- Compare four different types of method to find the best similarity. Cons- Bangla WordNet is weak compared to English WordNet.	Proposed model proves that word2vec is the state of the art model.
3	Shahjalal. et al. (2018) [16]		Texts	Word embedding (Word2Vec) + Pearson Correlation Coefficient	Pros- SemEval STS 2017 used as an evaluation metric. Cons- Need pre-trained word embedding model	Proposed algorithm gives 61% semantic similarity
4	Rudrapal et al.(2015) [14]		Tweets	WordNet + Partial Textual Entailment(PTE)	Pros- Simple method to measure similarity Cons- NLP resources are poor for Bangla Language.	Accuracy is good
5	Nandi. et al.(2018) [12]		Documents	Latent Semantic Analysis(LSA) & doc2vec & Latent Dirichlet Allocation (LDA)	Pros- Give comparative results. Cons- Traditional techniques give poor results.	doc2vec-91% LDA- 85% LSA- 84%
6	Masum. et al.(2019) [10]	Sentence Similarity	Summary (human given summary compare to machine given summary)	Word Mover's Distance (WMD) & Cosine Similarity & Jaccard Similarity	Pros- For better results compare three similarity approaches to find accurate similarity. Cons- Jaccard similarity technique gives the poor result compared to other techniques.	WMD gives better accuracy of the similarity.
7	Abujar et al.(2017) [1]		Sentences	Path measure + Levenshtein Distance + Wu and Palmer measure (WUP) + Lin measure	Pros- Proposed sentence similarity measuring model can measure English Bangla Language. Cons- Bangla WordNet is not stable.	Stable Bangla WordNet will give better results through this model.
8	Asadullah et al.(2007) [2]	Spelling Similarity	Strings	Levenshtein Distance(LD)	Pros- Use a finite state machine to build the algorithm. Cons- Lacking of handling multiple character errors.	Gives 70% accurate result.

If Bangla WordNet was rich in NLP then much accurate similarities among different types could be identified. Despite limitation on WordNet, researchers find some techniques to measure the similarities.Word2Vec , Jaccard coefficient similarity , Cosine similarity , WMD ,Path based similarity are the most effective mechanisms to find the similarity of different Bangla entries.

VI. CONCLUSION

This paper shows the findings of a systematic literature review on Bangla similarity analysis. Since there are few very works on Bangla similarity analysis, we investigated a little amount of studies. We presented a complete description of our systematic literature review process with findings and discussion about these findings. We discussed the current status of this research topic so that interested researchers can be benefited from this work. Our findings indicate that developing an approach for Bangla similarity analysis would address all challenges of complex rules of Bangla which is very much essential. For the practical usage of similarity analysis currently, various new word embedding techniques are being used which are hybrid in nature and they use some combinations of techniques described in this paper. The state-of-the-art word embedding techniques are Word2Vec, GloVe, BERT, ELMo, fastText, GPT-3 etc. where deep neural network is used. There is still a huge scope to work on Bangla language using these new

word embedding techniques. We hope to work on Neural Network (NN) based word embedding techniques for Bangla similarity analysis in the future to enhance the probability of accuracy of Bangla similarity analysis along with other areas of NLP domain.

ACKNOWLEDGMENTS

We are very much grateful to United International University for providing us the opportunity to work on this project.

REFERENCES

[1]. Sheikh Abujar, Mahmudul Hasan, and Syed Akhter Hossain. 2019. Sentence similarity estimation for text summarization using deep learning. In *Proceedings of the 2nd International Conference on Data Engineering and Communication Technology*. Springer, 155–164.

[2]. Munshi Asadullah. 2007. *Finite state recognizer and string similarity based spelling checker for Bangla*. Ph.D. Dissertation. BRAC University.

[3]. Jeffrey Hays. [n.d.]. BENGALIS. http://factsanddetails.com/india/Minorities_Castes_and_Regions_in_India/sub7_4b/entry-4198.html

[4]. Mustakim Al Helal. 2018. *Topic Modelling and Sentiment Analysis with the Bangla Language: A Deep Learning Approach Combined with the Latent Dirichlet Allocation*. Ph.D. Dissertation. Faculty of Graduate Studies and Research, University of Regina.

[5]. Sabir Ismail and M Shahidur Rahman. 2014. Bangla word clustering based on N-gram language model. In *2014 International Conference on Electrical Engineering and Information & Communication Technology*. IEEE, 1–5.

[6]. Mohammad Shibli Kaysar and Mohammad Ibrahim Khan. 2018. Word sense disambiguation for bangla words using apriori algorithm. In *International Conference on Recent Advances in Mathematical and Physical Sciences*. 61.

[7]. Yuhua Li, David McLean, Zuhair A Bandar, James D O'shea, and Keeley Crockett. 2006. Sentence similarity based on semantic nets and corpus statistics. *IEEE transactions on knowledge and data engineering* 18, 8 (2006), 1138–1150.

[8]. Goutam Majumder, Partha Pakray, Alexander Gelbukh, and David Pinto. 2016. Semantic textual similarity methods, tools, and applications: A survey. *Computación y Sistemas* 20, 4 (2016), 647–665.

[9]. Prianka Mandal and BM Mainul Hossain. 2017. A systematic literature review on spell checkers for bangla language. *International Journal of Modern Education and Computer Science* 9, 6 (2017), 40.

[10]. Abu Kaisar Mohammad Masum, Sheikh Abujar, Raja Tariqul Hasan Tusher, Fahad Faisal, and Syed Akhter Hossain. 2019. Sentence Similarity Measurement for Bengali Abstractive Text Summarization. In *2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*. IEEE, 1–5.

[11]. Abu Mohammad Masum, Sheikh Abujar, and Syed Hossain. 2019. Sentence Similarity Measurement for Bengali Abstractive Text Summa- rization. https://doi.org/10.1109/ICCCNT45670.2019.8944571

[12]. Rabindra Nandi, M. Zaman, Tareq Muntasir, Sakhawat Sumit, and Md. Jamil-Ur Rahman. 2018. Bangla News Recommendation Using doc2vec. https://doi.org/10.1109/ICBSLP.2018.8554679

[13]. Rajat Pandit, Saptarshi Sengupta, Sudip Kumar Naskar, Niladri Sekhar Dash, and Mohini Mohan Sardar. 2019. Improving Semantic Similarity with Cross-Lingual Resources: A Study in Bangla—A Low Resourced Language. In *Informatics*, Vol. 6. Multidisciplinary Digital Publishing Institute, 19.

[14]. Dwijen Rudrapal, Amitava Das, and Baby Bhattacharya. 2015. Measuring semantic similarity for bengali tweets using wordnet. In *Proceed- ings of the International Conference Recent Advances in Natural Language Processing*. 537–544.

[15]. Nafiz Sadman, Akib Sadmanee, Md Tanveer, Md Ashraful Amin, and Amin Ali. 2019. Intrinsic Evaluation of Bangla Word Embeddings. 1–5. https://doi.org/10.1109/ICBSLP47725.2019.201506

[16]. Md Shajalal and Masaki Aono. 2018. Semantic textual similarity in bengali text. In *2018 International Conference on Bangla Speech and Language Processing (ICBSLP)*. IEEE, 1–5.

[17]. Manjira Sinha, Abhik Jana, Tirthankar Dasgupta, and Anupam Basu. 2012. A new semantic lexicon and similarity measure in bangla. In *Proceedings of the 3rd Workshop on Cognitive Aspects of the Lexicon*. 171–182.