Volume 64, Issue 1, 2020
Journal of Scientific Research
Institute of Science,
Banaras Hindu University, Varanasi, India.

Accountability of NLP Tools in Text Summarization for Indian Languages

Pradeepika Verma[1*] and Anshul Verma[2]

[1*] Dept. of CSE, Indian Institute of Technology (ISM) Dhanbad, India. pradeepikav.verma093@gmail.com
[2] Dept. of CS, Banaras Hindu University, Varanasi, India. anshulverma87@gmail.com

Abstract: **In the era of digital world, online information is growing exponentially. It leads to emergence of inconvenient searching of relevant information in relevant time. In this regard, automatic text summarizer proves to be a good tool. It helps in creating a brief and meaningful form of the given text using natural language tool kit so that users can access the information in quick manner. Today, a lot of summarization tools are available for rich resource languages such as English. But, it seems difficult to summarize the text for Indian languages (low resource languages) due to limited availability of NLP tools and techniques for Indian languages. In this paper, we present a survey on existing text summarization methods and NLP tools for Indian languages. We also discuss about the issues associated with the Indian languages that are the bottlenecks for summarizing Indian language text.**

Index Terms: **Text summarization, Natural language processing, Indian languages, Language dependency.**

I. INTRODUCTION

The area of natural language processing gained much attention since the emergence of online information. It includes processing with Human generated language in either form to facilitate the user, for example, information extraction. In this regard, several challenges are introduced such as human language understanding, human language generation etc. The task of text summarization comes with these challenges. To make understandable the human languages, several NLP tools are available such as stemmer, PoS tagger, parser, named entity recognition system etc., but very limited for low resource languages.

Generally, Text summarization task can be classified into two categories, extractive summarization and abstractive summarization (Verma and Om 2016a,b,c). Extractive summarization extracts the most relevant sentences as it appears from the document while abstractive summarization generates new sentences from the set of concepts or topics residing in the document using natural language generation tools. It can also categories into several other domains such as single and multi-document summarization, query and topic focused summarization, monolingual and multilingual summarization etc. Single document summarization is simply defined as summary of text from a single document and summary from more than one document is called multi document. Multi document summarization is more difficult than single document in the sense of redundancy of text, compression of text from multiple documents, collection of significant information etc. Next, the summary process which involves summarization on the basis of Interrogating phrase is called query based summarization. Here, the system itself generate the topic according to given query and create summary of topic related documents. The topic based summarization involves summarization of topic related documents. It will help in finding a quick view of several documents in less time. Monolingual summarization involves summary for a particular language domain whereas multilingual summarization refer to generate summary for more than one language. Gupta (3) proposed a summarization system for Hindi and Punjabi language both so this system is called multilingual summarization system.

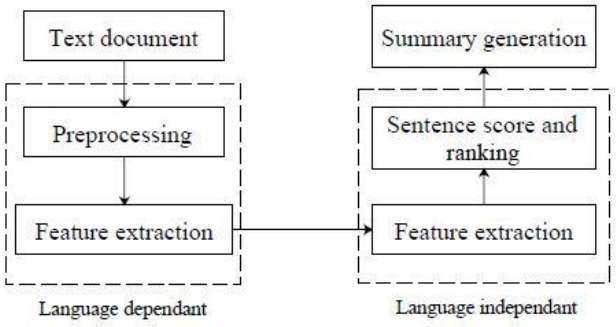

Fig. 1. Basic methodology for text summarization

Rest of the paper is organized as follows. Section II reviews

the state-of-the-art related to automatic text summarization for Indian languages. Section III briefly describes about issues in text summarization techniques for Indian languages. Section IV illustrates the impact of different NLP tools in the performance of summarization techniques. Finally, the paper concludes with the direction for future works.

II. Summarization Techniques for Indian Languages

There are several papers introduced related to the automatic text summarizer in the literature on different extractive and abstractive techniques for different Indian languages. Few of them are discussed here.

A. Hindi

Kumar et al. (2015) proposed a graph based technique for Hindi text summarization. The fundamental concept to present this approach is to find important information from a document of Hindi language. Graph based approach is used to find the relation between two sentences and find the importance of the sentence with respect to document. Here, they used the concept of semantic similarity to find the relevancy between sentences.

They assumed that the sentence with high relevance consist same information. Only that sentence should be added in summary if the importance of that sentence is high. (Kumar and Yadav 2015) proposed technique that is based on thematic words. Thematic words are generated by evaluating frequency of terms and their respective inverse frequency in the document. It generates a list of thematic words and creates summary by using these words. The generated summary is further processed by Hindi WordNet. (Gupta 2013) proposed an algorithm which summarizes the documents written in Hindi and Punjabi both. It is based on statistical approach. These are key phrases, cue phrases, nouns and verbs, negative terms, font feature, named entities, sentence position,sentence length and numerical data. System uses regression function to weight each feature.

B. Panjabi

Gupta and Lehal (2012) proposed a summarization method for punjabi text. It calculates the sentences based on nine weighted text features such as named entities, title words, keywords etc. They have used a rule based and dictionary based approaches to recognise the Punjabi words related to text features. (Gupta and Kaur 2016) proposed another punjabi text summarization method based on hybrid model of support vector mechine and simple text features. They have used entropy based approach for discovering important words in the document.

C. Bengali

Abujar et al. (2017) proposed a heuristic approach for Bengali text summarization. Different linguistic rules for extrction of each text feature has been used for obtaining better results. For example, they find the effect rate of every word in the document by several parameters like appearance of words in number of sentences, appearance of words in paragraphs, repeating nature etc. (Efat et al. 2013) also focused on finding text features scores to summarize the document in Bengali text. (Akter et al. 2017) proposed a Bengali summarizer based on

K-means clustering. They clustered the document into two according to their features' scores and top scored sentences from each clusters are extracted as summary sentences. (Sarkar 2012) also proposed a Bengali text summarizer based on text features scores of sentences.

D. Marathi

Rathod (2018) has proposed a marathi text summarizer based on text rank technique proposed by (Mihalcea and Tarau 2004). It is graph based approach where Pagerank algorithm has been used to obtain the significance of sentences. Also, it includes two unsupervised method for keywords and sentence extraction. (Gaikwad 2018) rule based Marathi text summarization method where noun words based a set of questions for each sentence is generated. Thereafter, each question is ranked according to their importance and top ranked questions are extracted to obtain their answers. The collection of answers of these questions are considered as the summary of the document.

E. Tamil

Devi et al. (2011) proposed a graph based summarization approach which is tested on Tamil text. It ranks each sentences based on their words frequencies and Levenshtein distance with other sentences. The average of these ranks are taken as final ranking of these sentences and top ranked sentences are extracted for generating summary. However, this method is language domain independent. Next, (Banu et al. 2007) also introduced a method for Tamil document summarization using semantic graph. A set of linguistic rules has been used to create semantic graph for the document. Moreover, support vector machine has been used to extract the sub graphs from the graph of document. An LF parser has been used to find the semantic similarity features.

F. Kannad

Jayashree et al. (2012) proposed a kannad text summarizer using keywords extraction. In this regard, they had combine GSS (Galavotti, Sebastiani, Simi) coefficient and TF-IDF method to extract keywords from the document. A list of keywords for each category using GSS and TF-IDF has been discovered and weight of each sentence is calculated by sum of the weights of these keywords. (Geetha and Deepamala 2015) also proposed a kannad text summarizer based on latent semantic analysis. In this work, they have find the semantic relationship between sentences using LSA. Moreover, the concept of SVD is used to generate summary of documents. (Kallimani et al. 2010) proposed a kannad text summarizer 'KanSum'which is based on the concept of AutoSum summary system for single Kannada document summarization. AutoSum is based on the features of

First line, sentence position, numerical data, keywords and combination function.

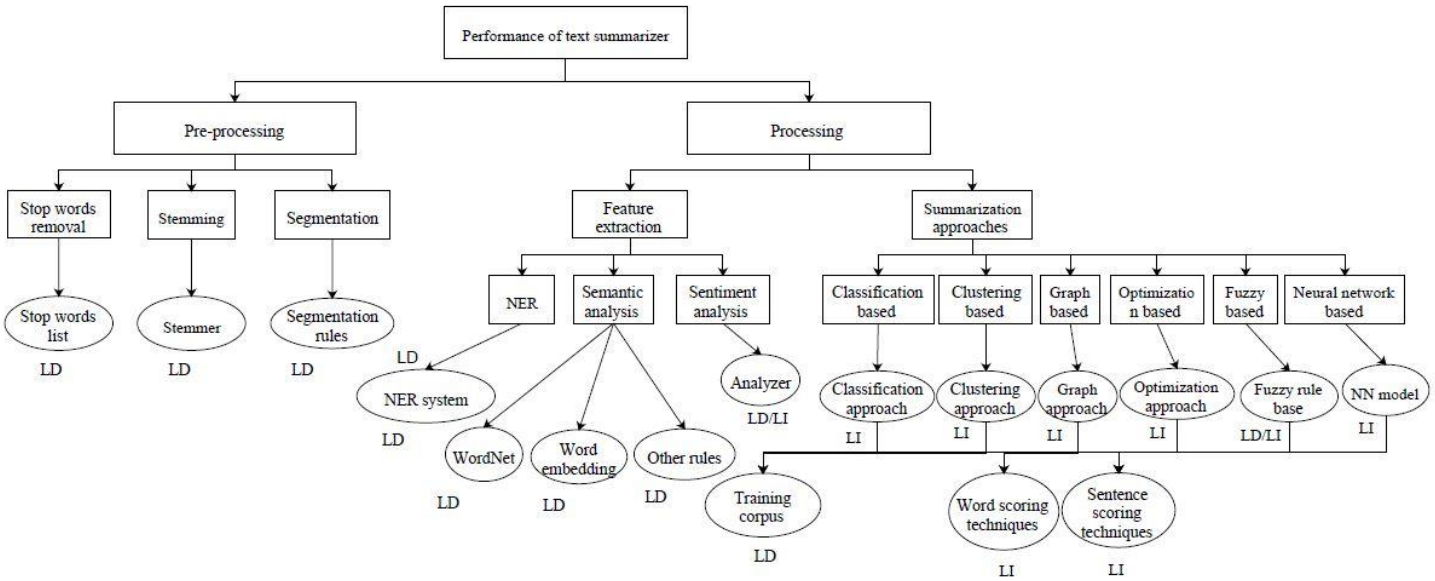

Fig. 2. Dependency in the performance of extractive text summarization (LD represents language dependent and LI represents language independent).

III. ISSUES IN AUTOMATIC TEXT SUMMARIZATION FOR INDIAN LANGUAGES

As illustrated in fig. 1, ATS process is partitioned into two major parts. One part is language dependent in the summarization process and another part is language independent. The later process is common for all summarizers in the context of language domain while previous process is dependent on resources and tools available for the language of the document. Therefore, the main challenge with Indian languages summarizers is to do accurate preprocessing and feature extraction. Fig. 2 shows the dependency of the performance of text summarization tools. It illustrates that availability of stop words list, stemmer, segmentation rules, named entity recognition system, wordnet and word vector dictionaries, sentiment analyzer and training corpus are basic essential tools and resources for text summarization. However, these are easily available for rich resource languages but are limited for Indian languages. Here, we highlight some resources that are available in the Indian languages contexts.

A. Stop words list

The first process for summary generation is to remove stop words from the document and to recognize the unique keywords. In this regard, some of the stop words lists are available by IIIT Hyderabad[1], TDIL[2], kaggle[3] and Ranks NL[4] where kaggle and Ranks NL provides stop words only for Hindi, Bengali, and Marathi. Some of the researchers are also proposed methods for self generation of stop words list from corpus. (Rani and Lobiyal 2018) proposed a method for construction of Hindi stop words list using statistical and information based methods. The information model which is based on entropy is used to find the significance of the terms in the corpus while statistical model which is based on TF-IDF feature is exploited for weighting the terms. According to these two models, every term is ranked by two ranks and final ranking of a term is given by summation of these two ranks. They found a total of 1475 stop words which is a big amount in comparison to other existing lists. Although, this technique can be applied on other languages corpus to generated their stop words. In the similar way, (Raulji and Saini 2017) also proposed a method for generating Sanskrit stop words based on frequency of term. However, No standard stop words list is available for Indian languages which causes in reduction in the performance of text summarizers.

B. Stemmer

Stemming is the process of normalizing the inflected words in the natural language text. A few work is done on this research area. (Ramanathan and Rao 2003) introduces a light weighted Hindi stemmer which works on longest match stripping using human generated list of total 65 suffixes. (Islam et al. 2007) also proposed a light weighted Bengali stemmer based on same approach as proposed by (Ramanathan and Rao 2003). They used 72 suffixes of verbs, 22 for nouns and 8 for adjectives. (Majumder et al. 2007) proposed a corpus based stemmer which works effectively for the primarily suffixing languages such as Bengali. It clusters the words with same stem but different variants using distance function to find out the stem word. (Saharia et al. 2014) proposed a rule based approach for

[1] https://ltrc.iiit.ac.in/showfile.php?filename=ltrc/internal/nlp/corpus/index.html
[2] http://tdil-dc.in/index.php?option=com_download&task=showresourceDetails&toolid=1637&lang=en
[3] https://www.kaggle.com/rtatman/stopword-lists-for-19-languages
[4] https://www.ranks.nl/stopwords

stemming the words of Assamese, Bengali, Bishnupriya Manipuri, and Bodo languages. They introduced a dictionary of frequent words to reduce the over-stemming and under-stemming errors and an HMM model to prevent the errors in special cases. (Dasgupta and Ng 2006) proposed a Bengali stemmer which is based on segmenting the words according to morphemes discovered in a large annotated corpus. (Pandey and Siddiqui 2008) introduces a Hindi stemmer based on finding probabilities for occurrences of suffixes and stem together using EMILLE corpus. (Majgaonker 2010) introduced a Marathi stemmer based on suffix stripping rules generated by human experts. (Suba et al. 2011) introduced two versions of Gujrati stemmer. One is light weighted, hybrid approach based stemmer and another is heavy weighted rule based stemmer. However, except all these methods, No standard tool to stem Indian language words with their effective performance in comparison of rich languages is available which causes in reduction in the performance of text summarizers.

C. Sentence boundary detection rules

Sentence boundary detection or segmentation process is the primary step of text summarization task. It is the task of detecting every sentence in the document. Four punctuation marks: period (.), exclamation mark (!), question mark (?), and pipe (|) are used to end the sentences in Indian languages. Hindi, Bengali, Punjabi languages use pipe to end a declarative sentence while other languages uses a period for the same. This punctuation mark has ambiguous definition as it is also used to represent an abbreviation. As we know, the English language also uses a period to end the sentences, but consists of other features such as 'capitalization of character at the beginning of every sentence' which is very helpful in detecting the sentence boundary. This is not an option for Indian languages which places an extra burden in segmentation. Moreover, a very limited work has been done in this area. In this regard, (Wanjari et al. 2016) proposed a rule based segmentation method for Marathi language, (Parakh et al. 2011) proposed a rule based segmentation method for Kannad language, (Ghosh et al. 2010) proposed a method for Bengali language, and (Devi and Lakshmi 2013) reported for Malayalam. No other work has been reported for other Indian languages.

D. Feature Extraction

To summarize the text, feature extraction is an essential part which requires a lot of processing with text. Feature extraction is mainly used to find the relevant or important sentence in the document. (Oliveira et al. 2016; Verma and Om 2018, 2019a,b,c,d; Verma et al. 2019) introduces 18 text features which requires the processing of named entity recognition, semantic analysis, sentiment analysis, cue-phrases recognition etc. in natural language text. Named entity recognition in text summarization can help in finding the centrality of the text. Moreover, the sentence appears with a number of these entities can be considered as significant sentence. However, a number of NER methods are available for some Indian languages but still their performances are limited to available rules and corpus for the language. A very limited work has been done on these areas which affects in generating summary for the text of Indian languages. Moreover, the available wordnets consist of limited synsets in comparison to English language and word vectors are also limited.

IV. IMPACT OF NLP TOOLS IN THE PERFORMANCE OF TEXT SUMMARIZATION

In this section, we have taken four techniques which were implemented for Indian languages text summarization. To show the impact of NLP tools, we have implemented these methods for English language and compare the results. The considered techniques are graph based technique for Hindi text summarization (Kumar et al. 2015), a hybrid model for Punjabi text summarization (Gupta and Kaur 2016), Text rank based technique for Marathi language (Rathod 2018), and semantic graph based Tamil summarizer (Banu et al. 2007). Here, we have experimented on 100 news articles for each language.

Methods	Language	Precision	Recall	F1 score
Kumar et al. (2015)	Hindi	0.44	0.32	0.37
	English	0.46	0.38	0.41
Gupta and Kaur (2016)	Punjabi	0.45	0.21	0.29
	English	0.49	0.28	0.35
Rathod (2018)	Marathi	0.43	0.27	0.33
	English	0.47	0.31	0.37
Banu et al. (2007)	Tamil	0.42	0.31	0.35
	English	0.45	0.36	0.40

Table 1. Results of precision, recall, and F1 measures for summarization methods for different languages

As illustrated in Table 1, the results for all summarization methods show that they performs better with English language in comparison to other languages in all cases. It proves the maturity of English language NLP tools better than other languages.

CONCLUSION

In this paper, we described briefly about existing text summarization methods for Indian texts. We have also discussed about the need of NLP tools during text summarization and their

importance. We have showed the results of existing techniques of text summarization for Indian languages with English language and found that the NLP tools affects the performance of any summarizer. Here, we have analyzed that although many summarizers are proposed previously but there are still lack of Indian context summarizers as most of the tools are not easily available or that are not performing satisfactory. Most of the proposed summarizers are based on statistical approaches or the combination of statistical and semantic models. There are other learning based, fuzzy based and neural network based approaches for text summarization. We can applied these approaches for better performance for Indian languages. Also, NLP tools can also be matured with these techniques for low resource languages.

REFERENCES

Sheikh Abujar, Mahmudul Hasan, MSI Shahin, and Syed Akhter Hossain. 2017. A heuristic approach of text summarization for Bengali documentation. In Computing, Communication and Networking Technologies (ICCCNT), 2017 8th International Conference on. IEEE, 1–8.

Sumya Akter, Aysa Siddika Asa, Md Palash Uddin, Md Delowar Hossain, Shikhor Kumer Roy, and Masud Ibn Afjal. 2017. An extractive text summarization technique for Bengali document (s) using K-means clustering algorithm. In Imaging, Vision & Pattern Recognition (icIVPR), 2017 IEEE International Conference on. IEEE, 1–6.

M Banu, C Karthika, P Sudarmani, and TV Geetha. 2007. Tamil document summarization using semantic graph method. In | iccima. IEEE, 128–134.

Sajib Dasgupta and Vincent Ng. 2006. Unsupervised morphological parsing of Bengali. Language Resources and Evaluation 40, 3-4 (2006), 311–330.

Sobha Lalitha Devi et al. 2011. Text Extraction for an Agglutinative Language. Language in India 11, 5 (2011).

Sobha Lalitha Devi and S Lakshmi. 2013. Malayalam clause boundary identifier: Annotation and evaluation. In Proceedings of the 4th Workshop on South and Southeast Asian Natural Language Processing. 83–90.

Md Iftekharul Alam Efat, Mohammad Ibrahim, and Humayun Kayesh. 2013. Automated Bangla text summarization by sentence scoring and ranking. In Informatics, Electronics & Vision (ICIEV), 2013 International Conference on. IEEE, 1–5.

Deepali Kailash Gaikwad. 2018. Rule Based Text Summarization for Marathi Text. Journal of Global Research in Computer Science 9, 5 (2018), 19–21.

JK Geetha and N Deepamala. 2015. Kannada text summarization using Latent Semantic Analysis. In Advances in Computing, Communications and Informatics (ICACCI), 2015 International Conference on. IEEE, 1508–1512.

Aniruddha Ghosh, Amitava Das, and Sivaji Bandyopadhyay. 2010. Clause identification and classification in bengali. In Proceedings of the 1st Workshop on South and Southeast Asian Natural Language Processing. 17–25.

Vishal Gupta. 2013. Hybrid algorithm for multilingual summarization of Hindi and Punjabi documents. In Mining Intelligence and Knowledge Exploration. Springer, 717–727.

Vishal Gupta and Narvinder Kaur. 2016. A novel hybrid text summarization system for Punjabi text. Cognitive Computation 8, 2 (2016), 261–277.

Vishal Gupta and Gurpreet Lehal. 2012. Automatic Punjabi text extractive summarization system. Proceedings of COLING 2012: Demonstration Papers (2012), 191–198.

Md Islam, Md Uddin, Mumit Khan, et al. 2007. A light weight stemmer for Bengali and its Use in spelling Checker. (2007).

R Jayashree, Srikanta Murthy, and Basavaraj S Anami. 2012. Categorized Text Document Summarization in the Kannada Language by Sentence Ranking. In Intelligent Systems Design and Applications (ISDA), 2012 12th International Conference on. IEEE, 776–781.

Jagadish S Kallimani, KG Srinivasa, et al. 2010. Information retrieval by text summarization for an Indian regional language. In Natural Language

Processing and Knowledge Engineering (NLP-KE), 2010 International Conference on. IEEE, 1–4.

K Vimal Kumar and Divakar Yadav. 2015. An Improvised Extractive Approach to Hindi Text Summarization. In Information Systems Design and Intelligent Applications. Springer, 291–300.

K Vimal Kumar, Divakar Yadav, and Arun Sharma. 2015. Graph Based Technique for Hindi Text Summarization. In Information Systems Design and Intelligent Applications. Springer, 301–310.

Mudassar M Majgaonker. 2010. Discovering suffixes: A case study for Marathi language. (2010).

Prasenjit Majumder, Mandar Mitra, Swapan K Parui, Gobinda Kole, Pabitra Mitra, and Kalyankumar Datta. 2007. YASS: Yet another suffix stripper. ACM transactions on information systems (TOIS) 25, 4 (2007), 18.

Rada Mihalcea and Paul Tarau. 2004. Textrank: Bringing order into text. In Proceedings of the 2004 conference on empirical methods in natural language processing.

Hilário Oliveira, Rafael Ferreira, Rinaldo Lima, Rafael Dueire Lins, Fred Freitas, Marcelo Riss, and Steven J Simske. 2016. Assessing shallow sentence scoring techniques and combinations for single and multi-document summarization. Expert Systems with Applications 65 (2016), 68–86.

Amaresh Kumar Pandey and Tanveer J Siddiqui. 2008. An unsupervised Hindi stemmer with heuristic improvements. In Proceedings of the second workshop on Analytics for noisy unstructured text data. ACM, 99–105.

Mona Parakh, N Rajesha, and M Ramya. 2011. Sentence Boundary Disambiguation in Kannada Texts. Language in India, www.languageinindia.com, Special Volume: Problems of Parsing in Indian Languages (2011), 17–19.

Ananthakrishnan Ramanathan and Durgesh D Rao. 2003. A lightweight stemmer for Hindi. In the Proceedings of EACL.

Ruby Rani and DK Lobiyal. 2018. Automatic Construction of Generic Stop Words List for Hindi Text. Procedia Computer Science 132 (2018), 362–370.

Yogeshwari V Rathod. 2018. Extractive Text Summarization of Marathi News Articles. (2018).

Jaideepsinh K Raulji and Jatinderkumar R Saini. 2017. Generating Stopword List for Sanskrit Language. In 2017 IEEE 7th International Advance Computing Conference (IACC). IEEE, 799–802.

Navanath Saharia, Utpal Sharma, and Jugal Kalita. 2014. Stemming resource-poor Indian languages. ACM Transactions on Asian Language Information Processing (TALIP) 13, 3 (2014), 14.

Kamal Sarkar. 2012. Bengali text summarization by sentence extraction. arXiv preprint arXiv:1201.2240 (2012).

Kartik Suba, Dipti Jiandani, and Pushpak Bhattacharyya. 2011. Hybrid inflectional stemmer and rule-based derivational stemmer for gujarati. In Proceedings of the 2nd Workshop on South Southeast Asian Natural Language Processing (WSSANLP). 1–8.

Nagmani Wanjari, GM Dhopavkar, and Nutan B Zungre. 2016. Sentence Boundary Detection for Marathi Language. Procedia Computer Science 78 (2016), 550–555.

Verma, P., & Om, H. (2016a). Extraction based text summarization methods on user's review data: A comparative study. In International Conference on Smart Trends for Information Technology and Computer Communications (pp. 346-354). Springer, Singapore.

Verma, P., & Om H. (2016b). Theme driven Text Summarization using k-means with gap statistics for Hindi Documents. In International conference on Computing, Communication and Sensor Network (pp. 90-94), IASTM Kolkata.

Verma, P., & Om, H (2016c). A survey on Indian Language Text Summarization Techniques, Evaluation, and Existing Tools. In International conference on Innovative Systems (pp. 32), IRO Bangalore.

Verma, P., & Om, H. (2018). Fuzzy Evolutionary Self-Rule Generation and Text Summarization. In 15th International Conference on Natural Language Processing (p. 115).

Verma, P., & Om, H. (2019a). MCRMR: Maximum coverage and relevancy with minimal redundancy based multi-document summarization. Expert Systems with Applications, 120, 43-56.

Verma, P., & Om, H. (2019b). Collaborative ranking-based text summarization using a metaheuristic approach. In Emerging Technologies in Data Mining and Information Security (pp. 417-426). Springer, Singapore.

Verma, P., & Om, H. (2019c). A variable dimension optimization approach for text summarization. In Harmony Search and Nature Inspired Optimization Algorithms (pp. 687-696). Springer, Singapore.

Verma, P., & Om, H. (2019d). A novel approach for text summarization using optimal combination of sentence scoring methods. Sādhanā, 44(5), 110.

Verma, P., Pal, S., & Om, H. (2019). A Comparative Analysis on Hindi and English Extractive Text Summarization. ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP), 18(3), 30.

***