Marathi Text Analysis using Unsupervised Learning and Word Cloud

Prafulla B. Bafna, Jatinder kumar, R. Saini

Abstract*: *Managing a large number of textual documents is a critical and significant task and supports many applications ranging from information retrieval to clustering search engine results. Marathi is one of the oldest of the regional languages in the Indo-Aryan language family, dating from about AD 1000. Abundance of Marathi literature has generated a big corpus and need of summarization of information. The objective of this study is to overcome the scalability problem while managing the documents and summarize the Marathi corpus by extracting tokens. The work is better in terms of scalability and supports the consistent quality of cluster for incremental data set. Most of the past and contemporary research works have targeted English corpus document management. Marathi corpus has been mostly exploited by the researchers for exploring stemming, single-document summarization and classifier design on Marathi corpus. Implementing unsupervised learning on the Marathi corpus for summarization of multiple documents through Word Cloud is still an untouched area. Technically speaking, the current work is an application of TF-IDF, cosine-based document similarity measures and cluster dendrograms, in addition to various other Natural Language Processing (NLP) activities. Entropy and precision are used to evaluate the experiments carried on different datasets and results prove the robustness of the proposed approach for Marathi Corpus.

Keywords: Classification, Clustering, Document Management, Marathi, Summarization, Word Cloud.

I. INTRODUCTION

In online and offline systems, documents are continuously generated, stored, and accessed every day in large volumes. Categorizing documents according to the contents present in it will help to retrieve documents based on a particular topic [41-43]. The maximum work is done in document management focuses on English corpus, but text in Marathi on the web has come of age since the advent of Unicode standards in Indic languages. information technology generated huge data on the internet [44]. Initially this data is mainly in English language so majority of data mining research work is on the English text documents. As the internet usage increased, data in other languages like Marathi, Tamil, Telugu and Punjabi etc. increased on the internet. Similarly, in case of India, Maharashtra,Gujarat and Tamil Nadu are highly industrialized states. Mumbai being the financial capital has its administrative work going on only in Marathi. Also, the largest recent indigenous empire of India was the Maratha empire. The literature of Maratha empire is present in Marathi [1].Very few researchers have focused on Marathi text clustering, classification, but none of them have evaluated results of clustering in terms of its compactness or purity [2]. Various data mining techniques like clustering, classification can be applied once the data is in a structured format. Document term matrix (DTM) allows to convert text files into table form, where rows are represented by documents and terms are placed as columns [45-47]. But DTM causes dimension curse because all the terms present in the corpus are considered while constructing DTM. Term Frequency–Inverse Document Frequency (TF-IDF) allows selecting significant terms based on the token weights of terms. The cosine similarity measure is a prerequisite for applying the hierarchical algorithm on documents [48-51]. Once the clusters are formed they can be summarized through a word cloud. Word clouds enables to visualize and understand the text information in an easy way It represents the words from higher to lower frequency from big to the small font size. It's a way of text summarization [3]. The proposed approach imitates removal stop words and finds out top N frequent terms using TF-IDF weights. The N value is called a threshold which is 50 % of maximum TF-IDF weight. It effectively removes all unuseful words. Considering these keywords, cosine similarity measure and hierarchical clustering are applied to get document clusters. Entropy is used to validate cluster quality and in turn, N-value. The dataset having predefined classes are used to decide the precision of the experimental setup. Proving the betterment of the technique, it is applied to the live dataset. Once the clusters are formed, the word cloud is applied to summarize the clusters. The approach is unique because

1. Multi-document Word Cloud based Summarization through Unsupervised Learning and its performance analysis on Marathi corpus is an untouched topic.
2. Various applications, for example, information retrieval will benefit from the proposed approach by saving time and effort required to read and manage an entire corpus.
3. The approach can process 300 documents having more than 15000 words and hence proves betterment in scalability.
4. Entropy is consistent even for large data size.

Document Management System (DMS) facilitates to access the documents in a fast and easy way and turn increases the productivity of the work [52-53].

Grouping of documents is one of the most important steps towards document management, which helps in identifying replicas of documents, clustering search engine results, and so on [54-55]. There are different document management systems existing for English corpus, DMS for Marathi corpus though significant, is not explored yet. In the paper terms, dimensions, words and tokens are used as synonyms, interchangeably.

Revised Manuscript Received on February 18, 2020.
* Correspondence Author

Prafulla B. Bafna*, Research, Symbiosis Institute of Computer Studies and Symbiosis International Deemed University-Maharashtra, India.

Jatinder kumar, Research, Symbiosis Institute of Computer Studies and Symbiosis International Deemed University-Maharashtra, India.

R. Saini, Research, Symbiosis Institute of Computer Studies and Symbiosis International Deemed University-Maharashtra, India.

The organization of the paper is as mentioned. The work done by other researchers on the topic is presented as a background in the next that is second section. The third section presents the methodology, along with the experimental setup and results. The paper ends with the fourth section that is the conclusion and future directions.

II. LITERATURE REVIEW

This section illustrates existing techniques of text preprocessing, text stemming, summarization, word cloud and so on.

For text analysis of Indian languages, Punjabi has been explored through stop words identification [22] and categorization [23] as well as poetry corpus creation [24] and classification [25-26]. Gujarati has been, similarly explored through diacritic extraction technique [27], information retrieval [28], stop words identification [29] and categorization [30], Machine Translation System (MTS) [31-32] and classification [33]. Sanskrit has been explored through stop word generation [34] and analysis [35], bilingual dictionary [36], constituency mapper [37], lemmatizer development [38] and comparison of its morphological analyzers [39]. Hindi text analysis of poetry was presented through an automated system for generation of metadata [40]. Extracting an original word from token is termed as stemming. It is the type of preprocessing and improves the performance of the algorithm to be used NLP tasks. [2]. The stemming problem has been addressed in many contexts, and by researchers in many disciplines, the main purpose of stemming is to reduce different grammatical forms/word forms to its root form. Stemming is widely used in an Information Retrieval system and reduces the size of the index files. The goal of stemming is to reduce inflectional forms and sometimes derivationally related forms of a word to a common base form. Stemming is used in information retrieval systems to improve performance. Additionally, this operation reduces the number of terms in an information retrieval system, thus decreasing the size of the index files.

Different stemming algorithms for non-Indian and Indian language, methods of stemming, accuracy, and errors are explained [3]. Authors have created intelligent system to get desired documents in Marathi language. The approach allows user specific documents to be retrived which are based on their personal interest [4].Most of stemming algorithms are based on a rule-based approach. The performance of a rule-based stemmer is superior to some well-known methods like brute force. Dictionary-based algorithms, including natural language processing approaches, are used to build stemmers. The purpose of stemmers is to encode a wide range of language related rules evolved across a period. Such stemmers with comprehensive rules are language-dependent. Marathi is a morphologically rich language. Abstracting of the document and presenting it to the user is achieved through text summarization, it extracts the significant information from the given text. Manual summarization of the document is costly in terms of time, efforts, etc. Automatic Text summarization is a more accurate way than manual summarization. A deeper analysis of the text needs to carry out summarization[5] [6]. Hierarchical document clustering

Generally, clustering is preferred to group the similar type of documents and multiple algorithms are suggested by various researchers. For retrieving documents in a clustered manner, Hierarchical techniques are being used. There are various kinds of distance measures listed by several researchers for hierarchical clustering that are a single link, average link, etc. [7].

The textual data is accumulating and increasing tremendously. Different types of reviews, webpages, etc. are examples of text data. Documents are used to store this information. To apply any clustering techniques the data should be in the tabular form. Various techniques are available to store such types of data for example bag of words [8]. But it creates dimension curse, as all terms in the corpus are considered. High dimensions affect the performance of the algorithm. To reduce high dimensions, only significant words need to be considered. The document clustering process will execute in less time if the top significant words are selected.

To improve the clustering process, the text is preprocessed by removing stop words, etc. [9-10]. Generally (TF-IDF) is a popularly used technique that transforms text data into matrix form. The measure represents the significance of the token with respect to text documents considering the entire corpus. In document processing, it acts as a weighting unit. In spite of increasing word count proportional to the number of documents in which it is present, The TF-IDF ignores the most commonly occurring words, by offsetting count of the words in the entire corpus [11]. Entropy and precision are popularly used parameters to evaluate clustering. To validate the purity of the clustering results, entropy is used and the accuracy of the clusters is measured by precision. Word Cloud

To represent the textual data graphically in the form of words, Word clouds are used. The words with higher frequency appear prominent. The font size reduces as the word count lowers. It's very easy to understand and interpret the word cloud [12]. Word cloud is referred to as a simple and effective visualization and summarization technique. It helps in the domain of text mining, visualization techniques and contextual data. A word cloud can be effectively used for focusing on the needs and problems of customers and in turn, to increase the business without reading the text. Research scholars can use word cloud for interpretation of qualitative data in a faster manner [13-14]. User comments entered on social media about service/product, or political party can be analyzed through word cloud, and the overall essence of the product or service can be understood without going through comments and thus to save time and effort [15-17].

Word cloud on Marathi text is not attempted yet. It will be benefited to the people who prefer to use their regional language for doing day to day activities like reading news articles, summarizing government documents and so on [18-20].

III. PROPOSED METHODOLOGY

Corpus containing Marathi Text is processed to remove the stop words, TF-IDF is used on a set of documents, and to weight is calculated. Terms having a weight greater than or equal to the threshold are considered and termed as modified TF-IDF and followed by computing cosine measure similarity matrix. The dendrogram is constructed using single link hierarchical clustering applied to get the clusters."

Entropy and precision are calculated to validate the experiment, and the word cloud is created for each cluster to summarize the cluster. Figure 2A depicts the flow chart of proposed method

3.1 Creating a corpus of Marathi text

The dataset consisting of 300 stories is created by collecting different stories belonging to different domains. To avoid bias, stories are collected from different websites. The sourced stories are uploaded from years between 2000-2019. [http://marathi.webdunia.com.][https://www.britannica.com/art/Marathi-literature][https://www.marathisahityadarpan.in] The dataset is created in Excel having a mixture of stories. The encoding="UTF-8" is used to store data.

3.2 Clustering of the corpus using unsupervised learning technique and evaluation

The corpus is tokenized to get a bag of words and preprocessed to remove stop words like "जो" (jo) or (that), "ते" (te) or (it) [12] and so on. Lemmatization is purposely avoided in the pre-processing to preserve the semantics of words. TF-IDF is applied to calculate the weight of each term. Instead of applying stemming, the top N terms are selected for further processing. N value depends on the maximum TF-IDF weight and TF-IDF weights of all other terms. To select the terms, modified TF-IDF weights are used. The cosine similarity measure between documents is calculated to generate a dendrogram. The algorithm is implemented on the corpus of stories and reviews. The stories are clustered and each cluster represents the topic of the story. Figure 1 shows the block diagram of a proposed technique. Reviews are clustered based on sentiments present in it. For example, for a sample statement from the story, POS tagging and its stemming is depicted in Figure 2. Table 1 shows the algorithm with package details.

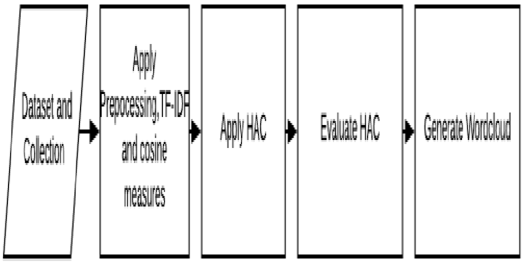
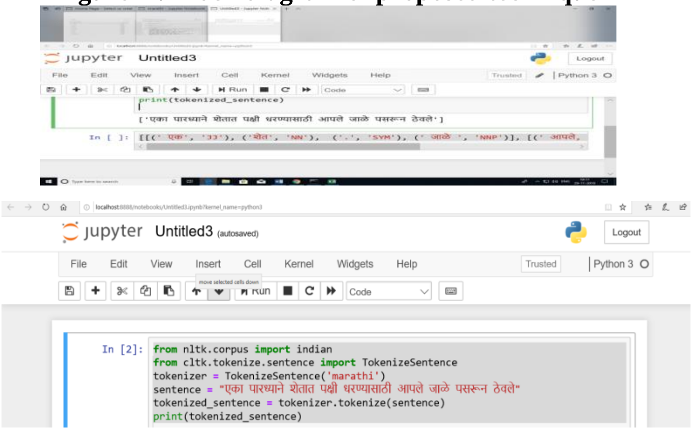


Figure 1 : Block diagram of proposed technique



```
print(tokenized_sentence)

['एका पारध्याने शेतात पक्षी धरण्यासाठी आपले जाळे पसरून ठेवले']
In [ ]: [[(' एक', 'JJ'), ('शेत', 'NN'), ('.', 'SYM'), (' जाळे ', 'NNP')], [(' आपले,
```

```
In [2]: from nltk.corpus import indian
        from cltk.tokenize.sentence import TokenizeSentence
        tokenizer = TokenizeSentence('marathi')
        sentence = "एका पारध्याने शेतात पक्षी धरण्यासाठी आपले जाळे पसरून ठेवले"
        tokenized_sentence = tokenizer.tokenize(sentence)
        print(tokenized_sentence)
```

Figure. 2: Output in the form of stemming of tokens and POS tagging

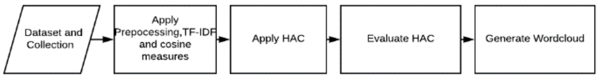


Figure 2A: Flow chart of proposed method

Table 1 Algorithm with package details

Step No.	Step Description	Library/Package/Function
1	Text preprocessing of stories	NLTK , Indian corpus
2	Term weights are calculated	DocumentTermMatrix_tfidf
3	Select terms having token weights greater than threshold	User defined function to select the terms, document_term_matrix(dtm_threshold)
4	Calculate cosine similarity matrix and Apply hierarchical agglomerative clustering	User-defined function to calculate cosine similarity and hclust(dist.mat)
5	Validation of clustering process	Package Entropy
6	Apply word cloud for each cluster to summarize the cluster	Package WordCloud

Generated dendrogram after applying steps mentioned in table 1 categorizes stories based upon the contents of the story. Entropy values are stated in Table 2 proves that the purity of the cluster is consistent for an incremental dataset that is from 10 stories to 100 stories. It means that stories belonging to the same topic are grouped. One story belongs to nearly about 150 to 200 words, including stopwords and other unuseful words. Significant token selection using modified TF-IDF attempts to have appropriate clusters of stories.

Figure 3 shows the dendrogram for 60 stories. The purpose is to extract stories of a particular topic. These are the moral stories used to inculcate ethics in children. These stories contain belonging to different domains like "Vikram Vetal" , "Panchatantra" and so on. There are two clusters representing two sets of stories.

Table 2 represents Entropy values for an incremental dataset of stories along with total words, words after removal of stop words and Top K words based on modified TF-IDF weight.
As there is no standard algorithm, a comparative analysis couldn't be performed.

Table 2 : Summary of Words

Sr. No.	Dataset Size	Entropy	Total Words	Words after preprocessing	Significant Words using modified TF-IDF
1	15	0.12	1990	1500	500
2	30	0.23	5244	2789	800
3	50	0.33	8242	4,212	1000
4	100	0.66	18282	8,213	3000

Table 3: Entropy Comparison

Sr.	Documents	Entropy		
No.		FKM	KM	HAC
1	10	0.11	0.14	0.12
2	50	0.14	0.23	0.33
3	100	0.15	0.31	0.66
4	200	0.15	0.41	0.68
5	300	0.16	0.55	0.68

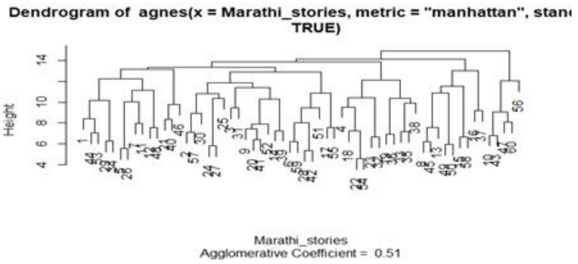


Figure 3 dendrogram of sixty stories

IV. RESULTS ANALYSIS

There is a considerable difference in entropy values of 50 to100 (0.66-0.33 =0.33) stories than 15 to 30 (0.23-0.12 =0.11). That is difference between the entropy values of 50 to 100 is 0.33 Which is greater than the difference between entropy values of 15 to 30 documents 0.11. It was decided to change the algorithm and check the entropy values.
The steps mentioned in Table 1 are carried out on a dataset of reviews using Fuzzy k-means (FKM) and K-means(KM) and Hierarchical Agglomerative Clustering (HAC) algorithms and Table 3 presents a comparative analysis of entropy values produced by three popular clustering techniques. It's clear that Fuzzy K-means gives better entropy for incremental data size that is for 300 documents entropy produced by FKM is improved by 40 %. Extracting cluster terms shows that stories are clustered based on characters present in it one cluster depicts stories of animals, other cluster depicts stories of human characters and third cluster indicates stories of human beings along with animals.
Figure 4 shows vector space of significant terms in the corpus. Figure 5 shows the fuzzy k-means plot applied on a dataset of stories, the intersection of two clusters depicts the stories of animals and human. For example, three sample sentences existing in the respective clusters of stories are stated. Sample sentence of the story belonging to the cluster1 (only humans) is “एके दिवशी राजा पुरंदरदासांना म्हणाला, भक्तराज, लोभ मनुष्याला आध्यात्मिक प्राप्तीपासून दूर करतो”, transliterated as “Eke divasi raja purandaradasanna mhaṇala, bhaktaraja, lobha manuṣyala adhyatmika praptipasuna dura karato” and translated as “One day the king said to Purandaras, Bhaktaraja, that greed removes man from spiritual pursuit”, that of cluster 2 (only animals) is “एकदा एके ठिकाणी एक गाढव आणि एक कोंबडा चरत असता, तेथे एक सिंह आला” (“Once upon a time there was a donkey and a hen grazing, there came a lion”)or(“Ēkadā ēkē ṭhikāṇī ēka gāḍhava āṇi ēka kōmbaḍā carata asatā, tēthē ēka sinha ālā”) and sample sentence of the story having both charcters that is human and animals is “द्रोण ब्राह्मणाच्या त्या दोनही सुंदर गायींवर एका चोराची नजर होती” that is “ Droṇa brahmaṇachya tya donhi sundara gāyīnvara ēkā cōrāchī najara hōtī” It means “A thief had a look at those two beautiful cows of Drona Brahmin”.

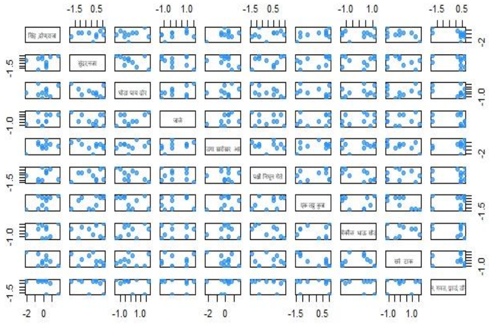


Figure 4 : Document representing significant terms in a vector space Legend

a.The words present diagonally in the matrix: नजरएक गवत

b. Dots present in matrix : Feature weights of significant terms

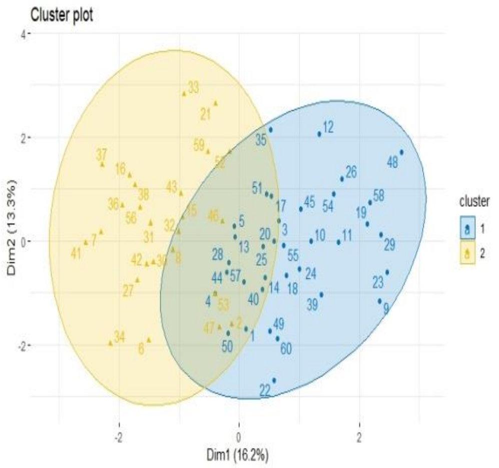


Figure 5 Fuzzy plot of stories data set

3.2.2 Summarization of each cluster

Visualization in the form of word cloud leads to fast and measurable (words having more frequency are shown in the big font) representation of the cluster.

Fig. 6 shows a word cloud generated for one of the clusters of stories dataset and contains very specific terms related to stories having both animals and humans. Input to word cloud is a token weight matrix that is formulated by considering the modified TF-IDF measure. Table 4 shows the sample of words/tokens along with weights, translation and transliteration of words. These words act as an input to the word cloud. The highest weight is 20, for the term "राजा" (*Raja*) translated as"King", words having TF-IDF weight more than 6 , are considered for word cloud.

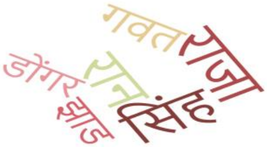


Figure. 6 Word cloud for a cluster of movie

Table 4: A sample of words and their weights

Sr.No.	Word	Feature Weight	Transliteration	Translation
1	राजा	10	Rājā	King
2	रान	9	rāna	Forest
3	सिंह	9	sinha	Lion
4	..	..	..	..
5	झाड	7	Zad	Tree

V. CONCLUSION AND FUTURE WORK

The current study achieves the summarization of the clusters of Marathi corpus, unlike the other published research works which have focused only on single-document summarization. Additionally, the contribution of this study is also to work on the application of unsupervised learning with the Marathi corpus. The formation of the clusters was achieved through the measures of cosine similarity, the summarization of the clusters was achieved through the modified TF-IDF and the representation was achieved through Word Cloud. This approach will be useful to researchers who need to explore Marathi literature. Summarization of the cluster of Marathi literature will act as stepstone towards future research and to understand state of the art in fast and efficient way. Experiments are conducted on data set of Marathi moral stories. In the absence of any other technique which achieves document clustering on Marathi corpus, precision and entropy prove the betterment of the technique. The current work is the first of its kind in the world which employs unsupervised learning for Marathi corpus followed by the summarization of the multi-document clusters. Also, it is the first of its kind work which represents the summarization results through Marathi Word Cloud.

REFERENCES

1. Altuncu, M. T., Yaliraki, S. N., & Barahona, M. (2018). Content-driven, unsupervised clustering of news articles through multiscale graph partitioning. arXiv preprint arXiv:1808.01175.
2. Amelia Walkley, Jatin Nagpal ,2015 available at https://www.thinkwithgoogle.com/intl/en-apac/trends-and-insights/marathi-matters-digital-age/ on 29/7/2019
3. Audichya M.A., Saini J.R., "Computational Linguistic Prosody Rule-based Unified Technique for Automatic Metadata Generation for Marathi Poetry", proceedings of ICAIT-2019, in press, IEEE, USA
4. Burch, M., Lohmann, S., Beck, F., Rodriguez, N., Di Silvestro, L., & Weiskopf, D. (2014, July). RadCloud: Visualizing multiple texts with merged word clouds. In 2014 18th International Conference on Information Visualisation (pp. 108-113). IEEE.
5. Chiranjibi Sitaula "Semantic text clustering using enhanced vector space model using nepali language" GESJ: Computer Science and pp 41-46 , 2012
6. Cui, W., Wu, Y., Liu, S., Wei, F., Zhou, M. X., & Qu, H. (2010, March). Context preserving dynamic word cloud visualization. In 2010 IEEE Pacific Visualization Symposium (PacificVis)(pp. 121-128). IEEE.]
7. Garg, Amita & Saini, Jatinderkumar. (2019). A Systematic and Exhaustive Review of Automatic Abstractive Text Summarization for Marathi Language.]
8. Hanyurwimfura, Damien, Liao Bo, Dennis Njagi, and Jean Paul Dukuzumuremyi. "A Centroid and Relationship based Clustering for Organizing." International Journal of Multimedia and Ubiquitous Engineering 9, no. 3 (2014): 219-234.
9. Heimerl, F., Lohmann, S., Lange, S., & Ertl, T. (2014, January). Word cloud explorer: Text analytics based on word clouds. In 2014 47th Hawaii International Conference on System Sciences (pp. 1833-1842). IEEE.]
10. Md Shad Akhtar, Asif Ekbal, Pushpak Bhattacharyya; Aspect Based Sentiment Analysis in Marathi: Resource Creation and Evaluation; In proceedings of the 10th International Conference on Language Resource and Evaluation (LREC 2016); 23-28; Portoroz, Slovenia; 2016
11. Mishra, U., & Prakash, C. (2012). MAULIK: an effective stemmer for the Marathi language. International Journal on Computer Science and Engineering, 4(5), 711.
12. Muhammad Zubair Asghar, Aurangzeb Khan , Shakeel Ahmad ,Fazal Masud Kundi , A review of feature extraction in sentiment analysis, Journal of Basic and Applied Scientific Research" J. Basic. Appl. Sci. Res., 4(3), pp.181-186, 2014
13. Ramanathan, Ananthakrishnan, and Durgesh D. Rao. "A lightweight stemmer for Marathi." In the Proceedings of EACL. 2003.
14. Rasmussen, E. M. (1992). Clustering algorithms. Information retrieval: data structures & algorithms, 419, 442.
15. Saini J.R., Desai A.A., "Identification of Hindi Words Used in Pornographic Unsolicited Bulk Emails", The IUP Journal of Systems Management, ISSN: 0972-6896, vol. 9, issue 2, May 2011, pages 53-60;
16. Sindhuja, B., & Trivedi, V. (2014). Usage of cosine similarity and term frequency count for textual document clustering. International Journal of Innovative Research in Computer Science & Technology (IJIRCST), 2(5), 9-12.
17. Thangarasu, M., & Manavalan, R. (2013). A literature review: stemming algorithms for Indian languages. arXiv preprint arXiv:1308.5423.
18. Vispute, S. R., Kanthekar, S., Kadam, A., Kunte, C., & Kadam, P. (2014, April). Automatic Personalized Marathi Content Generation. In 2014 International Conference on Circuits, Systems, Communication and Information Technology Applications (CSCITA) (pp. 294-299). IEEE.
19. Vispute, S. R., & Potey, M. A. (2013, September). Automatic text categorization of Marathi documents using clustering technique. In 2013 15th International Conference on Advanced Computing Technologies (ICACT) (pp. 1-5). IEEE.].
20. Dangre, N., Bodke, A., Date, A., Rungta, S., & Pathak, S. S. (2016). System for Marathi News Clustering. Procedia Computer Science, 92, 18-22.
21. Singh, S., & Siddiqui, T. J. (2012, March). Evaluating effect of context window size, stemming and stop word removal on Hindi word sense disambiguation. In 2012 International Conference on Information Retrieval & Knowledge Management (pp. 1-5). IEEE.
22. Kaur J. and Saini J.R., 2016, "Punjabi Stop Words: A Gurmukhi, Shahmukhi and Roman Scripted Chronicle", proc. of Symposium on ACM Women in Research (ACM-WIR-2016), Indore, India, vol. 01188, pages 32-37, Available online: http://dl.acm.org/citation.cfm?id=2909073

23. Kaur J. and Saini J.R., 2015, "POS Word Class based Categorization of Gurmukhi Language Stemmed Stop Words", proc. of International Conference on ICT for Intelligent Systems (ICTIS-2015), Ahmedabad, India, vol. 51(2), pages 3-10, Available online: http://link.springer.com/chapter/10.1007/978-3-319-30927-9_1
24. Saini J.R. and Kaur J., 2020, "Kāvi: An Annotated Corpus of Punjabi Poetry with Emotion Detection Based on 'Navrasa'", Procedia Computer Science, in press with Elsevier
25. Kaur J. and Saini J.R., 2017, "Punjabi Poetry Classification: The Test of 10 Machine Learning Algorithms", proc. of 9th International Conference on Machine Learning and Computing (ICMLC 2017), Singapore, pages 1-5
26. Kaur J. and Saini J.R., 2020, "Designing Punjabi Poetry classifiers using machine learning and different textual features", International Arab Journal of Information Technology, Jordan, vol. 17(3), in press, Available online: http://iajit.org/PDF/May%202020,%20No.%203/16024.pdf
27. Rakholia R.M. and Saini J.R., 2015, "The Design and Implementation of Diacritic Extraction Technique for Gujarati Written Script using Unicode Transformation Format", proc. of IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT-2015), Coimbatore, India, vol. 2, pages 654-659, Available online: https://ieeexplore.ieee.org/document/7226037
28. Rakholia R.M. and Saini J.R., 2017, "Information Retrieval for Gujarati Language using Cosine Similarity based Vector Space Model", proc. of The 5th International Conference on Frontiers of Intelligent Computing: Theory and Applications (FICTA-2016), Bhubaneshwar, India, vol. 516, pages 1-9, Available online: https://link.springer.com/chapter/10.1007/978-981-10-3156-4_1
29. Rakholia R.M. and Saini J.R., 2017, "A Rule-based Approach to Identify Stop Words for Gujarati Language", proc. of The 5th International Conference on Frontiers of Intelligent Computing: Theory and Applications (FICTA-2016), Bhubaneshwar, India, vol. 515, pages 797-806, Available online: https://link.springer.com/chapter/10.1007/978-981-10-3153-3_79
30. Rakholia R.M. and Saini J.R., 2016, "Lexical Classes Based Stop Words Categorization for Gujarati Language", proc. of 2nd International Conference on Advances in Computing, Communication & Automation (ICACCA-2016), Bareilly, India, pages 1-5, Available online: http://ieeexplore.ieee.org/document/7749005/
31. Saini J.R. and Modh J.C., 2017, "GIdTra: A Dictionary-based MTS for Translating Gujarati Bigram Idioms to English", proc. of IEEE sponsored 4th International Conference on Parallel, Distributed and Grid Computing (PDGC-2016), Solan, India, pages 192-196, Available online: http://ieeexplore.ieee.org/document/7913143/
32. Raulji J.K. and Saini J.R., "A Rule Based Architecture for Sanskrit to Gujarati Machine Translation System", proc. of International Conference on Emerging Trends in Engineering, Science and Technology (ICRISET-2018), Anand, India, in press with IEEE
33. Rakholia R.M. and Saini J.R., 2017, "Classification of Gujarati Documents using Naïve Bayes Classifier", Indian Journal of Science and Technology, vol. 10(5), pages 1-9, Available online: http://indjst.org/index.php/indjst/article/view/103233/78147
34. Raulji J.K. and Saini J.R., 2017, "Generating Stopword List for Sanskrit Language", proc. of 7th IEEE International Advance Computing Conference (IACC-2017), Hyderabad, India, pages 799-802, Available online: http://ieeexplore.ieee.org/document/7976898/
35. Raulji J.K. and Saini J.R., "Sanskrit Stopword Analysis through Morphological Analyzer and its Gujarati Equivalent for MT System", proc. of International Conference on ICT for Sustainable Development (ICT4SD-2019), Panaji, Goa, India, in press with Springer
36. Raulji J.K. and Saini J.R., "Bilingual Dictionary for Sanskrit – Gujarati MT Implementation", proc. of International Conference on ICT for Sustainable Development (ICT4SD-2019), Panaji, Goa, India, in press with Springer
37. Raulji J.K. and Saini J.R., "Sanskrit-Gujarati Constituency Mapper for Machine Translation System", proc. of IEEE Bombay Section Signature Conference (IBSSC-2019), Mumbai, India, in press with IEEE
38. Raulji J.K. and Saini J.R., 2019, "Sanskrit Lemmatizer for Improvisation of Morphological Analyzer", Journal of Statistics and Management Systems, vol. 22(4), pages 613-625, Available online: https://tandfonline.com/doi/abs/10.1080/09720510.2019.1609186
39. Saini J.R. and Raulji J.K., 2020, "Peer Analysis of "Sanguj" with Other Sanskrit Morphological Analyzers", proc. of 2nd International Conference on Computing Analytics and Networking (ICCAN-2019), Bhubaneshwar, India, in press with Springer
40. Audichya M.A. and Saini J.R., 2020, "Computational Linguistic Prosody Rule-based Unified Technique for Automatic Metadata Generation for Hindi Poetry", 1st IEEE International Conference on Advances in Information Technology, Karnatka, India, in press with IEEE
41. Bafna P.B., Saini J.R.,2019, "Identification of Significant Challenges in the Sports Domain using Clustering and Feature Selection Techniques", 9th International Conference on Emerging Trends in Engineering and Technology on Signal and Information Processing (ICETET-SIP-19), Nagpur, India, in press with IEEE.
42. Bafna P.B., Saini J.R.,2019, "Hindi Multi-document Word Cloud based Summarization through Unsupervised Learning", ", 9th International Conference on Emerging Trends in Engineering and Technology on Signal and Information Processing (ICETET-SIP-19), Nagpur, India, in press with IEEE.
43. Bafna P.B., Saini J.R., 2019, "Scaled Document Clustering and Word Cloud based Summarization on Hindi Corpus, 4th International Conference on Advanced Computing and Intelligent Engineering, Bhubaneshwar, India, in press with Springer.
44. Bafna P.B., Saini J.R., 2020, On Readability Metrics of Goal Statements of Universities and Brand-promoting Lexicons for Industries, 4th International Conference of Data Management, Analytics and Innovation (ICDMAI 2020), Delhi,India
45. Bafna P.B., Saini J.R., 2020, Identification of Significant Challenges Faced by Tourism and Hospitality Industry Using Association rules", 4th International Conference of Data Management, Analytics and Innovation (ICDMAI 2020), Delhi,India
46. Bafna, P., Pramod, D., & Vaidya, A. (2017, August). Precision based recommender system using ontology. In 2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS) (pp. 3153-3160). IEEE.
47. Bafna, P., Shirwaikar, S., Pramod, D., & Vaidya, A. Review based Feature Matrix for Predicting ratings in Recommender System.
48. Bafna, P., Pillai, S., & Pramod, D. (2016). Quantifying performance appraisal parameters: a forward feature selection approach. Indian J Sci Technol, 9(21), 1-7.
49. Bafna, P., Kaur, A., & Choudhary, N. (2015). Cluster based quantification to identify significant ERP critical failure factors. International Journal of Applied Engineering Research, 10(17), 37592-37594.
50. Bafna, P., Pramod, D., Shrwaikar, S. and Hassan, A. (2019), "Semantic key phrase-based model for document management", Benchmarking: An International Journal, Vol. 26 No. 6, pp. 1709-1727. https://doi.org/10.1108/BIJ-04-2018-0102
51. Bafna, P., Metkewar, P., & Shirwaikar, S. (2014). Novel Clustering approach for Feature selection. American International Journal of Available online at http://www. iasir. net Research in Science, Technology, Engineering & Mathematics, 62-67.
52. Bafna, P. B., Shirwaikar, S., & Pramod, D. (2016). Multi-Step Iterative Algorithm for Feature Selection on Dynamic Documents. International Journal of Information Retrieval Research (IJIRR), 6(2), 24-40. doi:10.4018/IJIRR.2016040102
53. Bafna, P., Shirwaikar, S. and Pramod, D. (2019), "Task recommender system using semantic clustering to identify the right personnel", VINE Journal of Information and Knowledge Management Systems, Vol. 49 No. 2, pp. 181-199. https://doi.org/10.1108/VJIKMS-08-2018-0068
54. Prafulla Bafna, Shailaja Shirwaikar, and Dhanya Pramod. 2016. Semantic Clustering Driven Approaches to Recommender Systems. In Proceedings of the 9th Annual ACM India Conference (COMPUTE '16). ACM, New York, NY, USA, 1-9. DOI: https://doi.org/10.1145/2998476.2998487
55. Bafna, P., Pramod, D., & Vaidya, A. (2016, March). Document clustering: TF-IDF approach. In 2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT) (pp. 61-66). IEEE.