Marathi Document: Similarity Measurement using Semantics-based Dimension Reduction Technique

Prafulla B. Bafna[1], Jatinderkumar R. Saini[2]
Symbiosis Institute of Computer Studies and research
Symbiosis International Deemed University, Pune, India

***Abstract*—Textual data is increasing exponentially and to extract the required information from the text, different techniques are being researched. Some of these techniques require the data to be presented in the tabular or matrix format. The proposed approach designs the Document Term Matrix for Marathi (DTMM) corpus and converts unstructured data into a tabular format. This approach has been called DTMM in this paper and it fails to consider the semantics of the terms. We propose another approach that forms synsets and in turn reduces dimensions to formulate a Document Synset Matrix for Marathi (DSMM) corpus. This also helps in better capturing the semantics and hence is context-based. We abbreviate and call this approach as DSMM and carry out experiments for document-similarity measurement on a corpus consisting of more than 1200 documents, consisting of both verses as well as proses, of Marathi language of India. Marathi text processing has been largely an untouched area. The precision, recall, accuracy, F1-score and error rate are used to prove the betterment of the proposed technique.**

***Keywords*—Cosine similarity; marathi; synset; term matrix; wordnet**

I. INTRODUCTION

India is a diverse country having around 23 different official languages and this has opened a wide area for natural language processing researchers. Indian language domains have lots of data accumulated in recent years and thus provide opportunities to mine this data. The Marathi language is not only popular in the world but also it is used as an official language in Maharashtra still it's a resource scare language. Marathi text gets generated day by day due to multilingual options provided by different websites. To process this data, natural language processing (NLP) techniques along with machine learning algorithms are available in the literature. To find out the similarity between text data, the corpus of Marathi verses and proses is being used. Verses and proses are part of the literature [1]. Proses and verses act as a guide to children about their behavior and manners and connect with elders to interconnect ideas and visualize life's opportunities, entertainment and so on. The use of rhyme and meter gives musical sense to the poetry, which is termed as literary elements whereas proses include a set of incidents and characters. Nouns, adjectives, adverbs are prominently used to construct a story or a poem [1]. To retrieve the required information from the text different NLP techniques are used. [23] The document term matrix is one of the ways on which different techniques could be applied to retrieve information. India is a diverse country having around 23 different official languages and this has opened a wide area for natural language processing researchers. Indian language domains have lots of data accumulated in recent years and thus provide opportunities to mine this data. The Marathi language is not only popular in the world but also it is used as an official language in Maharashtra still it‟s a resource scare language. Marathi text gets generated day by day due to multilingual options provided by different websites. To process this data, natural language processing (NLP) techniques [18] along with machine learning algorithms are available in the literature. To find out the similarity between text data, the corpus of Marathi verses and proses is being used. Verses and proses are part of the literature [2]. Proses and verses act as a guide to children about their behavior and Document Similarity determines how close the two text pieces are in a semantic and lexical way. In term vector space, suppose in document di, term k does not exist then wik= 0 and if kth term in document di, does exist with wik > 0, then wik in document di is called the weight of term [3]. Similarity Measure is quantitative and qualitative. Qualitative deals with the sentiment, general meaning of the corpus. Numerical measures such as the total number of tokens, size of the document, are considered in the quantitative approach. There are two steps, to find out document similarity, the first step is vectorization in which vector of numbers is obtained from documents, the second step is distance computation. It computes the distance or similarity between the document vectors Cosine value is one of the measures to compute the similarity between the document vectors [26]. Same vectors has the cosine dot product as zero and dissimilar or perpendicular vectors has dot product 1. Cosine measure always lies between zero and one.

In Document term matrix, documents are represented in the form of rows and columns are represented as frequent words. The matrix entry represents the frequency of the term for particular document. To decide frequent words only count of the token or word is considered. DTM do not consider meaning of the term [7].

Context based NLP options are available to involve sense and semantics of the words Polysemi is one of the options to detect sense of the word. One word with different meaning is termed as polysemy. For e.g. “right” has two meanings one is correct and other suggests direction. Same way, in Marathi “कर” (“kar”) it means hand as well as do something [8]. Dimension reduction means selecting only important attributes and removing noisy attributes from the data [9]. It improves the speed and accuracy of the algorithm, which is implemented on

the data. Singular value decomposition, latent semantic analysis, are some of the techniques of dimension reduction.

Different thesauruses like WordNet [6] are available to find out relevant terms. There are several applications once Document Term Matrix (DTM) is formulated. One of the applications is to find out document similarity, plagiarism detection and so on. The proposed approach is first of its kind to design sysnet document matrix for Marathi corpus. Fig. 1 shows the flow from textual data to document similarity.

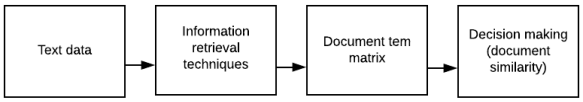

Fig. 1. Flow of Text Data to Decision Making.

The paper is organized as follows. The next section details literature review the third section depicts research methodology followed by results and discussions in the fourth section and the paper ends with the conclusions.

This research is unique because

1) Semantics and context is involved in Marathi documents' similarity process using dimension reduction.

2) Performance analysis of document similarity with existing technique is performed using four parameters.

3) Verses and proses are considered together for performing document similarity.

4) Polysemi problem is solved by identifying sense of the word.

II. Literature Review

Sentiment analysis [10] for Indian languages has become significant due to data present in Indian languages has expanded online and offline. The Marathi language is a resource scare language. The growth of Indian languages over a period in the area of sentiment mining is stated along with the taxonomy of Indian languages. A model is proposed for carrying out a sentiment analysis on Hindi tweets. It also focuses on the challenges of sentiment mining for Hindi tweets. The accuracy of the model is calculated [11]. It will provide sources of datasets with annotation for linguistic analysis and suggest the appropriate technique for sentiment analysis in a specific domain.

Different types of stemming techniques for Indian and Non-Indian languages are explained. The algorithm is proposed to retrieve the set of Marathi documents based on the users' requirements. The rule-based approach is followed by stemming techniques, which always performs better Brute force. Stemmers are build using NLP techniques along with Dictionary-based algorithms. The stemmers allow encoding different language-related rules. These stemmers are suitable for a specific language. A text summary of Marathi documents is performed by extracting tokens present in the data. It is done by abstracting documents and using morphological rules of language. It reduces the time and effort invested in reading the documents.

Due to large text available on different applications like travel aggregator, google assistant [12] the need for text summarization is evolved across the period. Summarization gives an abstract view of data in fewer words without changing its meaning. Different challenges of text mining are explored such as context-based analysis [16][17] and so on.

Generation of stop words in different Indian languages is also an evolving area [19][22].

Different Indian languages [20] such as Hindi [27][28][29] are explored by different researchers and NLP elements explored for each language are stated. A new framework that is bag of synset is proposed for multilingual document classification using synset document matrix and BabelNet knowledge base [15] Poetry corpus creation along with preprocessing of the corpus is achieved by Punjabi corpus and classifiers are executed [25]. Diacritic extraction methods are used for the Gujarati language along with information retrieval, stop word identification and classification and machine translation. List of stop word its analysis building dictionary, constituency mapping, development of lemmatizers and morphological analysis are developed in Sanskrit [5] [24]. Metadata is generated related to poetry and Hindi text analysis was performed. Stemming is used to improve the performance of the algorithm and it is a preprocessing technique. It removes tagging of the word and reduces it and used in information retrieval [21].

The sensitivity performance of negative news articles is implemented [13]. News articles are classified as positive, negative and neutral. The articles formed different domains that are sports, politics and so on. Local administration cannot take action against such news. Some news may be urgent to treat can be focused on a proposed approach. TF-IDF is used on unigrams and bigrams of 1000. Morphologically similar words present in the corpus are clustered using text stemming methods. NLP processing on the corpus is carried out after the collection of data and the creation of a corpus. Steps implemented on a corpus are tokenization, noise removal, normalization. Cognitive-inspired computing is used to discover morphologically related words. Document similarity determines the degree of closeness between two documents based on lexical and semantic similarity. Several measures are Manhattan, Word2Vec, cosine similarity, latent semantic indexing are suggested. Human intervention or language-specific knowledge is not used by the technique. Evaluation of the experiments for lemmatization and information retrieval is carried out. Four languages are chosen to carry out experiments [14].

Accuracy, precision, recall and F1 score [13] are popularly used parameters to evaluate document similarity. The ratio of correctly predicted similar documents to the total documents. Higher accuracy indicates the goodness of a model. Precision is defined as the ratio of correctly predicted similar documents to the total similar documents. More precision states the betterment of the model. The recall is the rightly predicted positive readings to all readings in the actual class and the F1 score is a weighted mean of precision and recall.

III. Research Methodology

This section details the metadata of the corpus as well as steps and the packages used to measure document similarity using DSMM. Library Udpipe [4] available in R programming along with different packages like tm, quanteda, spacyr are used to carry out experiments. Fig. 2 states different steps in research methodology and the description of each step is described in subsequent sections.

A. Data Collection, Corpus Creation and Preprocessing

Data is collected in the form of 713 verses and 493 proses [16] claimed that proses and verses are morphologically identical to process, so one can take either proses or verse or both to carry out NLP tasks. The corpus was not readily available so the data is collected using different websites [30-32]. Total 1206 Marathi text documents are collected near about 20 MB are processed. Separating the text strings into smaller units is known as tokenization. Paragraphs can be tokenized into sentences and sentences can be tokenized into words. The total number of tokens is 1,256,721. In the first step, tokenization is achieved by considering space as a delimiter, different tokens are identified. Removal of noise or stop word removal is carried out after tokenization. Stop words are those which need to be deleted from the corpus to remove noise. These are the words, which are not important and increase attributes, e.g. " " ("hai") meaning "this", punctuations, numbers, etc. Special characters and stop words identified and removed to get a total number of tokens as 56,345. The next step is lemmatization. It reduces the word to its base form. Lemmatization is said to be more accurate than stemming. It reduces word to a meaningful form. E.g. Lemma of studies is study and stem is "studi". Lemmatization uses morphological analysis while Stemming removes inflectional ending only. After lemmatization, unique terms are generated, Total unique tokens retrieved are 35,167.

B. Document Similarity Calculation

Total 1206 documents were processed, The Frequency of each term is calculated and significant terms are decided based on a threshold. The total significant terms which are retrieved are 8,123. The threshold is 60 % of maximum frequency which gives the best precision. Carrying out experiments on varied values of the threshold from 10 to 90, a 60 % threshold was found to have the highest precision. That is, suppose the maximum frequency of words is found to be 100, then tokens having frequency more than 60 are considered to be significant tokens. The document term matrix is formulated considering significant tokens which are placed in rows and cosine measure is used to find out the similarity between documents.

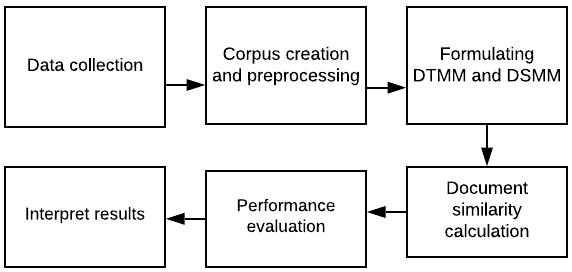

Fig. 2. Diagrammatic Representation of Research Methodology.

C. Synset based Vectorization

Similar tokens are identified and synset groups are formed. To identify similar tokens word net is used. To decide significant tokens synset group frequency is used. Total Group frequency is calculated by summing frequencies of terms present in the group. Unlike the traditional approach, synset groups are considered and the synset document matrix is constructed. Total synsets are 4110. For e.g. 'पात्र' means river bank and eligible too. If 'पात्र' (patra) and 'योग्य' (yogya) means proper, come together as synonyms it means the sense of the word is eligible. By using synset based vectorization not only dimensions are reduced, but the words which are significant with respect to corpus but were ignored due to their low frequency are being included in the form of a synset. The entire feature vector of the synset represents the context of the corpus. Thus proposed method attempts to involve semantics along with dimension reduction which justifies the title.

D. Distance Computation Matrix

It calculates the similarity between document vectors. Value "1" indicates the documents are the same and value closer to zero indicates the degree of dissimilarity. Cosine measure is value between 0 and 1. The similarity threshold is 75% which gives the best precision amongst the varied values of threshold, ranging from 10 to 100. If the cosine similarity measure between two documents is, more than 0.75 then those two documents are said to be similar.computation: Cosine measure is used to formulate the matrix

IV. Results and Discussions

Table I shows the morphological description of all extracted tokens. Table I also depicts the document position, paragraph number and sentence number for each and every token in the corpus. It assigns a unique number to each token and identifies other morphological information such as lemma, parts of speech that is a noun, verb and so on. "ADJ" means adjective, the column feats represents gender that is masculine or feminine, also the form of words that is singular or plural are been reelected in the table. For verbs, the voice of the verb is identified that is active or passive voice. The table depicts a total of 500 documents, containing 2567 paragraphs, 13,123 sentences and 4,213 unique tokens.

Only Nouns, Adjectives, Adverbs and Verbs are selected and unique terms are identified to formulate DTMM for 500 proses. Documents with their position in the corpus is depicted in the first column that is document id, the selected frequent terms, are placed as columns. It can be clearly seen in Table II that the frequency of 'गवत' in document 1 is 2 which is also called as term weight and when term weight is zero, it shows the absence of that term in the document

Table III shows the formulation of DTM using the proposed approach means DSMM. After identifying the frequency of each term, similar terms are grouped, and their frequencies are summed up. The group of similar terms is called synset and thus proposed approach considers synset as column heads. For e.g. 'राजा' and 'नृप' both of them means king. In Table 2 by existing approach frequency 'राजा' and 'नृप' is 1 and being similar treated as separate tokens but the

proposed approach considers mentioned synset group as a single token and synset frequency is 2 (Table III). It has effectively reduced dimensions and also considered the terms which are significant in the corpus but were ignored due to its less frequency. The proposed approach involves semantics in this way, by observing other synsets context of the corpus can be understood.

The next step is to identify the similarity between the documents based on the formulated synset document matrix. This step is also termed as a quantitative assessment of documents. A cosine measure is used to reflect the document similarity. Its value is between zero or one, a value near to one indicates the documents are more similar. Value zero indicates the documents do not share any token or synset group. Table IV shows the similarity between 1206 documents. The threshold value is 75 %, which means D2 is similar to D1206 because it has a value greater than 0.75, but D2 is not similar to D1 as its cosine measure is 0.64 which is less than decide threshold value

Fig. 3 shows a similarity between all verses with each other using the synset document term matrix, Y-axis represents a cosine similarity measure. V1 is the first verse. V1 is similar to V2 with 0.81 measure and it is the most similar to V9. In Fig. 4, the existing method (DTMM) is used to show the similarity between the verses and V1 is similar to V2 with less than 0.81 measure, that 0.56, the same case is observed for multiple instances, for example, V1 to V7 similarity, etc. The same experiment is carried out for proses also. Fig. 5 indicates a comparison of the similarity of V1 with all remaining verses using both techniques. Using the proposed approach degree of similarity is more accurate. To evaluate the performance of the proposed technique confusion matrix is generated and different evaluation parameters are considered.

TABLE I. MORPHOLOGICAL ANALYSIS OF MARATHI TEXT DISTANCE

Sr.No	doc_id	paragraph_id	sentence_id	sentence	token_id	token	lemma	pos	feats
1	doc1	1	1	**एक लांडगा खरोखर आला होता**	1	**एक**	**एक**	ADJ	type=Num, Number=Sing
2	doc1	1	1	**एक लांडगा खरोखर आला होता**	2	**लांडगा**	**लांडगा**	NOUN	Gender=Masc
3	doc1	1	1	**एक लांडगा खरोखर आला होता**	3	**खरोखर**	**खरोखर**	ADJ	

TABLE II. DOCUMENT TERM MATRIX FOR FREQUENT TERMS OF THE CORPUS

Document	रान	गवत	झाड	डोंगर	सिंह	राजा, नृप
doc1	2	2	1	1	1	3
doc 2	0	0	0	0	0	0
doc 500	0	0	1	0	0	0

TABLE III. DOCUMENT SYNSET MATRIX FOR FREQUENT TERMS OF THE CORPUS

Document	रान	गवत	झाड	डोंगर	सिंह	राजा,	नृप
doc1	2	2	1	1	1	2	1
doc 2	0	0	0	0	0	0	0
doc 500	0	0	1	0	0	0	0

TABLE IV. DOCUMENT SIMILARITY MATRIX USING COSINE MEASURE

Documents	D1	D2		D1205	D1206
D1	1	0.643041		0.783057	0.759827
D2	0.643041	1		0.904734	0.904734
D1205	0.710974	0.904734		1	0.781287
D1206	0.759827	0.799583		0.781287	1

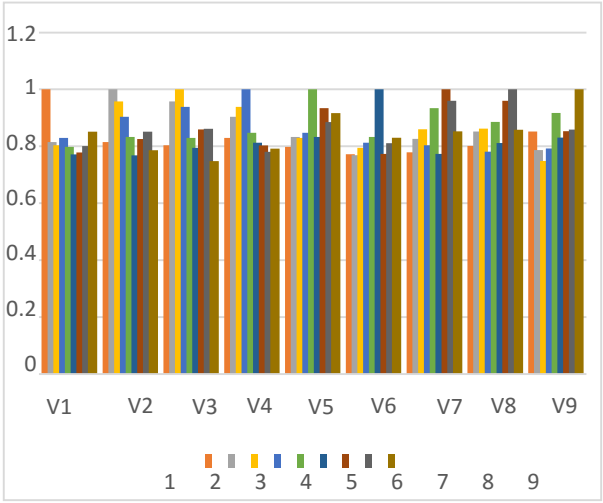

Fig. 3. Document Similarity using DSMM.

Table V shows the confusion matrix by the proposed method for 1206 documents. Positive indicates documents similar and negative indicates documents are not similar. For instance, out of 1009 similar documents 922 observed to be similar and 87 observed to be non-similar. Fig. 6 shows different evaluation parameters which prove the betterment of the technique, Accuracy, precision and F1 score is high and the error rate is low by 0.1 % than the existing method.

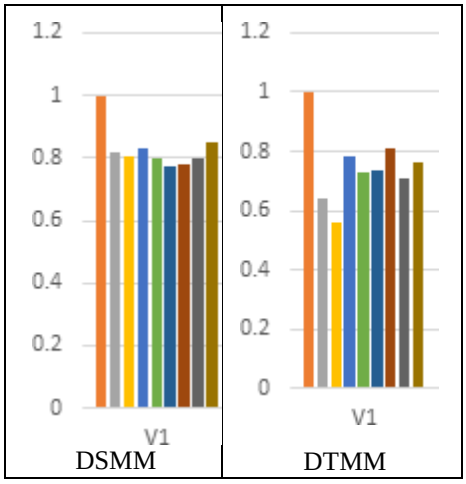

Fig. 4. Document Similarity using DTMM.

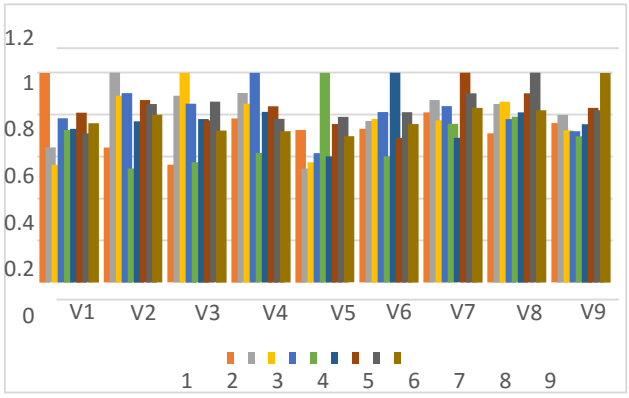

Fig. 5. Comparison between Vectors by DSMM and DTMM.

TABLE V. CONFUSION MATRIX

	Predicted	
Observed	Positive	Negative
Positive	922	87
Negative	59	138

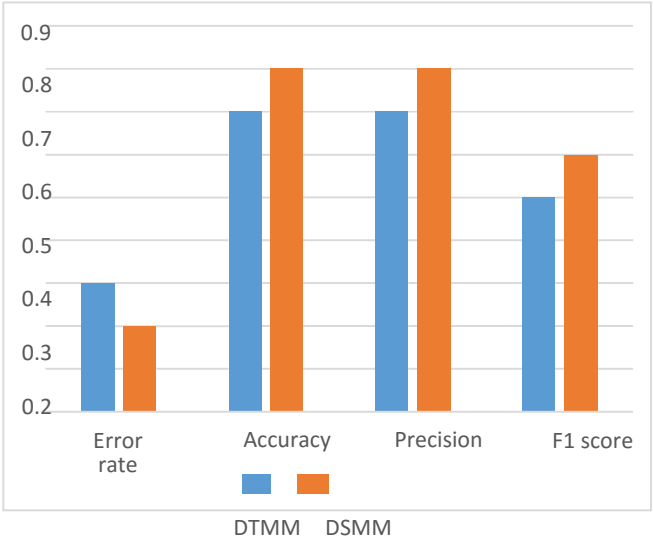

Fig. 6. Comparison between the Proposed and Existing Approach.

The similarity between documents using traditional approach such as DTMM and proposed approach that is DSMM is evaluated using different measures like precision, recall, accuracy and error. A confusion matrix is designed to show the results retrieved by DSMM and DTMM techniques. The proposed approach proved to be better.

V. CONCLUSION

The similarity of more than 1000 Marathi documents including proses and verses is calculated using DSMM. It uses the semantic relationship between words and forms a synset group of similar terms to form DSMM. Not only dimension reduction is achieved but the context of the documents also gets involved while formulating DSMM. DSMM identifies the sense of the word and solves the problem of polysemy. Comparative analysis between DSMM and DTMM using multiple evaluation parameters like error, precision, accuracy and F1 measure proves the betterment of DSMM. The technique can be used to detect the plagiarism of Marathi documents. Also, DSMM will act as a stepping stone towards NLP of regional languages.

REFERENCES

[1] Harikrishna, D. M., & Rao, K. S. (2015, September). Classification of children stories in Hindi using keywords and POS density. In 2015 International Conference on Computer, Communication and Control (IC4) (pp. 1-5). IEEE.

[2] Harikrishna, D. M., Reddy, G., & Rao, K. S. (2015, August). Multi-stage children story speech synthesis for Hindi. In 2015 Eighth International Conference on Contemporary Computing (IC3) (pp. 220-224). IEEE.

[3] SivaKumar, A. P., Premchand, P., & Govardhan, A. (2011). Application of latent semantic indexing for hindi-english clir irrespective of context similarity. In Trends in Network and Communications (pp. 711-720). Springer, Berlin, Heidelberg.

[4] https://rdrr.io/cran/udpipe/man/udpipe_download_model.html

[5] Murali, N., Ramasreee, D. R., & Acharyulu, D. K. (2014). Kridanta Analysis for Sanskrit. Int. Journal on Natural Language Computing, 3(3), 33-49.

[6] Kim, Y. B., & Kim, Y. S. (2008, July). Latent semantic kernels for WordNet: Transforming a tree-like structure into a matrix. In 2008 International Conference on Advanced Language Processing and Web Information Technology (pp. 76-80). IEEE.

[7] Mahmoud, A., & Zrigui, M. (2019, September). Similar Meaning Analysis for Original Documents Identification in Arabic Language. In International Conference on Computational Collective Intelligence (pp. 193-206). Springer, Cham.

[8] Baruah, N., Sarma, S. K., & Borkotokey, S. (2019, February). Text Summarization in Indian Languages: A Critical Review. In 2019 Second International Conference on Advanced Computational and Communication Paradigms (ICACCP) (pp. 1-6). IEEE.

[9] St, J. (2019, March). Reduction of Dimensionality of Feature Vectors in Subject Classification of Text Documents. In Reliability and Statistics in Transportation and Communication: Selected Papers from the 18th International Conference on Reliability and Statistics in Transportation and Communication, RelStat‟18, 17-20 October 2018, Riga, Latvia (Vol. 68, p. 159). Springer.

[10] Nakov, P., Ritter, A., Rosenthal, S., Sebastiani, F., & Stoyanov, V. (2019). SemEval-2016 task 4: Sentiment analysis in Twitter. arXiv preprint arXiv:1912.01973.]

[11] Sharma, Y., Mangat, V., & Kaur, M. (2015, September). A practical approach to sentiment analysis of Hindi tweets. In 2015 1st International Conference on Next Generation Computing Technologies (NGCT) (pp. 677-680). IEEE.]

[12] Dhawale, A. D., Kulkarni, S. B., & Kumbhakarna, V. (2019, October). Survey of Progressive Era of Text Summarization for Indian and

Foreign Languages Using Natural Language Processing. In International Conference on Innovative Data Communication Technologies and Application (pp. 654-662). Springer, Cham.],

[13] Jena, M. K., & Mohanty, S. (2019, December). Predicting Sensitivity of Local News Articles from Odia Dailies. In International Conference on Biologically Inspired Techniques in Many- Criteria Decision Making (pp. 144-151). Springer, Cham.

[14] Alotaibi, F. S., & Gupta, V. (2018). A cognitive inspired unsupervised language-independent text stemmer for information retrieval. Cognitive Systems Research, 52, 291-300.

[15] Romeo, S., Ienco, D., & Tagarelli, A. (2015, March). Knowledge-based representation for transductive multilingual document classification. In European Conference on Information Retrieval (pp. 92-103). Springer, Cham.

[16] Bafna P.B., Saini J.R., 2020, An Application of Zipf's Law for Prose and Verse Corpora Neutrality for Hindi and Marathi Languages, in press

[17] Bafna P.B., Saini J.R., 2020,"Marathi Text Analysis using Unsupervised Learning and Word Cloud", International Journal of Engineering and Advanced Technology,9(3),in press

[18] Rakholia, R. M., & Saini, J. R. (2015, March). The design and implementation of diacritic extraction technique for Gujarati written script using Unicode Transformation Format. In 2015 IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT) (pp. 1-6). IEEE.

[19] Kaur, J., & Saini, J. R. (2016). POS Word Class Based Categorization of Gurmukhi Language Stemmed Stop Words. In Proceedings of First International Conference on Information and Communication Technology for Intelligent Systems: Volume 2 (pp. 3-10). Springer, Cham.

[20] Rakholia, R. M., & Saini, J. R. (2017). Automatic Language Identification and Content Separation from Indian Multilingual Documents Using Unicode Transformation Format. In Proceedings of the International Conference on Data Engineering and Communication Technology (pp. 369-378). Springer, Singapore.

[21] Saini, J. R., & Rakholia, R. M. (2016). On continent and script-wise divisions-based statistical measures for stop-words lists of international languages. Procedia Computer Science, 89, 313- 319.

[22] Rakholia, R. M., & Saini, J. R. (2017). A rule-based approach to identify stop words for Gujarati language. In Proceedings of the 5th International Conference on Frontiers in Intelligent Computing: Theory and Applications (pp. 797-806). Springer, Singapore.

[23] Rakholia, R. M., & Saini, J. R. (2016). Lexical classes based stop words categorization for Gujarati language. In 2016 2nd International Conference on Advances in Computing, Communication, & Automation (ICACCA)(Fall) (pp. 1-5). IEEE.

[24] Raulji, J. K., & Saini, J. R. (2017, January). Generating Stopword List for Sanskrit Language. In2017 IEEE 7th International Advance Computing Conference (IACC) (pp. 799-802). IEEE.

[25] Kaur, J., & Saini, J. R. (2020). Designing Punjabi Poetry Classifiers Using Machine Learning and Different Textual Features. INTERNATIONAL ARAB JOURNAL OF INFORMATION TECHNOLOGY, 17(1), 38-44.

[26] Rakholia, R. M., & Saini, J. R. (2017). Information Retrieval for Gujarati Language Using Cosine Similarity Based Vector Space Model. In Proceedings of the 5th International Conference on Frontiers in Intelligent Computing: Theory and Applications (pp. 1-9). Springer, Singapore.

[27] Venugopal-Wairagade, G., Saini, J. R., & Pramod, D. (2020). Novel Language Resources for Hindi: An Aesthetics Text Corpus and a Comprehensive Stop Lemma List. arXiv preprint arXiv:2002.00171.

[28] 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. (Nov-2019)

[29] 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.(December)

[30] www.marathi.webdunia.com accessed on 18-04-2020

[31] www.britannica.com/art/ Marathi-literature accessed on 18-04-2020

[32] www.marathisahityadarpan.com accessed on 18-04-2020