An Application of Zipf's Law for Prose and Verse Corpora Neutrality for Hindi and Marathi Languages

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

***Abstract*—Availability of the text in different languages has become possible, as almost all websites have offered multilingual option. Hindi is considered as official language in one of the states of India. Hindi text analysis is dominated by the corpus of stories and poems. Before performing any text analysis token extraction is an important step and supports many applications like text summarization, categorizing text and so on. Token extraction is a part of Natural language processing (NLP). NLP includes many steps such as preprocessing the corpus, lemmatization and so on. In this paper the tokens are extracted by two methods and on two corpora. BaSa, a context-based term extraction technique having different NLP activities, e.g. Term Frequency Inverse Document Frequency (TF-IDF) and Zipf's law are used to count and compare extracted tokens. Further token comparison between both of the methods is achieved. The corpus contains proses and verses of Hindi as well as the Marathi language. Common tokens from corpora of verses and proses of Marathi as well as Hindi are identified to prove that both of them behave same as per as NLP activities are concerned. The betterment of BaSa over Zipf's law is proved. Hindi Corpus includes 820 stories and 710 poems and Marathi corpus includes 610 stories and 505 poems.**

Keywords—Marathi; NLP; Synset; Zipf's law

I. INTRODUCTION

Hindi and Marathi languages are not only popular in the world but also are used as an official language in North India and Maharashtra, respectively [1]. So, abundant Hindi and Marathi text get generated day by day. To process this data NLP techniques along with machine learning algorithms are available in the literature. Generally, to analyze the behavior of algorithms, the corpus of Hindi or Marathi poems and stories is being used. Poems and stories are part of the literature. Stories and poems act as a guide to children about their behavior and manners [2-3] and connect with elders to interconnect ideas and visualize life's opportunities. The use of rhyme and meter gives musical sense to the poetry, which is termed as literary elements whereas stories include a set of incidents and characters. Nouns, adjectives, adverbs are prominently used to construct a story or a poem [4].

NLP processing on this corpus is carried out after the collection of data and the creation of a corpus. There are three steps implemented on a corpus which are tokenization, noise removal and normalization [5-8]. Separating the text strings into smaller units is known as tokenization. Paragraphs can be tokenized into sentences and sentences can be tokenized into words [9-11].

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 increases attributes. Eg. 'में' (mei) in Hindi and 'या' (ya) in Marathi, that is in, punctuations, numbers, etc. The next step is normalization, stemming and lemmatization are part of normalization. It reduces the word to its base form. Lemmatization [12-15] 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. Lemma uses morphological analysis. Stem removes inflectional ending only.

After lemmatization, generally, term frequency-inverse document frequency is used. It is based on a total number of terms present in the corpus. The importance of the term increases as its count is increased but it is offset by inverse document frequency. Terms present in almost all of the documents are ignored. TF-IDF measure is assigned to the significant terms in the corpus.

Zipf's law is another measure to decide the significance of terms [16]. When applied to the language it states that the top 20% of the most frequently used words in a corpus large enough will make up 80% of it.

To make it clearer, say a novel contains 5000 different words. According to the rule, 80% of the novel will be the most frequently used 1000 words. It allows us to extract all terms/ words and states that the rank of a word is inversely related to its frequency. Mathematically, terms having frequency 40% of maximum frequency are significant. For e.g., If the maximum frequency of the term in the corpus is 100 the terms having frequency >=40 are significant.

In this paper statistical analysis of the tokens present in both the corpus is depicted visually along with the common tokens at each stage. It will provide guidelines to researchers working in the metalinguistic domain.

Corpora containing more than 1500 documents, more than 3 Million terms words are processed. Statistical visual analysis of the terms is carried out using BaSa [1]. Zipf's law is applied to the same corpus and common tokens are identified. Similarly, Marathi corpus is created to apply BaSa and common tokens are extracted for both of the corpora, also Zipf's law is applied on Marathi corpus common tokens are identified for Marathi corpus too. Finally, Common tokens extracted by Zipf's law and BaSa are compared for both of the corpora. This research is unique because:

1) More than 3 Million terms are processed

2) Context-based token comparison with lemmatization and its visualization is done the first time

3) 4 types of corpora with multilinguistic context-based approach are processed

Processing proses and verses are one and the same with respect to NLP activities.

II. Background

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 provided opportunities to mine this data.

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 [17].

Sentiment analysis [18-20] for Indian languages has become significant due to data present in Indian languages has expanded online and offline. The growth of Indian languages over a period in the area of sentiment mining is stated along with the taxonomy of Indian languages.

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 [2][21].

Due to large text available on different applications like travel aggregator, google assistant, 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 and so on.

Different Indian languages are explored by different researchers and NLP elements explored for each language are stated. Poetry corpus creation along with preprocessing of the corpus is achieved by Punjabi corpus and classifiers are executed. 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 [22]. 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.

The sensitivity performance of negative news articles is implemented. 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 by a proposed approach. TF-IDF is used on unigrams and bigrams of 1000 news collected from websites and performance of the classifier were evaluated [3][21].

To predict about the occurrences of the terms, Zipf's Law is used as base-line rule. Word's frequency decides its role in the entire corpus. The semantic influence of a word and probability of significance of the word is expressed by Zipf's law. Zipf's law allows to asses the relevance of the terms and identifies their patterns for the corpus [4]. There are some drawbacks to Zipf's law, the formulation of Zipf's law is ambiguous, from statistical perspectives, also it is not suitable for the big corpus. Three versions of Zipf's law are designed. The versions are tested on more than 30, 000 words. Statistical tests are used for fitting of functions. It's resulted in the fitting of more than 60 % terms at 0.05 significance level. [5].

III. Research Methodology

The first step in the proposed approach is data collection and corpus creation. Different type's poems and stories written by different authors were collected from various websites [22-24]. BaSa [1] is applied to identify common tokens present in both Hindi verses and proses. Similarly, Zipf's law is applied to extract tokens for Hindi language and Common token extracted by Zipf's law for proses and verses are identified. Similarly same process is repeated for Marathi corpus, that is, Basa and Zipf's law is applied on Marathi corpus having prose and verses and common tokens are extracted. Finally, compare the tokens retrieved by Zipf's law and Basa for both language corpora.

Library Udpipe present in "R" programming language is used to perform different NLP operations such as tokenization, tagging, lemmatization and so on. Udpipe is language-agnostic and can be trained given annotated data. Udmodel is a function of udpipe to load language type that is either Hindi or Marathi. Fig. 1 shows a diagrammatic representation of research methodology.

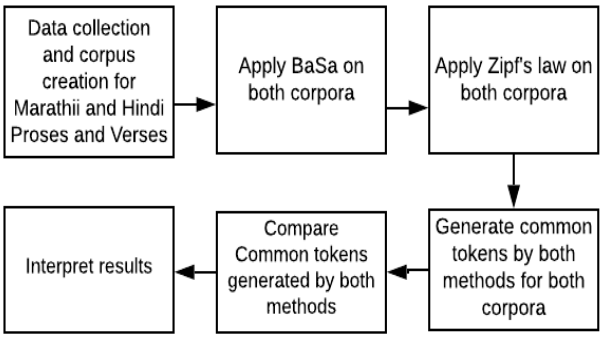

Fig 1. Diagrammatic Representation of Research Methodology.

Data collection and corpus creation: Different types of poem and stories written by different authors were collected from various websites. Hindi Corpus includes 820 stories and 710 poems and Marathi corpus includes 610 stories and 505 poems.

Implement BaSa on Marathi as well as Hindi corpus and identify common tokens.

BaSa includes tokenization followed by stop words and noise removal, normalization, TF-IDF and formation of synsets. Thus it involves context-based identification of common terms. It means that if the word 'raat' means night is present in the story and nisha is present in the poem , it identifies that these are synonyms and considers it as one synset group, thus in a final step comparison between synsets groups of verses and poems is being carried out. Similarly, synset groups of Marathi corpus are being constructed for e.g. 'Aayusha' and 'Jeevan' means life is being considered in a synset group and accordingly common tokens are identified.

Apply Zipf's law to extract tokens on Hindi as well as Marathi corpus. Zipf's law does not follow preprocessing and other NLP steps. It is based on the frequency and rank of the term. For eg. एकदा भक्त पुरंदरदास राजवाड्यात गेले होते, (Ēkadā bhakta purandaradāsa rājavāḍyāta gēlē hōt) (Once, the devotee, Purandaradas went to the palace). भक्ताने त्या तांदळात थोडे हिरे मिसळले होते. (Rājānē tyā tāndaḷāta thōḍē hirē misaḷalē hōtē) (The king had mixed little diamonds in the rice.). Zipf's law will consider 'होते' as the maximum frequency word that is 2. Rest all words have frequency 1 and thus the rank of 'होते 'is 1. Thus Frequency and Rank are inversely proportional.

Identify common token extracted by Zipf's law on the corpora (proses and verses) of Marathi and Hindi Zipf's law is applied on the corpus of Marathi and tokens which are present in both verse and proses are found out.

Compare common tokens identified by Zipf's law and BaSa on the corpora (proses and verses) of Marathi and Hindi: In the last step, common tokens identified by both of the methods are compared. It is observed that common tokens generated using BaSa are slightly more than tokens extracted by Zipf's law.

IV. Results and Discussions

Table I presents the detailed token analysis used by BaSa. The first column shows the sample statement of a Marathi poem, by considering space as delimiter tokenization is achieved, that is separating all words, punctuations numbers and so on. The third column removes stopwords, "ते", "म्हणजे" and so on removed. In the next step of lemmatization, the word is converted into its root form that is "माणसाच्या" means man's is reduced to "माणूस means man. The effect of TF-IDF can be reflected for multiple documents. Combining TF-IDF with synset construct the group of similar terms together and treated as one synset group.

TABLE I. Steps in BaSa Approach

Sample statement in the corpus of Marathi Poem	Tokenization	Removings topwords	Lemmatization	TF-IDF+Synset
माणसाच्या सुखाचं व आनंदी रहायचं एकमेव रहस्य ते म्हणजे हास्य,	"माणसाच्या",: सुखाचं", "व ","आनंदी" ,"रहायचं", "एकमेव", "रहस्य", "ते" ,"म्हणजे" ",हास्य", ","	माणसाच्या", :सुखाचं", "आनंदी""रहायचं", "एकमेव","रहस्य", ",हास्य"	"माणूस","सुख" "आनंद", "राहणे" , "एकमेव" , "रहस्य", ,हास्य"	**सुख, आनंद हास्य** ", "राहणे" , "एकमेव", "रहस्य"," माणूस"

Table II states the number of extracted tokens and the common tokens at each level of BaSa for Marathi as well as Hindi corpus. The corpus consists of 505 verses and 610 proses. At each stage, 50 to 60 % tokens are reduced. For both languages. NLP results are almost the same for both languages' proses and verses and it proves that prose and verse behave same that is neutral as per as NLP activities are concerned.

Zipf's law is used to decide important tokens of the corpus. It is executed on the corpus of Hindi stories/proses. It is based on the frequency and rank of the token. Table III shows the token frequency and its corresponding rank. It's clear that rank and frequency are inversely proportional. The last column specifies the significance level which shows importance of the term based on probability measure.

Rank Frequency graph allows to visualize word/token rank versus token frequencies based on Zipf's law. It is clear from the graph that the rank of the token is minimum for the highest frequent word and rank increases as the frequency decreases. The tokens are extracted from corpus of Hindi stories. Fig. 2 shows Rank-Frequency plot for the corpus of the Hindi proses Same way Zipf's law is executed on Marathi corpus. The sample of tokens extracted from Marathi Poems is presented in the table. It is observed that the maximum word frequency of the term is 500.

TABLE II. Summary of sample corpus at each stage of BaSa for 505 poems and 610 Stories

Sr.No	Corpus	Verses		Proses		Common tokens	
		Hindi	Marathi	Hindi	Marathi	Hindi	Marathi
1	Total number of tokens	77,282	75,144	1,29,260	88,505	45,157	43,891
2	After removing stop words	35,181	32,234	54,239	45,321	22,123	19,234
3	Lemmatization	18,282	16,123	23,202	22.123	11,231	10,340
4	TF-IDF	7,139	7,139	8,102	7,123	2,123	1,123
5	Synset (groups)	4034	3,543	4,203	3.912	916	910

TABLE III. ZIPF'S LAW MEASUREMENT FOR THE CORPUS OF HINDI STORIES

Sr.no	word	Frequency	Rank	Significance level
1	घुस	4	1	0.6688
2	चिल्ला	3	2	0.5033
3	छाँह	2	3	0.4133
4	रिक्शे	1	4	0.3722
5	बच्चों	1	5	0.3613

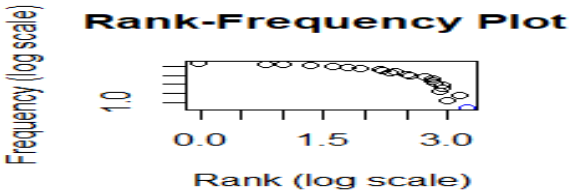

Fig 2. Rank Frequency plot by Zipf's law.

So according to Zipf's law, top and bottom 20 % tokens are discarded. So words having a frequency between 100 to 400 are considered. Eg. "तू","सगळ्यात", "तिला", won't be considered as major tokens and will be discarded. Table IV shows tokens with their frequencies with respect to the entire corpus.

Fig. 3 shows sample tokens extracted using Zipf's law for Marathi corpus. It can be observed that terms above 20 % threshold are being considered. X-axis shows the terms and Y axis shows the frequency of the term. The line shown in the graph is known as the Pareto line which is used to represent a cumulative percentage to show the importance of the term.

TABLE IV. TOKEN FREQUENCY FOR SELECTED TOKENS BY ZIPF'S LAW

Word	तू	सगळ्यात	आधी	आजी	बोलायचं	तिला
Frequency	500	459	410	389	210	90

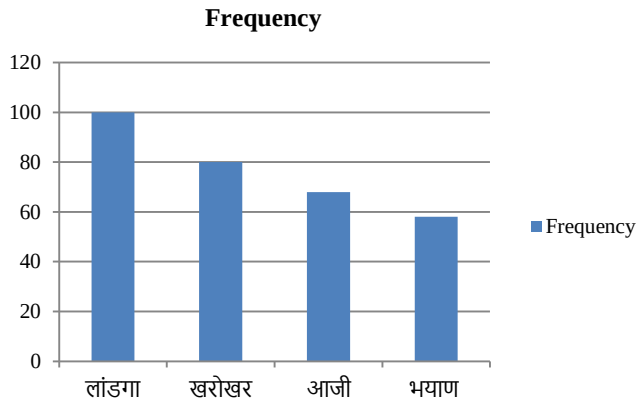

Fig 3. Common Tokens Identification by Zipf's law.

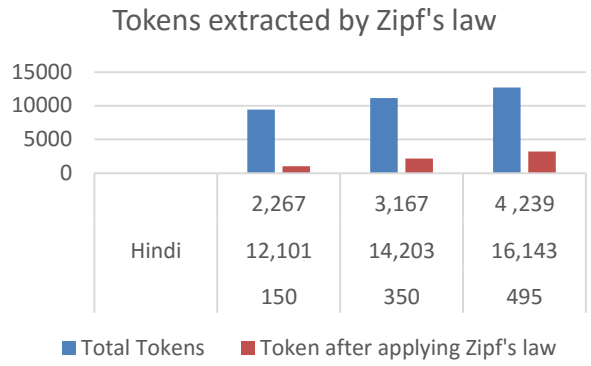

Fig 4. Token Count Extracted by Zipf's law.

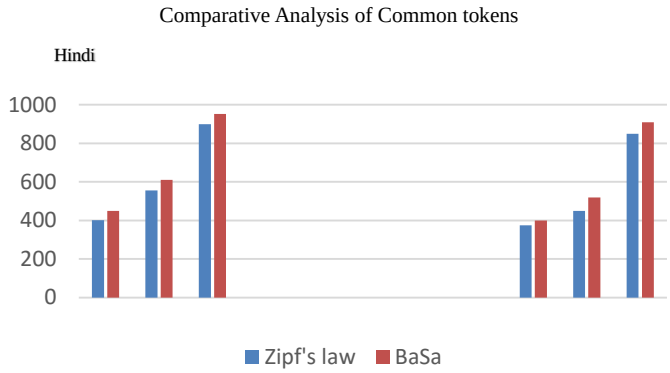

Fig 5. A Comparative Analysis of Common Tokens Extracted by Zipf's law and BaSa.

Fig. 4 shows a total number of tokens extracted and tokens retrieved after applying threshold using Zipf's law on Marathi and Hindi corpora for varied data size. There is more than 60% reduction in terms after applying Zipf's law. For the Hindi corpus size of 150 verses, total tokens are 12,101 and selected tokens are 2,267.

Fig. 5 shows a comparative analysis of the common tokens on varied corpus size of both the languages. X axis shows the sample data of verses and proses. Y axis shows common tokens. The number of common tokens extracted by BaSa is slightly more than Zipf's law. Not only the count of the token is more, but also the quality of tokens is better with respect to the context of the term. This is due to synset grouping varied out by BaSa.

V. CONCLUSIONS

Zipf's law is applied for corpora of proses and verses. NLP activities were carried out using BaSa. BaSa proved to be better than Zipf's law. Prose and verse give same results as per as NLP activities are concerned, so researchers can take either proses or verses or both for performing NLP activities. Hindi and Marathi corpora were considered and more than 3 Million documents were processed to identify common tokens between verses and proses. Considering Hinglish words (English words written in Hindi) can be incorporated in future.

REFERENCES

[1] Bafna P.B., Saini J.R., 2020, BaSa: A Context based Technique to Identify Common Tokens for Hindi Verses and Proses, under review

[2] Mishra, D., Venugopalan, M., & Gupta, D. (2016). Context specific Lexicon for Hindi reviews. Procedia Computer Science, 93, 554-563

[3] 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]

[4] Wyllys, R. E. (1981). Empirical and theoretical bases of Zipf's law.

[5] Moreno-Sánchez, I., Font-Clos, F., & Corral, Á. (2016). Large-scale analysis of Zipf's law in English texts. PloS one, 11(1).

[6] Milan Straka and Jana Straková. Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe(http://ufal.mff.cuni.cz/~straka/papers/2017-conll_udpipe.pdf). In Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universaldependencies, Vancouver, Canada, August 2017

[7] 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

[8] 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.

[9] 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.

[10] 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.

[11] 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

[12] 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

[13] 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

[14] Bafna P.B., Saini J.R., 2020, "Hindi Verse Class Predictor Using Eager Machine Learning Algorithms" International Conference On Emerging Smart Computing And Informatics 2020(IEEE-ESCI-2020), Pune, India.2018 (March)

[15] Bafna P.B., Saini J.R., 2020,"On Exhaustive Evaluation of Eager Machine Learning Algorithms for Classification of Hindi Verses", International Journal of Advanced Computer Science and Applications, in press

[16] Bafna P.B., Saini J.R., 2020, Hindi Verse Class Predictor using Concept Learning Algorithms, International Conference on Innovative Mechanisms for Industry Applications (ICIMIA 2020)

[17] Venugopal G., Saini J.R., Dhanya P., Novel Language Resources for Hindi: An Aesthetics Text Corpus and a Comprehensive Stop Lemma List, International Journal of Advanced Computer Science and Applications, vol. 11(1), Jan. 2020, in press

[18] 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.

[19] Harrag, F., & El-Qawasmah, E. (2009, August). Neural Network for Arabic text classification. In 2009 Second International Conference on the Applications of Digital Information and Web Technologies (pp. 778-783). IEEE.

[20] Yu, B., Xu, Z. B., & Li, C. H. (2008). Latent semantic analysis for text categorization using neural network. Knowledge-Based Systems, 21(8), 900-904.

[21] Zhang, D., & Lee, W. S. (2003, July). Question classification using support vector machines. In Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval (pp. 26-32).

[22] https://proudtobeprimary.com/reasons-teach-poetry-classroom/

[23] https://www.livehindustan.com/nandan/good-morning-story/news

[24] http://www.cfilt.iitb.ac.in