Journal of Engg. Research EMSME Special Issue pp. 122-132 DOI : 10.36909/jer.EMSME.13893

Keyword Extraction Methodologies Based on Rényi Entropy and Tsallis Relative Entropy

Aakanksha Singhal and D.K. Sharma

Department of Mathematics, Jaypee University of Engineering and Technology, A.B. Road, Raghogarh, Dist. Guna (M.P)-473226, India.
Corresponding Author : dilipsharmajiet@gmail.com

ABSTRACT

Textual and other forms of data is accumulating at a very fast pace through various social, academic and economic activities. Keyword extraction technique plays a significant role in analyzing, integrating, interpreting this amorphous data and exploiting its potential. Keyword extraction assists in making better sense of these vast information and data resources and leverage their value. In this article we propose and analyze keyword extraction methodology using Rényi entropy and Tsallis relative entropy. The proposed methods being statistics-based methods, these are language and domain independent method. The proposed methodology may find applications in dynamic text collection, information retrieval, natural language processing etc.

Keywords: Text mining; Keyword extraction; Rényi entropy; Tsallis relative entropy; Dynamic text collection.

INTRODUCTION

Keyword extraction plays a significant role in text mining. Text mining aims at extricating high-quality information from the given pool of textual data. In simpler terms it is the process of identifying key or relevant terms from a document, that supports in representing an suitable subject of the text. Huge quantity of new data and text is being generated through various social, economic and academic activities, with a substantial potential economic and social value. It is the techniques like text and data mining that assists in exploiting this potential. Text mining aids in assimilating, analyzing, interpreting the unstructured pool of data and puts the inferences for the apt and prompt use. Organizing and maintaining this dynamic text collection is challenging, time-consuming, expensive and tedious. And this is where keyword extraction, especially domain independent statistical keyword extraction plays an important role by helping in determining the relevant documents from a large pool of available data (Rossi, Marcacini and Rezende, 2014).

Keyword extraction has applications in text mining, information retrieval, web page retrieval, natural language processing, incremental clustering applications (Beliga, 2014; Rossi, Marcacini and Rezende, 2014) . Various methodologies for keyword extraction are broadly classified as: Linguistic, Machine learning and Statistical (Beliga, 2014). Linguistic approaches are usually derived from the linguistic attributes and are based on syntactic and semantic structures in the text. The linguistic approach comprises of the lexical analysis, syntactic analysis, discourse analysis and so on (Siddiqi and Sharan, 2015). Machine Learning approaches normally consider supervised learning methods. In these methods a model is trained based on a set of keywords extracted from training documents, and then the model is tested for performance through a testing module. Based on performance evaluation a satisfactory model is chosen for keyword extraction from new documents. This approach uses Naïve Bayes, Support Vector Machine, etc (Beliga, 2014; Jamaati and Mehri, 2018). Statistical methods require neither the training data nor the prior domain knowledge. Additionally, the statistical methods are language independent. In these methods the statistics of the words from the document forms a basis for keyword identification. The statistical methods are based on the following characteristics of text like term frequency (Luhn, 1958), standard deviation (Ortuno et al., 2002),

centrality measure, spatial distribution (Carpena et al., 2009), entropy (Herrera and Pury, 2008), word co-occurrences (Matsuo and Ishizuka, 2004), etc. The statistical methods are computationally more efficient than other methods but sometimes they may exhibit lower accuracy for some health and medical texts where the most substantial keywords may have low frequency. Statistical models may inadvertently not consider such words as keywords (Chen and Lin, 2010).

ENTROPY

Entropy in general is a measure of uncertainty or randomness of a system. C.E. Shannon (Shannon, 1948) introduced the concept of entropy to information theory. Let X be a discrete random variable with possible values $\{x_1,x_2,\ldots x_n\}$ and probability mass function $P_d(X)=\{p_1,p_2,\ldots,p_n\}$, then the Shannon's entropy (Shannon, 1948) is given by

$$S_S(P_d)=-\sum_i^N = 1 p_i \log p_i \tag{1}$$

Using the exponential mean, Rényi proposed a new generalized information measure commonly termed as Rényi's information measure or Rényi's entropy (Rényi and others, 1961):

$$S_R(P_d,q)=\frac{1}{(1-q)}\log_b \sum_i^N = 1 p_i^q \tag{2}$$

where b is the logarithm base and $q>0$ is a real parameter.

Considering applications to non-extensive systems, Tsallis (Tsallis, 1988) proposed a generalized entropy given by:

$$S_T(P_d,q)=\frac{1}{(1-q)}\left(\sum_i^N=1 p_i^q - 1\right) \tag{3}$$

As $\rightarrow 1$, both Rényi's $S_R(P_d,q)$ and Tsallis entropy $S_R(P_d,q)$ approaches Shannon's entropy

$S_S(P_d)$ i.e. $S_R(P_d,1)=S_S(P_d)$.

In information theory, The (Kullback and Leibler, 1951) relative entropy is a measure of how one probability distribution varies or differs from another distribution and is given by

$$D_{KL}(P\|Q)=\sum_i^n=1 p_i \log\left(\frac{pi}{qi}\right) \tag{4}$$

Tsallis relative entropy (Furuichi, Shigeru, Kenjiro Yanagi, 2004) is a generalization to relative entropy and is given by:

$$D_{Tsallis}(P\|Q)=-\sum_i^n=1 p_i \ln_q\left(\frac{pi}{qi}\right) \tag{5}$$

where $\ln_q(x) = (x^{1-q}-1)/1-q$, $q\neq 1$. As $q\rightarrow 1$, $\ln_q(x)\rightarrow \ln(x)$ and $D_{Tsallis}(P\|Q)\rightarrow D_{KL}(P\|Q)$.

Keyword extraction using Tsallis entropy has been explored in (Jamaati and Mehri, 2018). Rényi entropy has been successful in exploiting multifractal systems (Jizba and Arimitsu, 2004) and systems with mixed population of random variable. Moreover, Rényi entropy has successful application in the complex systems (Bashkirov, 2006) and the fact that the maximum entropy principle results in the same form of q-exponential distribution function for both Tsallis (Tsallis, 1988) and Rényi entropies (Johal and Tirnakli, 2004), we aim at exploring the use of Rényi entropy for keyword extraction. We have also explored the keyword extraction methodology using Tsallis relative entropy in this manuscript.

METHODOLOGY

This section aims at defining the methodology used for extrication relevant words from the given text using Rényi entropy and Tsallis relative entropy. The various notations used in the paper are as follows:

Table 1. List of common notations used in methodology for extracting relevant words from a given text

Notation	Description
N	Difference between the positions of last and first occurrences of word w
d_i (w)	Distance between i^{th} and $(i+1)^{th}$ occurrences of word w
p_i (w)	Probability distribution is obtained by dividing d_i (w), the distance between i^{th} and $(i+1)^{th}$ occurrences of word w by N
F_w	Frequency of intended word w
R(w,q)	Rényi entropy-based word ranking parameter
ΔS_T (w,q)	Tsallis entropy-based word ranking measure [(Jamaati and Mehri, 2018)]
$D_{Tsallis}$ (P||Q)	Tsallis relative entropy

To start with, a probability distribution is obtained as p_i (w)=d_i (w)/N. The probability distribution satisfies the normalization condition: $\sum_i^{FW=1} p_i(W) = 1$ as $\sum_i^{FW=1} d_i(W)=N$. Rényi entropy-based word ranking measure is evaluated for all word types in the given text and a word ranking measure using Rényi entropy is defined as:

$$R(w,q)=|\log_2 F_w - S_R(w,q)| \quad (6)$$

where $S_R(w,q)=\frac{1}{(1-q)} \log_2 \sum_i^{FW=1} p_i^q$. For even distribution of the word type w in the text and $0<q\leq 2$ and $q\neq 1$, S_R (w,q) approaches $\log_2 N_w$. It is observed that grammatical words are more homogeneously distributed in the text, and therefore have lower values for the word relevance ranking parameter R(w,q). In contrast, the relevant words are not so evenly spread and hence have greater values for parameter R(w,q). Words are sorted in descending order of R(w,q) value. The most relevant words appear in the upper part of the sorted list, which can be considered as the retrieved set or index.

Since grammatical words are more homogeneously spread in the text, they should have lower values of relative entropies when evaluated with respect to a completely homogeneous distribution. The homogeneous probability distribution is obtained as q_i (w)= $\frac{1}{F_w}$. The methodology for extracting relevant words from a given text based on Tsallis, Rényi entropy and Tsallis relative entropy are as shown in Figure 1 and Figure 2 respectively.

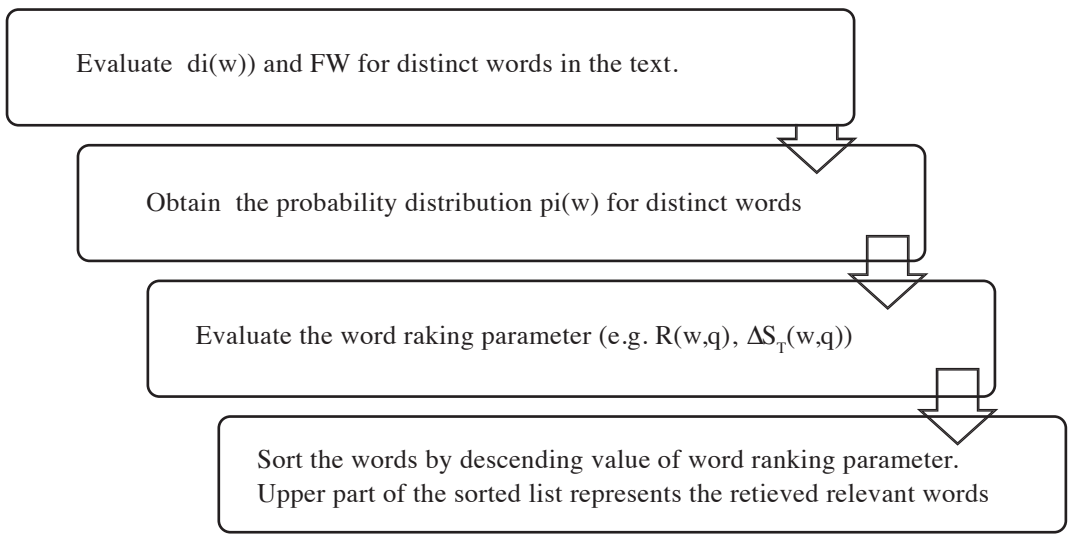

Figure 1. Methodology for extracting retrieved list from given text using entropy

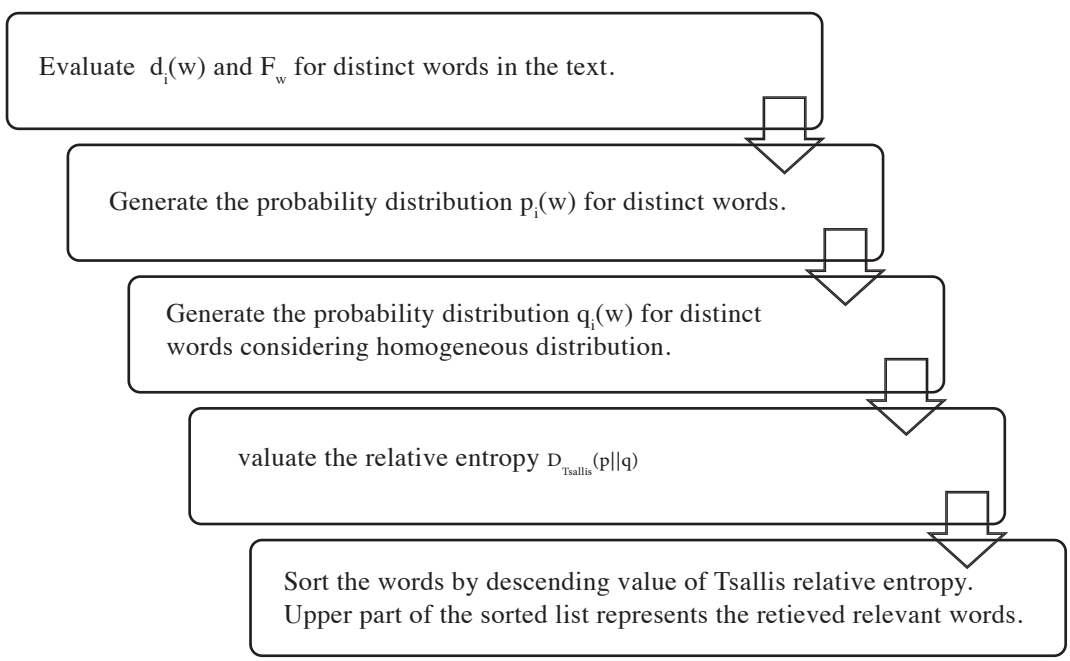

Figure 2. Methodology for extracting retrieved list from given text using Tsallis relative entropy

RESULTS AND DISCUSSIONS

Both the actual and relative performance of proposed Rényi entropy-based keyword extraction method has been shown. The relative performance of Rényi entropy-based word ranking parameter has been evaluated with respect to Tsallis entropy-based word ranking parameter [4]. The books used for evaluating the performance of proposed word ranking method are: The essence of Hinduism (Gandhi, 1987)(Book 1) and Global warming: the complete briefing (Houghton, 2009) (Book 2). Firstly, the books are obtained as text files. Post deleting the punctuations & numerals and converting the complete text to alphabetical lower case, space character in text is used as a separator to extricate consecutive terms as individual words. A set of relevant words have been extracted from the index of given books. The cardinality of this relevant set is 244 and 551 for Book 1 and Book 2 respectively. These sets are used to evaluate the actual performance of proposed word ranking metric.

Since no stemming and lemmatization process have been applied as pre-processes, all inflectional and derivationally related forms of a word are considered as different words.

Let N_{rel} and N_{ret} respectively represent the cardinality of relevant and retrieved sets and $N_{rel \cap ret}$ denote the cardinality of the relevant and retrieved sets, then recall R, precision P and f-measure F are given by:

$$R = \frac{N_{rel \cap ret}}{N_{rel}}, \quad P = \frac{N_{rel \cap ret}}{N_{rel}}, \quad F = \frac{2RP}{R+P}$$

The above measures have been calculated and analysed considering different sets as relevant and retrieved as mentioned in Table 2.

Table 2. Recall (R), Precision(P) and f-measure for Book 1 and Book 2

	Step 1	**Step 2**	**Step 3**
Relevant set	***Generated using index***	***Generated using index***	***Extracted using ΔST (w,0.8)***
Retrieved set	***Extracted using ΔS_T (w,0.8)***	***Extracted using R(w,2)***	***Extracted using R(w,2)***

Book	1	2	1	2	1	2
N_rel	244	551	244	551	244	551
N_ret	244	551	244	551	244	551
N_(rel ∩ ret)	60	222	60	201	188	458
R	0.2459	0.4029	0.2459	0.3648	0.7705	0.8312
P	0.2459	0.4029	0.2459	0.3648	0.7705	0.8312
F	0.2459	0.4029	0.2459	0.3648	0.7705	0.8312

Table 3 and Table 4 displays the list of top 20 words extracted from Book 1 and Book 2 respectively using entropy-based keyword extraction methodology. Words have been categorized as relevant based on their existence in the respective indexes. It is interesting to note that some words like hate, adhikara, son, dasharatha, in Book 1 are indicated as irrelevant in Table 3 as per their unavailability in Index, but are essentially not irrelevant and seem relevant to Book 1. These words being quite familiar might not have been included in the index. Likewise, few terms like commercial, God, cost, wec (abbreviation of World Energy Council), Cambridge are classified as irrelevant in Table 4, but appears pertinent in accordance with comprehensive text in Book 2. This unveils the exactitude of the defined keyword extraction methodology using Rényi entropy.

Table 3. Top 20 words retrieved from Book 1 based on keyword extraction methodology using entropy

Book1	**Top 20 words using methodology based on Rényi entropy**				**Top 20 words using methodology based on Tsallis entropy**			
Rank	**Word**	**Frequency**	**R(w,2)**	**Relevance**	**Words**	**Frequency**	**ΔS_T (w,0.8)**	**Relevance**
1	prayer	293	4.452	Relevant	prayer	293	4.457	Relevant
2	congregational	27	3.967	Relevant	gita	227	3.943	Relevant
3	gita	227	3.915	Relevant	ramanama	94	3.880	Relevant
4	renunciation	31	3.872	Relevant	hinduism	228	3.826	Relevant
5	mantra	38	3.858	Relevant	renunciation	31	3.644	Relevant
6	ramanama	94	3.700	Relevant	mantra	38	3.603	Relevant
7	son	13	3.675	Irrelevant	cow	31	3.374	Relevant
8	hinduism	228	3.674	Relevant	ashram	58	3.250	Relevant
9	dasharatha	12	3.559	Irrelevant	hate	24	3.215	Irrelevant
10	maths	12	3.527	Relevant	you	425	3.213	Irrelevant
11	cow	31	3.470	Relevant	son	13	3.207	Irrelevant
12	vows	15	3.380	Relevant	rama	65	3.154	Relevant
13	ashram	58	3.379	Relevant	congregational	27	3.102	Relevant
14	ishopanishad	14	3.349	Relevant	dasharatha	12	3.078	Irrelevant
15	rama	65	3.346	Relevant	god	674	3.002	Relevant
16	hate	24	3.313	Irrelevant	maths	12	2.982	Relevant
17	sanatani	13	3.204	Relevant	students	36	2.936	Relevant
18	students	36	3.151	Relevant	he	589	2.898	Irrelevant
19	englishmen	9	3.136	Relevant	i	1161	2.873	Irrelevant
20	adhikara	9	3.115	Irrelevant	evil	69	2.822	Relevant

Book2	Top 20 words using methodology based on Rényi entropy				Top 20 words using methodology based on Rényi entropy			
Rank	**Word**	**Frequency**	**R(w,2)**	**Relevance**	**Words**	**Frequency**	**ΔS_T (w,0.8)**	**Relevance**
1	stabilization	5.048	84	Relevant	j	5.189	4256	Irrelevant
2	cost	4.937	185	Irrelevant	cambridge	4.910	204	Irrelevant
3	god	4.412	24	Irrelevant	energy	4.755	666	Relevant
4	ice	4.368	162	Relevant	stabilization	4.431	84	Relevant
5	renewable	4.282	105	Relevant	pp	4.261	125	Irrelevant
6	sustainable	4.119	61	Relevant	cost	4.187	185	Irrelevant
7	mitigation	4.052	68	Relevant	et	4.144	158	Irrelevant
8	commercial	4.041	25	Irrelevant	radiation	4.034	175	Relevant
9	cloud	4.023	56	Relevant	ice	4.025	162	Relevant
10	j	3.988	256	Irrelevant	cloud	3.921	56	Relevant
11	wec	3.947	33	Irrelevant	god	3.888	24	Irrelevant
12	cambridge	3.938	204	Irrelevant	al	3.857	165	Irrelevant
13	environment	3.902	111	Relevant	renewable	3.813	105	Relevant
14	runoff	3.892	21	Relevant	carbon	3.677	631	Relevant
15	costs	3.864	89	Irrelevant	press	3.549	140	Irrelevant
16	savings	3.847	33	Relevant	ppm	3.492	75	Irrelevant
17	energy	3.844	666	Relevant	ocean	3.476	215	Relevant
18	cores	3.843	20	Relevant	climate	3.469	1162	Relevant
19	radiative	3.822	86	Relevant	religious	3.459	28	Irrelevant
20	particles	3.800	39	Relevant	forcing	3.448	100	Relevant

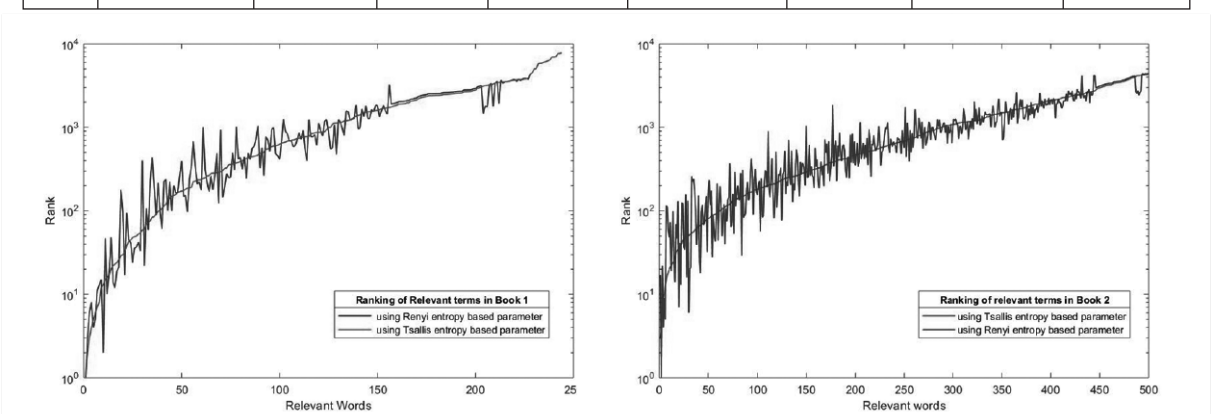

Figure 3. Shows the comparison of word rankings using Rényi and Tsallis entropy-based parameter for relevant words from Book 1 and Book 2.

Figure 3 shows the comparisons of word ranking using Rényi and Tsallis entropy-based parameter for relevant words from Book 1 and Book 2. As visible in Figure 3 the ranks of relevant words are closely related. The correlation coefficient between word ranking using parameter based on Rényi and Tsallis entropy are 0.9850 and 0.9779 for Book 1 and Book 2 respectively, which indicate a strong statistical association between the two variables.

Further, we have used the Tsallis relative entropy for extracting keywords from given text using the methodology based on relative entropy. Since Tsallis relative entropy is a generalization to Kullback & Leibler divergence, for entropic index q=1 , results correspond to Kullback & Leibler divergence. Figure 5 exhibits the Tsallis relative entropy values for various entropic index values ranging from 0.2 to 4.0 for top 30 words retrieved from Book 1 using methodology based on Tsallis relative entropy for entropic index q=3. The entropic index can be chosen based on the non-extensivity of the system. To improve the accuracy of keyword extraction process machine learning approach may be employed for deciding on the value of entropic index. Figure 6 shows the variation of Tsallis relative entropy values for varying values of entropic index q for top 30 words retrieved from Book 1 using methodology based on Tsallis relative entropy for entropic index q=3. It can be observed that higher values of entropic index q provides more leverage to words with higher frequencies.

String	Frequency	q=0.2	q=0.6	q=1	q=1.4	q=1.8	q=2.2	q=2.6	q=3	q=3.4	q=3.8	q=4
maths	12	1.042	1.504	2.331	3.893	6.986	13.354	26.902	56.491	122.517	272.411	409.276
hate	24	0.570	0.834	1.318	2.263	4.203	8.360	17.587	38.646	87.801	204.619	314.771
definitions	15	0.760	1.088	1.667	2.743	4.829	9.032	17.766	36.384	76.907	166.606	247.059
atheists	15	0.721	1.027	1.562	2.544	4.422	8.148	15.768	31.746	65.934	140.312	206.211
thee	11	0.872	1.224	1.822	2.883	4.836	8.553	15.827	30.395	60.128	121.783	174.579
englishmen	9	0.999	1.390	2.043	3.179	5.224	9.022	16.267	30.393	58.443	114.998	162.469
dedication	10	0.922	1.289	1.904	2.984	4.948	8.635	15.745	29.770	57.955	115.477	164.182
atheist	13	0.715	1.002	1.489	2.348	3.924	6.910	12.725	24.314	47.850	96.403	137.827
wealth	9	0.901	1.240	1.796	2.739	4.394	7.384	12.917	23.372	43.476	82.702	114.864
renunciation	31	0.399	0.579	0.903	1.521	2.760	5.344	10.917	23.263	51.217	115.625	175.052
conditions	14	0.667	0.935	1.391	2.198	3.681	6.500	12.005	23.008	45.424	91.809	131.472
cow	31	0.392	0.567	0.881	1.480	2.672	5.145	10.446	22.118	48.378	108.496	163.713
bihar	13	0.673	0.938	1.381	2.153	3.547	6.144	11.114	20.835	40.206	79.399	112.380
shlokas	8	0.863	1.167	1.648	2.434	3.755	6.031	10.041	17.238	30.361	54.616	73.743
blazer	7	0.938	1.258	1.759	2.564	3.895	6.147	10.037	16.878	29.092	51.186	68.341
ambedkar	7	0.938	1.258	1.758	2.564	3.894	6.145	10.033	16.869	29.074	51.151	68.293
habits	7	0.935	1.254	1.752	2.552	3.874	6.109	9.966	16.742	28.829	50.672	67.621
vegetarianism	7	0.934	1.253	1.750	2.550	3.870	6.100	9.950	16.711	28.769	50.555	67.456
enjoyment	7	0.933	1.251	1.747	2.545	3.861	6.084	9.920	16.654	28.659	50.340	67.155
deny	11	0.683	0.935	1.343	2.028	3.215	5.327	9.176	16.335	29.875	55.854	76.901
ahimsa	39	0.288	0.414	0.638	1.056	1.872	3.533	7.015	14.511	30.987	67.828	101.108
thy	7	0.873	1.163	1.609	2.316	3.465	5.375	8.610	14.186	23.939	41.210	54.417
creed	10	0.662	0.894	1.260	1.856	2.855	4.570	7.580	12.960	22.730	40.713	54.852
untouchability	33	0.306	0.435	0.659	1.067	1.842	3.366	6.456	12.876	26.484	55.806	81.609
quetta	9	0.688	0.922	1.284	1.865	2.820	4.428	7.188	12.012	20.572	35.957	47.849
congress	10	0.637	0.856	1.200	1.755	2.676	4.242	6.961	11.767	20.395	36.087	48.320
relief	6	0.889	1.163	1.573	2.201	3.184	4.752	7.297	11.490	18.492	30.319	39.057
renounce	6	0.881	1.151	1.555	2.173	3.138	4.674	7.160	11.248	18.057	29.528	37.987
puranas	7	0.790	1.041	1.421	2.010	2.946	4.463	6.967	11.164	18.301	30.579	39.773
temples	12	0.545	0.736	1.038	1.530	2.356	3.777	6.274	10.744	18.875	33.864	45.663

Figure 5. Frequency and Tsallis relative entropy values of top 30 words extracted from Book 1 using relative Tsallis entropy-based methodology for entropic index q=3

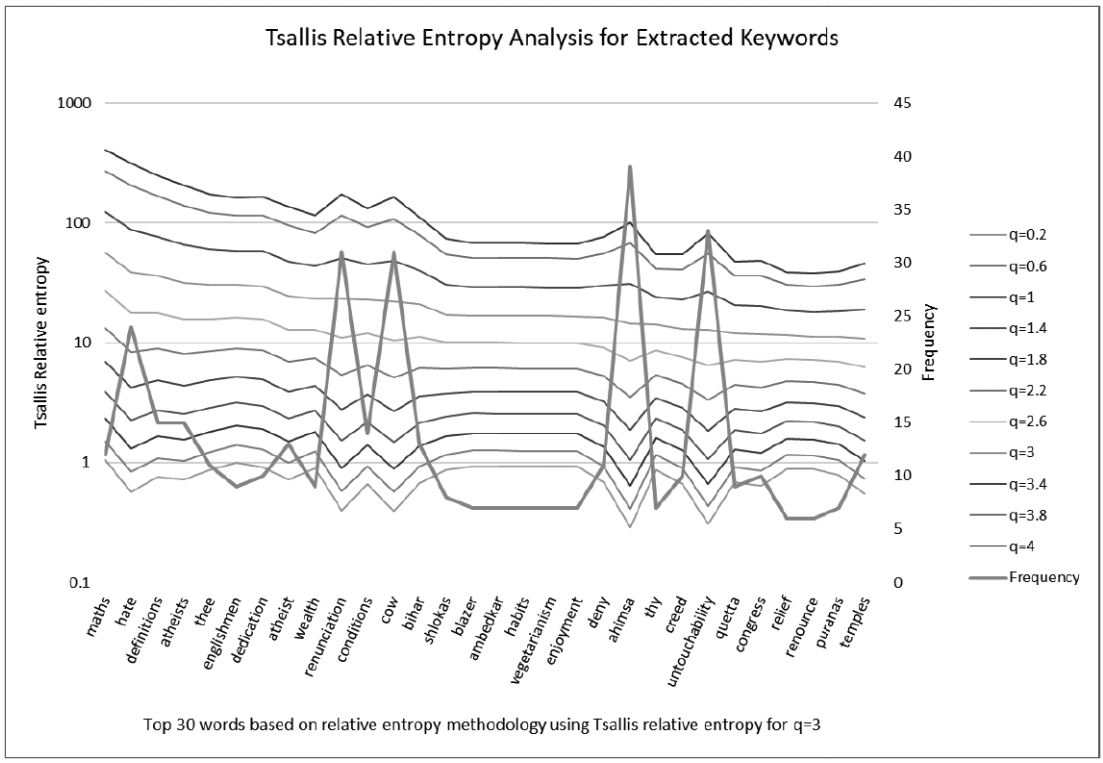

Figure 6. Shows the variation of Tsallis relative entropy values for varying values of entropic index q for top 30 words retrieved from Book 1 using methodology based on Tsallis relative entropy for entropic index q=3.

CONCLUSION

The proposed Rényi entropy and Tsallis relative entropy-based methodologies are successful in extricating pertinent words from the text. It is interesting to note that proposed methodology was effective in even extracting the terms that being quite common and familiar were missing in the index or glossary but were pertinent to the text. A reliable performance in keyword extraction was achieved using entropic index of q=2 for Rényi entropy and q=3 for Tsallis relative entropy-based methodology. Better results can be obtained by using lemmatization and stemming as pre-processing tools. Being a statistical method, proposed keyword extraction method are domain and language independent and therefore may have applications in organizing dynamic text collection, incremental clustering applications and many related applications. Also, for improvised performance machine learning approaches can be employed to evaluate the value of entropic index best suitable for the text.

REFERENCES

Bashkirov, A. G. 2006. Renyi entropy as a statistical entropy for complex systems, Theoretical and Mathematical Physics, 149(2), pp. 1559–1573.

Beliga, S. 2014. Keyword extraction: a review of methods and approaches, University of Rijeka, Department of Informatics, pp. 1–9.

Carpena, P. et al. 2009. Level statistics of words: Finding keywords in literary texts and symbolic sequences, Physical Review E, 79(3), p. 35102.

Chen, P.-I. and Lin, S.-J. 2010. Automatic keyword prediction using Google similarity distance, Expert Systems with Applications, 37(3), pp. 1928–1938.

Furuichi, Shigeru, Kenjiro Yanagi, and K. K. 2004. Fundamental properties of Tsallis relative entropy., Journal of Mathematical Physics, 45(12), pp. 4868–4877.

Gandhi, M. 1987. The essence of Hinduism. Navajivan Publishing House Ahmedabad.

Herrera, J. P. and Pury, P. A. 2008. Statistical keyword detection in literary corpora, The European Physical Journal B, 63(1), pp. 135–146.

Houghton, J. 2009. Global warming: the complete briefing. Cambridge university press.

Jamaati, M. and Mehri, A. 2018. Text mining by Tsallis entropy, Physica A: Statistical Mechanics and its Applications, 490, pp. 1368–1376.

Jizba, P. and Arimitsu, T. 2004. The world according to Rényi: thermodynamics of multifractal systems, Annals of Physics, 312(1), pp. 17–59.

Johal, R. S. and Tirnakli, U. 2004. Tsallis versus Renyi entropic form for systems with q-exponential behaviour: the case of dissipative maps, Physica A: Statistical Mechanics and its Applications, 331(3–4), pp. 487–496.

Kullback, S. and Leibler, R. A. 1951. On information and sufficiency, The annals of mathematical statistics, 22(1), pp. 79–86.

Bashkirov, A. G. 2006. Renyi entropy as a statistical entropy for complex systems, Theoretical and Mathematical Physics, 149(2), pp. 1559–1573.

Beliga, S. 2014. Keyword extraction: a review of methods and approaches, University of Rijeka, Department of Informatics, pp. 1–9.

Carpena, P. et al. 2009. Level statistics of words: Finding keywords in literary texts and symbolic sequences, Physical Review E, 79(3), p. 35102.

Chen, P.-I. and Lin, S.-J. 2010. Automatic keyword prediction using Google similarity distance, Expert Systems with Applications, 37(3), pp. 1928–1938.

Furuichi, Shigeru, Kenjiro Yanagi, and K. K. 2004. Fundamental properties of Tsallis relative entropy., Journal of Mathematical Physics, 45(12), pp. 4868–4877.

Gandhi, M. 1987. The essence of Hinduism. Navajivan Publishing House Ahmedabad.

Herrera, J. P. and Pury, P. A. 2008. Statistical keyword detection in literary corpora, The European Physical Journal B, 63(1), pp. 135–146.

Houghton, J. 2009. Global warming: the complete briefing. Cambridge university press.

Jamaati, M. and Mehri, A. 2018. Text mining by Tsallis entropy, Physica A: Statistical Mechanics and its Applications, 490, pp. 1368–1376.

Jizba, P. and Arimitsu, T. 2004. The world according to Rényi: thermodynamics of multifractal systems, Annals of Physics, 312(1), pp. 17–59.

Johal, R. S. and Tirnakli, U. 2004. Tsallis versus Renyi entropic form for systems with q-exponential behaviour: the case of dissipative maps, Physica A: Statistical Mechanics and its Applications, 331(3–4), pp. 487–496.

Kullback, S. and Leibler, R. A. 1951. On information and sufficiency, The annals of mathematical statistics, 22(1), pp. 79–86.

Luhn, H. P. 1958. The automatic creation of literature abstracts, IBM Journal of research and development, 2(2), pp. 159–165.

Matsuo, Y. and Ishizuka, M. 2004. Keyword extraction from a single document using word co-occurrence statistical information, International Journal on Artificial Intelligence Tools, 13(01), pp. 157–169.

Ortuno, M. et al. 2002. Keyword detection in natural languages and DNA, EPL (Europhysics Letters), 57(5), p. 759.

Rényi, A .1961. On measures of entropy and information, in Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics.

Rossi, R. G., Marcacini, R. M. and Rezende, S. O. 2014. Analysis of domain independent statistical keyword extraction methods for incremental clustering, Learning and Nonlinear Models, 12(1), pp. 17–37.

Shannon, C. E. 1948. A mathematical theory of communication, Bell system technical journal, 27(3), pp. 379–423.

Siddiqi, S. and Sharan, A. 2015. Keyword and keyphrase extraction techniques: a literature review, International Journal of Computer Applications, 109(2).

Tsallis, C. 1988. Possible generalization of Boltzmann-Gibbs statistics, Journal of statistical physics, 52(1–2), pp. 479–487.