


A Comparative Evaluation of Different Keyword Extraction Techniques

Raj Kishor Bisht, Department of Mathematics and Computing, Graphic Era Hill University, Dehradun, India

 <https://orcid.org/0000-0002-2691-8210>

ABSTRACT

Retrieving keywords in a text has been attracting researchers for a long time as it forms a base for many natural language applications like information retrieval, text summarization, document categorization, etc. A text is a collection of words that represent the theme of the text naturally and to bring the naturalism under certain rules is itself a challenging task. In the present paper, the author evaluate different spatial distribution-based keyword extraction methods available in the literature on three standard scientific texts. The author choose the first few high-frequency words for evaluation to reduce the complexity as all the methods are somehow based on frequency. The author find that the methods are not providing good results particularly in the case of the first few retrieved words. Thus, the author propose a new measure based on frequency, inverse document frequency, variance, and Tsallis entropy. Evaluation of different methods is done on the basis of precision, recall, and F-measure. Results show that the proposed method provides improved results.

KEYWORDS

Inverse Document Frequency, Keyword Extraction, Spatial Distribution, Term Frequency, Tsallis Entropy, Variance

1. INTRODUCTION

A text is a collection of words. A major part of the text is covered with function words that are necessary to make a sentence meaningful and grammatically correct. The author finds many other words in the text related to the theme of the topic. These words carry important information about the text and this information is useful in many tasks like information retrieval, natural language processing, text summarization, document categorization, etc. These words can be described as keywords. Thus, the automatic extraction of keywords is an important research direction in the field of text mining. The process of extracting keywords is to find the words that are sufficiently informative to represent the text. It is a challenging task to define a generalized rule for every text as different texts may have different linguistic features. To uncover these challenges, researchers have been making continuous efforts to establish the relationship among linguistic features, laws of Mathematics and Physics. The keyword extraction methods can be categorized under three broad categories: linguistics, machine learning, and statistical methods. In linguistics methods, the main focus is to observe syntactic, semantic aspects of words, morphological features, and linguistic relationships among words like synonym, hypernym, hyponym, etc. In machine learning methods, first, the learning algorithm is trained using a tagged training set and then its performance is evaluated through a tagged test set. The weighting of words in a text plays an important role in information retrieval. Initially, weighting

DOI: 10.4018/IJIRR.289573

This article published as an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>) which permits unrestricted use, distribution, and production in any medium, provided the author of the original work and original publication source are properly credited.

schemes are defined in the term of the frequency of words in a text. Term frequency (tf) and inverse document frequency (idf) were the weighting schemes firstly used for the weighting of words.

Luhn (1958) introduced an early idea of the importance of words in a text by analyzing Zip's analysis of the word's frequency in a text. Since then, a number of approaches for measuring the importance of words in a text appeared in the literature. The details of weighting schemes in information retrieval can be found in the books of Dominich (2008) and Manning and Schütze (1999). Earlier methods were based on the frequency of words in a text, later on, many other aspects were considered by different researchers. Turney (2000) performed a supervised learning approach for keyword extraction. The standard deviation of the distance between successive occurrences of a word is considered as a parameter to extract keywords by Ortuño et al. (2002). In their work, they found that the relevant words have greater standard deviation as their spatial distribution is more inhomogeneous in comparison to irrelevant words. Hulth (2003) suggested a keyword extraction method based on linguistics knowledge like syntactic features. A study on the fractal structure of the text can be found in Andres et al. (2010) and Andres et al. (2011). Yang et al. (2013) used Shannon's entropy difference between the intrinsic and extrinsic modes for determining the relevance of words in a text. Najafi and Darooneh (2015) used the concept of fractal dimension for keyword extraction. Jamaati and Namaati and Mehri (2018) used Tsallis entropy for ranking of the relevance of terms taking advantage of the spatial correlation length. Mehri et al. (2019) used distorted entropy for word ranking.

In addition to the spatial distribution based techniques, a number of other different approaches have been utilized for keyword extraction with applications in various fields. Florescu and Caragea (2017) suggested a graph-based unsupervised approach for extracting keyphrases from online texts. Wang and Zhang (2017) utilized the recurrent neural network method for extracting keywords from reviews of online products. Horita et al. (2016) used morphological analysis tools for the extraction of keywords for linking them to related Wikipedia articles. Lahiri et al. (2017) used supervised and unsupervised learning methods for extracting keywords from emails to explore the topics in the mails and avoiding excessive information. Thushara et al. (2019) provided a comparative study of unsupervised methods Position Rank, TextRank, Rapid automatic keyword extraction (RAKE) for keyphrase extraction. Ying et al. (2017) proposed a graph based method for keyphrase extraction considering important sentences in mind and word-sentence relationships. Rabby et al. (2018) proposed a domain independent tree based extraction method for keyphrases. Sterckx (2018) applied different supervised and unsupervised keyphrase extraction techniques including opinions about documents to increase efficiency. A detailed discussion on keyword extraction methods and issues like 'keyness' can be seen in the work of Firoozeh et al. (2020).

1.1 Standard Deviation Using Spatial Distribution

Ortuño et al. (2002) suggested the statistical analysis of the spatial distribution of words in a text to overcome the dependency on the frequency of words. They calculated the standard deviation of the distribution of successive occurrences of a word. The method can be summarized as follows: Let the number of words in a text is n . and let the occurrence of a word w . be denoted by a time sequence $L_w = \{t_1, \dots, t_f\}$. where f . is the frequency of the word w . in the text. Let $d_i = t_{i+1} - t_i$. denotes the distance between two successive appearances (waiting time) of the word w . at i^{th} . place.

The average distance is given by

$$\mu = \frac{1}{f-1} \sum_{i=1}^{f-1} d_i = \frac{t_f - t_1}{f-1} \quad (1)$$

and the standard deviation is given by

$$\sigma = \sqrt{\frac{1}{f-2} \sum_{i=1}^{f-1} (d_i - \mu)^2} \quad \cdot \quad (2)$$

To remove the dependence on frequency, they proposed $\hat{\sigma} = \sigma / \mu$.as the measure to calculate the relevance of a word.

Zhou and Slater (2003) suggested an improvement in the above model by adding boundary conditions to the time sequence, that is, $L_w = \{t_0, t_1, \dots, t_f, t_{f+1}\}$. where $t_0 = -1$.and $t_{f+1} = n$. The revised average distance is $\hat{\mu} = (n + 1) / (f + 1)$.and the revised standard deviation is

$$\sigma_2 = \sqrt{\frac{1}{f} \left[(t_1 + 1 - \hat{\mu})^2 + \sum_{i=1}^{f-1} (d_i - \hat{\mu})^2 + (n - t_f - \hat{\mu})^2 \right]} \quad (3)$$

and modified normalized standard-deviation is $\tilde{\sigma} = \sigma_2 / \hat{\mu}$.

They defined that arrival time t_i .is a cluster point if $d(t_i) < \hat{\mu}$. where $d(t_i)$.is the average separation at arrival t_i .given as

$$d(t_i) = \frac{w_{i+1} + w_i}{2} = \frac{t_{i+1} - t_{i-1}}{2} \quad (1 \leq i \leq f) \quad (4)$$

They further defined a local cluster index

$$\gamma(t_i) = \begin{cases} \frac{\hat{\mu} - d(t_i)}{\hat{\mu}} & \text{if } t_i \text{ is a cluster point} \\ 0 & \text{otherwise} \end{cases} \quad (5)$$

and suggested a new metric weighted measure of the number of cluster points given as

$$\Gamma(w) = \frac{1}{f} \sum_{i=1}^f \gamma(t_i) \quad (6)$$

A word w_1 . is more relevant than w_2 .if $\Gamma(w_1) > \Gamma(w_2)$.

1.2 Degree of Fractality (DF)

A text can be considered as an arrangement of words in one dimension array. The spatial pattern of occurrence of a word forms a fractal dimension Najafi and Darooneh (2015). In their work, they enhanced the concept of fractal dimension of words and defined degree of fractality of words for extracting content word in a text. Utilizing the box-counting method, the text is divided into boxes of different sizes and the number of filled boxes, the boxes in which a particular word under consideration appears, is counted. They further calculated the number of filled boxes in a shuffled version of the text

and defined the degree of fractality as the sum of ratios of the number of filled boxes in the original version and shuffled version for different box sizes.

Let s .is the box size and N .is the text length, then the number of boxes $N_s = N / s$. A box is said to be filled if the word under consideration appears in the box. For a word w . the number of filled boxes is denoted by $N_b(s, w)$,. which can be obtained from the given text and the box size. The number of filled boxes in the shuffled version of the text is defined as

$$N_b^{sh}(s, w) = \frac{f}{1 + \left(\frac{f-1}{N-1}\right)(s-1)} . \text{ where } f .\text{ is the frequency of the word } w . \text{ The degree of}$$

fractality is defined as

$$d_f(w) = \sum_s \log \left| \frac{N_b^{sh}(s, w)}{N_b(s, w)} \right| . \quad (7)$$

The higher degree of fractality shows the importance of a word in the text.

1.3 Jensen- Shannon Divergence

Jensen-Shannon Divergence (JSD) (Endres & Schindelin 2003, Österreicher & Vajda 2003) is the measure of (dis)similarity between two probability distributions. Mehri et al. (2015) utilized JSD for extracting keywords in a text. They defined a spatial distribution of words in a text by applying box-counting method. Let L be the length of the text. The text is partitioned into boxes of equal size. Let N_l .denotes the number of boxes with size l . $N_l = \lceil L/l \rceil$. where $\lceil x \rceil$.represents the integral part of x .and $n_l(w)$.denotes the number of boxes that contain the word w . The spatial probability $P(w)$ of a word w .is defined as $P(w) \propto n_l(w) / N_l$. They considered the spatial distribution of a word w .in two ways; one in the original text $P(w) = \{p_1(w), \dots, p_l(w)\}$.and another in the shuffled version of the text $Q(w) = \{q_1(w), \dots, q_l(w)\}$.when the text is partitioned with box size l .The value of $q_l(w)$.is calculated theoretically with the help of frequency $f(w)$.of the word w .in the text without getting a randomly shuffled version of the text as

$$q_l(w) = \begin{cases} \frac{f(w)}{N_l} & \text{if } f(w) < N_l . \\ 1 & \text{otherwise} \end{cases} . \quad (8)$$

Finally JSD is defined as

$$JSD(P(w) \parallel Q(w)) = \frac{1}{2} \sum_{l=2}^{L/2} p_l(w) \log \left(\frac{p_l(w)}{m_l(w)} \right) + \frac{1}{2} \sum_{l=2}^{L/2} q_l(w) \log \left(\frac{q_l(w)}{m_l(w)} \right) , . \quad (9)$$

$$\text{where } m_i(w) = \frac{p_i(w) + q_i(w)}{2}.$$

1.4 Tsallis Entropy

Entropy is the measure of disorder or uncertainty in the physical system. Shannon (1948) defined a formal measure of entropy known as Shannon entropy. Further, Renyi (1970) and Tsallis et al. (1998) proposed generalized entropy. Maszczyk and Duch (2008) compared the three entropies Shannon, Renyi and Tsallis used in decision trees. Let the probability distribution $P = \{p_1, p_2, \dots, p_n\}$ contains the occurrence probabilities of all microstates. The non-additive entropy established by Tsallis et al. (1998) is given by

$$S_q(P) = \frac{1}{q-1} \left(1 - \sum_{i=1}^n p_i^q \right) \quad (10)$$

where q . a real number, is non-extensivity parameter for the physical system that represents long-range interactions. Tsallis entropy attains it maximum value for homogeneous distribution, that is $p_i = 1/n$. and the maximum Tsallis entropy S_q^{Max} . is given by

$$S_q^{Max}(P) = \frac{1}{q-1} (1 - n^{1-q}) \quad (11)$$

Jamaati and Mehri (2018) suggested the use of Tsallis entropy for extracting keywords. They considered the spatial probability distribution of a word w .that appears n_w .times in the text as $P(w) = \{p_1, p_2, \dots, p_{n_w}\}$. The probability $p_i(w)$.is calculated as $p_i(w) = d_i(w) / N$. where $d_i(w)$ is the distance between i^{th} .and $(i+1)^{th}$.occurrence of the word w .and N . is the length of the text. They found the difference

$$D = |S_q^{Max}(w) - S_q(w)| \quad (12)$$

between maximum Tsallis entropy and the actual Tsallis entropy of a word. They argued that the higher difference shows more relevance.

1.5 Distorted Entropy

Mehri et al. (2019) suggested the application of distorted entropy in word ranking. In their work, they defined distorted probability and distorted entropy as follows: Let Ω .be a non empty set and \mathcal{F} .be a collection of subsets of Ω .A set function v .on (Ω, \mathcal{F}) is called a distorted probability if there exists a probability measure \mathbb{P} .on (Ω, \mathcal{F}) .and a non decreasing function $f : [0,1] \rightarrow [0,1], f(0) = 0$.and $f(1) = 1$.such that

$$v(A) = f(\mathbb{P}(A)) \quad \forall A \in \mathcal{F} \quad (13)$$

The distorted entropy of v .is defined as

$$E(v) = \sum_{i=1}^m v_i \log(v_i^{-1}), \quad (14)$$

where $v_i = f o p_i$.such that there are total m .microstates and p_i .is the probability of i^{th} .microstate.

They used three different functions $f_1(x) = \frac{1 - e^{-\alpha x}}{1 - e^{-\alpha}}$, $\alpha \neq 0$. $f_2(x) = x^\alpha$, $\alpha > 0$.and

$f_3(x) = \frac{\sinh \alpha x}{\sinh x}$, $\alpha > 0$. and hence three different forms of distorted entropy are defined as

$$DE_1(\alpha) = E_1^\alpha(\mathbb{P}) = -\sum_{i=1}^m \frac{1 - e^{-\alpha p_i}}{1 - e^{-\alpha}} \log\left(\frac{1 - e^{-\alpha p_i}}{1 - e^{-\alpha}}\right).$$

$$DE_2(\alpha) = E_2^\alpha(\mathbb{P}) = -\alpha \sum_{i=1}^m p_i^\alpha \log(p_i). \text{ and}$$

$$DE_3(\alpha) = E_3^\alpha(\mathbb{P}) = -\alpha \sum_{i=1}^m \frac{\sinh \alpha p_i}{\sinh \alpha} \log\left(\frac{\sinh \alpha p_i}{\sinh \alpha}\right). \quad (15)$$

In their calculations, they observed that the performance of the methods depends on distorted parameter α .the F-measure reached its global maximum at $\alpha(DE_1) = 4.5$, $\alpha(DE_2) = 0.9$.and $\alpha(DE_3) = 0.1$. Higher values of distorted entropy show the presence of keywords.

The present work adopts a little different approach for evaluating different methods and extracting keywords in a text. The previous methods consider all the words in the text, the author considers the first few high-frequency words to reduce the complexity of the methods. The author first evaluate different keyword extraction methods for different standard texts and check the performances of these methods using precision, recall, and F-measure by comparing the results with the given list of keywords. The author then proposes a new method for keyword extraction based on frequency, inverse document frequency, variance and Tsallis entropy to define the new measure. Function words/ stop words are excluded and singular and plural forms of a word given in keyword are also considered as keywords.

The organization of the paper is as follows: In Section 2, different methods of keyword extraction for three standard texts are evaluated. In Section 3, a relevance measure is proposed based on frequency, inverse document frequency, variance and Tsallis entropy and a comparative evaluation of different methods is conducted Finally, Section 4 concludes the works.

2. EVALUATION OF DIFFERENT METHODS

In this section, the author evaluates the methods of keyword extraction discussed in section 1. For this purpose, three standard scientific texts are chosen at random from <https://github.com/zelandiya/keyword-extraction-datasets> (text no. 11415952 (T1), 10984465 (T2), 10984466(T3)). Keywords for each text are given. The texts are converted into plain texts by removing all punctuations marks, numbers, etc. All the words are converted into lower case and function/stop words like articles, propositions, helping verbs, etc are excluded. One or two-letter words are not considered. The words are arranged according to their frequency (f) from highest to lowest and rank is assigned in the same order. The author considers the first 15 top-ranked words having a frequency greater than 2. Text 1, Text 2, and Text 3 have 17, 20 and 28 such words respectively. Table 1 shows the different

15 top-ranked words having a frequency greater than 2 in T1, T2, and T3. The author then applies the various methods discussed in section 1 to the top ranked words in each text. For evaluation purpose, the basic measures Precision, Recall and F measure (Manning and Schütze 1999) are used. Let Ret and Rel denote the set of retrieved and relevant words for a text, then the three measures are defined as follows:

$$\text{Precision (P)} = \frac{|Rel \cap Ret|}{|Ret|} . \text{ Recall (R)} = \frac{|Rel \cap Ret|}{|Rel|} . \text{ and } F = \frac{2PR}{P + R} .$$

F-measure combines Precision and Recall. It is the harmonic mean of the two measures, thus F-measure can be used to compare different methods.

Table 2 shows the top 10 ranked words in three texts according to the values of $\hat{\sigma}$. The number of keywords are 24, 11, and 28 in Text1, Text 2 and Text 3 respectively. From Table 2, if we observe first five words, then we see that only 1 word is extracted as keyword in Text 1 and no words are extracted from Text 2 and Text 3, thus providing 20%, 0%, 0% precision for Text1, Text 2, and Text 3 respectively and 4%, 0%, 0% recall for Text 1, Text 2, and Text 3 respectively. If we consider Top 10 words, then one word is extracted as keyword in every text. Thus, providing 10%, 10%, 10% precision for Text1, Text 2, and Text 3 respectively and 4%, 9%, 4% recall for Text 1, Text 2, and Text 3 respectively. Similarly we can calculate the Precision, Recall and F-measure for each of the method.

The values of Γ for different words in three texts are given in Table 3. Table 4 shows the top 10 ranked words in three texts according to the values of degree of fractality (DF). Table 5 shows the top 10 ranked words in three texts according to the values of Jensen- Shannon Divergence (JSD). Table 6 and 7 show the top 10 ranked words in three texts according to the values of Tsallis entropy (TE) and distorted entropy (DE). Keywords are highlighted in these tables.

The values of Precision, Recall and F-measure for different methods and different texts are shown in Table 8 and Table 9 based on the top five and top 10 retrieved words respectively. From Table 8, we observe that only TE and DE values for Text 1 are non zero, showing the variability in results for different texts and low performance of Γ , DF, and JSD. From Table 9, we observe that the performances of TE and DE are higher than other methods but these measures do not provide good results for the chosen standard scientific texts. In general, the performances of these measures is not good in the case of selecting words from the top five words, however, if we consider top 10 words, then a slight improvement is shown yet there may be a chance of further improvement.

In order to search a new method that can provide good results for scientific texts also based on the frequency or spatial distribution, the author considers various different measures and different possible combinations of these methods. Since every measure has its own characteristic, thus if we can combine different approaches, then we can accumulate different characteristic and the method can perform well. After going through various experiments, the author finds that the measures ‘frequency’, ‘inverse document frequency’, ‘variance’ and ‘Tsallis entropy’ work very well if applied in a certain way as a combination of these measures provides a holistic information for being a keyword. In the next section, the author proposes a method based on these measures and compare the results of the proposed method with the other methods.

3. RELEVANCE MEASURE

In this section, the author defines a relevance measure to find the relevance of the words in a text. After going through various measures, the author considers frequency, Idf and variance (Var) and Tsallis entropy to define the new measure through a different approach of applying these measures.

Table 1. Top ranked words in three texts

Text 1 (11415952)								
Word	f	Rank	Word	f	Rank	Word	f	Rank
angina	190	1	disease	61	7	study	29	12
Men	97	2	ischemic	44	8	possible	27	13
Pain	87	3	risk	42	9	infarction	26	14
heart	73	4	questionnaire	38	10	table	26	14
chest	72	5	event	30	11	myocardial	24	15
percent	62	6	symptoms	29				
Text 2(10984465)								
Word	f	Rank	Word	f	Rank	Word	f	Rank
software	56	1	speech	36	7	study	21	13
error	54	2	percent	34	8	word	20	14
recognition	53	3	vocabulary	33	9	three	20	14
medical	45	4	ibm	28	10	participants	17	15
Rate	40	5	each	25	11	general	17	15
dictation	37	6	dragon	22	12	scoring	17	15
package	37	6	number	22				
Text 3(10984466)								
Word	f	Rank	Word	f	Rank	Word	f	Rank
computer	101	1	perceived	16	10	specific	10	14
students	40	2	opinions	16	10	examinees	9	15
experience	35	3	study	15	11	administration	9	15
medical	33	4	between	15	11	feel	9	15
usmle	31	5	content	12	12	year	9	15
percent	26	6	paper	12	12	literature	9	15
based	21	7	scale	12	12	expertise	9	15
preparedness	20	8	school	11	13	used	9	15
Cbt	19	9	gender	11				
variables	19	9	tests	10				

The author divides the text into small equal parts called boxes. Weighting of words through mean, inverse document frequency, and variance has almost no effect on the length of boxes (Bisht and Dhimi 2008). The author chooses box length 25 for calculations. The last box may not contain exactly 25 words. The author checks the frequency of a word in each box and count the number of boxes that contain the word. Let f_w denotes the frequency of the word w in a text, N denotes the total number of boxes, N_w denotes the number of boxes in which the word w appears and L denotes the text length. Then, $N = \frac{L}{25}$ where $\frac{L}{25}$ is the ceiling function. Mean, inverse document frequency and variance are calculated as follows:

Table 2. Value of $\hat{\sigma}$ for ten top ranked words in three texts

Sr. No.	Text 1		Text 2		Text 3	
	Word	$\hat{\sigma}$	Word	$\hat{\sigma}$	Word	$\hat{\sigma}$
1	percent	2.21	percent	2.46	percent	2.94
2	symptoms	2.17	scoring	2.23	scale	2.45
3	methods	2.11	participants	2.00	tests	2.15
4	abstract	1.91	rate	1.91	paper	2.04
5	Chest	1.76	ibm	1.71	preparedness	1.81
6	Table	1.76	medical	1.59	examinees	1.76
7	subsequent	1.69	word	1.59	variables	1.51
8	British	1.62	vocabulary	1.56	students	1.49
9	variability	1.62	dragon	1.55	based	1.41
10	Heart	1.61	each	1.40	used	1.39

Table 3. Value of Γ for ten top ranked words in three texts

Sr. No.	Text 1		Text 2		Text 3	
	Word	Γ	Word	Γ	Word	Γ
1	percent	0.57	percent	0.59	percent	0.73
2	Table	0.55	scoring	0.57	examinees	0.68
3	abstract	0.51	participants	0.53	tests	0.58
4	event	0.49	word	0.45	paper	0.57
5	methods	0.48	ibm	0.45	scale	0.56
6	possible	0.46	vocabulary	0.44	feel	0.51
7	myocardial	0.45	dragon	0.44	preparedness	0.48
8	Risk	0.44	rate	0.44	based	0.47
9	chest	0.44	number	0.43	variables	0.40
10	variability	0.43	medical	0.40	students	0.37

$$M = \frac{f_w}{N} \quad Idf = \log \left(\frac{N}{N_w} \right), \quad Var = \frac{\sum_{i=1}^N (f_i - M)^2}{N}, \quad (16)$$

where f_i is the frequency of word w in i^{th} box.

The author calculates these values for different high-frequency words in a text. Since these measures have different ranges, the author normalizes mean and variance values to get the values in same range. For normalization, each value of a measure is divided by the maximum value of the measure in the list of different words so that we get the normalized values in the interval [0, 1]. Let $\hat{M}(w)$ denotes the normalized value of mean. Since Idf is already in the range [0, 1], thus we need

Table 4. Value of df or ten top ranked words in three texts

Sr. No.	Text 1		Text 2		Text 3	
	Word	df	Word	df	Word	df
1	Table	75.00	participants	111.50	examinees	400.96
2	percent	44.63	percent	96.85	percent	217.72
3	methods	40.65	ibm	62.29	tests	120.23
4	persistent	-7.03	scoring	48.95	paper	97.47
5	symptoms	-42.21	dragon	47.67	feel	87.48
6	subsequent	-47.15	number	32.35	scale	81.23
7	Event	-47.96	vocabulary	6.88	variables	18.03
8	persistence	-91.21	general	-3.00	gender	0.72
9	myocardial	-105.83	rate	-12.07	expertise	-9.48
10	ischemic	-111.59	word	-42.21	administration	-64.79

Table 5. Value of JSD measure for ten top ranked words in three texts

Sr. No.	Text 1		Text 2		Text 3	
	Word	JSD	Word	JSD	Word	JSD
1	methods	44.51	participants	38.16	examinees	93.22
2	Table	42.97	scoring	32.79	percent	47.54
3	subsequent	34.16	percent	28.78	tests	37.97
4	persistent	33.01	dragon	23.62	scale	34.12
5	percent	29.75	ibm	23.34	feel	33.96
6	symptoms	27.85	number	21.23	paper	33.43
7	Event	26.30	general	19.12	expertise	19.05
8	persistence	26.21	word	16.92	gender	17.38
9	variability	24.40	vocabulary	15.90	variables	16.76
10	abstract	23.36	rate	13.30	used	13.75

not normalize Idf . Hence, $\widehat{Idf}(w) = Idf$. Inverse document frequency provides zero weight to a word if it appears in all boxes and the highest weight to the word which appears in only one box, while mean provides the highest weight to the word with highest frequency, thus we define a new measure called frequency weight (f_w) of a word w which is the harmonic mean of $\hat{M}(w)$ and $\widehat{Idf}(w)$.

Functions words are spread across the document in a symmetric way as they are always accompanied with keywords, thus function words have low variability in comparison to keywords. The author takes variance of a word w $Var(w)$ as a second measure. In order to make the range of variance values in the interval $[0,1]$, variance is also normalized using the previous method. Let $\widehat{Var}(w)$ denotes the normalized variance. The author defines the normalized variance as the variability

Table 6. Value of TE measure or ten top ranked words in three texts

Sr. No.	Text 1		Text 2		Text 3	
	Word	TE	Word	TE	Word	TE
1	Percent	2.44	percent	2.25	percent	2.82
2	Table	1.74	participants	1.69	examinees	1.98
3	methods	1.73	scoring	1.69	scale	1.57
4	Chest	1.61	vocabulary	1.50	preparedness	1.44
5	abstract	1.48	rate	1.41	paper	1.32
6	possible	1.43	dictation	1.39	tests	1.30
7	Event	1.35	word	1.37	feel	1.19
8	Pain	1.35	ibm	1.35	based	1.09
9	Heart	1.33	medical	1.35	variables	1.08
10	questionnaire	1.32	number	1.33	students	1.03

Table 7. Value of DE measure or ten top ranked words in three texts

Sr. No.	Text 1		Text 2		Text 3	
	Word	DE	Word	DE	Word	DE
1	symptoms	0.012	participants	0.029	feel	0.049
2	subsequent	0.012	scoring	0.022	scale	0.040
3	british	0.012	number	0.014	tests	0.036
4	abstract	0.011	dragon	0.007	paper	0.033
5	Pain	0.010	error	0.006	percent	0.031
6	Chest	0.010	general	0.006	examinees	0.024
7	Table	0.010	medical	0.006	used	0.020
8	methods	0.010	software	0.006	expertise	0.018
9	angina	0.009	ibm	0.005	literature	0.014
10	Heart	0.008	recognition	0.005	preparedness	0.014

Table 8. Precision, Recall and F- measure (in %) of different methods in first 5 words

Method	Text 1			Text 2			Text 3		
	P	R	F	P	R	F	P	R	F
$\hat{\sigma}$	20	4	7	0	0	0	0	0	0
Γ	0	0	0	0	0	0	0	0	0
DF	0	0	0	0	0	0	0	0	0
JSD	0	0	0	0	0	0	0	0	0
TE	20	4	7	0	0	0	0	0	0
DE	20	4	7	0	0	0	0	0	0

Table 9. Precision, Recall and F- measure (in %) of different methods in first 10 words

Method	Text 1			Text 2			Text 3		
	P	R	F	P	R	F	P	R	F
$\hat{\sigma}$	10	4	6	10	9	10	10	4	5
Γ	30	13	18	10	9	10	10	4	5
DF	20	8	12	0	0	0	0	0	0
JSD	0	0	0	0	0	0	0	0	0
TE	30	13	18	10	9	10	10	4	5
DE	30	13	18	20	18	19	0	0	0

weight (v_w) of a word w . The author assumes that both the measures are equal important for calculating the relevance of a word, thus weight (T) of a word w is defined as the sum of frequency weight and variability weight, that is, $T(w) = f_w + v_w$. Since entropy contains the amount of information a variable contains, Tsallis entropy difference (D) as given by equation (12) is used to know the amount of information a word contained for being a keyword. The author proposes the relevance of a word as the weight raised to the power of Tsallis entropy difference, that is,

$$Rel(w) = T^D \quad (17)$$

as it gives more strength to weight and provides the total amount of information a word contained for being relevant.

The method can be summarized as follows:

1. Arrange the words in the text in descending order of their frequencies and assign rank accordingly.
2. Select top 15 ranked words for processing.
3. Calculate text length L and number of boxes $N = \frac{L}{25}$.
4. For a word w find the frequency f_i in each i^{th} box and the number of boxes N_w in which the word w appears
5. Calculate $M(w) = \frac{f_w}{N}$, $Idf(w) = \log\left(\frac{N}{N_w}\right)$ and $Var(w) = \frac{\sum_{i=1}^N (f_i - M)^2}{N}$.
6. Calculate $\hat{M}(w) = \frac{M(w)}{Max(M(w_i))}$, $\widehat{Idf}(w) = Idf(w)$ and $\widehat{Var}(w) = \frac{Var(w)}{MaxVar(w_i)}$.
7. Calculate $f_w = \frac{2 \cdot \hat{M}(w) \cdot \widehat{Idf}(w)}{(\hat{M}(w) + \widehat{Idf}(w))}$ and $v_w = \widehat{Var}(w)$.
8. Calculate $T(w) = f_w + v_w$.
9. Calculate D .
10. Calculate $Rel(w) = T^D$.

11. Arrange the words in decreasing order of their *Rel* .values and assign ranks.

The proposed method is the amalgamation of different measures frequency weight, variability weight and Entropy weight. Since frequency weight and inverse frequency weight are inversely proportional, thus the author has defined the harmonic mean of $\hat{M}(w)$.and $\hat{Idf}(w)$.to get frequency weight which gives the weight due to frequency. Variability is another important aspect to measure relevance for being keyword, thus the author has taken variability weight as another weight. The sum of these two weights provides complete information regarding importance due to frequency and variability. Further, as Entropy provides the information contained by a variable for being a content word, thus raising the weight due to frequency and variability up to the powers of entropy gives additional strength to describe total information contained by a word and hence the relevance of a word. Since previous methods are based on some particular aspects and proposed method considers different aspects, thus provides a better measure than the previous methods. The proposed method of extracting keywords is applied to the three scientific texts. Table 10 shows the values of the relevance measure for the three texts.

Table 10. Value of Relevance or ten top ranked words in three texts

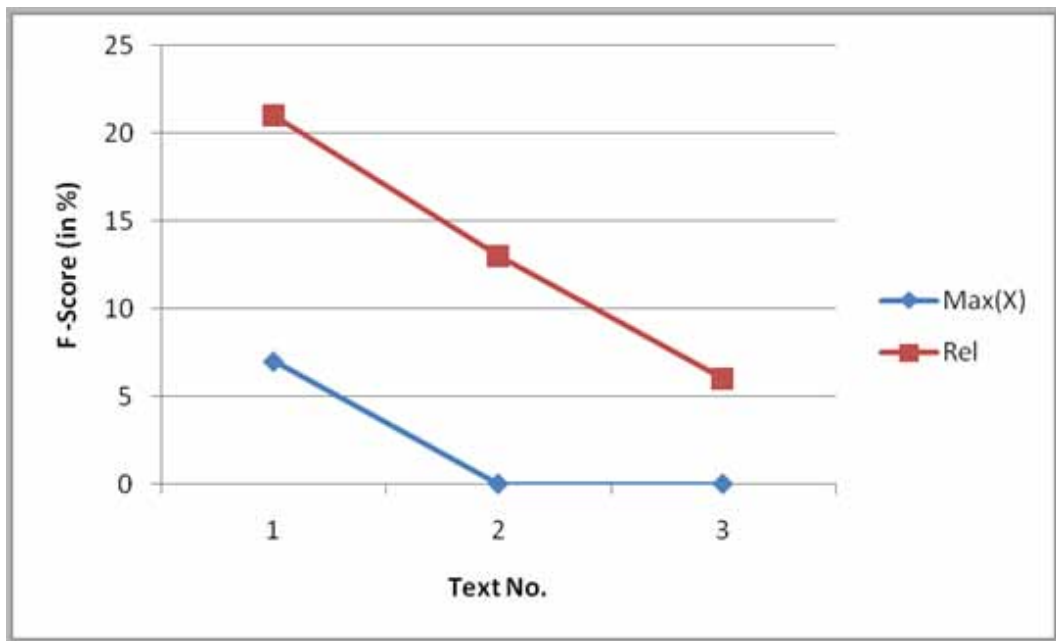
Sr. No.	Text 1		Text 2		Text 3	
	Word	Rel	Word	Rel	Word	Rel
1	angina	1.14	percent	2.71	percent	2.47
2	percent	1.08	error	1.13	computer	0.94
3	chest	0.90	rate	1.12	specific	0.62
4	pain	0.90	vocabulary	1.08	content	0.59
5	men	0.64	medical	1.02	experience	0.57
6	heart	0.59	software	0.96	administration	0.55
7	reported	0.49	speech	0.85	students	0.52
8	disease	0.48	word	0.84	between	0.50
9	risk	0.46	each	0.74	usmle	0.49
10	ischemic	0.44	package	0.73	based	0.48

The comparison of the results of the proposed method with previous methods shows that a significant improvement is achieved. Table 11 shows the comparison of the values of Precision, Recall and F measure between the proposed method and the maximum achieved values of any of the previous methods in the set $X=\{ DF, JSD, TE, DE\}$ denoted by $Max(X)$. Figure 1 and Figure 2 show the values of F measures for the proposed method and the maximum of any previous method. It is clear that the proposed method is superior to the other measures. The results shows that the proposed method provides improved results for each text, thus it provides consistent results also. Thus, the proposed method works well in comparison to previous methods particularly in retrieving first few top ranked words.

Table 11. Comparison of results between proposed and best of other methods

Up to top five words									
Method	Text 1			Text 2			Text 3		
	P	R	F	P	R	F	P	R	F
Max(X)	20	4	7	0	0	0	0	0	0
Rel	60	13	21	20	9	13	20	4	6
Up to top 10 words									
Method	Text 1			Text 2			Text 3		
	P	R	F	P	R	F	P	R	F
Max(X)	30	13	18	20	18	19	10	4	5
Rel	50	21	29	30	27	29	20	7	11

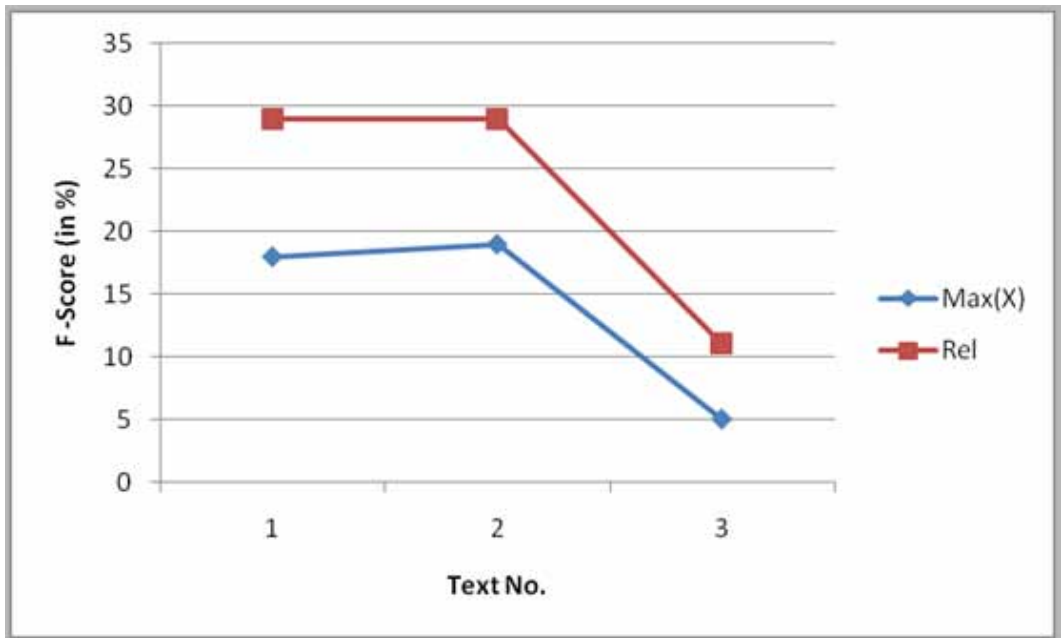
Figure 1. Comparison of F-Score (For top 5 words)



4. CONCLUSION

In the present works, first, different keyword extraction techniques are evaluated on certain standard scientific texts and then a new keyword extraction technique is proposed based on some previous measures. The present work summarises the various keyword extraction techniques available in the literature, thus it provides a detailed description of various developments in the direction of keyword extraction. Experimental results proved that the previous methods were not working well in terms of finding keywords in the first few top-ranked words. Then, a relevance measure is defined based on frequency, inverse document frequency, variance, and Tsallis entropy. The proposed measure contains holistic information about the relevance of a word. The experimental results proved the superiority

Figure 2. Comparison of F-Score (For top 10 words)



of the proposed method over the existing methods. The previous methods considered all the words in the text while the author considered only a few high-frequency words, thus the complexity is reduced up to a great extent which is an additional advantage of the proposed measure.

In spite of the improvement in previous methods, keyword extraction itself is a challenging task, for example, in some of the texts keywords may not be a part of the text. In this situation, frequency or spatial distribution methods need some additional criteria to apply and this may be a future direction of work.

In spite of a number of researchers worked in the direction of keyword extraction but still, no method is perfect. For different texts, the result of a particular method may vary. A number of different factors may work for a particular kind of text. Firoozeh et al. (2020) discussed such issues in keyword extraction. Thus, it is interesting to know the results of different methods for some particular kinds of texts. This motivated the author to conduct research work in this direction. Here the objective is to provide a comparative evaluation of different spatial distribution based keyword extraction methods, particularly for scientific text. Some standard texts are available on the Web with the list of keywords. Thus, different methods can be evaluated for their performances with respect to the scientific texts. First, the author provides a brief discussion of each of these methods. The various existing methods are as follows:

REFERENCES

- Andres, J. (2010). On a conjecture about the fractal structure of language. *Journal of Quantitative Linguistics*, 17(2), 101–122. doi:10.1080/09296171003643189
- Andres, J., Benešová, M., Kubá, L., & Vrbková, J. (2012). Methodological Note on the Fractal Analysis of Texts. *Journal of Quantitative Linguistics*, 19(1), 1–31. doi:10.1080/09296174.2011.608604
- Bisht, R. K., & Dhimi, H. S. (2008). On some properties of content words in a document. *Proc. 6th Annual conference of Information Science and Technology Management*, 51, 1–19.
- Dominich, S. (2008). *The Modern Algebra of Information Retrieval (Information Retrieval Series)*. Springer-Verlag.
- Endres, D. M., & Schindelin, J. E. (2003). A new metric for probability distributions. *IEEE Transactions on Information Theory*, 49(7), 1858–1860. doi:10.1109/TIT.2003.813506
- Firoozeh, N., Nazarenko, A., Alizon, F., & Daille, B. (2020). Keyword extraction: Issues and methods. *Natural Language Engineering*, 26(3), 259–291. doi:10.1017/S1351324919000457
- Florescu, C., & Caragea, C. (2017). A Position-Biased PageRank Algorithm for Keyphrase Extraction. *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, 4923–4924.
- Horita, K., Kimura, F., & Maeda, A. (2016). Automatic Keyword Extraction for Wikification of East Asian Language Documents. *International Journal of Computer Theory and Engineering*, 8(1), 32–35. doi:10.7763/IJCTE.2016.V8.1015
- Hulth, A. (2003). Improved automatic keyword extraction given more linguistic knowledge. *Proceedings of the Conference Empirical Methods in Natural Language Processing*, 216–223. doi:10.3115/1119355.1119383
- Jamaati, M., & Mehri, A. (2018). Text mining by Tsallis entropy. *Physica A*, 490, 1368–1376. doi:10.1016/j.physa.2017.09.020
- Lahiri, S., Mihalcea, R., & Lai, P.-H. (2017). Keyword Extraction from Emails. *Natural Language Engineering*, 23(2), 295–317. doi:10.1017/S1351324916000231
- Luhn, H. P. (1958). The automatic creation of literature abstracts. *IBM Journal of Research and Development*, 2(2), 159–165. doi:10.1147/rd.22.0159
- Manning, C. D., & Schütze. (1999). *Foundations of Statistical Natural Language Processing*. The MIT Press.
- Maszczyk, T., & Duch, W. (2008). Comparison of Shannon, Renyi and Tsallis Entropy used in Decision Trees. *Lecture Notes in Computer Science*, 5097, 643–651. doi:10.1007/978-3-540-69731-2_62
- Mehri, A., Agahi, H., & Mehri, D. H. (2019). A novel word ranking method based on distorted entropy. *Physica A*, 521, 484–492. doi:10.1016/j.physa.2019.01.080
- Mehri, A., Jamaati, M., & Mehri, H. (2015). Word ranking in a single document by Jensen–Shannon divergence. *Physics Letters. [Part A]*, 379(28–29), 1627–1632. doi:10.1016/j.physleta.2015.04.030
- Najafi, E., & Darooneh, A. H. (2015). The Fractal Patterns of Words in a Text: A Method for Automatic Keyword Extraction. *PLoS One*, 10(6), e0130617. doi:10.1371/journal.pone.0130617 PMID:26091207
- Ortuño, M., Carpena, P., Bernaola-Galván, P., Muñoz, E., & Somoza, A. M. (2002). Keyword detection in natural languages and DNA. *Europhysics Letters*, 57(5), 759–764. doi:10.1209/epl/i2002-00528-3
- Österreicher, F., & Vajda, I. (2003). A new class of metric divergences on probability spaces and its statistical applications. *Annals of the Institute of Statistical Mathematics*, 55, 639–653. doi:10.1007/BF02517812
- Rabby, G., Azad, S., Mahmud, M., Zamli, K. Z., & Rahman, M. M. (2018). A Flexible Keyphrase Extraction Technique for Academic Literature. *Procedia Computer Science*, 135, 553–563. doi:10.1016/j.procs.2018.08.208
- Renyi, A. (1970). *Probability Theory*. North-Holland.
- Shannon, C. E. (1948). The Mathematical Theory of Communication. *The Bell System Technical Journal*, 27(3), 379–423. doi:10.1002/j.1538-7305.1948.tb01338.x

- Sterckx, L., Demeester, T., Deleu, J., & Develder, C. (2018). Creation and evaluation of large keyphrase extraction collections with multiple opinions. *Language Resources and Evaluation*, 52(2), 503–532. doi:10.1007/s10579-017-9395-6
- Thushara, M. G., Mownika, T., & Mangamuru, R. (2019). A Comparative Study on different Keyword Extraction Algorithms. *Proceedings of 3rd International Conference on Computing Methodologies and Communication (ICCMC)*, 969-973. doi:10.1109/ICCMC.2019.8819630
- Tsallis, C., Mendes, R. S., & Plastino, A. R. (1998). The role of constraints within generalized nonextensive statistics. *Physica A*, 261(3-4), 534–554. doi:10.1016/S0378-4371(98)00437-3
- Turney, P. D. (2000). Learning algorithms for keyphrase extraction. *Information Retrieval*, 2(4), 303–336. doi:10.1023/A:1009976227802
- Wang, Y., & Zhang, J. (2017). Keyword extraction from online product reviews based on bi-directional LSTM recurrent neural network. *Proceedings of IEEE International Conference on Industrial Engineering and Engineering Management (IEEM)*, 2241–2245. doi:10.1109/IEEM.2017.8290290
- Yang, Z., Lei, J., Fan, K., & Lai, Y. (2013). Keyword extraction by entropy difference between the intrinsic and extrinsic mode. *Physica A*, 392(19), 4523–4531. doi:10.1016/j.physa.2013.05.052
- Ying, Y., Qingping, T., Qinzhen, X., Ping, Z., & Panpan, L. (2017). A graph-based approach of automatic keyphrase extraction. *Procedia Computer Science*, 107, 248–255. doi:10.1016/j.procs.2017.03.087
- Zhou, H., & Slater, G. W. (2003). A metric to search for relevant words. *Physica A*, 329(1-2), 309–327. doi:10.1016/S0378-4371(03)00625-3