iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 9 **Issue:** IV **Month of publication:** April 2021

DOI: https://doi.org/10.22214/ijraset.2021.33625

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue IV Apr 2021- Available at www.ijraset.com

Using Term Frequency - Inverse Document Frequency to find the Relevance of Words in Gujarati Language

Tripti Dodiya[1], Dr. Ali Yawar Reha[2]
[1]*Research Scholar, Pacific Academy of Higher Education and Research University, Udaipur, Rajasthan*
[2]*Pacific Institute of Management, Pacific University, Udaipur, Rajasthan*

Abstract: ***TF-IDF, term frequency-inverse document frequency, is used to evaluate the relevance of a word to a document in a collection. It is used in information retrieval and text mining field. In this technique, the weight of the word is calculated which signifies the importance of the word in the document. This paper discusses about the approach to determine TF-IDF for documents in Gujarati language, which is a morphologically rich Indo-Aryan language. The approach calculates TF-IDF, based on the method to find the frequency of the words in the document. The system shows significant results, which is discussed in detail.***

Keywords: TF-IDF, Gujarati, Information retrieval, Text mining

I. INTRODUCTION

Gujarati is the language of Gujarat, a western Indian state, and is spoken by 70% of the state's population. Apart from Gujarat, it is widely spoken in the states of Maharashtra, Rajasthan, Karnataka and Madhya Pradesh and also around the world. Gujarati is morphologically rich Indo-Aryan language and derived from Devnagari script.

TF-IDF is an information retrieval technique wherein a word's term frequency (TF) and its inverse document frequency (IDF) is calculated [1][2]. The product of TF and IDF, is the tf-idf weight of that word. The weight signifies the importance of the word based on the number of times it appears in the document. Generally, to calculate the tf-idf the following steps are required:

1) *Tokenization:* the text is divided into smaller tokens. Further, the tokenization can be done at sentence level and word level.
2) *Term Frequency (TF):* it calculates the number of times each word appears in the document.
3) *Document Frequency (DF):* it specifies the number of documents that has the term.
4) *Inverse Document Frequency (IDF):* it signifies the importance of the term. It helps to find the important terms that can provide specific document context.

II. LITERATURE REVIEW

The term-weighting function IDF was proposed in 1972, and has since been widely used, usually as part of a TF*IDF function [3]. A lot of work has been done in English language for NLP tasks. Apart from English, research on other Indian languages is also seen but is at its initial stage.

Abu-Errub et al. [4] proposed a method for Arabic text classification. The system compares the document with pre-defined document categories based on its contents using the TF-IDF measure. The document is further classified into the appropriate sub-category using Chi Square measure. The experimental results shows the classification of the tested documents to its appropriate sub category.

Chan et al. [5] proposed a method to enhance the effectiveness of news classification. They used term frequency in news segment to train the weighting of each category of each term. Further, they classified the test news based on the weighting.

Dhar et al. [6] worked on categorizing Bangla text documents of online web corpus using standard features as well as machine learning approaches. They used TF-IDF feature with dimensionality reduction technique (40% of TF) for precision.

Jayashree et al. [7] proposed an algorithm that extracts keywords from pre-categorized Kannada documents. They combined GSS [8] coefficients and TF-IDF for extracting key words for summarization.

Rakholia et al. [9] proposed a system wherein Naïve Bayes (NB) statistical machine learning algorithm is used along with TF-IDF approach to classify Gujarati documents.

As seen from the literature review, TF-IDF is an important technique used by researchers for document classification as well as language processing tasks.

©IJRASET: All Rights are Reserved

III.DISCUSSION

A. Term Frequency

Generally, the term tf-idf is the combination of two terms: Term Frequency (tf) which specifies the number of times a word appears in a document [10][11][12]. For example, document 'D1' is having 1000 words, and the word "અસ્થમા" appears 20 times in the document. The frequency of the word, highly depends on the length of the document. If the document is large, the term "અસ્થમા" may appear more number of times compared to the document of small size. This does not make the longer document more important than the smaller one. To resolve this issue, the number of times a word in the document is divided by the total words in the document.

tf(t)=(Number of times term "t" appears in a document d)/(Total number of terms in the document d) (1)

In our example, the term frequency of the word "અસ્થમા" will be:

tf = 20/1000 = 0.02

B. Inverse Document Frequency

As discussed, term frequency gives how frequently the word appears in the document [1][10][12]. However, it treats all terms in the document equally, so the occurrence of less important terms such as "*the*", "*is*", "*of*" and so on (also known as stop words) in the document is also calculated. The inverse document frequency helps in resolving this issue. It decreases the weight of terms that occur frequently while increases the weight of terms that occur rarely.

For example, we have 3 documents and the term "અસ્થમા" is present in two of the documents, so the inverse document frequency can be calculated as:

idf (t) = log_e((Total number of documents)/(Number of documents with term t in it)) (2)

idf (asthma) = log_e(3/2) = 0.651

Finally, the tf-idf is calculated by:

TF-IDF = tf × idf (3)

IV.METHODOLOGY

The proposed methodology is implemented using the standard approach to find the tf-idf [2]. As discussed, the steps to calculate the tf-idf are tokenization, term frequency and inverse document frequency. Figure 1 shows the approach.

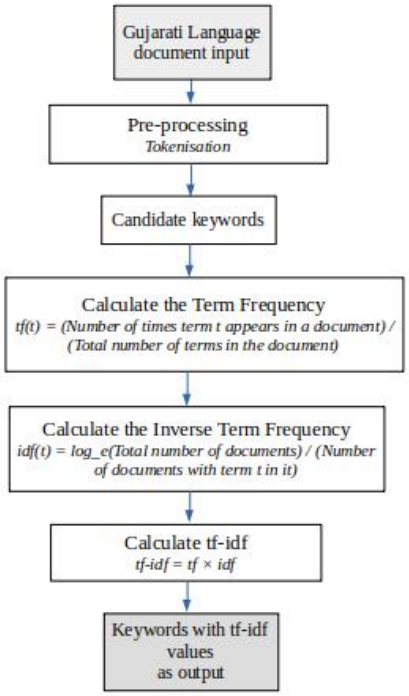


Fig. 1 Approach to find tf-idf

©IJRASET: All Rights are Reserved

A. Algorithm for tf-idf Generation

1) Input : Gujarati text

2) Output: tf-idf of each word in the document

Begin

a) For each Document D in the corpus

b) Tokenize the document into words

c) For each term in the document, calculate the term frequency

d) For each term in the document, calculate the inverse document frequency

e) Calculate the tf-idf parameter

f) Generate the output showing the tf-idf of each term

End

In the pre-processing step, the words in the document are tokenized using iNLTK, natural language toolkit for Indic languages. In the next stage, a function to calculate the term frequency (tf), takes the dictionary containing the pre-processed words and creates the term frequency of the words. Term frequency is calculated as per the formula (1). Next, the function, inverse document frequency (idf) calculates the *idf* as per the formula (2). Finally, after having the tf and idf values for each word, tf-idf is calculated using the formula (3).

V. EXPERIMENTAL RESULTS

The algorithm is implemented using Python language. The performance of the tf-idf algorithm has been evaluated using the documents on medical domain collected from different medical resources. As the medical literature is not easily available in Gujarati language, the documents were translated from English to Gujarati. For the experiment, we started with simple sentences in Gujarati and then with set of paragraphs. The results of the experiments are given in Figure 2 and Figure 3.

```
આ એક શા ળા  હતી
('આ', 0.0)
('એક', 0.02567211779851649)
('શા ળા ', 0.0)
('હતી ', 0.02567211779851649)
આ બી જી  શા ળા  છે
('આ', 0.0)
('બી જી ', 0.02567211779851649)
('શા ળા ', 0.0)
('છે', 0.02567211779851649)
```

Fig. 2 tf-idf values of two simple sentences

```
લ્યુકેમિયા  એ બ્લડ કેન્સર છે
('લ્યુકેમિયા ', 0.0)
('એ', 0.0)
('બ્લડ', 0.01019334089058743)
('કેન્સર', 0.0)
('છે', 0.01019334089058743)



લ્યુકેમિયા  એ લો હી ના  કો ષો નું એક કેન્સર છે.
('લ્યુકેમિયા ', 0.0)
('એ', 0.0)
('લો હી ના ', 0.01019334089058743)
('કો ષો નું', 0.01019334089058743)
('એક', 0.01019334089058743)
('કેન્સર', 0.0)
('છે.', 0.01019334089058743)
```

Fig. 3 tf-idf values of two medical domain sentences

©IJRASET: All Rights are Reserved

The experimental results show that the tf-idf values for the words in the document given as output of the code, works efficiently for a set of inputs. The program gives results as expected and satisfactory for a small size dataset, but needs to be refined for a large dataset. In addition, some pre-processing techniques like stop-word remover, stemmer needs to be implemented for better results. The improvement in tf-idf algorithm will enhance the accuracy of text classification results and other NLP tasks.

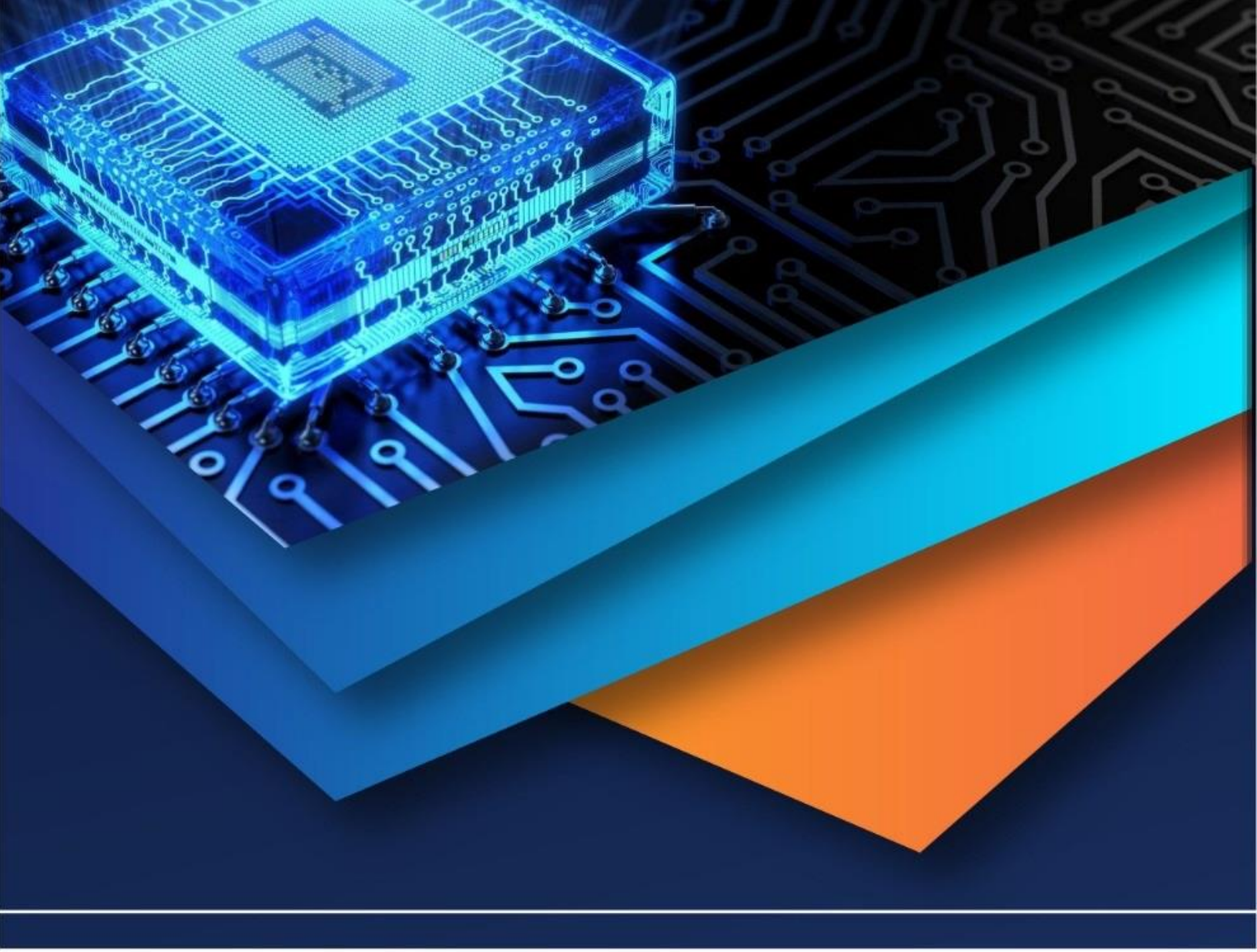
VI. CONCLUSIONS

In today's world of big data, tf-idf technique is simple and powerful tool which can further help in NLP tasks. This paper we have discussed the approach to find the tf-idf feature vector for Gujarati texts. This can further help in many applications of linguistic computing. The research in Gujarati language is in its initial stage as compared to other languages such as English and Hindi and a lot of work needs to be done.

REFERENCES

[1] Bafna, Prafulla, Dhanya Pramod, and Anagha Vaidya. "Document clustering: TF-IDF approach." In 2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT), pp. 61-66. IEEE, 2016.
[2] Borkakoty, Hsuvas, Chandana Dev, and Amrita Ganguly. "A Novel Approach to Calculate TF-IDF for Assamese Language." In Electronic Systems and Intelligent Computing, pp. 387-393. Springer, Singapore, 2020.
[3] Robertson, Stephen. "Understanding inverse document frequency: on theoretical arguments for IDF." Journal of documentation (2004).
[4] Abu-Errub, Aymen. "Arabic text classification algorithm using TFIDF and chi square measurements." International Journal of Computer Applications 93, no. 6 (2014).
[5] Chan, Tzu-Yi, and Yue-Shan Chang. "Enhancing classification effectiveness of Chinese news based on term frequency." In 2017 IEEE 7th International Symposium on Cloud and Service Computing (SC2), pp. 124-131. IEEE, 2017.
[6] Dhar, Ankita, Niladri Sekhar Dash, and Kaushik Roy. "Application of tf-idf feature for categorizing documents of online bangla web text corpus." In Intelligent Engineering Informatics, pp. 51-59. Springer, Singapore, 2018.
[7] Jayashree, R., K. M. Srikanta, and K. Sunny. "Document summarization in Kannada using keyword extraction." Proceedings of AIAA 11 (2011): 121-127.
[8] Galavotti, Luigi, Fabrizio Sebastiani, and Maria Simi. "Experiments on the use of feature selection and negative evidence in automated text categorization." In International Conference on Theory and Practice of Digital Libraries, pp. 59-68. Springer, Berlin, Heidelberg, 2000.
[9] Rakholia, Rajnish M., and Jatinderkumar R. Saini. "Classification of Gujarati documents using Naïve Bayes classifier." Indian Journal of Science and Technology 5 (2017): 1-9.
[10] Hakim, A. A., Erwin, A., Eng, K. I., Galinium, M., & Muliady, W. (2015). "Automated document classification for news article in Bahasa Indonesia based on term frequency inverse document frequency (TF-IDF) approach," 6th International Conference on Information Technology and Electrical Engineering: Leveraging Research and Technology, (ICITEE), 2014
[11] Ramos, Juan. "Using tf-idf to determine word relevance in document queries." In Proceedings of the first instructional conference on machine learning, vol. 242, pp. 133-142. 2003.
[12] Fan, Huilong, and Yongbin Qin. "Research on text classification based on improved tf-idf algorithm." In 2018 International Conference on Network, Communication, Computer Engineering (NCCE 2018). Atlantis Press, 2018.

©IJRASET: All Rights are Reserved



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)