JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue

JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)

An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Automatic Keyword Mining Model for Gujarati Text

[1] **Janvi Sheta,** [2] **Barkha Wadhvani,** [1]**Janvi Savani**
[1]M.Sc.(DS) Student, [2]Assistant Professor
[1]School of Science, P P Savani University, Surat, Gujarat, India
[2]Computer Engineering Department, School of Engineering, P P Savani University, Surat, Gujarat, India

Abstract — Manual assessment of textual information on the internet is now not possible because to the huge increase in the number of these publications on the internet. Keywords and keyword phrases are very useful for quickly and efficiently evaluating massive volumes of information for internet searches, and they give a summary of the material. A document's keywords are a group of descriptive words that provide prospective readers with the most frequently used and important words and expressions from the text. Keyword extraction is widely used in computer science, notably in information extraction and processing natural language. In recent years, many keyword extraction techniques have developed for the English language. However, it is still a new method for Indian regional languages, and development is too sluggish. This study provides a model for extractive summarization in Gujarati using the TF-IDF approach.

Index Terms—Keyword Extraction, Natural Language Processing, Text Summarization, Unsupervised learning, Gujarati Language

I.INTRODUCTION

Due to the ever-increasing need for the internet in today's society, readers have access to a variety of e-materials, which makes it easier for them to read and glean new knowledge. These e-materials include things like electronic books, journals, articles, and newspapers, among other things. People could have a hard time reading this full text in the allotted amount of time if they try to do so. Therefore, they want a method that will enable them to locate the keywords inside the text in order to save both their time and their energy [1]. There are many different human languages, and numerous strategies and procedures are being developed to extract keywords from them [2]. Currently, our focus is on the Gujarati language as we try to identify significant keywords within Gujarati text in order to make it simpler to get data from Gujarati text. It is necessary to have some kind of technology that can extract important data from Gujarati text in order to make this method workable. Text mining is a method that involves extracting large volumes of information in order to deliver high-quality data. This procedure is referred to as "text mining." Text mining is a kind of natural language analysis that offers a variety of technologies for use in text analysis. These techniques include automated phrase mining and text summarization.

The use of keywords allows readers to decide whether or not a text is relevant to their needs. According to one definition, a keyword is "a term that correctly characterizes the topic, or an element of the subject, addressed in a text in a short and precise manner" [3]. Keyword assignment can be done both manually and automatically, with the manual process taking much more time and costing much more money. As a result of this, a process that can automatically extract keywords from documents is necessary. In text mining, one of the most important tasks is called keyword extraction. The process of extracting keywords may be carried out in a number of different ways, such as via supervised and unsupervised forms of machine learning, statistical methods, and linguistic methods. Automated feature extraction is the process of selecting phrases and words from a formal document that, based on the model, are at their absolute best and also depict the fundamental emotion of the content. This process is primarily focused on information organization without the additional expenses of human annotators while making use of the speed and accuracy of existing calculation skills to solve the problem of connectivity and recovery. If we could find a tiny fraction of words—keywords—that might disclose the primary qualities, ideas, subjects, etc., of the text, then it would be much simpler to analyze such enormous volumes of data.

The remaining parts of the paper are organized as follows: Different keyword extraction methods were explored in Section II. In Section III, we give the results of the analysis and any other relevant preparations. This section laid the groundwork for the proposed procedure in Section IV. Final thoughts and suggestions for further research were presented in Section V.

II.KEYWORD EXTRACTION TASK

The purpose of Keyword Extraction may shift depending on the context in which it is used. The fundamental objective of media analysis and surveillance is to extract from the text of each news piece the most important topics discussed, the most significant episodes discussed, the parties involved, as well as the outcome, impact, and significance of such incidents. Text mining makes use of a number of Natural Language Processing (NLP) methods, such as part-of-speech (POS) tagging, parsing, N-grams, text categorization, and so on, in order to carry out the text analysis [4]. Keyword extraction strategies may be broken down into two primary categories: domain-dependent and domain-independent methods [5]. The techniques of keyword extraction that are reliant on the domain need to retain a record of every word in the textual collection, while the methods of keyword extraction that are independent of the domain do not need to do an analysis of the full text collection. In the process of manually assigning keywords, phrases from lexical features are chosen to serve as key words and phrases, and documents are categorized into classes according to the material they contain, which matches to terms from the vocabulary. Therefore, in order to save time and effort, a number of automated keyword extraction strategies were developed and categorized as follows: fundamental statistics, linguistic, machine learning, and other approaches [6]. Figure 1 displays the four basic ways for extracting relevant keywords and key phrases. The methods are defined below:

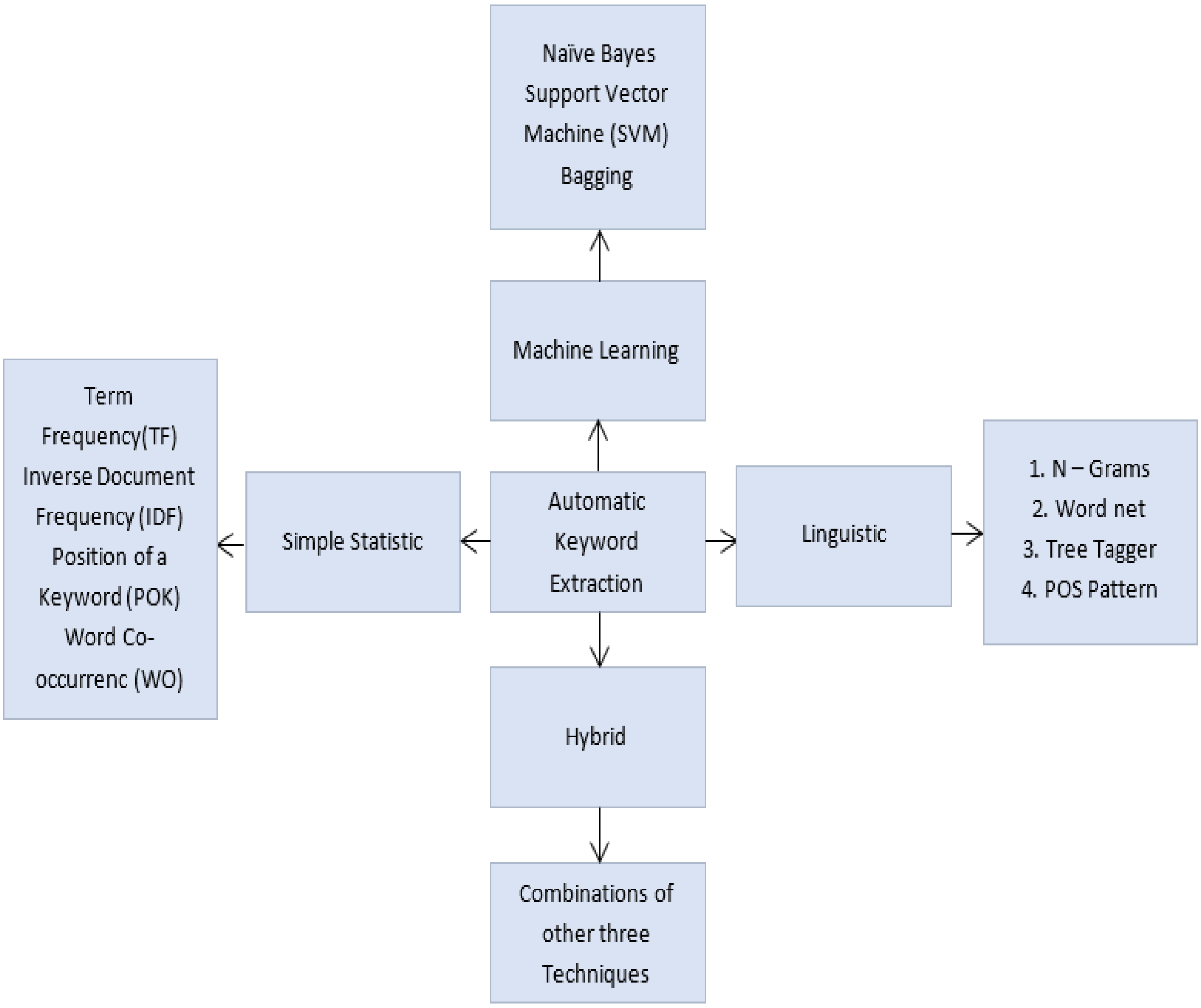

Figure 1. Automatic keyword Extraction Approaches

2.1. Statistical Method:

The field of basic statistical approaches consists of straightforward procedures that don't call for the use of training data. Furthermore, techniques are not limited to a single domain or language. The statistics of the words in the text may be used to find keywords. These statistics include n-gram statistics, word frequency, TFIDF, word co-occurrences, and the PAT tree (also known as the Patricia tree), which is either a suffix tree or a position tree. One potential drawback is that the most significant keyword could only show up once in the whole article in some types of professional writing, such as those dealing with health and medicine. When statistically enabled models are used, it is possible that certain terms will be omitted by mistake [7]. The availability of a large number of datasets has made it feasible to perform statistical analysis and create useful findings, notwithstanding the possibility that the results produced by these techniques will not be as accurate as those produced by other approaches.

2.2. Machine learning methods:

In the majority of cases, the machine learning algorithms for the keyword extraction task will make use of supervised learning techniques. Retrieving keywords from instructional material is one of the approaches used in machine learning. Once the model has been developed, it is put through its paces by way of a testing module. When a model has been finished to the point where it is considered good, it is put to use to harvest keywords from new texts. Techniques such as support vector machines and naive bayes are used in the execution of this method. This approach necessitates the collection of training data, and its results are often contingent on the domain. Every time the domain was changed, the system would need to re-learn everything and construct a new model [8].

2.3. Linguistic methods

Techniques like this often adhere to standards and take cues from language facets as well as professional knowledge. These approaches, although having a higher probability of producing correct results, are computationally expensive and need not just linguistic competence but also a comprehension of the topic matter. The primary focus of these methodologies is on the linguistic properties of the sentences and texts [9]. Analysis of speech, lexicon, and syntactic structure are some of the components that make up the linguistic approach.

2.4. Hybrid approach

In order to identify the automated Keyword extraction Task, a range of different strategies might be combined.

III Related Work

The technique for keyword extraction is decided upon according to the nature of the domain in which it will be implemented. The authors of [10] offered a lightweight technique for extractive summarization and rankings based on an uncontrolled methodology to choose the keywords that are most relevant to a particular text. This approach was founded on the idea that the most relevant keywords should be selected first. They came to the conclusion that their technique was superior to other methods such as RAKE, Text Rank, Single Rank methods, and TF IDF. This was the conclusion reached by the researchers. [11] referred to a keyword extraction method that makes use of lexical chains to achieve positive results. The authors of [12] described a multi-featured automated keyword extraction method that required human supervision. They employed important semantic characteristics that are descriptive of potential key phrases in order to produce useful findings and train the Random Forest classifier that was used. Deep neural networks have the potential to be used in order to enhance the performance of the model [13]. [14].

[15] proposed a novel keyword extraction approach for Chinese newspapers that was built on TF/IDF with many strategies. This method was developed through an analysis of the linguistic features of news documents. A stemmer is a fundamental linguistic resource that is necessary to construct any sort of activity in Natural Language Processing (NLP) with good accuracy. Some examples of these applications include machine translation, document categorization, document grouping, text information retrieval, topic tracking, text summarization, and keyword extraction, among others. A stemmer may be used for any language in the world. For efficient results, the authors of [16] advocated a fully automated stemming of all Punjabi words, including nouns, verbs, adjectives, adverbs, pronouns, and proper names. This would cover all Punjabi word categories. Sometimes TF-IDF will provide findings that are surprising since this method is unable to recognize words that have even the slightest change in their tenses, including such words as go and go [17]. The author [18] used TF-IDF and Naive Bayes in order to correctly classify the information while taking into account the connections between the different categories. The TF-IDF Technique was suggested by the authors in [19] for the purpose of text summarization in the Hindi language. [20] Proposed methods for classifying texts specified a wide variety of Indian languages, including Assamese, Bengali, Hindi, Kannada, Oriya, Gujarati, Punjabi, and Telugu, among others, and employed TF-IDF as a feature vector.

IV Proposed Model

In order to make things easier for end users, the work that was proposed focused on the process of extracting keywords from long Gujarati texts in order to get a better understanding of the topics discussed in such texts. Figure 2 illustrates the process of extracting keywords from a document. The TF-IDF approach is used in order to investigate the significance of individual words within the provided text [21]. The TF-IDF approach is one that we have incorporated in our suggested strategy in order to extract relevant keywords from the Gujarati language. The TF-IDF algorithm determines how relevant each phrase in a corpus is to a particular text by calculating the relevance index. The TF-IDF approach involves analyzing the words that are present in the text in order to identify which ones are likely to be significant based on factors such as frequency and length [22], [23].
The work that is being suggested may be broken down into these four stages:

Step-1: Data Collection

In order to make the effort more manageable, we have compiled a sample of Gujarati textual data from a number of different websites that are devoted to the language. The material of a website that is open to the general public and may be accessed for free through the internet is what was utilized in this investigation. After collecting the data in CSV format, it was cleansed and highly processed for later use.

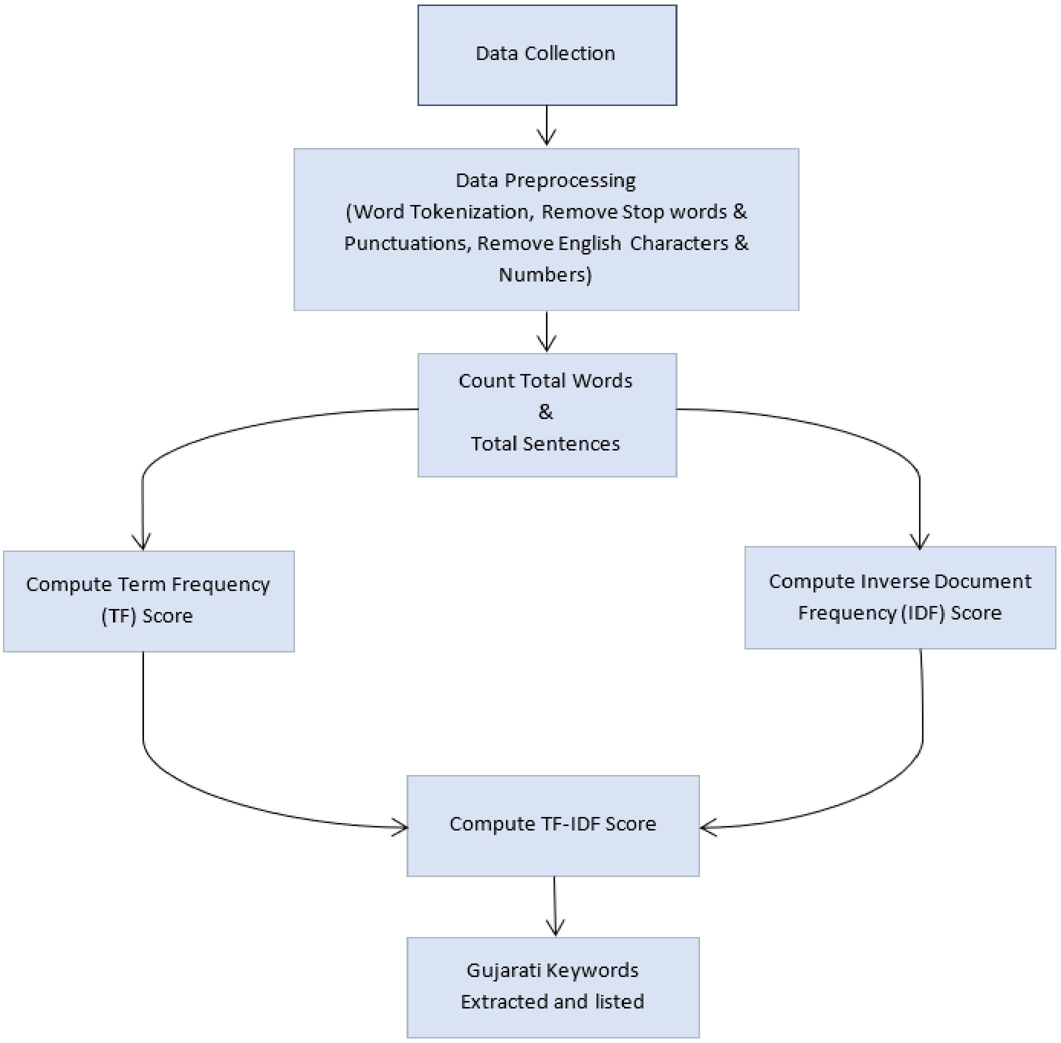
Data Collection
Data Preprocessing (Word Tokenization, Remove Stop words & Punctuations, Remove English Characters & Numbers)
Count Total Words & Total Sentences
Compute Term Frequency (TF) Score
Compute Inverse Document Frequency (IDF) Score
Compute TF-IDF Score
Gujarati Keywords Extracted and listed

Figure 2. Keyword Extraction process for Gujarati Text

Step-2 : Data Preprocessing:
The data were preprocessed in order to get them ready for the application of the proposed model. The data has been cleaned of any stop words, punctuation, English characters, and numerical characters. The term "stop words" refers to a collection of terms that are often used in writing and speaking any language. First things first, we did some tokenization of the text in order to get rid of any stop words [24] and symbols that weren't essential. In addition to that, characters from the English language have been taken out of the text that was provided. The conversion from the numerical format to text will take place. After deleting any null values from the Gujarati text, the tokenized words should be joined together. We have been making use of the Natural Language Toolkit library in order to work with natural languages (NLTK).

Step-3 : Total Words and Sentences Count
The TF-IDF algorithm, which stands for "Term Frequency Inverse Document Frequency of Words," raises the proportionality to the number of times in the text a word occurs, but it is balanced out by the word frequency in the provided text. Following the steps for data preparation, the first step is to calculate the count of the total number of words and sentences by using a variety of library resources.

Step-4 : Calculate TF-IDF Score
There are several different methods through which the significance of a word within a text or corpus may be ascertained. Through the use of term weighting algorithms, a text corpus may be utilized to identify stop words and keywords. The following is a list of the significant word weighting measures:

i. **Term count**: The frequency with which a certain word occurs in a given text is what is meant by the phrase "term count," which is a shorthand way of referring to this statistic. This metric assigns a greater weight to frequently occurring words, in recognition of the fact that it is not uncommon for a significant term to turn up a number of times within a single piece of writing.

ii. **Term Frequency:** The frequency of a term is measured by counting the total number of occurrences it has in a certain corpus. To calculate the overall term frequency, we compile the number of times each word occurs throughout all of the corpus texts. The TF metric, also known as term frequency, may be used to ascertain which words are the most significant inside a certain text or document. Normalizing the frequency of that term and then divide by the total amount of words used in the text is something that has to be done.

$$\boldsymbol{TF(T,D)} = \frac{Count\ the\ T\ in\ D}{Number\ of\ Terms\ in\ D} \quad (1)$$

Where T=Term and D= Document

iii. **Document Frequency:** If there are more than N documents in a text, and there are less than n of those documents, then the number of times that Term T appears in documents is referred to as the document frequency of Term T.

$$\boldsymbol{DF}(\boldsymbol{T}) = Occurrence\ of\ T\ in\ N\ number\ of\ Documents \quad (2)$$

iv. **Inverse document frequency (IDF):** IDF is a measurement of comprehensibility that embodies the notion that the scarcity of a term increases its chance of being pertinent to the text in which it appears. This principle underpins the IDF concept. It is recommended that the document frequency be normalized by dividing it by the total number of documents. The IDF value, however, skyrockets once we work with a huge corpus. Therefore, in order to mitigate the impact, IDF may be computed by taking the log of IDF.

$$\mathbf{IDF = log(\frac{N}{d_T})} \quad (3)$$

Where N = Number of documents in the corpus, d_T = Number of documents containing the Term T.

v. **Term Length (TL):** The actual population of tokens in a keyword word is denoted by the abbreviation TL. The purpose of using longer sentences is to convey things in a way that is clearer and more important.

vi. **TF-IDF:** On the basis of the collected TF and IDF, a frequency-inverse document frequency weight (TF-IDF weight) is calculated. This weight indicates how important a Term is to a document included inside a Text. The frequency with which a phrase occurs in a document has a direct correlation to its level of significance; however, this correlation is cancelled out by the total number of papers that contain that term.

$$\boldsymbol{TF - IDF = TF(T,D) * IDF} \quad (4)$$

Where TF(T,D) = Term Frequency, IDF = Inverse document frequency

Step : 5 Gujarati Keywords extracted:

The supplied Gujarati text was analyzed using TF-IDF scores, and the top fifteen words with the highest TF-IDF scores were selected to serve as Gujarati keywords.

V. Conclusion and Future work

The authors of this study offered a model in which the TF-IDF approach might be used to extract keywords from the Gujarati text that was provided. Since TF-IDF is independent of both language and domain, this would be a useful technique. In the future, other approaches to the extraction of keywords, including supervised and unsupervised learning, may be used in order to improve the degree of precision achieved by Gujarati text. It is difficult to go further with Gujarati text since there are not enough libraries, modules, or datasets in Gujarati. Because there are not a lot of precise stemmer and lemmatization datasets or libraries available for the Gujarati language, it is possible that in the future, that dataset will be enlarged based on customized rules for the Gujarati language in order to serve as a preprocessing step for the purpose of achieving more efficient and precise outcomes for Gujarati text.

Acknowledgement

The authors would like to express their gratitude to everyone who assisted them in gathering the essential data for this study, supported them in their efforts to finish the study, and assisted them in any manner, direct or indirect. The authors would also like to take this occasion to express their appreciation to P P Savani University for providing the necessary environment for them to carry out their research.

References

[1] Barkha A. Wadhvani, Sameer A. Chauhan, "Hike the Performance of Collaborative Filtering Algorithm with the Inclusion of Multiple Attributes", International Journal of Information Technology and Computer Science(IJITCS), Vol.10, No.4, pp.73-80, 2018. DOI: 10.5815/ijitcs.2018.04.08

[2] Nomoto, T. Keyword Extraction: A Modern Perspective. *SN COMPUT. SCI.* **4**, 92 (2023). https://doi.org/10.1007/s42979-022-01481-7.

[3] S. Lahiri, S. R. Choudhury, C. Caragea, "Keyword and Keyphrase Extraction Using Centrality Measures on Collocation Networks", arXiv preprint arXiv:1401.6571, 2014.

[4] Rose, S., Engel, D., Cramer, N., & Cowley, W. (2010). Automatic Keyword Extraction from Individual Documents. Text Mining, 1–20. https://doi.org/10.1002/9780470689646.ch1

[5] Onan, A., Korukoğlu, S., & Bulut, H. (2016). Ensemble of keyword extraction methods and classifiers in text classification. Expert Systems With Applications, 57, 232–247. https://doi.org/10.1016/j.eswa.2016.03.045

[6] Bharti, Santosh Kumar, and Korra Sathya Babu. "Automatic keyword extraction for text summarization: A survey." *arXiv preprint arXiv:1704.03242* (2017).

[7] P. Chen, S. Lin, "Automatic keyword prediction using Google similarity distance", presented at Expert Syst. Appl., pp. 1928- 1938, 2010.

[8] F. Sebastiani, "Machine learning in automated text categorisation", ACM Computing Survays, 34(1), 1-47, 2002.

[9] Liao, Huchang, Xiaomei Mi, and Zeshui Xu. "A survey of decision-making methods with probabilistic linguistic information: bibliometrics,

preliminaries, methodologies, applications and future directions." *Fuzzy Optimization and Decision Making* 19.1 (2020): 81-134.

[10] Campos, Ricardo, et al. "A text feature based automatic keyword extraction method for single documents." *European conference on information retrieval*. Springer, Cham, 2018.

[11] Ercan, Gonenc, and Ilyas Cicekli. "Using lexical chains for keyword extraction." *Information Processing & Management* 43.6 (2007): 1705-1714.

[12] John, Adebayo Kolawole, Luigi Di Caro, and Guido Boella. "A supervised keyphrase extraction system." *Proceedings of the 12th International Conference on Semantic Systems*. 2016.

[13] Zhang, Yu, et al. "Keywords extraction with deep neural network model." *Neurocomputing* 383 (2020): 113-121.

[14] Nasar, Zara, Syed Waqar Jaffry, and Muhammad Kamran Malik. "Textual keyword extraction and summarization: State-of-the-art." *Information Processing & Management* 56.6 (2019): 102088.

[15] Li, Juanzi, Qi'na Fan, and Kuo Zhang. "Keyword extraction based on tf/idf for Chinese news document." *Wuhan University Journal of Natural Sciences* 12.5 (2007): 917-921.

[16] Gupta, Vishal. "Automatic stemming of words for Punjabi language." *Advances in signal processing and intelligent recognition systems*. Springer, Cham, 2014. 73-84.

[17] Ramos, J. (2003). "Using TF-IDF to Determine Word Relevance in Document Queries," Proceedings of the First Instructional Conference on Machine Learning, pp. 1–4.

[18] Fan, H., and Qin, Y. (2018). "Research on Text Classification Based on Improved TF-IDF Algorithm," International Conference on Network, Communication, Computer Engineering (NCCE 2018), vol. 147.

[19] Kumar, Atul, Vinodani Katiyar, and Bhavesh Kumar Chauhan. "Text Summarization in Hindi Language Using TF-IDF." *Cognitive Informatics and Soft Computing*. Springer, Singapore, 2022. 319-331.

[20] Raghuveer, K., and Kavi Narayana Murthy. "Text Categorization in Indian Languages using Machine Learning Approaches." IICAI. 2007.

[21] Qaiser, Shahzad, and Ramsha Ali. "Text mining: use of TF-IDF to examine the relevance of words to documents." *International Journal of Computer Applications* 181.1 (2018): 25-29.

[22] Liu, Cai-zhi, et al. "Research of text classification based on improved TF-IDF algorithm." *2018 IEEE International Conference of Intelligent Robotic and Control Engineering (IRCE)*. IEEE, 2018.

[23] Havrlant, Lukáš, and Vladik Kreinovich. "A simple probabilistic explanation of term frequency-inverse document frequency (tf-idf) heuristic (and variations motivated by this explanation)." *International Journal of General Systems* 46.1 (2017): 27-36.

[24] Dataset, Jan, 2021. Retrived from, https://www.kaggle.com/datasets/heeraldedhia/stop-words-in-28-languages.