

e-ISSN: 2319-8753 | p-ISSN: 2347-6710

TEDIA

#### International Journal of Innovative Research in SCIENCE | ENGINEERING | TECHNOLOGY

# INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 11, Issue 10, October 2022

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

## Impact Factor: 8.118

9940 572 462

 $\bigcirc$ 

IJIRSET

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| <u>www.ijirset.com</u> | Impact Factor: 8.118|

Volume 11, Issue 10, October 2022

DOI:10.15680/IJIRSET.2022.1110042

### An Extractive Knowledge-based Automatic Summarizer System for Gujarati Text

#### Nikita P. Desai

Assistant Professor, Department of Information Technology, Dharmsinh Desai University,

Nadiad, Gujarat, India

**ABSTRACT**: Automatic Text Summarization (ATS) is an Artificial Intelligence (AI) based Natural Language Processing (NLP) technique of shortening long pieces of text. Gujarati is an Indian language with more than sixty million users across the globe. Manually generating a summary from piles of electronic text accumulated due to advent of social media and Internet can be a time-consuming, tedious and context sensitive task. This article presents the design, implementation and results obtained thereof for a Gujarati ATS system. The proposed ATS system extracts the important sentences from a given document using rich set of knowledge bases. A corpus of 50 stories along-with their gold summaries were designed to evaluate the system efficiency. The article discusses results obtained after automated and manual evaluation of proposed system. The aspects of non-redundancy, referential clarity, focus, structure and coherence were considered to judge the summaries manually. While the automated measures used were ROUGE, recall, precision and F\_score. The proposed system performance was also compared with a random selection-based baseline model. The ATS system designed was found to be nearly 35% better than the baseline model.

KEYWORDS: Vector Space Model, Gujarati Text Summarization, Knowledge Based, Extractive Summarization

#### I. INTRODUCTION

The Natural Language Processing systems (NLP) use computational techniques to analyze, process and generate input and output in human languages for use of an Artificially Intelligent (AI) system. Nowadays with wide penetration of Internet, social media, messaging services, news websites and blogs, there has been exponential growth in the textual electronic data in natural languages. Amongst many applications of NLP, one prominent is the Automated Text Summarization (TS/ATS) of vast amount of text written in human language.

ATS is widely used in text analytic applications like information extraction, information retrieval, question answering, etc. There can be two types of ATS - extractive or abstractive, based on the methodology adopted to summarize given data. The abstractive ATS rephrases whereas extractive reuses the input sentences, to form the gist of input text(s).

The Gujarati language is one of 22 major Indian languages having millions of users worldwide (see [1]). However, in a most recent survey by Desai et al [2], it was found that the NLP systems development for the language is still challenging due to issues like unavailability of tools and resources needed and lack of annotated data-sets. This article portraits the design, implementation and result analysis; in attempt of building knowledge based ATS system for such a low resource language. The rest of paper is organized as follows. The section 2 would discuss the necessary theory of ATS. The section 3 would discuss the work done in ATS with special mention of work done in Gujarati. The design aspects of proposed model would be elaborated in section 4 which would be followed by the implementation details of system in next section. A discussion of results obtained is in section6. And in section7 we finally conclude. The major contribution of this article is a first of a kind ATS system for Gujarati text. The proposed system works on electronic Gujarati text to generate extractive type of summary.

#### **II. BACKGROUND THEORY**

The main steps in ATS systems are pre-processing, feature extraction/ identification for sentence scoring and ranking and finally summary generation with some needed post-processing. The pre-processing does the needed modification on text to



| e-ISSN: 2319-8753, p-ISSN: 2347-6710| <u>www.ijirset.com</u> | Impact Factor: 8.118|

Volume 11, Issue 10, October 2022

DOI:10.15680/IJIRSET.2022.1110042

prepare it for summarization. The modification is done to clean the text of extra unwanted text like stop-words, punctuation marks, HTML tags, header and footer tags. Also, it would be helpful to do stemming or lemmatization to get the lemma or base form of given words in text. The normalization is required so that the different variations of same word can be identified. Next feature extraction can be done, based on language dependent linguistic resources like Named Entity Recognizer (NER), wordnet, word embeddings, sentiwordnet etc. The TS can also use the language independent features like word counts and sentence score. Techniques like chunking, parsing and semantic role labelling can also prove useful especially in-case of graph-based TS systems or when abstractive summary are to be generated. Methods like word sense disambiguation, anaphora resolution, textual entailment, lexical chain can aid in identifying the semantics of the source text. For discourse identification predefined rhetorical relations can be used to explain the relationships amongst a wide range of texts [3]. Subsequent phases of ATS are language independent as they involve the mathematical operations of calculating the sentence counts and ranking them based on their comparative values. Finally, some post-processing like elimination of redundant entailing sentences, can be applied to further improvise the generated summary.

#### A. Classification of text summarization systems

If the sentences of original text are re-framed in the summary text, then it is known as abstractive TS, while the extractive TS fetches most important sentences from original text to generate summary.

TS can also be categorized based on parameters [4] like the following - :

- Focus Query based, Generic.
- Output types Indicative, Informative, Critical evaluation abstracts, highlights, graphs, charts.
- Language used in text Mono lingual, Multi lingual, Cross lingual.
- Source of text- Web pages, Emails, Survey, Microblogs, Social media.
- Structure of input text- Structured, Unstructured.

#### B. Evaluation approaches

For evaluation of TS systems two broadly identified methodologies are manual and automated. It was pointed out by Lloret et al [5], the task is highly challenging due to the subjectivity and different perceptions of assessors, leading to different judgements for same summary.

1) Evaluation Using reference summary: In this technique, focus is to identify the co-occurrence of terms or sentences, in system summary (peer summary) and human-written reference summary (gold summary). This approach needs special handling for identification of semantically equivalent fragments while matching the fragments between peer and gold summaries. Widely used parameters for co-selection-based evaluation are recall, precision, and F\_measure. These measures use the count of RS (Reference Summary Sentences) and SS (System Generated Sentences) in calculation.

i. Recall: As show in equation 1, it is the ratio of total retrieved correct sentences to the total number of the retrieved and non-retrieved correct sentences of a text document.

$$Recall = \frac{RS \cap SS}{RS} \tag{1}$$

ii) Precision: Recall evaluates which portion of the sentences selected by a human are also identified by a summa-rization system, whereas precision is the fraction of these sentences identified by the summarization system that are correct.

$$Precision = \frac{RS \cap SS}{SS} \tag{2}$$

iii) F\_measure: The equation given in 3 is used to find the harmonic mean between precision and recall. It attaches beta times more importance to recall as compared to precision  $0 \le \beta \le \infty$ . The F-score is computed as follows. Generally, beta is selected to be 1.

e-ISSN: 2319-8753, p-ISSN: 2347-6710 www.ijirset.com | Impact Factor: 8.118



Volume 11, Issue 10, October 2022

DOI:10.15680/IJIRSET.2022.1110042

$$F\_measure = \frac{(1+\beta^2)Recall*Precision}{Recall+\beta^2Precision}$$
(3)

The ROUGE (Recall Oriented Understanding for Gisting Evaluation) given by Lin, Chin-Yew in 2004 [6], is a widely used algorithm using sequences of words (N-grams) to identify the overlap between ideal human generated gold summary sentences and system generated sentences. ROUGE is reported to be a low-cost choice for obtaining similar results as human evaluations, as was found by Lloret and others [5]. The various version used are ROUGE-1,2 and ROUGE-LCS indicating one gram, two gram and longest common sequence methodology respectively. The ROUGE1 recall score of 41% is considered acceptable in ATS [5].

2) *Manual Evaluation:* The manual evaluation of summaries is highly subjective task. To overcome the subjectivity issue, one widely accepted method is to assign fixed set of criteria for evaluation to human annotators. And also provide them a set of evaluation values to be used for grading the summary. As per the mechanism set up by Document understanding conference [7], grammar of text, non-redundancy, referential clarity, focus and structure coherence can be used, for the assessment of readability aspects.

#### **III. LITERATURE REVIEW**

Reviews on Indian Languages Text Summarization (IL-TS) systems by Baruah et al [8] reported Indian TS systems still work at syntactic level and that IL- TS still struggle to get good results due to issues related to nonstandard tools like stopword list, stemmer, NER and underdeveloped wordnet. While English has better tools and hence improved TS systems as found by Verma et al [9]. However, a comprehensive survey in 2020 by El-Kassas et al. [10], discovered the system generated summaries across the languages, are still far away from human-generated ones.

Few works on Hindi, one of official Indian language were reported in last decade. An extraction-based TS for Hindi language, using statistic and linguistic features having 73% precision was reported in 2017 by Gulati et al [11]. Kumar et al. [12], had given a method for single document summarization using semantic similarity to get recall of 69% at 60% compression but at higher compression of 80% the recall reduced to 45%. The method of using Hindi wordnet for POS tagging to facilitate feature extraction followed by using genetic algorithm for selection and ranking is suggested by Thaokar, Chetana and Malik, Latesh [13] but they had not disclosed the results.

One recent research is reported in IL-TS by Rani and D.K. Lobiyal [14]. They have applied extractive summarization on novel written in Hindi as well has on the counterpart English version. The authors faced issue of inadequacy of tools and data-sets needed for the ATS for the resource poor language. They used self-designed Part-of-Speech tagger to identify topics in document to help the ATS system.

#### A. Automatic Text Summarization for Gujarati

In 2007, Patel, Alkesh et al., [15] had given language independent approach for extractive TS. They suggested to use only structural and statistical features and no semantic features so the method can be generalized for any language TS. They claimed that their system was giving 80% degree of representative summaries when tested on 50 Gujarati literature. However, they have not mentioned the details of the dataset size or source used to do the experimentation. One toolkit is also released by Jhaveri [16]in 2019, which provides cross language summarization (CLS). The toolkit has modules for sentence simplification, inbuilt graph-based models for summarization, machine translation, bootstrap code to develop new model and evaluation module using ROUGE implementation. The English summary found in 2004 DUC dataset is manually converted to Gujarati and subsequently English to Gujarati CLS is tested using ROUGE measure. The results obtained had average Rouge-1 F-score of 21.44 and F-score of 3.76 for Rouge-2.

#### B. Automatic Text Summarization for Short Stories

In 2010 a work by Kazantseva and Szpakowicz, [17] was reported to identify extraction-based summary for short fictional stories. Total of 47 stories with average 3333 word tokens in each were used in experiments. The researchers had designed and tested rule based and machine learning based system using decision tree induction. The main objective of the research was to generate appropriate summary which can help a reader decide whether she would be interested in reading the complete story.



| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.118|

#### Volume 11, Issue 10, October 2022

| DOI:10.15680/IJIRSET.2022.1110042 |

The system developed used parser and named entity recognizer. They had mainly used the manual evaluation measures to finally conclude the effectiveness of their system.

#### **IV. PROPOSED MODEL**

From above literature survey it can be concluded that not much work is done for Gujarati text summarization. Specifically for short story summarization, no work was found in literature. Also lack of annotated data-sets was major challenge in designing and implementing TS for low resource Gujarati text. The figure 1 shows the architecture of our proposed system.

The features of the proposed system are: -

- Extractive, single document, non- structured document, monolingual summary generation.
- Vector space model for text representation.
- Rigorous Pre-processing using 1) sentence segmentation with abbreviations handling (2) stopword removal (3) stemming along with lemmatization.
- Co-reference resolution.
- Statistical and position/ location features used.
- Knowledge based- Using multiple expert knowledge-based lists at various phases in the model.
- Model Evaluation in both manual and automated fashion.

a) Text Representation: The following mechanisms were used to vectorize the text.

- Term Frequency (TF) It is a statistical feature which associates a value to word based on total times that word appears in document divided by total words in document.
- Term Frequency Inverse Document Frequency (TF-IDF)- It reflects how important a word is to a document in the collection or corpus.

e-ISSN: 2319-8753, p-ISSN: 2347-6710 www.ijirset.com | Impact Factor: 8.118



Volume 11, Issue 10, October 2022

| DOI:10.15680/IJIRSET.2022.1110042 |



Figure 1. Architecture: Proposed System

b) Knowledge Bases: Based on more than 40 years the usage knowledge of Gujarati language, following knowledge bases were designed and subsequently used in the proposed system -

- Boundary Words List of all words occurring at sentence endpoints in sentences. Ex: "ຢູ່"," &"," ອີ"," ອີ","",""," ອີ","",""," ອີ","",""," ອີ",","","","
- Stop Words It is a list of words that can be omitted in a sentence without affecting sentence meaning. Ex:"dul","dl","%","ul"... Dictionary List of standard base words originally from Gujarati or which were derived from other languages (loan words). Sources like [18], [19] and [20] were used to generate dictionary for proposed system. Ex:"uj33,","urletat"...

• Prefix Words- List of words attached in front of words to form the morphological variant of a basic word. Ex: "બન"."ગેર"."સવર". "અધર"

- Exchange words- Some words don't follow the commonly used suffix strings to derive the morphological variant from their base word. Such exception strings are enlisted in this knowledge base. Ex:"will":"u"," ol":"u"," ol":"u","
- . Salute words- Words used in Gujarati, to indicate respect to the target person. Ex:" સાહેબ"," બાઈ",..



| e-ISSN: 2319-8753, p-ISSN: 2347-6710| <u>www.ijirset.com</u> | Impact Factor: 8.118|

Volume 11, Issue 10, October 2022

| DOI:10.15680/IJIRSET.2022.1110042 |

• Actor list- The main entities involved in a text are given by this list. This list is fed as input to the proposed system along-with the text to be summarized.Ex: "HHG,", "ERU,",...

For stemming, the longest matching, affix-based Gujarati stemmer given by Desai and Dalwadi [21] is used.

#### **V. IMPLEMENTATION DETAILS**

#### A. Data Sets and Tools/ Modules Used

Owing to unavailability of tagged data-sets for summarization of short stories written in Gujarati, annotated data sets were generated, with aid of native Gujarati speakers. A total of 50 stories were collected from newspapers and

blogs (see [22], [23], [24]). The stories had an average 122 sentences and 5618 words. Next 15 native speakers having experience of more than 18 years were asked to identify the essential and non-essential sentences in the given story. Each story was independently annotated by 3 annotators. On average the stories were compressed between 25% to 65%. The categorical response is determined by annotators based on the understanding of whether dropping that sentence would have effect on the overall meaning of the text. Each judge was asked to assign only "needed" or "not needed" category to the sentences. The evaluators were not in contact or connection of each other and hence interference in decision making was nil. The group of 10 stories given to each judge were selected at random from the dataset. The Fleiss' kappa (FK) measure [25] and pairwise agreement was also calculated to determine inter-rater agreement. The average pairwise agreement was observed to be 71.50%. The Table I gives details of the agreement for the 5 sets. Finally, a single gold summary for each story was formed by combining the individual summaries framed

Table I						
Inter-Rater Agreement Statistics						
Story Set	Total sentences	FK	Avg agreement			
1 to 10	505	0.2241	80.59%			
11 to 20	1073	0.1278	68.93%			
21 to30	536	0.1935	72.76%			
31 to 40	1445	0.2140	70.80%			
41 to 50	2457	0.1061	64.40%			

by all annotators. The sentences which were identified as "needed" by ALL the evaluators, were retained in the gold text while the rest sentences were dropped. The resultant gold summary was again manually checked for referential clarity and proper structure and cohesion.

The proposed model utilizes few open-source tools. From the tool of IndicNLP, the trivial tokenize module was used for finding tokens from lines [26]. The set of packages utilized for word scoring were the Sklearn Count Vectorizer and TFIDF Vectorizer [27] available in Python. Also, an implementation of ROUGE metric in Python<sup>1</sup> was utilized for automated evaluation of summaries.

#### B. Baseline Model

Due to lack of any existing work in Gujarati TS system, one basic system named baseline model was designed. The design of this basic model was inspired by human behaviour wherein the reader skims the pages or lines randomly from start to end of story, to read the story in short time. In similar fashion, the designed baseline model picks lines randomly from given story. The number of lines extracted by baseline model is directly proportional to number of lines in the gold summary.

#### C. Evaluation Setup

1) Manual Evaluation Settings: To aid the evaluation of proposed system, three annotators had volunteered. The annotators were native speakers of Gujarati language who had language usage experience of 2 to 4 decades. The

e-ISSN: 2319-8753, p-ISSN: 2347-6710 www.ijirset.com | Impact Factor: 8.118



#### Volume 11, Issue 10, October 2022

#### DOI:10.15680/IJIRSET.2022.1110042

Table II Manual Evaluation of Models						
Referential clarity	82.67%	84.00%				
Focus	85.00%	79.00%				
Structure and Coherence	83.33%	79.00%				

Table III							
Automated Evaluation of Models - ROUGE1							
Model	Fmeasure	Precision	Recall				
Baseline Model	0.6628	0.6747	0.6509				
Proposed Model-TF	0.6919	0.7005	0.6887				
Proposed model-TFIDF	0.6668	0.7002	0.6444				

annotators were not given access to the original text. They were asked to use a five-point scale to rate the summaries. The various criteria used are listed below, which are largely based on the DUC and TAC conference evaluation measures, as discussed in paper by Lloret and others [5]. The results obtained are reported in Table II as the average percentage of evaluation criteria achieved.

· Evaluation criteria Used-

- 1) Non-redundancy: Unnecessary repetitions like the repeated use of a noun or noun phrase when a pronoun would suffice or repeating facts again like (e.g., He is strong. His strength is good.) etc, indicates redundancy in text.
- Referential clarity: When pronouns or references are made to some entity, it should be easy to identify who they are referring to. Especially for Gujarati summary this criterion would be interesting to check as unlike in English, Gujarati has all genders using same pronouns.
- 3) Focus: The story summary should have sentences which are related to the main theme of story.
- 4) Structure and coherence: The summary should be well-structured and well-organized. The summarized story should build the story in proper flow.

2) Automated Evaluation Measures Used: For automatically evaluating by comparing with the gold standard summary, the measures used were recall, Precision and F\_measure for ROUGE 1-gram, 2-gram and LCS. The results obtained for ROUGE1 are shown in Table III.

*3)* Statistical Significance Tests: In order to establish fact that the proposed ATS model did not get results by chance, it was applied the statistical significance test. As we had 50 same set of stories in both baseline system and proposed system, the two tailed paired statistical t-test was applied. The obtained P-value of 0.0001 obtained was conclusive for accepting hypothesis indicating proposed model is statistically significant.

#### VI. RESULTS AND DISCUSSION

The chart in figure3 indicates TF model outperforms the TF-IDF based text representation model owing to better recall capacity. One major factor might be the fact that in stories, the sentences mentioning the central acting entities more essential in-contrast to the sentences describing them in more detail. As generally any story can be understood without adding additional elaborations about the main actors in story. But on the other hand, we lose the precise sentences expected as the TF-model gives more importance to the lines having main actor mentioned, which might not always have some concrete contribution to the overall meaning of story. It was found that from almost 5700 sentences in totality, in 18% cases the important information containing sentences were ignored while almost 21% non-essential sentences were included in the resulting summaries. The few main reasons identified in these failure cases were:



| e-ISSN: 2319-8753, p-ISSN: 2347-6710| <u>www.ijirset.com</u> | Impact Factor: 8.118|

Volume 11, Issue 10, October 2022

| DOI:10.15680/IJIRSET.2022.1110042 |

- Sentences having actor words are not always necessary but due to higher score end up in the top listed sentences. In dialogue-based stories, for rhetorical effect cataphora is used. But such references are not resolved and hence if sentence does not have any other good scoring word, the sentence is dropped from summary, even though it was essential for context identification.
- Fictional prose uses wide vocabulary. Due to this characteristic of literature writing the words composing a sentence end up being on lower frequency band and hence not contributing to increasing the overall sentence value. In such cases, the frequency feature vector is found to be of very small size after dimensionality reduction. And hence many sentences are judged poorly.





#### Figure 3. Model Comparison



**ROUGE Parameter Applied** 

e-ISSN: 2319-8753, p-ISSN: 2347-6710 www.ijirset.com | Impact Factor: 8.118



Volume 11, Issue 10, October 2022

DOI:10.15680/IJIRSET.2022.1110042



Figure 4. Compression Size Effect on Model Performance (TFmodel)

Figure 5. Feature(dimension) reduction done by proposed model



The chart in fig.3 shows the proposed model using the TF representation exhibits average 6.53% improvement over the baseline model, across all of ROUGE1,2 and LCS. The best performance of proposed model is the ROUGE2 Fscore with nearly 11% better results than baseline model. The minimum improvement observed by TF based proposed model is 2% for recall of ROUGE-LCS while maximum is the recall of 12% for the ROUGE-2 measure. On evaluation by human judges also, the proposed TF model was found to be better by almost 18% 3 as compared to the baseline model. The major positive effect of TF-based proposed model was on the structure and coherence of the extracted summary. The judges found proposed model to be almost 35% better than the baseline. As per the human raters, the lowest improvement of proposed model was for the non-redundancy criteria which was 5% better than the baseline model. In totality the proposed model using Term frequency for text vector representation, outperforms all the other setups.

The chart in figure 4 indicates that for text compression in range of 20 to 60%, proposed TF model performs better than baseline. The proposed model's precision is adversely affected at other compression ratios.

As is evident from chart in figure 5, the proposed model is helpful to decrease the unique words (features) in given input original text. Also, it is seen that as the total words in text increases, more feature (dimension) reduction is possible by the proposed model.



e-ISSN: 2319-8753, p-ISSN: 2347-6710 <u>www.ijirset.com</u> | Impact Factor: 8.118

#### Volume 11, Issue 10, October 2022

DOI:10.15680/IJIRSET.2022.1110042

#### VII. CONCLUSION AND FUTURE WORK

On utilizing statistical and positional features, the proposed system was able to achieve significantly improved summary. By using the customized preprocessing pipeline considerable feature reduction was possible. The proposed system gives considerable improvement in recall in contrast to the baseline model. However, the proposed model showed marginal improvement of precision. Another noteworthy observation was the out-performance of TF over TFIDF mechanism for scoring of words in the stories dataset. To improvise the system further, addition of semantic and contextual knowledge to the proposed system can be explored.

The knowledge-based system would tend to be bound to the story domain and might not be adaptable to upcoming new requirements in the domain of storytelling. Hence other suitable methods which can support seamless adaptions to handle changes in story narration can be explored. In this regard, feasibility of machine learning, graphs and other techniques can be evaluated.

#### ACKNOWLEDGEMENT

We feel sense of high gratitude for the various Gujarati language speakers who selflessly helped us in extracting the gold summary and for evaluating the summaries generated by the system.

#### REFERENCES

- [1] P. Bhattacharyya, H. Murthy, S. Ranathunga, and R. Munasinghe, "Indic language computing," *Communications of the ACM*, vol. 62, no. 11, pp. 70–75, 2019.
- [2] N. P. Desai and V. K. Dabhi, "Resources and components for gujaratinlp systems: a survey," *Artificial Intelligence Review*, pp. 1–19, 2022.
- [3] W. C. Mann and S. A. Thompson, "Rhetorical structure theory: Toward a functional theory of text organization," *Text*, vol. 8, no. 3, pp. 243–281, 1988.
- [4] M. Gambhir and V. Gupta, "Recent automatic text summarization techniques: a survey," *Artificial Intelligence Review*, vol. 47, no. 1, pp. 1–66, 2017.
- [5] E. Lloret, L. Plaza, and A. Aker, "The challenging task of summary evaluation: an overview," *Language Resources and Evaluation*, vol. 52, no. 1, pp. 101–148, 2018.
- [6] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Text summarization branches out*. Barcelona, Spain: Association for Computational Linguistics, 2004, pp. 74–81.
- [7] "Document understanding conference," 2002. [Online]. Available: http://www-nlpir.nist.gov/projects/duc/.
- [8] N. Baruah, S. K. Sarma, and S. Borkotokey, "Text summarization in indian languages: A critical review," in 2019 Second International Conference on Advanced Computational and Communication Paradigms (ICACCP). Gangtok, India: IEEE, 2019, pp. 1–6.
- [9] P. Verma and A. Verma, "Accountability of nlp tools in text summarization for indian languages," *Journal of Scientific Research*, vol. 64, no. 1, 2020.
- [10] W. S. El-Kassas, C. R. Salama, A. A. Rafea, and H. K. Mohamed, "Automatic text summarization: A comprehensive survey," *Expert Systems with Applications*, vol. 165, p. 113679, 2020.
- [11] A. N. Gulati and S. Sawarkar, "A novel technique for multidocumenthindi text summarization," in 2017 International Conference on Nascent Technologies in Engineering (ICNTE), IEEE. Vashi, India: IEEE, 2017, pp. 1–6.
- [12] K. V. Kumar, D. Yadav, and A. Sharma, "Graph based technique for hindi text summarization," in *Information Systems Design and Intelligent Applications*. Kalyani, India: Springer, 2015, pp. 301–310.
- [13] C. Thaokar and L. Malik, "Test model for summarizing hindi text using extraction method," in 2013 IEEE Conference on Information & Communication Technologies, IEEE. Thuckalay, India: IEEE, 2013, pp. 1138–1143.
- [14] R. Rani and D. Lobiyal, "An extractive text summarization approach using tagged-lda based topic modeling," *Multimedia Tools and Applications*, vol. 80, no. 3, pp. 3275–3305, 2021.
- [15] A. Patel, T. Siddiqui, and U. Tiwary, "A language independent approach to multilingual text summarization," *Large scale semantic access to content (text, image, video, and sound)*, pp. 123–132, 2007.
- [16] N. Jhaveri, M. Gupta, and V. Varma, "clstk: The cross-lingual summarization tool-kit," in *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*. Melbourne, Australia: ACM, 2019, pp. 766–769.



e-ISSN: 2319-8753, p-ISSN: 2347-6710 www.ijirset.com | Impact Factor: 8.118

#### Volume 11, Issue 10, October 2022

| DOI:10.15680/IJIRSET.2022.1110042 |

- [17] A. Kazantseva and S. Szpakowicz, "Summarizing short stories," *Computational Linguistics*, vol. 36, no. 1, pp. 71–109, 2010.
- [18] Wikipedia, "the free encyclopedia," 2021. [Online]. Available: https://en.wikipedia.org/wiki/List\_of\_loanwords\_in\_Gujarati
- [19] "1000 mostcommonwords.com," 2021. [Online]. Available: https://1000mostcommonwords.com/1000-most-common-gujarati-words
- [20] S. Babu, "Gujarati-english learner's dictionary," 2021. [Online]. Available: https://ccat.sas.upenn.edu/plc/gujarati/gujengdictionary. pdf
- [21] N. Desai and B. Dalwadi, "An affix removal stemmer for gujarati text," in 2016 3rd international conference on computing for sustainable global development (INDIACom). IEEE, 2016, pp. 2296–2299.
- [22] "matrubharti.com," 2021. [Online]. Available: https://www.matrubharti.com [23] S. Joshi, https://swatisjournal.com/, 2020.
- [24] FIRE, http://fire.irsi.res.in/fire/static/data, 2020.
- [25] J. L. Fleiss, "Measuring nominal scale agreement among many raters." Psychological bulletin, vol. 76, no. 5, p. 378, 1971.
- [26] A. Kunchukuttan, "The indicnlp library," 2020. [Online]. Available: https://github.com/anoopkunchukuttan/indic\_nlp\_library/blob/ master/docs/indicnlp.pdf
- [27] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.





Impact Factor: 8.118





## INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY





www.ijirset.com