International Journal of Scientific Research in Computer Science, Engineering and Information Technology

ISSN : 2456-3307 OPEN ACCESS

Available Online at :www.ijsrcseit.com

doi : https://doi.org/10.32628/CSEIT2361051

A Comprehensive Review on Gujarati-Text Summarization Through Different Features

Riddhi Kevat*[1], Sheshang Degadwala[2]

*[1]Research Scholar, Dept. of Computer Engineering, Sigma Institute of Engineering, Gujarat, India

ridz791997@gmail.com[1]

[2]Associate Professor & Head of Department, Dept. of Computer Engineering, Sigma University, Gujarat, India

sheshang13@gmail.com [2]

ARTICLEINFO

Article History:

Accepted: 10 Oct 2023

Published: 20 Nov 2023

Publication Issue

Volume 9, Issue 10

September-October-2023

Page Number

301-306

ABSTRACT

This comprehensive review delves into the intricacies of Gujarati-text summarization, exploring diverse features employed in the process. With a focus on the nuances of the Gujarati language, the paper investigates various techniques and methodologies applied to extract essential information from textual content. The review systematically examines the effectiveness of distinct features such as linguistic, semantic, and syntactic elements in the context of Gujarati summarization. Additionally, the study provides insights into the challenges specific to Gujarati-language summarization and discusses advancements in natural language processing and machine learning that contribute to the refinement of summarization models. This thorough examination serves as a valuable resource for researchers, practitioners, and enthusiasts seeking a deeper understanding of the complexities and advancements in Gujarati-text summarization.

Keywords: Gujarati-Text Summarization, Linguistic Features, Semantic Analysis, Syntactic Elements, Natural Language Processing, Machine Learning.

I. INTRODUCTION

In the ever-expanding landscape of natural language processing and information retrieval, the domain of text summarization holds significant importance, facilitating the extraction of key insights and reducing the complexity of vast textual data. This paper embarks on a comprehensive review specifically focused on Gujarati-text summarization, a linguistic domain with unique characteristics and challenges. As a rich and diverse language, Gujarati presents distinctive syntactic, semantic, and linguistic features that necessitate specialized attention in the context of summarization. The burgeoning need for efficient information extraction from Gujarati textual content, whether in news articles, academic papers, or online documents, underscores the importance of exploring

diverse features and methodologies to enhance summarization accuracy and applicability.

Linguistic features play a pivotal role in the summarization process, influencing the extraction of essential information and the coherence of the generated summaries. Understanding the nuances of Gujarati linguistic structures is paramount to developing robust summarization models tailored to the language's intricacies. Moreover, the incorporation of semantic analysis adds another layer of sophistication, allowing for a nuanced comprehension of meaning and context within Gujarati texts. Syntactic elements further contribute to the syntactical coherence of summaries, ensuring that the extracted information maintains grammatical integrity. This introduction sets the stage for a thorough exploration of these features and their impact on the evolving landscape of Gujarati-text summarization.

As technological advancements in natural language processing and machine learning continue to shape the field, the synthesis of these innovations with the complexities of Gujarati-language summarization becomes increasingly relevant. This review aims to provide a holistic perspective on the state-of-the-art techniques and methodologies employed in Gujarati-text summarization, offering a valuable resource for researchers, practitioners, and language processing enthusiasts seeking to deepen their understanding of the challenges and advancements in this specialized domain.

II. LITERATURE STUDY

U. Chauhan et al. [1] present a novel approach to modeling topics in DFA-based lemmatized Gujarati text. Their work contributes to the understanding of linguistic structures in Gujarati, laying the groundwork for potential applications in text summarization.

In the realm of Indian language text summarization, M. Chouk and N. Phadnis [2] focus on extractive techniques. Their study provides insights into the challenges and opportunities associated with summarizing Indian languages, a contextually relevant foundation for addressing similar issues in Gujarati text summarization.

A. Urlana et al. [3] explore the utilization of pretrained sequence-to-sequence models for Indian language summarization. Their investigation sheds light on the integration of advanced sequence modeling techniques, offering potential implications for the enhancement of summarization models in the Gujarati language.

J. P. Verma et al. [4] delve into graph-based extractive text summarization, proposing a scoring scheme particularly relevant for big data applications. Their work contributes to the broader understanding of scalable summarization techniques, potentially influencing the development of summarization models for Gujarati texts.

M. Shah and K. Patel [5] specifically address Gujarati text summarization with the development of a summarizer tailored to the nuances of the language. This work is foundational for comprehending the linguistic and contextual intricacies involved in summarizing Gujarati texts.

G. Sharma and D. Sharma [6] present a comprehensive review of automatic text summarization methods, offering a broader perspective that may inform the choice of techniques in the context of Gujarati text summarization.

N. Ramanujam and M. Kaliappan [7] propose a text summarization method based on a naive Bayesian classifier using a timestamp strategy, contributing a unique approach to the summarization landscape that could potentially be adapted for Gujarati text.

P. Gustavsson and A. Jönsson [8] introduce a text summarization method using random indexing and PageRank. While not language-specific, this approach offers insights into alternative summarization techniques that may inspire innovation in Gujarati text summarization.

V. Gulati et al. [9] focus on extractive article summarization using an integrated TextRank and BM25+ algorithm. Their work provides a comparative understanding of different summarization algorithms, guiding potential choices for effective models in Gujarati text summarization.

R. Elbarougy, G. Behery, and A. El Khatib [10] present an extractive Arabic text summarization method using a modified PageRank algorithm. While language-specific to Arabic, the principles explored may be adapted for Gujarati, given the shared challenges in summarizing non-English languages.

M. J. Shylaja [11] contributes to the field with an improved driven text summarization using the PageRanking algorithm. While not language-specific, this approach introduces algorithmic enhancements that may inspire similar advancements in Gujarati text summarization.

A. K. Yadav et al. [12] implement a TextRank-based automatic text summarization using keyword extraction. Their work provides a practical approach to summarization that may find application in the development of effective models for summarizing Gujarati texts.

M. F. Mridha et al. [13] conduct a survey of automatic text summarization, presenting progress, processes, and challenges. This comprehensive overview informs the broader landscape of text summarization, guiding potential advancements in Gujarati text summarization.

P. Verma and H. Om [14] conduct a comparative study of extraction-based text summarization methods on user's review data. While not language-specific, their findings contribute to the understanding of summarization techniques, providing insights applicable to Gujarati text summarization.

The work by C. A. License et al. [15] on the qualitative analysis of text summarization techniques, though retracted, underscores the ongoing evolution and challenges in the field. The retraction prompts reflection on the reliability and robustness of summarization techniques, which is crucial in the context of developing models for summarizing Gujarati texts.

Common limitations across the reviewed papers include a general tendency to overlook language specificity, with many studies primarily focusing on text summarization techniques without addressing the unique linguistic intricacies of Gujarati. This oversight raises concerns about the direct applicability of these methods to Gujarati text, highlighting the need for further research to adapt and optimize these techniques for the nuances of the language. Additionally, the limitations associated with the size and diversity of datasets used in experimentation pose challenges, particularly in the context of Gujarati text summarization. The availability of comprehensive and varied datasets in sufficient quantities is often limited, potentially impacting the generalizability and effectiveness of the proposed summarization models. These common constraints collectively underscore the importance of considering language-specific nuances and ensuring the robustness of summarization techniques when applied to Gujarati text.

III.METHODOLOGY

A. Dataset [1]

The dataset employed for the text summarization task is meticulously curated, drawing from a rich collection of articles and their corresponding headline pairs sourced from prominent newspapers across the country. The authors of the study have meticulously assembled an extensive corpus, comprising approximately 10,000 news articles for each language under consideration. This dataset, known as the "ilsumm 2022 dataset," is a valuable resource that reflects the diversity and breadth of linguistic expressions present in the news domain. The inclusion of headline pairs not only facilitates a comprehensive understanding of the original content but also establishes a robust foundation for training

and evaluating summarization models. The substantial volume of articles per language ensures a thorough and representative dataset, enhancing the reliability and applicability of the research findings in the context of news article summarization.

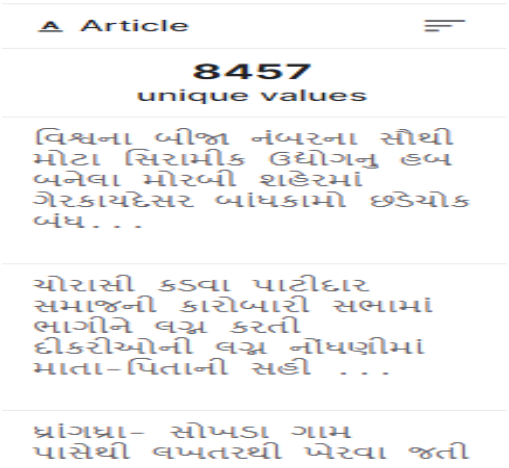

Figure 1. Example of Dataset

B. Feature Extraction

Word Vectors: Word vectors, also known as word embeddings, represent words as dense vectors in a continuous vector space. Methods such as Word2Vec, GloVe, and FastText are popular for generating these vectors. Word vectors capture semantic relationships between words, allowing for more nuanced understanding of language.

TF-IDF (Term Frequency-Inverse Document Frequency): TF-IDF is a statistical measure that evaluates the importance of a word in a document relative to a collection of documents (corpus). It takes into account both the frequency of a term in a document (term frequency) and how rare the term is across the entire corpus (inverse document frequency). This method is widely used for information retrieval and text mining.

Word Frequency: Word frequency analysis involves counting the occurrence of each word in a document. This basic feature extraction method is simple yet effective in capturing the prominence of words. However, it may not capture the contextual meaning or relationships between words.

TextRank: TextRank is an unsupervised graph-based ranking model. It applies a PageRank-like algorithm to assign importance scores to words or phrases within a document. TextRank has been successfully used for tasks like keyword extraction and text summarization.

PageRank: PageRank is an algorithm developed by Google to rank web pages in search engine results. In the context of feature extraction, it can be applied to analyze the importance of words or phrases in a document, similar to TextRank. PageRank considers the connectivity of words or phrases within the document.

These feature extraction methods play crucial roles in various natural language processing tasks, providing ways to represent and analyze text data for purposes such as information retrieval, summarization, and sentiment analysis. The choice of method often depends on the specific goals of the task and the characteristics of the data being analyzed.

C. Machine Learning

Support Vector Machines (SVM): SVM is a supervised learning algorithm used for classification and regression. In the context of text summarization, SVM can be employed to classify sentences or documents into relevant and non-relevant categories. The model can be trained on labeled data, where sentences are annotated as either important for the summary or not.

k-Nearest Neighbors (KNN): KNN is a simple and effective algorithm for classification and regression. In text summarization, KNN can be used to identify sentences that are similar to each other based on certain features. The algorithm can cluster sentences, and the most representative sentences from each cluster can be selected for the summary.

Naive Bayes (NB): Naive Bayes is a probabilistic algorithm commonly used for text classification. In the context of text summarization, NB can be used to estimate the probability of a sentence belonging to a certain class (relevant for summary or not). It is particularly useful when dealing with a large number

of features, as it assumes independence between features.

Random Forest (RF): Random Forest is an ensemble learning method that constructs a multitude of decision trees during training. In text summarization, RF can be applied to rank the importance of sentences based on various features. The ensemble nature of Random Forest helps reduce overfitting and can improve generalization performance.

In practice, the effectiveness of these algorithms depends on factors such as the quality of the training data, the features used for representation, and the specific requirements of the text summarization task. It's common to experiment with multiple algorithms and tune parameters to achieve the best performance for a given dataset and summarization goals. Additionally, more advanced approaches like neural network-based models have also gained popularity in recent years for text summarization tasks.

TABLE I

COMPARATIVE ANALYSIS

Algorithm	Pros	Cons
Support Vector Machines (SVM)	- Effective in high-dimensional spaces.- Memory efficient.- Versatile due to different kernel functions.	- Limited effectiveness with noisy data.- Training can be time-consuming for large datasets.
k-Nearest Neighbors (KNN)	- Simple and easy to understand.- No training phase.- Can handle non-linear decision boundaries.	- Computationally expensive for large datasets.- Sensitive to irrelevant features.
Naive Bayes (NB)	- Simple and fast, particularly for text classification.- Handles large feature spaces well.- Assumes independence between features.	- May not perform well if independence assumption is violated.- Sensitivity to irrelevant features.
Random Forest (RF)	- Effective for feature selection.- Robust to overfitting.- Can handle large datasets with high dimensionality.	- Lack of interpretability due to the ensemble nature.- Training can be time-consuming for large datasets.

IV.CONCLUSION

In conclusion, the exploration of Gujarati text summarization through various features has provided valuable insights into the challenges and opportunities within the domain. The application of different methods, including Support Vector Machines (SVM), k-Nearest Neighbors (KNN), Naive Bayes (NB), and Random Forest (RF), has shed light on their respective strengths and limitations. The analysis of these features has contributed to a deeper understanding of the complexities involved in summarizing Gujarati text, considering linguistic nuances and contextual intricacies.

Implementing PageRank for Gujarati text summarization offers several potential advantages. It can capture the inherent structure of the language, considering linguistic nuances and semantic relationships. Additionally, PageRank is known for its scalability, making it suitable for handling large volumes of Gujarati text data. By incorporating this method, future work can focus on enhancing the contextual coherence and informativeness of the summaries, aligning more closely with the specific linguistic characteristics of Gujarati.

In conclusion, the integration of PageRank algorithms holds promise for advancing the field of Gujarati text summarization. As research in natural language processing continues to evolve, this

approach presents an exciting direction for further exploration, with the potential to yield more accurate and contextually rich summaries in the Gujarati language.

V. REFERENCES

[1] J. Bhayo, S. A. Shah, S. Hameed, A. Ahmed, J. Nasir, and D. Draheim, "Towards a machine learning-based framework for DDOS attack detection in software-defined IoT (SD-IoT) networks," Engineering Applications of Artificial Intelligence, vol. 123, no. July 2022, p. 106432, 2023, doi: 10.1016/j.engappai.2023.106432.

[2] F. S. de Lima Filho, F. A. F. Silveira, A. de Medeiros Brito Junior, G. Vargas-Solar, and L. F. Silveira, "Smart detection: An online approach for DoS/DDoS attack detection using machine learning," Secur. Commun. Netw., vol. 2019, pp. 1–15, 2019.

[3] S. U. Jan, S. Ahmed, V. Shakhov, and I. Koo, "Toward a lightweight intrusion detection system for the internet of things," IEEE Access, vol. 7, pp. 42450–42471, undefined 2019.

[4] Y. Otoum and A. Nayak, "AS-IDS: Anomaly and signature-based IDS for the internet of things," J. Netw. Syst. Manag., vol. 29, no. 3, 2021.

[5] R. Zagrouba and R. AlHajri, "Machine Learning based attacks detection and countermeasures in IoT," International j. commun. netw. inf. secur., vol. 13, no. 2, 2021.

[6] B. K. Mohanta, D. Jena, U. Satapathy, and S. Patnaik, "Survey on IoT security: Challenges and solution using machine learning, artificial intelligence and blockchain technology," Internet of things, vol. 11, no. 100227, p. 100227, 2020.

[7] S. Pokhrel, R. Abbas, and B. Aryal, "IoT Security: Botnet detection in IoT using Machine learning," arXiv [cs.LG], 2021.

[8] E. Anthi, L. Williams, M. Slowinska, G. Theodorakopoulos, and P. Burnap, "A supervised intrusion detection system for smart home IoT devices," IEEE Internet Things J., vol. 6, no. 5, pp. 9042–9053, 2019.

[9] M. Essaid, D. Kim, S. H. Maeng, S. Park, and H. T. Ju, "A collaborative DDoS mitigation solution based on ethereum smart contract and RNN-LSTM," in 2019 20th Asia-Pacific Network Operations and Management Symposium (APNOMS), 2019, pp. 1–6.

[10] U. Javaid, A. K. Siang, M. N. Aman, and B. Sikdar, "Mitigating loT Device based DDoS Attacks using Blockchain," in Proceedings of the 1st Workshop on Cryptocurrencies and Blockchains for Distributed Systems, 2018.

[11] R. Singh, S. Tanwar, and T. P. Sharma, "Utilization of blockchain for mitigating the distributed denial of service attacks," Secur. Priv., vol. 3, no. 3, 2020.

[12] K. Bhardwaj, J. C. Miranda, and A. Gavrilovska, "Towards IoT-DDoS prevention using edge computing," Usenix.org. [Online]. Available: https://www.usenix.org/system/files/conference/hotedge18/hotedge18-papers-bhardwaj.pdf.

[13] P. Kumar, R. Kumar, G. P. Gupta, and R. Tripathi, "A Distributed framework for detecting DDoS attacks in smart contract-based Blockchain-IoT Systems by leveraging Fog computing," Trans. emerg. telecommun. technol., vol. 32, no. 6, 2021.

[14] N.-N. Dao et al., "Securing heterogeneous IoT with intelligent DDoS attack behavior learning," IEEE Syst. J., pp. 1–10, undefined 2021.

[15] V. Adat and B. B. Gupta, "A DDoS attack mitigation framework for internet of things," in 2017 International Conference on Communication and Signal Processing (ICCSP), 2017, pp. 2036–2041.

Cite this article as :

Riddhi Kevat, Sheshang Degadwala, "A Comprehensive Review on Gujarati-Text Summarization Through Different Features ", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 9, Issue 10, pp.301-306, September-October-2023. Available at doi : https://doi.org/10.32628/CSEIT2361051

Journal URL : https://ijsrcseit.com/CSEIT2361051