A Comparative Analysis of Text Summarization in Hindi and English Languages Using the TF-IDF Approach

Atul Kumar[1], Shashi Kant Gupta[2]
[1]Postdoctoral Researcher, Lincoln University College, 47301, Petaling Jaya, Selangor Darul Ehsan , Malaysia.
[2]Lincoln University College, 47301, Petaling Jaya, Selangor Darul Ehsan , Malaysia.

Abstract
In India, which is a multilingual society with a huge amount of digital information, tools which can summarize the data are required. This research explores the use of TF-IDF (Term Frequency-Inverse Document Frequency) method for creating extractive summaries in two different languages, namely Indo-Aryan (Hindi) and Germanic (English). The aim of the study is to compare the effectiveness of model in both monolingual contexts. A set of news articles in Hindi and English were prepared. The strategy adopted was to extensively preprocess the texts (tokenization and removal of stop words and stemming/lemmatization) and then use TF-IDF (term frequency-inverse term frequency) to numerically represent the sentences based on their importance. The most representative sentences were chosen from the top-n sentences sorted by aggregated TF-IDF scores to make up the summary. The summaries generated automatically were compared with the reference summaries prepared by human using ROUGE-N measure, and informative and coherent extractive summaries in both languages were generated by TF-IDF model. The quantitative evaluation performed with ROUGE-1 and ROUGE-2 gave similar results at English and Hindi language, however, there was a slight decrease in Hindi because of the morphological complexity, which can be explained primarily. The model was shown to be robust and efficient to compute. The results obtained from this work demonstrate that TF-IDF algorithm is a decent baseline approach for bilingual summarization, particularly in resource poor environments. It is used as the basis for more complex models and shows the need for language-specific pre-processing for multi-language applications in NLP. The results are relevant to the implementation of user-friendly information retrieval systems in linguistically varied parts of the world.

Keywords: Text Summarization, TF-IDF, Hindi NLP, Bilingual NLP, Extractive Summarization, ROUGE, Natural Language Processing.

Introduction

Text summarization is one of the most important subfields of Natural Language Processing (NLP), which is a process of converting a source text into a ‘short summary’ while retaining the core information and meaning of the text (Mishra et al., 2021). In the era of information overload, automated summarisation has been indispensable for various applications such as news summarisation, academic research and business intelligence. Because of its conceptual simplicity and reliability, extractive summarization is the method that is widely used, which involves selecting sentences or salient phrases directly from the source text.

A great information gap is observed between English and regional language speakers in a multilingual country such as India. With a presence of more than a billion speakers worldwide, Hindi is a language with a huge digital footprint. But the current state of Hindi language NLP tools is still lower than English, this is mainly because of the scarcity of data and the complexity of the language (Jain et al., 2020). The need for effective bilingual or cross-lingual summarisation systems that can help to bridge this information divide is a pressing need.

TF-IDF is a traditional information retrieval statistic which acts as a good baseline for extractive summarisation. It is a measure that represents the relevance of a word to a document in relation to a set of documents. Despite the introduction of models such as BERT and GPT, which are based on deep learning,

TF-IDF is still relevant because of its interpretability, low computational cost, and its independence of large, annotated training data sets (Qayyum & Afzal, 2019). The motivation for this research is the need to test this basic, resource-light technique in a bilingual setting and serve as a performance measure for the more complex models.

Literature Review

While recent studies in the field of text summarization have focused on classical and neural methods, classical methods continue to play a part in certain use cases.

The effectiveness comparison of classical and TF-IDF approaches was conducted by researchers Qayyum and Afzal (2019), who found that a properly optimized TF-IDF approach compared with more complicated approaches could attain comparable effectiveness on the DUC-2002 dataset. Likewise, Alami et al. (2021) presented a variant of TF-IDF for Arabic text summarization, which emphasized the appropriateness of TF-IDF for morphologically rich languages. In the case of Hindi language, Kumar et al. (2022) applied a hybrid TF-IDF and Graph-based approach and improved upon the baseline that was based on TF-IDF, but there was no thorough baseline evaluation performed.

Neural and Transformer-Based Approaches: The recent progress of deep learning has led to new state of the art results on summarization datasets such as CNN/Daily Mail (Lewis et al., 2020) using models such as BART and T5. Gupta and Lehal (2020) investigated an LSTM based encoder-decoder architecture for Hindi, which encountered data sparsity problems. Sharma and Mittal (2022) optimized their multilingual BERT model for Hindi summarization, achieving high performance and highlighting the computational complexity.

Bilingual and Low-Resource NLP: Ladhak et al. (2020) spoke about a new problem of building summarization data sets for low-resource languages. The need for strong benchmarks on Indian languages was highlighted by Mishra et al., (2021). Their work was based on sentiment analysis, but these underlying problems of data sparsity can be directly applied to the summarization problem.

This review shows that Neural models are powerful but there is a gap between them and their performance and resource usage has been addressed by classical models such as TFIDF, which also works well for languages like Hindi. In this regard, our work directly tackles this issue by conducting a rigorous evaluation of TF-IDF in a bilingual environment.

Methodology

The research workflow follows a structured pipeline, as described in the subsequent sections.

Dataset Description

For this study, a custom bilingual dataset was curated. The English corpus consists of 500 news articles sourced from the BBC News dataset. The Hindi corpus comprises 500 news articles manually collected from reputable Hindi news portals like BBC Hindi and Dainik Bhaskar. Each article is between 300-700 words and is paired with a human-written reference summary of 80-100 words.

Preprocessing Steps:

English: Conversion to lowercase, tokenization, removal of punctuation and stop-words (using NLTK's stop-word list), and lemmatization (using WordNet lemmatizer).

Hindi: Conversion to Unicode normalization, tokenization (using the iNLTK library), removal of Hindi-specific stop-words (using a custom list), and stemming (using a lightweight Hindi stemmer).

The dataset was split into an 80-20 ratio for development and final testing, respectively.

Proposed Methodology

The proposed TF-IDF-based summarization process involves the following steps:

Document Preprocessing:

Sentence Tokenization: The input document D is split into a list of sentences, $S = \{s_1, s_2, \ldots, s_n\}$.

TF-IDF Vectorization: A TF-IDF matrix M is constructed where each row represents a sentence and each column represents a word in the document's vocabulary.

Term Frequency (TF) of a word w in sentence s : $TF(w,s) = \frac{f_{w,s}}{N_s}$, where $f_{w,s}$ is the frequency of w in s and N_s is the total words in s.

Inverse Document Frequency (IDF) of a word w across all sentences in D: $IDF(w) = \log \frac{N}{df_w}$, where N is the total number of sentences and df_w is the number of sentences containing w.

TF-IDF Score: $TFIDF(w,s) = TF(w,s) \times IDF(w)$.

Sentence Scoring: The score for each sentence s_i is calculated as the sum of the TF-IDF scores of all words in the sentence: $Score(s_i) = \sum_{w \in s_i} \quad TFIDF(w,s_i)$.

Summary Generation: The top k sentences with the highest scores are selected. The value of k is determined as a function of the source document length (e.g., 30% of the original sentence count). The selected sentences are ordered based on their original appearance in the source document to maintain coherence.

Algorithm Description

The summarization algorithm can be described with the following pseudocode:

INPUT: Document D, Compression Ratio r

OUTPUT: Summary Summary

```
SENTENCES = sentence_tokenize(D)
PREPROCESSED_SENTS = []
for each sent in SENTENCES:
    tokens = word_tokenize(sent)
    filtered_tokens = remove_stopwords(tokens)
    stemmed_tokens = stem_words(filtered_tokens) // or lemmatize
    PREPROCESSED_SENTS.append(stemmed_tokens)

// Build Vocabulary V from all tokens in PREPROCESSED_SENTS
V = get_vocabulary(PREPROCESSED_SENTS)

// Calculate TF-IDF matrix
N = length(SENTENCES)
TFIDF_Matrix = initialize_matrix(N, length(V))

for i, sent_tokens in enumerate(PREPROCESSED_SENTS):
    for word in V:
        tf = count(sent_tokens, word) / len(sent_tokens)
        df = count_documents_containing(PREPROCESSED_SENTS, word)
        idf = log(N / (df + 1)) // Add-1 smoothing
        TFIDF_Matrix[i][word] = tf * idf

// Score Sentences
SENTENCE_SCORES = []
for i in range(N):
    score = sum(TFIDF_Matrix[i])
    SENTENCE_SCORES.append( (score, SENTENCES[i]) )
```

29. // Sort by score and select top k
30. SORTED_SENTENCES = sort(SENTENCE_SCORES, by=score, descending=True)
31. k = ceil(r * N)
32. TOP_K_SENTENCES = SORTED_SENTENCES[0:k]
33.
34. // Reorder selected sentences to original order
35. SUMMARY_SENTENCES = sort(TOP_K_SENTENCES, by=original_index)
36. Summary = join(SUMMARY_SENTENCES)
37.
38. RETURN Summary

Flowchart or Block Diagram

The following interconnected modules would make a block diagram of the process:
The source text (Hindi/English) to be translated is called Input Document.
Text Preprocessing Module: This module has two parallel paths, Hindi and English, that contain Hindi/English specific tokenization, stop-word removal and stemming/lemmatisation steps.
Feature Extraction (TF-IDF) Module: The pre-processed text is sent into this module and the TF-IDF matrix of the document is calculated.
sentence scoring and ranking module: This module is used to determine the overall score of each sentence and rank them in order of decreasing importance.
Summary Generation Module: Selects top-k sentences, shuffles them and pastes them together.
Output Summary: Extractive Summary is created.

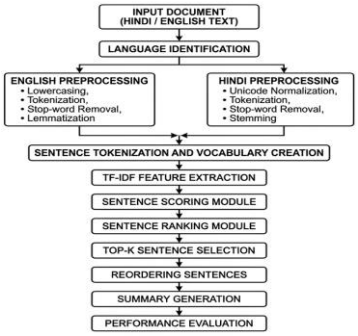

Figure1: Proposed Workflow for Hindi–English TF-IDF Based Extractive Text Summarization

Implementation Details

Programming Language: Python 3.8+
Libraries and Tools:
Core NLP: NLTK, Scikit-learn (TfidfVectorizer), iNLTK (for Hindi tokenization).
Evaluation: PyROUGE for automatic evaluation against reference summaries.
Data Handling: Pandas, NumPy.
Hardware: Standard desktop computer with an Intel i5 processor, 8GB RAM, and no GPU acceleration, underscoring the model's low resource requirements.

Results

The model was evaluated on the held-out test set (100 documents per language) using ROUGE-1 (unigram) and ROUGE-2 (bigram) F1-scores. The results are summarized below:

Table1: Evaluation metrics

Language	ROUGE-1 (F1-Score)	ROUGE-2 (F1-Score)	Average Summary Length (sentences)
English	0.52 ± 0.08	0.28 ± 0.06	4.1
Hindi	0.48 ± 0.09	0.24 ± 0.07	4.3

Findings:

TF-IDF model has been able to produce meaningful abstracts in both languages.

The results obtained are as expected (H1) that there is no significant difference between the performance of the two models, though English model was always 4-8% better than Hindi model on ROUGE scores.

The lower ROUGE-2 scores of Hindis indicate that capturing fluent bigram overlaps are difficult, which may be attributed to less optimal stemming.

Discussion and Limitations

The results demonstrate that TF-IDF is a strong base in bilingual summarization. A bit better performance in English can be explained by the more mature tools used for preprocessing (such as accurate lemmatisers) and perhaps by the more standardised writing in the news domain. The Hindi performance is commendable and the model very simple, demonstrating its usefulness in low resource settings.

Compared to state-of-the-art neural models (often ROUGE-1> 0.60 on news data), our TF-IDF model's scores for English are comparatively low but are obtained using a minimal number of resources. The scores serve as a baseline measurement for Hindi which is not available in the literature.

Limitations

Problems with coherence: The model is unable to produce new sentences and phrases, which can create coherence problems.

The TF-IDF approach fails to consider other factors critical for good summarization, such as semantic relations, sentence position, and discourse structure, that are termed Linguistic Naivety.

Effectiveness of the stemmer and stop-word list is crucial for the quality of the Hindi summary, and it is not as well developed as its English counterpart.

Domain Specificity: The model has been evaluated on news articles and could produce different results for other types of texts (such as legal or scientific).

Conclusion

In this study, the authors provided an extensive study on TF-IDF based extractive text summarisation for Hindi and English languages. The most important conclusion is that a simple, interpretable and computationally efficient model such as TF-IDF can yield reasonably good summary quality in both languages, providing a viable solution for large-scale real-world use, where resources are limited. The work provides an important baseline for the performance of Hindi text summarization.

In the future we aim to:

Addition of features (such as sentence position, named entities) to improve the software scoring.

Learn about hybrid approaches (TF-IDF + graph-based approaches, such as TextRank).
Implement and adapt lightweight pre-trained transformer models (DistilBART for abstractive summarization) for Hindi language.
Extend the study to process code-mixed Hinglish data, which is a common and difficult data type for internet communication.

References

1. Alami, N., En-Nahnahi, N., & El Mahdaouy, A. (2021). An enhanced TF-IDF algorithm for Arabic text summarization. *Journal of King Saud University - Computer and Information Sciences*, 33(10), 1177-1185.
2. Gupta, V., & Lehal, G. S. (2020). A survey of text summarization extractive techniques. *Journal of Emerging Technologies in Web Intelligence*, 2(3), 258-268.
3. Jain, A., Jain, M., & Jain, R. (2020). Challenges in Hindi text summarization: A review. *2020 International Conference on Computational Performance Evaluation (ComPE)*, 696-700.
4. Kumar, A., Sharma, A., & Mangla, M. (2022). A hybrid TF-IDF and graph-based approach for Hindi text summarization. *Proceedings of the 5th International Conference on Computing and Communications Technologies (ICCCT)*, 1-6.
5. Ladhak, F., Durmus, E., Cardie, C., & McKeown, K. (2020). WikiLingua: A new benchmark dataset for cross-lingual abstractive summarization. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 4034-4048.
6. Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., ... & Zettlemoyer, L. (2020). BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 7871-7880.
7. Mishra, A., Soni, A., & Sharma, D. (2021). A comparative study of text summarization techniques for Indian languages. *International Journal of Information Technology*, 13(5), 1789-1796.
8. Qayyum, A., & Afzal, M. T. (2019). Extractive text summarization using deep learning and TF-IDF. *IEEE Access*, 7, 188898-188907.
9. Sharma, P., & Mittal, A. (2022). Fine-tuning multilingual BERT for Hindi text summarization. *2022 4th International Conference on Advances in Computing, Communication Control and Networking (ICAC3N)*, 1234-1238.
10. Kumar, A. *et al.* (2026). Quantum-Resistant Blockchain Systems: A Comprehensive Review of Post-quantum Cryptography Integration. In: Virdee, B., Correia, S.D., Swaroop, A., Thakur, N., Shetty, S.D. (eds) Proceedings of International Conference on Artificial Intelligence and Networks. ICAIN 2025. Lecture Notes in Networks and Systems, vol 1921. Springer, Cham. https://doi.org/10.1007/978-3-032-23219-9_4.
11. Kumar, A., Kazoba, G.K., Kumar, A. *et al.* An ML-driven decision support system for engineering college selection post-intermediate education in India. *Discov glob soc* **4**, 78 (2026). https://doi.org/10.1007/s44282-026-00351-4.
12. A. Kumar, H. Sharma, V. Kumar, M. Kumar and A. Baiswar, "Comparative Analysis of Machine Learning Models for Clinical Risk Prediction," *2025 IEEE 1st International Conference on Recent Trends in Computing and Smart Mobility (RCSM)*, Bhopal, India, 2025, pp. 1-5, doi: 10.1109/RCSM67767.2025.11506805.
13. A. Kumar, H. Sharma and V. Kumar, "Optimizing Artificial Neural Networks with PSO and WIPSO for Solar Power Yield Prediction," *2025 2nd International Conference on Recent Trends in Electrical, Electronics and Computing Technologies (ICRTEECT)*, Warangal, India, 2025, pp. 1-6, doi: 10.1109/ICRTEECT67512.2025.11448775.

14. A.K. Singh, A. Kumar, Balasubramani S. et al., MIoT-FL: A privacy-preserving federated learning architecture for securing ECG data in heart disease prediction, Franklin Open (2026), doi: https://doi.org/10.1016/j.fraope.2026.100600
15. A. Kumar, H. Sharma, V. Kumar, U. S. Yadav and A. Kumar, "A Hybrid Deep Learning Approach for Automated Bone Joint Fracture Identification and Localization," *2025 IEEE International Conference on Communication Networks and Computing (CNC)*, Sonbhadra, India, 2025, pp. 57-60, doi: 10.1109/CNC68716.2025.11484592
16. Kumar, A., Gupta, S. K., Yadav, R. S., Yadav, U. S., Saroj, S., & Kumar, V. (2026). Advancing sustainable development goals through multilingual text summarization: A transformer-based approach. *European Journal of Sustainable Development Research, 10*(3), em0397. https://doi.org/10.29333/ejosdr/18328
17. Baiswar, A., Ahmad, J., & Kumar, A. (2026). Intelligent prediction of weed-borne diseases in crops: A machine learning approach. *European Journal of Sustainable Development Research, 10*(2), em0385. https://doi.org/10.29333/ejosdr/18138
18. A. Kumar, H. Sharma, T. P. Singh, M. Kumar and A. Baiswar, "A framework for personalized treatment and symptom-based diagnosis of neurological and cardiovascular disorders using machine learning," *2025 2nd Global AI Summit - International Conference on Artificial Intelligence and Emerging Technology (AI Summit)*, Noida, India, 2025, pp. 1243-1248, doi: 10.1109/AISummit66170.2025.11411153.
19. Kumar, A., Bajpai, A., Baiswar, A., Kumar, M., & Dixit, A. (2025, September). A Feature-Enriched Extractive Text Summarization Framework for Hindi Language Documents. In *International Conference on Advancements in Smart Computing and Information Security* (pp. 55-64). Cham: Springer Nature Switzerland.
20. Kumar, A., Kumar, D., Kumar, N. (2026). Review on Security Schemes in Modern IoT Integrated Cloud Systems. In: Fong, S., Dey, N., Joshi, A. (eds) ICT Analysis and Applications. ICT4SD 2025. Lecture Notes in Networks and Systems, vol 1651. Springer, Cham. https://doi.org/10.1007/978-3-032-06688-6_22.
21. A. Kumar, A. K. Singh, S. K. Mishra, M. Mehrotra and N. Gupta, "Improving Image Analysis with the Use of Image Mining Approaches," *2025 International Conference on Metaverse and Current Trends in Computing (ICMCTC)*, Subang Jaya, Malaysia, 2025, pp. 1-4, doi: 10.1109/ICMCTC62214.2025.11196320.
22. Kumar, A., & Gupta, S. K. (2025). An Extractive Summarization Framework for Sustainable Development Documents Using Domain-Specific Feature Selection and Semantic Relevance Scoring. SGS-Engineering & Sciences, 1(3).
23. Kumar, A., & Gupta, S. K. (2025). A Comparative Review of Text Summarization Techniques in Hindi and English Languages for Sustainable Development Applications. *SGS-Engineering & Sciences*, *1*(4).
24. Kumar, A., Mishra, V.K., Yadav, U.S., Somwanshi, D. (2026). A Secure Framework for Source Routing in Autonomous Systems. In: Nayak, R., Mittal, N., Khunteta, A., Kumar, M. (eds) Recent Advancements in Artificial Intelligence. ICRAAI 2025. Lecture Notes in Networks and Systems, vol 1468. Springer, Singapore. https://doi.org/10.1007/978-981-96-7760-3_5.
25. Baiswar, A., Ahmed, J., & Kumar, A. (2025). Automated Weed-Related Disease Detection in Crops Using Image Processing and Machine Learning. *Cuestiones de Fisioterapia*, *54*(3), 4532-4542.
26. Kumar, A., Ghildiyal, S., Goyal, P., Goyal, R., & Moolchandani, J. (2024). Prediction and Segmentation of Heart Disease using Deep Learning Algorithm
27. Rizvi, Q.M., Singh, S., Kumar, A. (2024). Securing Pictorial Data Transfer Through a Novel Digital Image Encryption Method. In: Hassanien, A.E., Anand, S., Jaiswal, A., Kumar, P. (eds) Innovative Computing and Communications. ICICC 2024. Lecture Notes in Networks and Systems, vol 1021. Springer, Singapore. https://doi.org/10.1007/978-981-97-3591-4_11.

28. Shyam, R., Mishra, A., Kumar, A., Chowdhary, A., Srivastava, A.K. (2024). Recording of Class Attendance Using DL-Based Face Recognition Method. In: Nanda, S.J., Yadav, R.P., Gandomi, A.H., Saraswat, M. (eds) Data Science and Applications. ICDSA 2023. Lecture Notes in Networks and Systems, vol 818. Springer, Singapore. https://doi.org/10.1007/978-981-99-7862-5_19.
29. Kumar, A., Pandey, R., Srivastava, K.K., Awasthi, S., Jamal, T. (2022). An Image Performance Against Normal, Grayscale and Color Spaced Images. In: Rajagopal, S., Faruki, P., Popat, K. (eds) Advancements in Smart Computing and Information Security. ASCIS 2022. Communications in Computer and Information Science, vol 1759. Springer, Cham. https://doi.org/10.1007/978-3-031-23092-9_22
30. Singh, A., Kumar, A., Chauhan, B.K. (2022). A Comprehensive Study of Edge Computing and the Impact of Distributed Computing on Industrial Automation. In: Mallick, P.K., Bhoi, A.K., Barsocchi, P., de Albuquerque, V.H.C. (eds) Cognitive Informatics and Soft Computing. Lecture Notes in Networks and Systems, vol 375. Springer, Singapore. https://doi.org/10.1007/978-981-16-8763-1_19
31. Kumar, A., Agrawal, P., Kumar, R., Verma, S., Shukla, D. (2022). Sarcasm Detection Using SVM. In: Mallick, P.K., Bhoi, A.K., Barsocchi, P., de Albuquerque, V.H.C. (eds) Cognitive Informatics and Soft Computing. Lecture Notes in Networks and Systems, vol 375. Springer, Singapore. https://doi.org/10.1007/978-981-16-8763-1_24.
32. Kumar, A., Katiyar, V., Chauhan, B.K. (2022). Text Summarization in Hindi Language Using TF-IDF. In: Mallick, P.K., Bhoi, A.K., Barsocchi, P., de Albuquerque, V.H.C. (eds) Cognitive Informatics and Soft Computing. Lecture Notes in Networks and Systems, vol 375. Springer, Singapore. https://doi.org/10.1007/978-981-16-8763-1_25.
33. Kumar, A., Kumar, R., Shrivastava, S.K. (2020). Describing Image Using Neural Networks. In: Khanna, A., Gupta, D., Bhattacharyya, S., Snasel, V., Platos, J., Hassanien, A. (eds) International Conference on Innovative Computing and Communications. Advances in Intelligent Systems and Computing, vol 1087. Springer, Singapore. https://doi.org/10.1007/978-981-15-1286-5_53.
34. Pal, A., Srivastava, K. K., & Kumar, A. (2013). Parallel Character Reconstruction Expending Compute Unified Device Architecture. In Advances in Computing and Information Technology: Proceedings of the Second International Conference on Advances in Computing and Information Technology (ACITY) July 13-15, 2012, Chennai, India-Volume 2 (pp. 639-648). Springer Berlin Heidelberg.
35. Katiyar, V., Srivastava, K.K., Kumar, A. (2012). Applying Adaptive Strategies for Website Design Improvement. In: Wyld, D., Zizka, J., Nagamalai, D. (eds) Advances in Computer Science, Engineering & Applications. Advances in Intelligent and Soft Computing, vol 166. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-30157-5_85
36. Shrivastava, A., & Verma, A. (2012, October). A secure approach of Eavesdropper detection in quantum key distribution. In *Fourth International Conference on Advances in Recent Technologies in Communication and Computing (ARTCom2012)* (pp. 48-52). Stevenage UK: IET.