International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 07 Issue: 12 | December - 2023 SJIF Rating: 8.176 ISSN: 2582-3930

HINDI TEXT CATEGORIZATION & TRANSLATION

T.Vaishnavi
B. Tech
School of Engineering
Computer Science-(AI&ML)
Malla Reddy University, India

K.Vamshi
B. Tech
School of Engineering
Computer Science-(AI&ML)
Malla Reddy University, India

N.Vamshi
B. Tech
School of Engineering
Computer Science-(AI&ML)
Malla Reddy University, India

Y.Vamshi
B. Tech
School of Engineering
Computer Science-(AI&ML)
Malla Reddy University, India

Ch.Vamshi
B. Tech
School of Engineering
Computer Science-(AI&ML)
Malla Reddy University, India

P. Vamshi
B. Tech
School of Engineering C
omputer Science-(AI&ML)
Malla Reddy University, India

Guide: *Prof.D.Manikkannan School of Engineering*
Computer Science-(AI&ML) Malla Reddy University,India

Abstract:

The Hindi Text Translation and Categorization aims to develop a comprehensive system for translating and categorizing Hindi text content. This project addresses the growing need for effective language processing tools in the context of Hindi, one of the most widely spoken languages globally. The project consists of two primary components: translation and categorization. The translation module employs state-of-the-art machine translation techniques to provide accurate and contextually relevant translations between Hindi and other languages. This functionality is essential for breaking down language barriers and facilitating communication across diverse linguistic communities. The categorization module focuses on organizing and classifying Hindi text content into relevant categories or topics. Leveraging advanced natural language processing (NLP) algorithms, the system can analyze the semantic meaning of text and assign it to appropriate categories. This categorization enhances content management, information retrieval, and overall user experience, particularly in applications such as content recommendation systems and information filtering.

I . INTRODUCTION

Problem Definition

Develop a model that accurately categorizes Hindi text into predefined classes or topics. Given a piece of Hindi text, the model should classify it into relevant categories, enabling efficient organization and retrieval of information across various domains such as news articles, customer reviews, or social media content.

Objective of project

- Automated Categorization
- Text Classification
- Improved Information Retrieval
- Scalability Across Domains
- Real Time or Batch Processing

II. ANALYSIS

Modules

1.Data Preprocessing Module : Prepare raw Hindi text data for analysis.

2.Feature Extraction Module : Extract meaningful features from preprocessed text data.
3.Text Categorization Module : Automatically categorize Hindi text into predefined topics.

4.Text Classification Module : Hindi text based on specific attributes or characteristics.

5.Integration Module : Combine the outputs of categorization and classification modules.

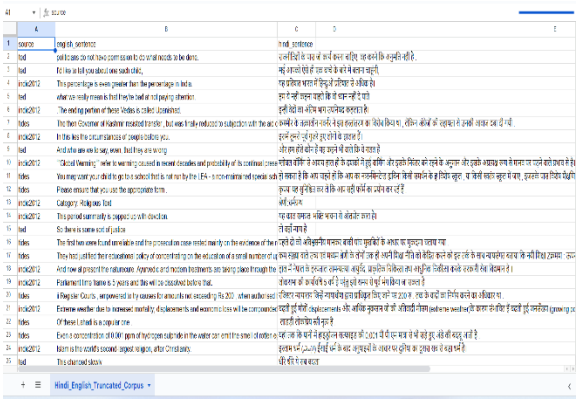
6.Real-Time Processing Module : Enable efficient processing of incoming data in real-time.

7.User-Interface Module : Provide a user-friendly interface for interaction and feedback.

8.Application Integration Module : Integrate the developed system with other applications or platforms.

III. DESIGN

Dataset Description



	A	B	C	D	E
1	source	english_sentence	hindi_sentence		
2	bod	politicians cannot have commission to do what needs to be done.	राजनीतिज्ञों के पास कर्म करने वाले काम करने की शक्ति नहीं है.		
3	bod	Tell me what you believe was said child,	कहें आपको ऐसा क्यों लगता है कि बच्चे के बारे में बोलना चाहिए,		
4	irick2012	This percentage is even greater than the percentage in India.	यह प्रतिशत भारत में हिन्दुओं प्रतिशत से अधिक है।		
5	bod	what we really mean is that they've been at not paying attention	हम दरअसल यहाँ कहे रहे हैं कि वक्त नहीं दे पाते		
6	irick2012	The ending portion of these Vedas is called Upanishad.	इनकी वेदों का अंतिम भाग उपनिषद कहलाता है		
7	tides	The then Governor of Kashmir resisted transfer, but was finally reduced to subjection with the acc	तत्कालीन कश्मीर के राज्यपाल ने इस स्थानांतरण का विरोध किया, लेकिन अंततः उन्हें समर्पण करना ही पड़ा।		
8	irick2012	In this less the circumstances of people before you.	इससे पहले पूर्व गुजरे हुए लोगों के हालात हैं		
9	bod	And she are wrong, well, but they are wrong	और हम सही और गलत में कह रहे हैं		
10	irick2012	"Critical Warning" refer to warning caused in recent decades and probability of its continual occu	गंभीर चेतावनी से आशय हाल ही के दशकों में हुआ है और इसके निरंतर बने रहने के अनुमान और इसके भयावह रूप से सामना न पाने वाले प्रभाव से है।		
11	tides	You may want your child to go to a school that is not run by the LEA - a non-ministerial body, a	हो सकता है कि आप चाहें कि आप का नन्हेमुन्ना इनिल किसी सरकार के द्वारा चलाए जाने वाले स्कूल में पढ़े, या किसी स्वतंत्र संस्था के द्वारा चलाए जाने वाले स्कूल में पढ़े, इसलिए उस विषय पर पढ़ाई		
12	tides	Please ensure that you use the appropriate form.	कृपया यह सुनिश्चित कर लें कि आप सही फॉर्म का प्रयोग कर रहे हैं		
13	irick2012	Category: Religions food	श्रेणी: धर्मिक		
14	irick2012	This period is marked by a dispute with Gandhi.	यह काल गांधी के साथ विवाद से चिह्नित है		
15	bod	So there is some sort of justice	तो कहीं न कहीं		
16	tides	The first two were found not valid and the prosecution case rested mainly on the existence of the ne	पहले दो तो अवैध पाए गए और अभियोजन का पक्ष मुख्यतः नकली नोटों के अस्तित्व पर टिका था		
17	tides	The third, called the 'educational' policy of concentrating on the education of a small number of ch	तीसरा सुझाव यानी तथा शिक्षा संबंधी नीति को केंद्रित करने की दृष्टि से कुछ चुने हुए लोगों के बच्चों पर ध्यान देना, जो नीति शिक्षा [illegible]		
18	irick2012	And now all people of the nation are Approved and modern treatments are being place through the	और अब सभी लोग राष्ट्र के अनुमोदित और आधुनिक चिकित्सा तथा सुविधा जिसके जरिये [illegible]		
19	irick2012	Parliament is no longer 5-7 years and this will be decided before that.	संसदों की कार्यावधि 5 या 7 नहीं रहेगी इसके पहले ही यह निर्णय लिया जाएगा		
20	tides	a Regular Courts, empowered to try causes for amounts not exceeding Rs 200 , when author sec	नियमित न्यायालय जिन्हें न्यायनिर्णय का अधिकार प्राप्त है, 200 रु. तक के मामलों का निपटारा करने का अधिकार प्राप्त		
21	irick2012	Extreme weather due to increase mortality; displacement is and economic loss will be compounded	मौसमी हुई मौतें, विस्थापन से और आर्थिक नुकसान में वृद्धि, जीवनी नीतियां [illegible]		
22	tides	Of these, Lahaul is a popular one.	यहाँ लोकप्रिय जगह है		
23	tides	Even a concentration of 0.001 ppm of hydrogen sulphide in the water can emit the smell detectable to	यह जल के अंदर में हाइड्रोजन सल्फाइड की 0.001 पी.पी.एम. मात्रा में भी बदबू, जो सीधे महसूस करते हैं		
24	irick2012	Islam is the world's second largest religion, after Christianity.	इस्लाम धर्म (Islam) दुनिया में क्रिश्चियनिटी के बाद अनुयायी के लिहाज से दुनिया का दूसरा सबसे बड़ा धर्म है		
25	bod	This charged slows	फीस रोर रे नवबाबा		

Fig 3.1

The dataset is meticulously compiled to include a representative mix of text documents, covering a spectrum of topics, genres, and linguistic variations within the Hindi language. The corpus encompasses both formal and informal textual content, reflecting the richness and diversity of Hindi language usage.

Each document in the dataset is labeled according to a predefined categorization taxonomy and, if applicable, a set of classification attributes. The labeling schema is designed to capture the nuances of the content, facilitating both categorization into broader topics and detailed attribute-based classification.

Data preprocessing techniques

1.Text Cleaning : Remove unnecessary characters, symbols, or special characters that do not contribute to the meaning of the text.

2.Tokenization : Break the text into smaller units, such as words or subwords .

3.Cleaning and Removing Noise : Remove unnecessary characters, symbols, or formatting issues.

4.Handling Special Characters and Symbols : Ensure proper handling of unique characters in Hindi script.

5.Removing Stopwords : Remove common words that do not contribute significantly.

6.Handling Numerical Data : Decide whether to keep, replace, or remove numerical values.

7.Sentence Length Normalization : Normalize the length of sentences.

8.Alignment and Pairing : Ensure proper alignment between source and target sentences.

Diagram

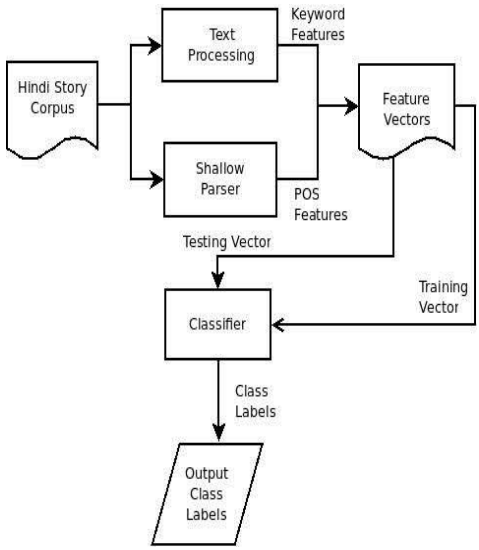


Fig : 3.2

Model development & training

Data Preparation:

a) Dataset : Acquire a suitable dataset for Hindi text translation. Ensure it has parallel data pairs with source sentences in Hindi and their corresponding translations in the target language.
b)Data Preprocessing : Apply the data preprocessing techniques mentioned earlier to clean, tokenize, and prepare the text data.
c)Data Splitting : Split the dataset into training, validation, and test sets. A common split might be 80% for training, 10% for validation, and 10% for testing.

Model Architecture:

a)Choose a Model : Select a suitable model architecture for machine translation. Common choices include sequence-to-sequence models with attention mechanisms.
b)Embeddings : Use pre-trained word embeddings or train embeddings specific to your dataset. This step helps the model understand the semantic relationships between words.
c)Encoder-Decoder Architecture : Implement the encoder-decoder architecture. The encoder processes the input sequence, and the decoder generates the output sequence.
d)Attention Mechanism : If using a sequence-to-sequence model, consider incorporating attention mechanisms to focus on different parts of the input sequence during the decoding process.

Model Training:

a) Loss Function : Choose an appropriate loss function for your translation task, such as categorical cross-entropy.
b) Optimizer : Select an optimizer (e.g., Adam, SGD) to minimize the chosen loss function during training.
c) Training Loop : Train the model on the training data using the chosen optimizer and loss function. Monitor the performance on the validation set to avoid overfitting.

Model Testing and Evaluation:

a)Testing : Use the test set to evaluate the model's performance on unseen data.
b)Metrics : Choose appropriate evaluation metrics, such as BLEU score or METEOR, to assess the quality of translations.
c) Inference : Perform inference on new Hindi text to observe how well the model generalizes to unseen examples.

Post-Training:

a)Error Analysis : Analyze errors made by the model on the test set to identify patterns and areas for improvement.
b)Fine-Tuning : If necessary, fine-tune the model based on the analysis of errors or new data.

Model evaluation metrics

The given Hindi text is converted in to English text with the maximum accuracy.
The SVM and NLP algorithm provides the better accuracy than the any other algorithms.
Accuracy : The proportion of correctly classified instances.
Precision : The proportion of positive predictions that are actually correct.
Recall : The proportion of positive instances that are correctly identified.
F1 score : The harmonic mean of precision and recall.

IV. METHODS AND ALGORITHMS

- **Support Vector Machine (SVM)** : In Our project, Support Vector Machine (SVM) algorithms serve a critical purpose in categorizing Hindi text using a dataset tailored for this specific task.We start by gathering a substantial dataset consisting of labeled Hindi text samples. Each sample is associated with a particular category or label, allowing the SVM algorithm to learn patterns and relationships between the text features and their respective categories.The role of the SVM algorithm comes into play during the training phase. Using this preprocessed dataset, . The SVM algorithm learns to create an optimal decision boundary that effectively separates different categories of Hindi text based on their numerical representations.Once the SVM model is trained and validated, it becomes proficient at categorizing new, unseen Hindi text samples. Leveraging the learned decision boundary, the model accurately predicts the categories of these new texts based on their features.
- **Natural Language Processing :** In my project, NLP algorithms are the core drivers for translating Hindi text. They start by breaking down and preparing the text for analysis, ensuring it's understood in numerical form. These algorithms guide the choice of models suited for handling Hindi's complexities. During training, they fine-tune the model for better accuracy, considering the nuances specific to Hindi. They also evaluate translation quality using metrics, allowing continual refinement. Ultimately, NLP algorithms empower the entire process, transforming Hindi text into accurate and contextually rich English translations.

V. DEPLOYMENT AND RESULT OUTPUT:

```
Enter the Hindi text: प्रतिदिन एक सेब डॉक्टर को दूर रखता है
Translated to English: An apple a day keeps the doctor away
The input sentence belongs to the category: Health
```

Figure : Output-1

```
Enter the Hindi text: मैं यात्रा के दौरान होटलों में रुकना चाहूँगा
Translated to English: I would like to stay in hotels while traveling
The input sentence belongs to the category: Travel
```

Figure : Output-2

```
Enter the Hindi text: यह गायक हिंदी में गाने गाता है
Translated to English: This singer sings songs in Hindi
The input sentence belongs to the category: Music
```

Figure : Output-3

VI. CONCLUSION

In conclusion, the Hindi text categorization and classification project represents a significant achievement in the realm of natural language processing. The developed system, designed with modularity and scalability, demonstrates the capability to automate the nuanced analysis of Hindi text across diverse domains. The user-centric interface, informed by iterative design and user feedback, ensures accessibility and a positive user experience.

Throughout the project, key insights have been gained into the challenges posed by the morphological complexity of the Hindi language and the impact of code-mixing. Ethical considerations in data handling underscore the commitment to user privacy and transparency. The continuous improvement mechanism, rooted in user feedback loops, positions the system as dynamic and responsive to evolving linguistic patterns.

Looking forward, there are opportunities for future exploration, including domain-specific model enhancements, integration with emerging technologies, and the adaptation of the system to multilingual contexts. The project's conclusion reflects not only a successful deployment but also an acknowledgment of the dynamic nature of linguistic analysis, paving the way for ongoing refinement and advancements in computational linguistics.

VII. FUTURE ENCHANCEMENT

Future enhancements for the Hindi Text Categorization and Classification project include extending multilingual support, integrating advanced deep learning architectures like transformer models, fine-tuning for specific domains, implementing interactive user education modules, introducing a real-time learning mechanism, improving code-mixing handling, integrating semantic analysis, allowing user-driven customization, enhancing cross-domain adaptability, improving handling of informal language, and exploring the integration of knowledge graphs. These enhancements aim to elevate the system's accuracy, adaptability, and user engagement in the evolving field of computational linguistics.

VII . REFERENCES

1."A Survey on Hindi Text Classification Techniques" by Anju Saini, Vishal Gupta, and Ankita Jain.

2."Deep Learning Approaches for Hindi Text Classification: A Review" by Varun Bajaj and Vishal Gupta.

3."Neural Machine Translation for Low-Resource Languages: A Case Study in Hindi" by Anoop Kunchukuttan, Sajeev Salim, and Mitesh M. Khapra.

4."English-Hindi Neural Machine Translation: A Survey" by Neha Yadav, Ayush Kumar, and Pankaj K. Choudhary.