CHAPTER 1
INTRODUCTION

INTRODUCTION

1.1 OPINION – SOLVE SCUFFLE

The realm of language identification, classification, and translation presents a complex yet fascinating challenge. At its core lies the quest to bridge linguistic divides, facilitating communication across cultures and borders. However, achieving seamless integration in these domains requires a multifaceted approach that considers various factors, including linguistic diversity, technological advancements, and cultural nuances. Language identification serves as the foundational step in this journey, enabling systems to discern the language of a given text input accurately. With the proliferation of multilingual content on the internet, robust language identification algorithms are indispensable for tasks ranging from content moderation to information retrieval. Leveraging techniques such as machine learning and natural language processing, these algorithms continuously evolve to handle diverse linguistic patterns and dialectal variations.

Classification amplifies the efficacy of language identification by categorizing text into predefined classes or categories. Whether it's sorting documents by topic, sentiment analysis of social media posts, or filtering spam emails, classification algorithms streamline information processing and decision-making processes. Yet, the challenge persists in devising classification frameworks that accommodate the fluidity of language and adapt to evolving semantics and contexts. Translation represents the pinnacle of linguistic technology, promising to dismantle language barriers and foster global connectivity. From machine translation systems like Google Translate to neural machine translation models, significant strides have been made in automating the process of converting text from one language to another. However, the pursuit of flawless translation remains elusive, as nuances in grammar, syntax, and cultural connotations defy simple rule-based approaches.

One of the critical issues plaguing language technology is the scarcity of resources for underrepresented languages. While major languages enjoy extensive linguistic datasets and research attention, smaller languages often languish in the shadows, underserved by mainstream language technologies. Addressing this disparity requires concerted efforts to collect and annotate data, develop language resources, and empower local communities to participate in language technology initiatives. Ethical considerations also loom large in the realm of language technology, particularly concerning privacy, bias, and cultural sensitivity. As algorithms wield increasing influence over communication channels, the need for transparency, accountability, and inclusivity becomes paramount. Striking a balance between technological innovation and ethical responsibility demands ongoing dialogue, collaboration, and vigilance from stakeholders across academia, industry, and policymaking spheres.

Moreover, the dynamic nature of language poses a perpetual challenge for language technology developers. Languages evolve over time, influenced by socio-cultural dynamics, technological advancements, and global interactions. Keeping pace with these changes necessitates agile methodologies, continuous learning mechanisms, and adaptive algorithms that can quickly adapt to emerging linguistic trends. Interdisciplinary collaboration emerges as a cornerstone in the quest to solve the scuffle of language identification, classification, and translation. Linguists, computer scientists, sociologists, and ethicists must join forces to navigate the intricate interplay between language, technology, and society. By fostering synergies between diverse disciplines, holistic solutions can emerge that not only address technical challenges but also uphold principles of linguistic diversity, cultural preservation, and societal equity.

In conclusion, the scuffle of language identification, classification, and translation embodies a multifaceted endeavor that transcends technological boundaries. It calls for innovation, inclusivity, and ethical stewardship to navigate the complexities of linguistic diversity and foster meaningful cross-cultural communication. As technology continues to reshape the landscape of language, the quest for linguistic harmony remains an ongoing journey fueled by collaboration, compassion, and curiosity.

1.2 PURPOSE

Language identification, classification, and translation serve crucial roles in communication, facilitating understanding and interaction across linguistic boundaries. Firstly, language identification is essential for determining the language in which a piece of text or speech is written or spoken. This process enables systems to appropriately process and handle the input, whether it's for translation, analysis, or other purposes. Without accurate language identification, misinterpretation or errors in processing can occur.

Secondly, language classification involves categorizing languages into different groups based on various linguistic criteria such as syntax, morphology, phonology, and semantics. This categorization helps linguists and researchers better understand the structure and evolution of languages, as well as their relationships with one another. Additionally, language classification assists in the development of language technologies tailored to specific language groups, improving their accuracy and performance. Thirdly, translation plays a vital role in bridging linguistic barriers by converting text or speech from one language to another. Translation facilitates communication, enabling individuals, organizations, and nations to exchange information, ideas, and culture across languages. It fosters global cooperation, commerce, diplomacy, and cultural exchange, enhancing intercultural understanding and harmony.

Moreover, language identification, classification, and translation are fundamental components of various technological applications and services. For instance, machine translation systems like Google Translate rely on language identification to determine the source and target languages for translation. Similarly, natural language processing (NLP) applications use language classification to adapt their algorithms and models for different languages, improving their accuracy and effectiveness. Furthermore, these processes are indispensable in fields such as international business, diplomacy, academia, and tourism, where effective communication across languages is paramount. Language identification and classification aid in market research, localization, and cross-cultural communication strategies, while translation facilitates the dissemination of information and the expansion of global markets.

In addition to their practical applications, language identification, classification, and translation also have significant implications for preserving and promoting linguistic diversity and cultural heritage. By enabling the documentation and translation of minority and endangered languages, these processes contribute to their revitalization and preservation for future generations. Furthermore, language identification, classification, and translation are essential tools in combating misinformation, disinformation, and hate speech across languages. By accurately identifying and translating content, researchers and organizations can better monitor and counter harmful narratives and propaganda spread through various linguistic channels.

In summary, language identification, classification, and translation play multifaceted roles in facilitating communication, enabling technological advancements, fostering cultural exchange, preserving linguistic diversity, and combating misinformation. These processes are indispensable in our increasingly interconnected and multilingual world, contributing to mutual understanding, cooperation, and progress across linguistic boundaries.

CHAPTER 2
LITERATURE REVIEW

LITERATURE REVIEW

The literature survey advancements in language identification within short strings. Toftrup et al. (2021) replicated Apple's bi-directional LSTM models, offering insights into their efficacy. Meanwhile, Mathur et al. (2020) explored language identification in test documents, contributing to intelligent system and technology research. These works provide valuable methodologies and insights for further study in this field.

Author Priyank Mathur, Arkajyoti Misra, Emrah Budur 2020 paper published in the Journal of Information Science, Anna Avenberg addresses the automatic language identification of short text, providing valuable insights for research in this field. CNN, RNN. Researchers have explored diverse feature representations, including character n-grams, word embeddings, and phonetic information. Additionally, cross-lingual transfer learning and multilingual models

Rene Hass and Leno Derczynski's 2023 study, published in the International Journal of Medical Informatics, focuses on discriminating between similar Nordic languages, offering pertinent contributions to the field of language discrimination and informatics research.

The literature survey on language identification focuses on methods for automatically detecting the language of a given text or speech signal. It encompasses various approaches such as statistical methods, machine learning algorithms (e.g., SVM, neural networks), and deep learning techniques (e.g., CNN, RNN). Researchers have explored diverse feature representations, including character n-grams, word embeddings, and phonetic information. Additionally, cross-lingual transfer learning and multilingual models have emerged as promising directions to improve accuracy and robustness in language identification systems. The survey highlights the significant progress made in this field and identifies potential areas for further research and development. These works provide valuable methodologies and insights for further study in this field. and ethicists must join forces to navigate the intricate interplay between language, technology, and society. By fostering synergies between diverse disciplines, holistic solutions can emerge that not only address technical challenges

CHAPTER 3
AIM AND OBJECTIVE OF STUDY

3.1 AIM AND OBJECTIVES OF STUDY

- **Aim:**

  To accurately determine the language of a given text or speech sample.

- **Objectives:**

  1. Develop robust language identification algorithms.
  2. Improve multilingual communication and information processing.
  3. Enhance language-specific applications and services.

The aim of a study in language identification, classification, and translation is to develop methodologies and technologies that can accurately identify, categorize, and translate text or speech from one language into another. The primary objective is to enhance the understanding and processing of linguistic data, enabling seamless communication across different languages. This involves creating algorithms and models that can recognize language patterns, understand context, and adapt to new linguistic inputs. Such a study seeks to improve the accuracy of translation services, reduce language barriers in global communication, and support the preservation and understanding of diverse languages and dialects. Additionally, it aims to advance the field of computational linguistics, contributing to better natural language processing tools and applications for various purposes, including education, business, and international relations.

CHAPTER 4
METHODOLOGY

4.1 METHODOLOGY

Language Identification:

Data Collection: Gather a diverse dataset containing text samples from various languages.

Preprocessing: Clean the text by removing noise, special characters, and non-textual elements.

Feature Extraction: Extract relevant features from the text, such as character n-grams, word n-grams, or statistical properties.

Model Selection: Choose a suitable classification algorithm, such as Naive Bayes, SVM, or deep learning models like CNNs or RNNs.

Training: Train the model on the labeled dataset, optimizing parameters for accuracy.

Evaluation: Evaluate the model's performance using metrics like accuracy, precision, recall, and F1-score on a separate test dataset.

Deployment: Deploy the trained model for language identification tasks.

Language Classification:

Data Collection: Collect labeled data for different language classes.

Preprocessing: Clean and preprocess the text data.

Feature Extraction: Extract features, such as bag-of-words, TF-IDF, word embeddings, or character-level features.

Model Selection: Choose a classification algorithm (e.g., SVM, Random Forest, or neural networks).

Training: Train the model on the labeled dataset, adjusting hyperparameters for optimal performance.

Evaluation: Evaluate the model's performance using metrics like accuracy, precision, recall, and F1-score on a separate test dataset.

Deployment: Deploy the trained model for language classification tasks.

Translation:

Data Collection: Collect parallel corpora containing translated sentences.

Preprocessing: Tokenize, clean, and normalize the text data.

Alignment: Align corresponding sentences in the parallel corpora to create training data.

Model Selection: Choose a translation model architecture (e.g., statistical machine translation, neural machine translation).

Training: Train the translation model on the aligned parallel corpora, optimizing parameters for translation quality.

Evaluation: Evaluate the translation model's performance using metrics like BLEU score, METEOR, TER, or human evaluation.

Fine-tuning (Optional): Fine-tune the model on specific domains or languages to improve translation quality.

Deployment: Deploy the trained translation model for translation tasks.

Challenges and Considerations:

Data Quality: Ensuring high-quality, representative datasets for training and evaluation.

Resource Requirements: Training models, especially neural networks, may require significant computational resources.

Language Variability: Accounting for dialects, slang, and regional variations within languages.

Domain Adaptation: Adapting models to specific domains (e.g., technical, medical) for better performance.

Evaluation Metrics: Choosing appropriate metrics to evaluate model performance accurately.

4.2 ALGORITHMS

Text Rank Algorithm:

Text Rank Algorithm converts word into token with the help of spacy and NLTK (Natural Language Toolkit). Language identification, classification, and translation are fundamental tasks in natural language processing (NLP) and involve multiple techniques and algorithms, including the use of ranking mechanisms. Here's a breakdown of these tasks and where ranking algorithms can play a role:

Language Identification:

Language identification is the process of determining the language in which a given piece of text is written. This task is usually the first step in any multilingual NLP pipeline and is critical for routing text to corresponding language-specific processing modules.

Ranking in Language Identification:

Statistical Models: Early models often use n-gram frequency statistics. Text is analyzed, and the frequencies of character sequences (n-grams) are compared against a precomputed model of known languages. The language whose model ranks highest in similarity to the text's n-gram profile is chosen as the predicted language.

Machine Learning Models: Modern approaches use machine learning classifiers trained on large corpora of labeled text in various languages. These classifiers, such as neural networks, may use embeddings that encode text into numerical vectors. The model outputs probabilities for each language, essentially ranking them based on likelihood.

Text Classification:

Text classification involves categorizing text into predefined categories. This can include sentiment analysis, topic classification, and more.

Ranking in Text Classification:

Ranking Algorithms: In some scenarios, especially with hierarchical classification schemas, texts are passed through a decision process where each node in the hierarchy ranks potential categories and chooses the path with the highest rank.

Machine Learning Approaches: Similar to language identification, classifiers are used to assign probabilities to each category, ranking them based on how likely a text belongs to each category.

Translation:

Translation is converting text from one language to another. This is typically handled by machine translation systems such as statistical machine translation (SMT) or neural machine translation (NMT).

Ranking in Translation:

Beam Search: In both SMT and NMT, during the decoding phase (translating a sentence from the source language to the target language), a beam search algorithm is often used. This algorithm explores multiple translation hypotheses at each step and keeps a fixed number (beam width) of the most promising translations based on a scoring function.

Re-ranking: After initial hypotheses are generated, a re-ranking step can be employed where additional models (like language models or specialized scoring functions) evaluate and rank these hypotheses, refining the selection to produce the most fluent and accurate translation.

Summary:

Text Rank Algorithm converts word into token with the help of spacy and NLTK (Natural Language Toolkit). Ranking algorithms are integral to language processing tasks, helping to sift through multiple hypotheses or potential classifications to select the most probable or accurate outcome. In language identification and classification, ranking typically involves choosing the language or category with the highest probability. In translation, more complex ranking mechanisms, like beam search and re-ranking, are used to manage and optimize multiple translation possibilities to achieve better accuracy and fluency.

Naive Bayes Algorithm:

Naive Bayes Algorithm check the probability of token and generate meaning of it. Language identification classification and translation are two significant tasks in natural language processing (NLP) that can utilize algorithms like Naive Bayes for effective implementation. Let's explore how each of these tasks can be approached and the role that Naive Bayes plays in them.

Language Identification Classification

Language Identification aims to determine the language of a given piece of text. It's a fundamental task, especially in applications that handle multilingual datasets or user input.

How Naive Bayes is used:

Feature Extraction: First, the text is processed to extract features that are useful for language identification. Common features include character n-grams (sequences of 'n' characters in the text) and word n-grams.

Model Training: Naive Bayes classifiers are trained on these features. The training involves calculating the probability of each feature appearing in each language. This is based on Bayes' Theorem, which computes the probability of an event based on prior knowledge of conditions related to the event.

Classification: For classification, the Naive Bayes algorithm calculates the posterior probability of each language given the observed features in the text. The language with the highest probability is predicted as the language of the input text.

Advantages of Naive Bayes:

Simple and efficient, especially with large datasets.

Works well with text data and handles multiple classes effectively.

Translation:

Translation involves converting text from one language to another. While Naive Bayes is less commonly used for the translation task itself (which is now dominated by neural networks and deep learning models), it can still play a role in related tasks.

Role of Naive Bayes:

Pre-processing steps: For example, in classifying whether a segment of text needs translation based on language identification.

Post-processing: Such as correcting minor errors in a translation output based on probability models of language usage.

However, translation primarily utilizes models like sequence-to-sequence models (seq2seq), transformers, and other forms of deep learning architectures that better capture the complexities of language syntax and semantics beyond what Naive Bayes can typically handle.

Summary:

Naive Bayes Algorithm check the probability of token and generate meaning of it. While Naive Bayes is effective for tasks like spam detection, sentiment analysis, and language identification due to its simplicity and efficiency with statistical inference, it is less effective for complex sequence modeling tasks such as translation. Translation requires understanding and generating sequences, tasks for which models like transformers, which are capable of handling long-range dependencies and contextual nuances, are more suited.

4.3 Software Architecture

Front End:

Language identification, classification, and translation are complex processes involving various technologies and algorithms. Let's break down how the front-end trio of HTML, CSS, and JavaScript play a role in these tasks:

HTML (Hypertext Markup Language): HTML is the backbone of web pages, providing the structure and content. In the context of language identification and classification, HTML can contain metadata such as lang attribute, which specifies the language of the content. This metadata can be utilized by language identification algorithms to determine the language of the text.

CSS (Cascading Style Sheets): CSS is used for styling and presentation. While CSS itself doesn't directly contribute to language identification, it's essential for rendering text in different languages correctly. For example, CSS properties like font-family can be set to specific fonts suitable for different languages, ensuring proper display.

JavaScript: JavaScript is a powerful scripting language that can be used for various tasks, including language identification and translation. JavaScript libraries and APIs such as Google Translate API or Microsoft Translator API can be integrated into web applications to detect the language of text and translate it into other languages. JavaScript can also be used to manipulate DOM elements based on the identified language, such as dynamically changing the layout or loading localized resources.

In summary, HTML provides the structure and metadata, CSS ensures proper presentation, and JavaScript enables dynamic functionality including language identification and translation, making the web experience more accessible and user-friendly across different languages.

Back End:

Python: This refers to using Python for server-side programming, handling data processing, and managing server-side logic. Python is often chosen for its versatility and ease of use in developing backend systems.

Flask: Flask is a lightweight web framework for Python. It's used to build web applications, RESTful APIs, and more. Flask allows developers to quickly create web applications with minimal code and has a simple and easy-to-use syntax.

OCR (Optical Character Recognition): OCR is the technology that enables the conversion of different types of documents, such as scanned paper documents, PDF files, or images captured by a digital camera, into editable and searchable data.

NLP (Natural Language Processing): NLP is a field of artificial intelligence that focuses on the interaction between computers and humans through natural language. It involves tasks such as text parsing, sentiment analysis, language translation, and more.

NLTK (Natural Language Toolkit): NLTK is a Python library for NLP. It provides tools and resources for tasks such as tokenization, stemming, tagging, parsing, and more. NLTK is widely used in academia and industry for research and development in natural language processing.

Putting it all together, the explanation might be: "Language identification classification and translation using a Python backend with Flask framework. OCR is employed to extract text from images/documents. NLP tasks, facilitated by NLTK, are then applied to the extracted text for various purposes like sentiment analysis or language translation."

4.4 Flow Diagram

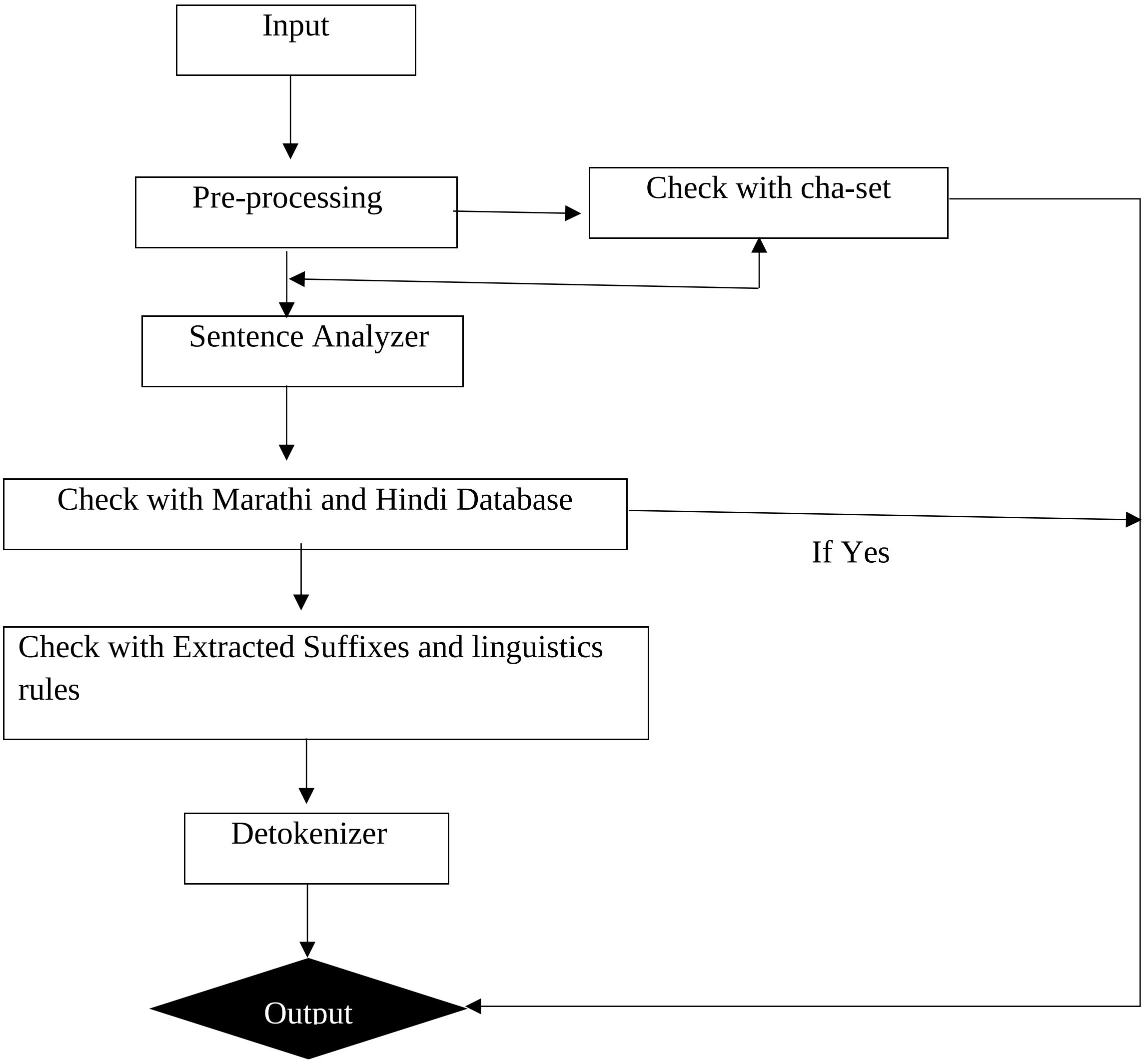

Input
Pre-processing
Check with cha-set
Sentence Analyzer
Check with Marathi and Hindi Database
If Yes
Check with Extracted Suffixes and linguistics rules
Detokenizer
Output

Fig No. 4.4.1

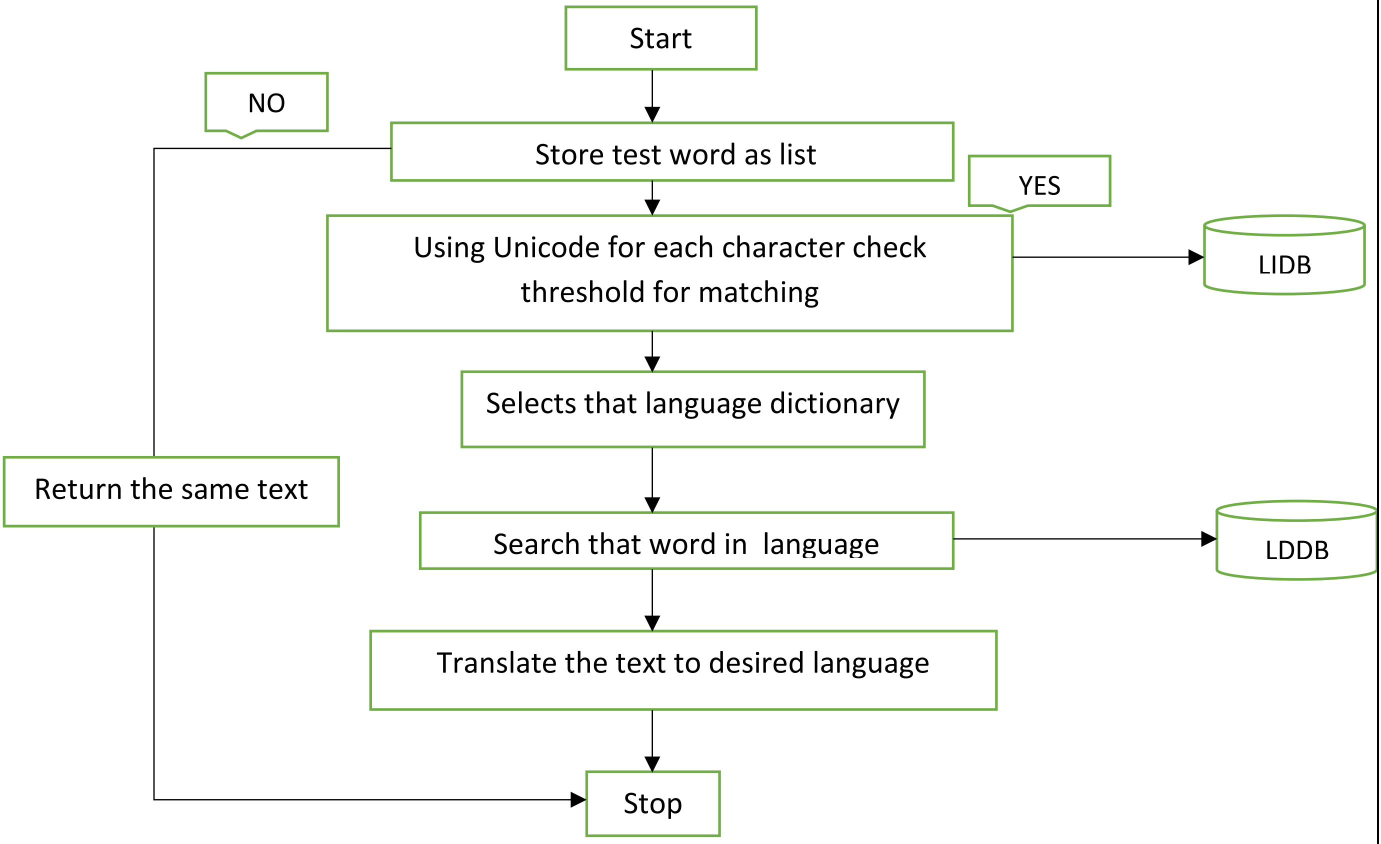

Start
NO
Store test word as list
YES
Using Unicode for each character check threshold for matching
LIDB
Selects that language dictionary
Return the same text
Search that word in language
LDDB
Translate the text to desired language
Stop

Fig. No. 4.4.2

CHAPTER 5

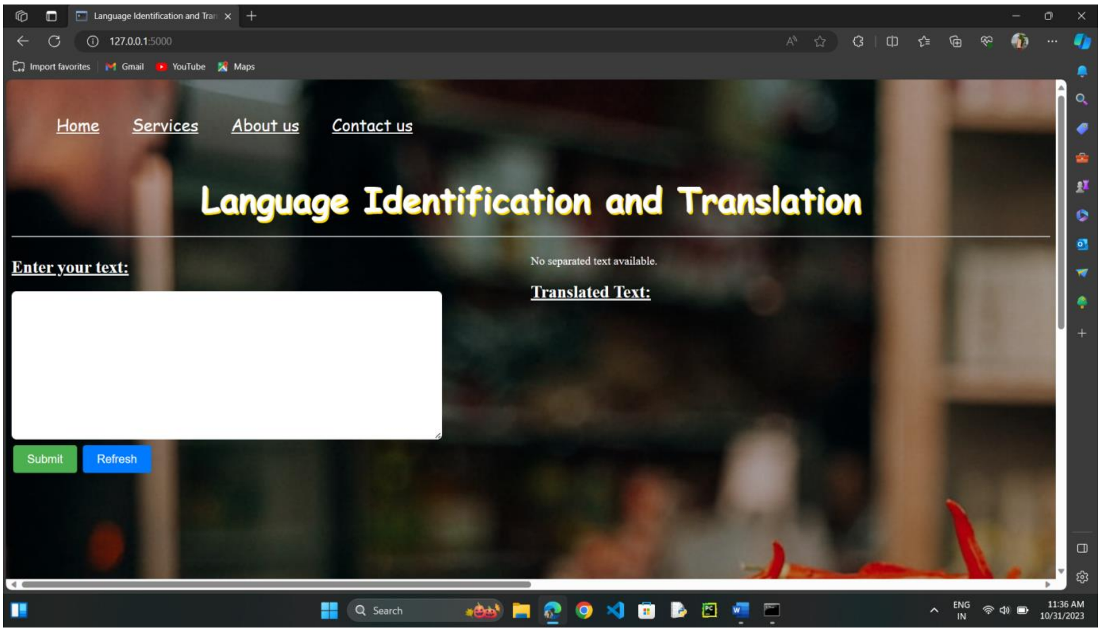
DESIGN AND IMPLEMENTATION

5.1 Design and Implementation

Designing a language identification, classification, and translation system involves several key steps:

1. Data Collection: Gather a diverse dataset of text samples in various languages to train the system. This dataset should cover a wide range of languages and dialects.
2. Preprocessing: Clean and preprocess the text data by removing noise, normalizing text, and tokenizing sentences.
3. Feature Extraction: Extract relevant features from the text data to represent each language. This could include n-grams, character frequencies, word embeddings, etc.
4. Model Selection: Choose appropriate machine learning models for language identification and classification tasks. Common models include Naive Bayes, Support Vector Machines (SVM), Random Forest, or deep learning models like recurrent neural networks (RNNs) or transformers.
5. Training: Train the selected models on the preprocessed data using appropriate training techniques such as cross-validation and hyperparameter tuning.
6. Evaluation: Evaluate the trained models using metrics like accuracy, precision, recall, and F1-score on a separate test dataset to assess their performance.
7. Translation: For translation, consider using pre-trained machine translation models like Google Translate API, MarianMT, or training your own translation model using sequence-to-sequence architectures like encoder-decoder models or transformer-based models.
8. Integration: Integrate the language identification and translation components into a cohesive system, ensuring efficient communication between them.
9. Deployment: Deploy the system in the desired environment, whether it's a web application, mobile app, or API, ensuring scalability, reliability, and security.
10. Continuous Improvement: Continuously monitor the system's performance and gather user feedback to improve accuracy and address any issues that arise over time.

Implementation details will vary based on the specific requirements, available resources, and constraints of the project. It's essential to iterate on the design and implementation based on real-world usage and feedback to create a robust and effective language identification and translation system.

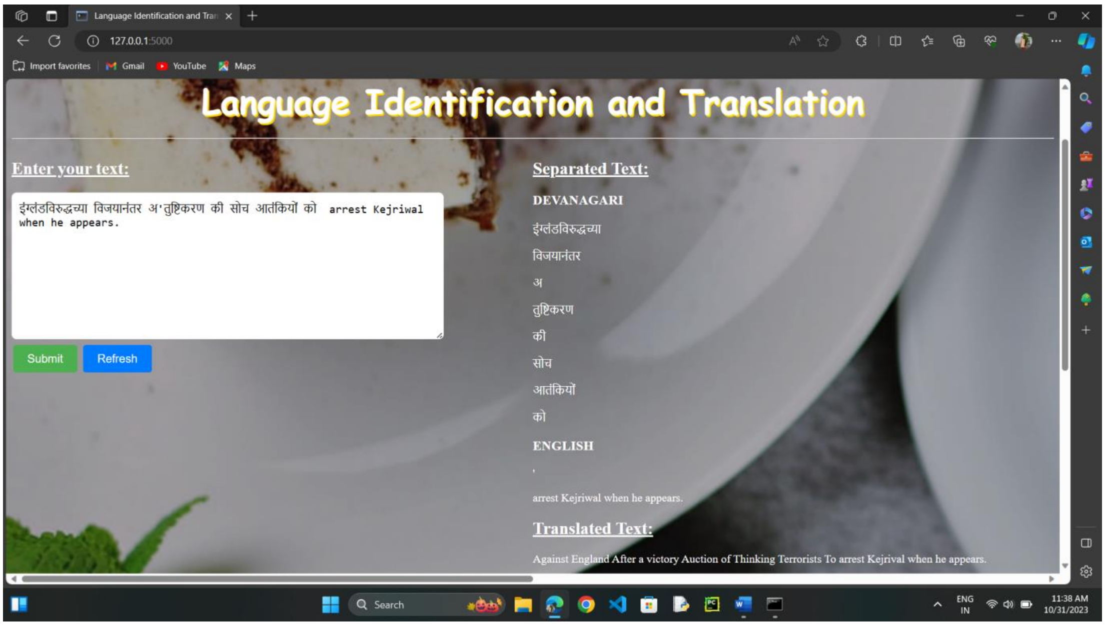


Language Identification and Translation

Enter your text:

No separated text available.

Translated Text:



Language Identification and Translation

Enter your text:

```
इंग्लंडविरुद्धच्या विजयानंतर अ'तुष्टिकरण की सोच आतंकियों को  arrest Kejriwal
when he appears.
```

Separated Text:

DEVANAGARI

इंग्लंडविरुद्धच्या

विजयानंतर

अ

तुष्टिकरण

की

सोच

आतंकियों

को

ENGLISH

'

arrest Kejriwal when he appears.

Translated Text:

Against England After a victory Auction of Thinking Terrorists To arrest Kejrival when he appears.

CHAPTER 6
CONCLUSION

6.1 CONCLUSION

Language identification classification and translation involves identifying the language of a given text, classifying it accordingly, and then translating it into another language if needed. The conclusion of such a process would typically summarize the accuracy and efficiency of the language identification and classification models used, as well as the quality of the translations produced. It might also discuss any challenges encountered during the process and potential areas for improvement in the future, such as enhancing the accuracy of language detection algorithms or improving the fluency and naturalness of translations.

CHAPTER 7
FUTURE SCOPE

FUTURE SCOPE

The future scope for language identification, classification, and translation is vast. With advancements in artificial intelligence and machine learning, we can expect more accurate and efficient language processing systems. This includes improved identification of languages, better classification of text based on context, and more seamless translation between languages in real-time. Additionally, there's potential for better integration of these technologies into various applications and services, making communication across different languages more accessible and natural.

Future scope for language identification classification and translation:

1. Introduction to Language Technology:

- Define language identification, classification, and translation.
- Highlight their importance in a globalized world.

2. Current State of Language Technology:

- Discuss existing language identification, classification, and translation systems.
- Mention common challenges such as accuracy, speed, and resource requirements.

3. Advancements in Machine Learning:

- Explore how machine learning, especially deep learning, has revolutionized language technology.
- Discuss the role of neural networks in improving accuracy and efficiency.

REFERENCES

1. Rene Hass, Leno Derczynski (2023). Discriminating Between Simila Nordic Languages. International journal of medical science.

2. Poutsma, Arjen. (2022) Applying Monte Carlo techniques to language identification. SmartHaven, Amsterdam. Presented at CLIN 2022.

3. Mads Toftrup, Soren Asger (2021). A reproduction of Apple's bi-directional LSTM models for language identification in short string. Journal of Natural Processing.

4. Radim Řehůřek and Milan Kolkus. (2021) "Language Identification on the Web: Extending the Dictionary Method" Computational Linguistics and Intelligent Text Processing.

5. Priyank Mathur, Arkajyoti Misra, Emrah Budur (2020). Language identification from Test document. ACM Transition on intelligent system and Technology.

6. Anna Avenberg (2020). Automatic Language indentification of short text. Journal of information science.

7. Lui, M., & Baldwin, T. (2012). langid.py: An Off-the -shelf Language Identification Tool. In Proceedings of the ACL 201 2 System Demonstrations, Jeju Island, Korea, 25 -30.

8. Nakagawa, T., & Mori, T. (2010). Gleaning Derivational Morphology from Unsegmented Phoneme Sequences. In Proceedings of the 23rd International Conference on Computational Linguistics (COL ING-2010), Beijing, China, 853-861.

9. Trenkle, J. M. (1996). Improving Language Identification Systems via Error Analysis. In Proceedings of the 16th International Conference on Computational Linguistics (COLING -96), Copenhagen, Denmark, 824-829.

ANNEXURE

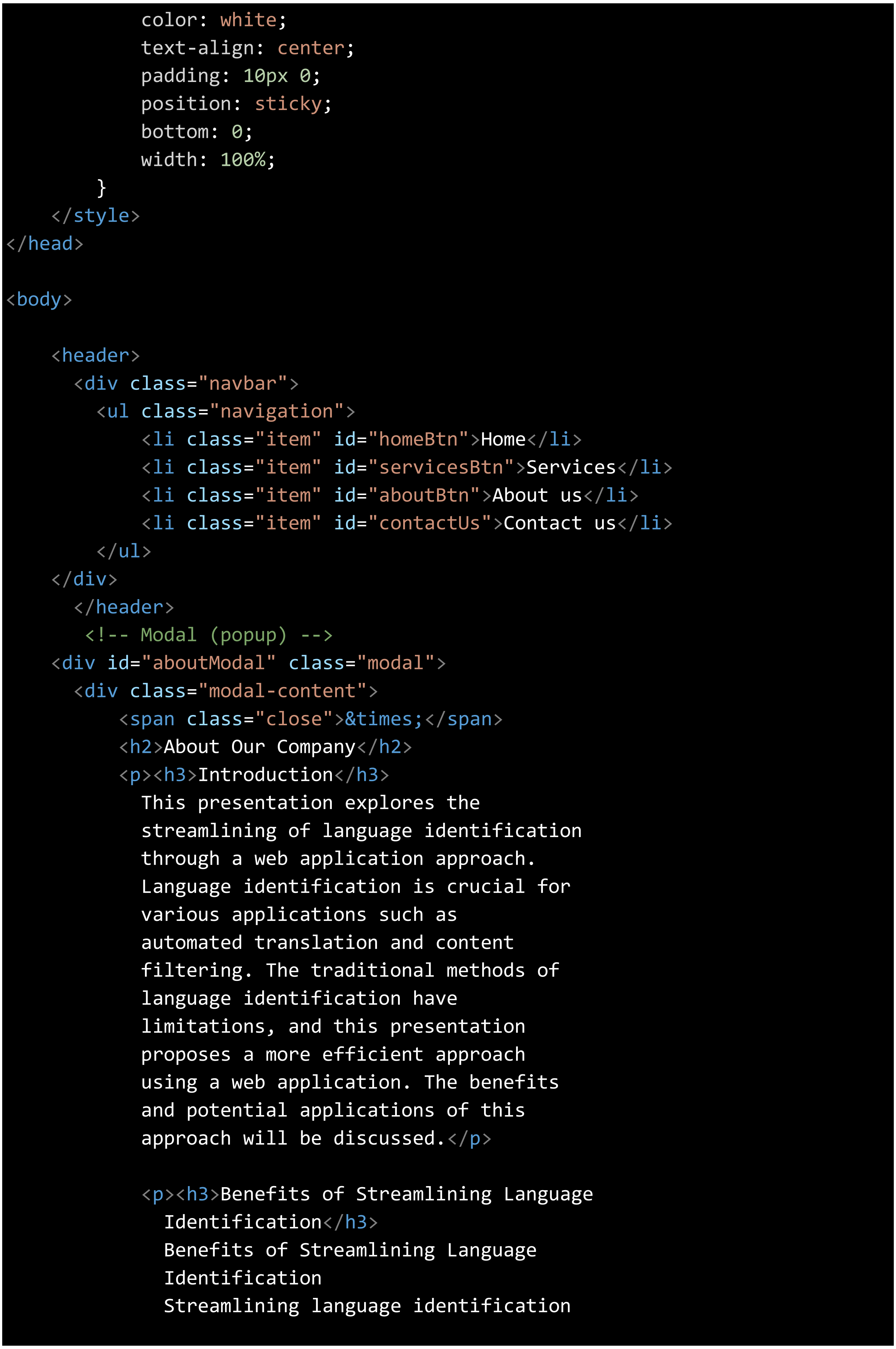
- ## Annexure

Code :-   Frontend :-   Html :-

```
<!DOCTYPE html>
<html lang="en">

<head>
    <meta charset="UTF-8">
    <meta name="viewport" content="width=device-width, initial-scale=1.0">
    <title>Language Identification and Translation</title>
    <style>
        body {

            background-color: black;
            color: white;
        }
        header::before{
      background:
url('https://source.unsplash.com/collection/190727/1600x900')no-repeat center
center/cover;
      content: " ";
      position: absolute;
      top: 0;
      left: 0;
      height: 200%;
      width: 200%;
      z-index: -1;
      opacity: 0.7;
    }
    .navigation{
      font-size: 23px;
      font-family: 'Baloo Bhai', cursive;
      display: flex;

    }
    .item{
      text-decoration: underline;
      cursor: pointer;

    }
    li:hover{
      background-color: grey;
      border: 2px sloid black;
      border-radius: 8px;
      padding: 10px;
      margin: 10px;
```

```
   color: blue;
 }

 li{
   list-style: none;
   padding: 23px 23px;

 }

 .modal {
         display: none;
         position: fixed;
         z-index: 1;
         left: 0;
         top: 0;
         width: 100%;
         height: 100%;
         overflow: auto;
         background-color: rgba(0, 0, 0, 0.7);
     }

     .modal-content {
         background-color: #fefefe;
         color: black;
         margin: 10% auto;
         padding: 20px;
         border: 1px solid #888;
         width: 80%;
     }

     .close {
         color: #aaa;
         float: right;
         font-size: 28px;
         font-weight: bold;
     }

     .close:hover,
     .close:focus {
         color: black;
         text-decoration: none;
         cursor: pointer;
     }
#right{
   position: fixed;
   right: 2px;
   top: 10px;
```

```
      margin: 23px;
      padding: 20px;}
        .right:hover{
        border: 2px solid red;
        background-color: cyan;
      }


    section{
      height: 80px;
      font-family: 'Baloo Bhai', cursive;
      display: flex;
      flex-direction: column;
      margin: 13px;
      align-items:center;
      justify-content: center;
    }
    h1 {

      font-size: 3rem;
      text-shadow: 2px 2px gold;

    }

    h2{
      text-decoration: underline;
    }

    #input_text {
    width: 80%; /* Set the width to 100% to fill the parent container */
    padding: 10px; /* Add padding for better readability */
    font-size: 16px; /* Set the font size */
    border: 1px solid #f1eeee;
 /* Add a border */
    border-radius: 7px; /* Add rounded corners */
    resize: vertical; /* Allow vertical resizing */
}

#input_text:focus {
    border-color: #007bff; /* Change border color when focused */
    outline: none; /* Remove the default focus outline */
}


    input[type="submit"] {
    background-color: #4CAF50; /* Set the background color */
    color: white; /* Set the text color */
    border: none; /* Remove border */
    padding: 10px 20px; /* Add padding */
    text-align: center; /* Center-align text */
    text-decoration: none; /* Remove underline */
    display: inline-block; /* Make it an inline element */
    font-size: 16px; /* Set the font size */
    margin: 4px 2px; /* Add margin */
    cursor: pointer; /* Add cursor pointer on hover */
    border-radius: 4px; /* Optional: Add rounded corners */
}

/* Change button background color on hover */
input[type="submit"]:hover {
    background-color: #45a049;
}

button[type="button"] {
    background-color: #007bff; /* Set the background color */
    color: white; /* Set the text color */
    border: none; /* Remove border */
    padding: 10px 20px; /* Add padding */
    text-align: center; /* Center-align text */
    text-decoration: none; /* Remove underline */
    display: inline-block; /* Make it an inline element */
    font-size: 16px; /* Set the font size */
    margin: 4px 2px; /* Add margin */
    cursor: pointer; /* Add cursor pointer on hover */
    border-radius: 4px; /* Optional: Add rounded corners */
}

/* Change button background color on hover */
button[type="button"]:hover {
    background-color: #0056b3;
}



    .row {
  display: flex;
}

.column {
  flex: 50%
}



        footer {
            background-color: #333;
```



```
            color: white;
            text-align: center;
            padding: 10px 0;
            position: sticky;
            bottom: 0;
            width: 100%;
        }
    </style>
</head>

<body>

    <header>
      <div class="navbar">
        <ul class="navigation">
            <li class="item" id="homeBtn">Home</li>
            <li class="item" id="servicesBtn">Services</li>
            <li class="item" id="aboutBtn">About us</li>
            <li class="item" id="contactUs">Contact us</li>
        </ul>
    </div>
      </header>
       <!-- Modal (popup) -->
    <div id="aboutModal" class="modal">
      <div class="modal-content">
          <span class="close">&times;</span>
          <h2>About Our Company</h2>
          <p><h3>Introduction</h3>
            This presentation explores the
            streamlining of language identification
            through a web application approach.
            Language identification is crucial for
            various applications such as
            automated translation and content
            filtering. The traditional methods of
            language identification have
            limitations, and this presentation
            proposes a more efficient approach
            using a web application. The benefits
            and potential applications of this
            approach will be discussed.</p>

            <p><h3>Benefits of Streamlining Language
              Identification</h3>
              Benefits of Streamlining Language
              Identification
              Streamlining language identification
              through a web application approach
              brings several benefits. It enables faster
              response times, making it suitable for
              real-time applications. The web
              application can be easily updated with
              new language models and improved
              algorithms. Furthermore, it provides
              accessibility across different devices
              and platforms, enhancing user
              experience.</p>
      </div>
  </div>
      <section>
        <h1>Language Identification and Translation</h1>
      </section>
      <hr/>

      <div class="row">
        <div class="column">
          <form method="post">
              <label for="input_text"><h2>Enter your text:</h3></label>
            <textarea id="input_text" name="input_text" rows="10" cols="40">{{
input_text }}</textarea><br>

            <div class="image-upload-container">
              <label for="image_upload">Upload Image:</label>
              <input type="file" name="image" id="image_upload"
accept="image/*">
          </div>

            <label for="target_language">Select Target Language:</label>
            <select id="target_language" name="target_language">
                <option value="1">English</option>
                <option value="2">Marathi</option>
            </select>

            <input type="submit" value="Translate">
            <input type="submit" value="Upload Image" class="upload-button">
            <button type="button" onclick="refreshForm()">Refresh</button>

        </form>
        </div>

        <div class="column">
          {% if separated_text %}
              <h2>Separated Text:</h2>
              {% for lang, segments in separated_text.items() %}
```

```
                  <h3>{{ lang|upper() }}</h3>
                  <p>{{ segments|join(', ') }}</p> <!-- Join segments with a
comma between words -->
              {% endfor %}
              {% else %}
                   <p>No separated text available.</p>
          {% endif %}

          <h2>Translated Text:</h2>
          <p>{{ translated_text }}</p>
      </div>




    <script>
        function refreshForm() {
            document.getElementById("input_text").value = "";
        }



        var contentDiv = document.getElementById('content');

// Function to handle Home button click
function loadHomePage() {
    contentDiv.innerHTML = '<h1>Welcome to the Home Page!</h1><p>This is the
home page content.</p>';
}

// Function to handle Services button click
function loadServicesPage() {
    contentDiv.innerHTML = '<h1>Our Services</h1><p>Explore our services
here.</p>';
}

// Function to handle About us button click
function loadAboutPage() {
    contentDiv.innerHTML = '<h1>About Us</h1><p>Learn more about our
company.</p>';
}

// Function to handle Contact Us button click
function openContactEmail() {
    var emailAddress = 'lokeshchiwarkar7057@gmail.com';
    window.location.href = 'mailto:' + emailAddress;
}
```

```
// Add event listeners to the buttons
var homeBtn = document.getElementById('homeBtn');
homeBtn.addEventListener('click', loadHomePage);
var servicesBtn = document.getElementById('servicesBtn');
servicesBtn.addEventListener('click', loadServicesPage);

var aboutBtn = document.getElementById('aboutBtn');
aboutBtn.addEventListener('click', loadAboutPage);

var contactUsBtn = document.getElementById('contactUs');
contactUsBtn.addEventListener('click', openContactEmail);

var contentDiv = document.getElementById('content');
        var aboutModal = document.getElementById('aboutModal');
        var aboutBtn = document.getElementById('aboutBtn');
        var closeBtn = document.getElementsByClassName('close')[0];

        // Function to open the about modal
        function openAboutModal() {
            aboutModal.style.display = 'block';
        }

        // Function to close the about modal
        function closeAboutModal() {
            aboutModal.style.display = 'none';
        }

        // Event listeners for About Us button and modal close button
        aboutBtn.addEventListener('click', openAboutModal);
        closeBtn.addEventListener('click', closeAboutModal);

        // Close the modal if the user clicks outside the modal content
        window.addEventListener('click', function(event) {
            if (event.target === aboutModal) {
                closeAboutModal();
            }
        });
    </script>


</body>

</html>
```

Css :-

```css
body {
    margin: 0;
    padding: 0;
    font-family: Arial, sans-serif;
}
header {
    background-color: #333;
    color: white;
    padding: 15px 0;
    display: flex;
    justify-content: space-between;
    align-items: center;
}

.image-upload-container {
    display: flex;
    align-items: center;
}

.upload-button {
    margin-left: 10px;
}

.logo {
    font-size: 24px;
    margin-left: 20px;
}

nav ul {
    list-style: none;
    display: flex;
}

nav ul li {
    margin: 0 15px;
}

nav ul li a {
    text-decoration: none;
    color: white;
    transition: 0.3s;
}

nav ul li a:hover {
    color: #f39c12;
}

.search-bar {
    margin-right: 20px;
}

.search-bar input[type="text"] {
    padding: 8px;
    font-size: 14px;

    border: none;
    border-radius: 5px 0 0 5px;
    margin-right: -4px;
}

.search-bar button {
    padding: 8px 15px;
    background-color: #f39c12;
    color: white;
    border: none;
    border-radius: 0 5px 5px 0;
    cursor: pointer;
    transition: 0.3s;
}

.search-bar button:hover {
    background-color: #e67e22;
}
```

BACKEND :- PYTHON+FLASK+OCR:

```
from flask import Flask, render_template, request
from googletrans import Translator

app = Flask(__name__)

def is_devanagari(char):
    return 0x0900 <= ord(char) <= 0x097F

def separate_languages(text):
    lang_segments = {}
    current_lang = None
    current_segment = ""

    for char in text:
        if is_devanagari(char):
            lang = "Devanagari"
        else:
            lang = "English"

        if current_lang is None:
            current_lang = lang
        elif lang != current_lang:
            if current_lang not in lang_segments:
                lang_segments[current_lang] = []

            lang_segments[current_lang].append(current_segment)
            current_segment = ""
            current_lang = lang

        current_segment += char

    if current_lang is not None:
        if current_lang not in lang_segments:
            lang_segments[current_lang] = []
        lang_segments[current_lang].append(current_segment)

    return lang_segments

def translate_words(text, target_language=1):
    translator = Translator()
    words = text.split()
    translated_words = []

    target_lang_code = 'en'  # Default to English
    if target_language == 2:
        target_lang_code = 'mr'  # Set target language to Marathi

    for word in words:
        try:
            translated_word = translator.translate(word,
dest=target_lang_code).text
            translated_words.append(translated_word)
        except Exception as e:
            print(f"Translation error: {e}")
            translated_words.append("Translation error")

    translated_text = ' '.join(translated_words)
    return translated_text


@app.route('/', methods=['GET', 'POST'])
def index():
    separated_text = {}  # Default value for separated_text
    translated_text = ""  # Default value for translated_text

    if request.method == 'POST':
        input_text = request.form['input_text']
        target_language = int(request.form['target_language'])  # Get selected target language
        separated_text = separate_languages(input_text)
        translated_text = translate_words(input_text, target_language)
        return render_template('index.html', separated_text=separated_text, translated_text=translated_text)


if __name__ == '__main__':
    app.run(debug=True)
```

PAPER PUBLICATION

e-ISSN: 2582-5208

International Research Journal of Modernization in Engineering Technology and Science

(Peer-Reviewed, Open Access, Fully Refereed International Journal)

Volume:06/Issue:02/February-2024 Impact Factor- 7.868 www.irjmets.com

LANGUAGE IDENTIFICATION, CLASSIFICATION AND TRANSLATION

Yugal Bihune*[1], Lokesh Hiwarkar*[2], Ankit Patil*[3], Madhulika Gajbhiye*[4], Minal Parkahd*[5], Prof. Aparitosh Gahankari*[6]

*[1,2,3,4,5,6]Department Of Computer Science & Engineering, Nagpur Institute Of Technology, Rashtrasant Tukadoji Maharaj Nagpur University, Mahurzari, Katol Road Nagpur, India.

ABSTRACT

The process of automatically determining whether language(s) is/are present in a document based on its content is called language recognition. In this study, we address the issue of multilingual documents—documents that include text in more than one language. We describe a method that allows one to determine the relative proportions of the languages present, as well as detect if a text is multilingual. We show our method's efficacy on both artificial and real-world multilingual documents gathered from the internet.

Keywords: Translate, Classification, Natural Language Processing (NLP), Steammig, Cross-Language Information Retrieval (CLIR), Part-Oof-Speech Tagger (POS), Parsing, Morphology.

I. INTRODUCTION

More and more, applications requiring machine translation, spell checks, and other tasks require sophisticated language recognition and multilingual data retrieval systems due to the rapidly growing number of lesser-known languages on the internet. For a number of reasons, this task is very complicated. First of all, character sets vary among languages, and many character sets might exist within a single language. The situation is further complicated by certain languages that use the same script. Applications that need natural language processing, such online indexing, querying, and reading assistance, depend on the automatic processing of these heterogeneous texts.

For processing to be effective, the language utilised must be identified early on. For example, morphologically based stemming has shown to be essential for improving retrieval of information. Language-specific algorithms are important, because their use requires a thorough knowledge of the language that is being processed. Language-specific lemmatization requires systems to determine the language in which they need access to dictionaries. In the field of text mining, which looks for patterns in textual data, one of the challenges is categorising documents, where each example corresponds to a certain type of document. Words are what define a document; the existence or lack of a certain word or character is considered a Boolean characteristic. These characteristics are similar to a "bag of words," stressing word occurrences over word frequency.

In India, 528 million people regard Hindi to be their mother tongue, making it one of the major languages spoken by 769 million people. 44 million people speak Hindi in the state of Maharashtra, 35 of whom use voiceless sounds and 9 of whom use voiced sounds. A variety of sounds, such as diphthongs, long and short vowels, are present in Hindi. It is important since it is one of the 22 languages included in the Eighth Schedule of the Indian Constitution and is the official language of Maharashtra for administrative purposes. To differentiate it from the Dravidian family, Hindi is a member of the Indo-Aryan language family, namely the South Dravidian Group. The languages of the Indo-Aryan family, which includes Hindi, are distinct from those of the Sino-Tibetan and Austro-Asiatic language families in that they have differences in phonology.

The official state language of India is Marathi, a language belonging to the Southern Indo-Aryan Zone that is the most commonly spoken in Maharashtra. With 74 million speakers, Marathi is the third most spoken mother tongue in India according to the 2001 census. Among the Indo-Aryan languages, it is the fifteenth most spoken language in the world. English employs the ASCII coding scheme for character specifications in the context of Indian languages. However, because of their distinctive characters, Indian languages require encoding using Unicode systems like UTF-8, UTF-16, UTF-32, and ISCII. In particular, UTF-8

II. METHODOLOGY

Since language identification for multilingual materials requires a multi-label classification task, it is not an easy process. A document in this scenario can be linked to any number of labels from a predefined closed set. We refer to the whole set of languages as L in this study. Languages Lx and Ly are included in a document D, which is represented as D → {Lx, Ly}, where Lx, Ly ∈ L. Languages that are absent from a document are normally deleted for brevity, however documents lacking a specific language are indicated as D → {Lx}. The notation '.' is used to signify classifier output; for example, D. {La, Lb} indicates that text in the languages La and Lb has been predicted to be present in document D.

2.1 Document Representation and Feature Selection:-

As seen in Table 1, represent each document D as a frequency distribution across a byte sequence of n g. Every document is transformed into a vector format. Each entry in this list indicates how many times a specific N-gram byte appears in the document. The bag-of-words concept, in which a vocabulary of "words" is a collection of byte sequences, is comparable to this.It has been selected to differentiate across languages.

The following options are used to choose the precise feature set:

training data with acquired knowledge (IG).

Decision trees now have a partitioning criteria in the form of an information theory metric (Quinlan, 1993). It has been demonstrated that the classifier is very successful at language recognition when IG-based feature selection is combined with Naive Bayes (Louis and Baldwin, 011).

2.2 Generative Mixture Models:-

For text modelling tasks where a variety of effects shape a document's content, such as multi-label document classification (McCallum, 1999; Ramage et al., 2009) and topic modelling (Blei et al., 2003), generative mixture models are well-liked.

These models typically assign a single discrete label to each token and assume complete exchangeability between tokens (also known as the bag-of-words assumption).

The essential representation of the latent structure of a document is shared by our approach for language identification in multilingual documents, topic modelling, and multi-label text classification. A probabilistic mixture of labels is used to model each document, and each label is represented as a probability distribution across tokens. Based on the information provided by Griffiths and Steyvers (2004), the likelihood that a particular collection of the token.

III. MODELING AND ANALYSIS

The model described in Section 3.2 can be used to compute the most likely distribution to generate an unlabeled document on a given set Languages for which we have monolingual training data, letting terms w be byte n-grams. We select sequences using per-language information gain (Section 3.1), and allowing label z. The set of all languages L. Using the training data, we compute φ^(w) j (Equation 3), and then we estimate P(L_j|D) for each L_j ∈ L for an unlabeled document, running the Gibbs sampler until .Samples for z_i converge and then tabulating z_i over d and |d|. Normalize by frankly, we can identify the languages present in a document by D. {Lx if ∃(zi = Lx or D. However, closely related languages have the same frequency distribution over byte n-gram features, so it is possible that some tokens will be incorrectly mapped to a language that is similar to a "real" language.

We solve this problem by finding the subset of languages from the λ training set L that is maximized P(λ|D) (a similar approach is adopted in McCallum (1999)). By applying Bayes' theorem, P(λ|D) ∝ P(D|λ)·P(λ), noting that P(D) is The normalizing constant and can be omitted. We assume that P(λ) is stationary (i.e.e. Any subset of languages is equally likely, a reasonable assumption absence of other evidence), and therefore maximize P(D|λ). For any given D = w1 · · wn and λ, we where both P(wi|zi = j) and P(zi = j) are estimated by their maximum likelihood estimation.

In practice, a complete evaluation of Powerset The value of L is prohibitively expensive, and so we greedily estimate the optimal λ using Algorithm 1.

In essence, we initially rank all candidate languages by computing the maximum likelihood distribution on the complete set of candidate languages. Then, for each of the top-n languages, top-n language."

In the evaluation phase, we aimed to assess the effectiveness of each method in two key aspects: firstly, accurately identifying the language(s) present in each test document, and secondly, for multilingual documents, estimating the relative proportion of the document written in each language. In the context of language identification, the task primarily constitutes a classification problem, adhering to standard metrics such as precision (P), recall (R), and F-n languages. We followed established practices in language recognition research, reporting both document-level micro-averages and language-level macro-averages. Consistent with Baldwin and Lui (2010a), we utilized macro-average F-scores, calculated as the harmonic average of macro-average precision and recall. It's important to note that this approach was chosen to ensure equal emphasis on both precision and recall, with β set to 1, aiming for a balanced evaluation.

To establish the statistical significance of performance differences among systems, we utilized an approximate randomization procedure (Yeh, 2000) with 10,000 repetitions. All reported differences between systems within the results tables (Tables 2, 3, and 4) were statistically significant at the $p < 0.05$ level.

When evaluating the predictions regarding the relative proportion of documents (D) written in each detected language (Li), we compared the predicted subject proportions to the gold-standard ratios determined by our model. This comparison was quantified in terms of a byte ratio, calculated as follows:

gs(Li|D)= Length of D in bytes length of Li part of D in bytes (7).

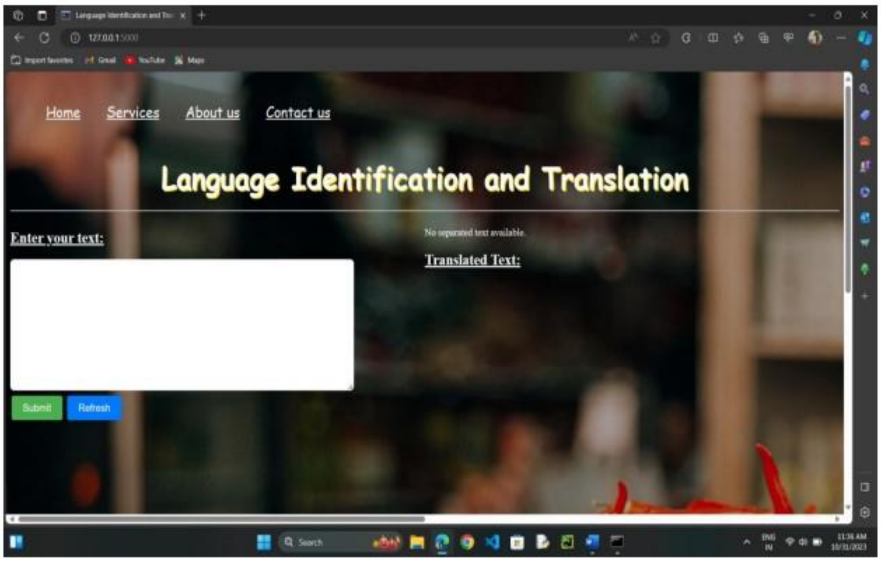
The correlation between predicted and actual proportions was reported using Pearson's R coefficient. Additionally, we calculated the mean absolute error (MAE) over all document-language pairs to provide a comprehensive assessment of the model's performance in estimating language proportions.

IV. RESULTS AND DISCUSSION

To estimate the parameters and evaluate the effectiveness of the suggested models and methodology, many corpora were put together. BUBShabda Sagara-2011 was one such corpus that included a lexicon with 9,000 words for Kannada/Telugu training. A multilingual word list utilised for training was included in this corpus. In our tests, recall rates and accuracy in bytes were used to assess word translation ability (7). In every trial, a single word from the source language and its equivalent translation from the target language were taken into account. A consistent set of 500 distinct sentences that were purposefully left out of the training corpus were used to assess the system's performance.

Our experiment findings, which are shown in Fig. 5, show a notable increase in performance. The systems show much improved accuracy.

Input:-



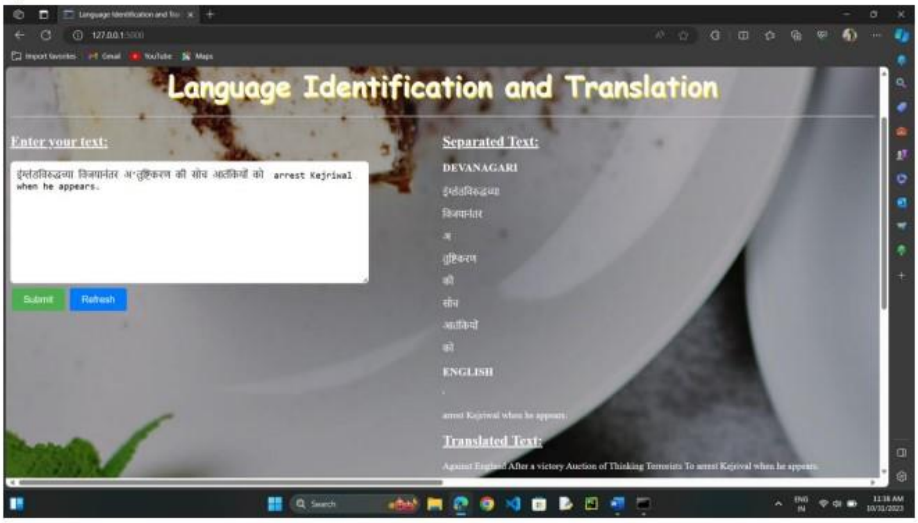
Language Identification and Translation

Enter your text:

No separated text available.

Translated Text:

Output:-



Language Identification and Translation

Enter your text:

इंग्लंडविरुद्धच्या विजयानंतर अ'तुष्टिकरण की सोच आतंकियों को arrest Kejriwal when he appears.

Submit Refresh

Separated Text:

DEVANAGARI

इंग्लंडविरुद्धच्या

विजयानंतर

अ

तुष्टिकरण

की

सोच

आतंकियों

को

ENGLISH

.

arrest Kejriwal when he appears.

Translated Text:

Against England After a victory Auction of Thinking Terrorists To arrest Kejriwal when he appears.

V. CONCLUSION

We present a deep neural network-based language identification system that achieves almost 100% accuracy in categorising various languages and about 90% accuracy in differentiating between substantially related languages. Of particular difficulty were the languages of the West Slavic peoples. There was little progress made in expanding the corpus of these languages from outside sources. This restriction was mostly caused by the lack of word n-grams that were exclusive to some languages and were not included in the larger corpus.

We used convolutional and recurrent neural network models to discover structures unique to each language in order to overcome this difficulty. But we were not able to develop further features because we were note experienced with these specific languages.

ACKNOWLEDGEMENTS

I owe my sincere feelings of gratitude to Prof. A. Gahankari for his valuable guidance and suggestions, which helped me a lot to write this paper. This paper is in continuation of the major research project entitled "Language identification web application" sanctioned to Prof. A. Gahankari, Department of Computer Science and Applications by the University Grant Commission carried out at the RTMNU Nagpur, India.

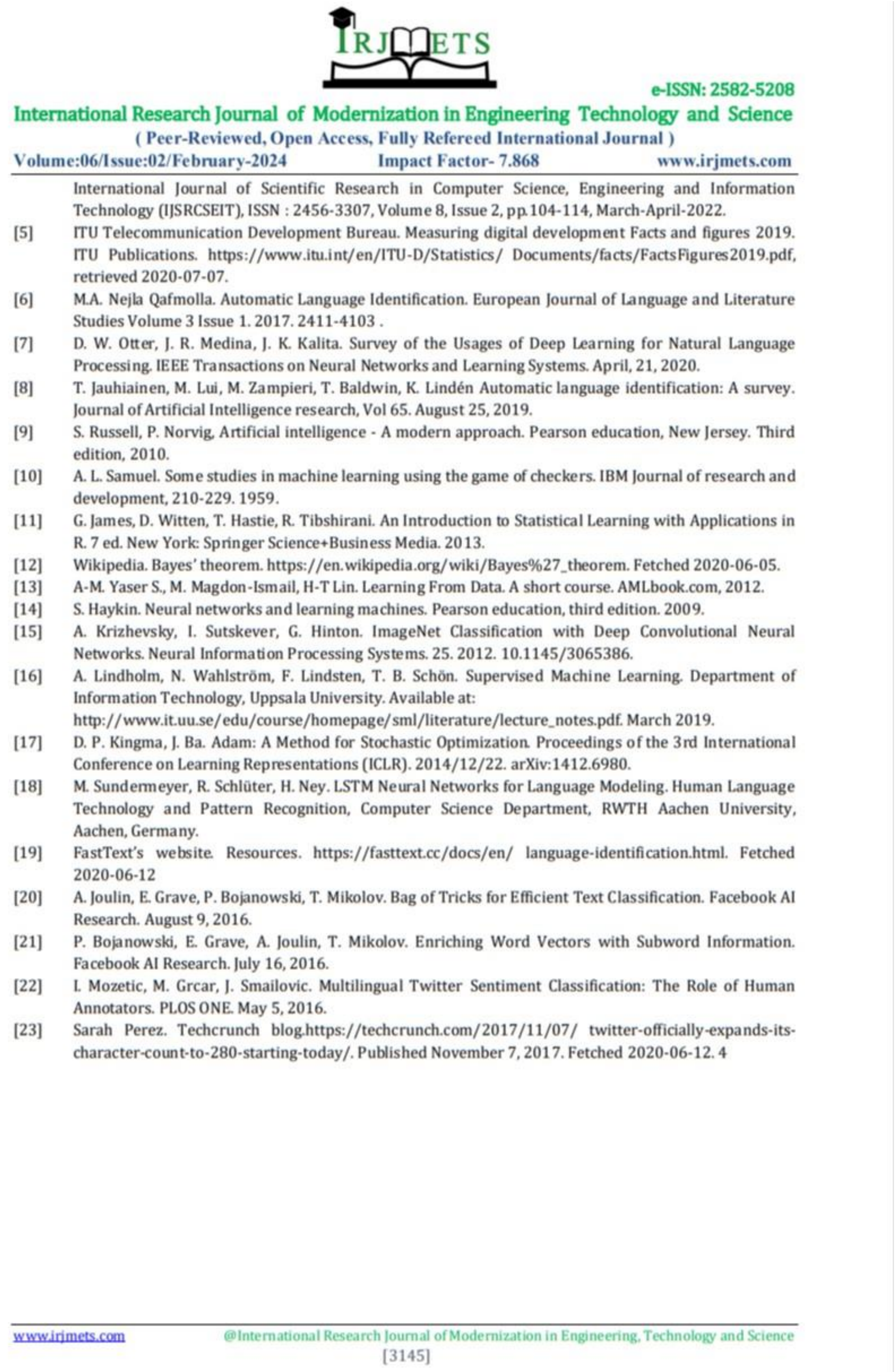
VI. REFERENCES

[1] Aparitosh Gahankari, Dr. Avinash S.Kapse, Dr. V. M. Thakre, " Word Sense Disambiguation - Supervised Approaches: Present Scenario", International Journal of Scientific Research in Science and Technology(IJSRST), Print ISSN : 2395-6011, Online ISSN : 2395-602X, Volume 5, Issue 6, pp.150-154, January-February-2020. Journal URL : https://ijsrst.com/IJSRST208230.

[2] Aparitosh Gahankari, Dr. Avinash S. Kapse, Dr. Mohammad Atique, Dr. V.M. Thakare & Dr. Arvind S. Kapse. (2022). "Word Sense Disambiguation in Marathi Language using Fast-Text model and Indo-Wordnet", Computer Integrated Manufacturing Systems, 28(11), 1607–1617. Retrieved from http://cims-journal.com/index.php/CN/article/view/376.

[3] Aparitosh Gahankari,etal,"A Review on Methods to Solve Polysemy Problem in WSD Occurring while Processing Marathi Text." International Journal of Advanced Research in Science, Communication and Technology (IJARSCT),vol. 2, issue 1, January 2022,ISSN (Online) 2581-9429, https://ijarsct.co.in/jani1.html, DOI: 10.48175/IJARSCT-2463

[4] Aparitosh Gahankari,Vaishnavi Wasankar, Yashika S. Nimje, Kajal Pardhi, Divyani C. Shende, "An Empirical Analysis of Word Sense Disambiguation through Machine Learning Approaches",



International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 8, Issue 2, pp.104-114, March-April-2022.

[5] ITU Telecommunication Development Bureau. Measuring digital development Facts and figures 2019. ITU Publications. https://www.itu.int/en/ITU-D/Statistics/ Documents/facts/FactsFigures2019.pdf, retrieved 2020-07-07.

[6] M.A. Nejla Qafmolla. Automatic Language Identification. European Journal of Language and Literature Studies Volume 3 Issue 1. 2017. 2411-4103 .

[7] D. W. Otter, J. R. Medina, J. K. Kalita. Survey of the Usages of Deep Learning for Natural Language Processing. IEEE Transactions on Neural Networks and Learning Systems. April, 21, 2020.

[8] T. Jauhiainen, M. Lui, M. Zampieri, T. Baldwin, K. Lindén Automatic language identification: A survey. Journal of Artificial Intelligence research, Vol 65. August 25, 2019.

[9] S. Russell, P. Norvig, Artificial intelligence - A modern approach. Pearson education, New Jersey. Third edition, 2010.

[10] A. L. Samuel. Some studies in machine learning using the game of checkers. IBM Journal of research and development, 210-229. 1959.

[11] G. James, D. Witten, T. Hastie, R. Tibshirani. An Introduction to Statistical Learning with Applications in R. 7 ed. New York: Springer Science+Business Media. 2013.

[12] Wikipedia. Bayes' theorem. https://en.wikipedia.org/wiki/Bayes%27_theorem. Fetched 2020-06-05.

[13] A-M. Yaser S., M. Magdon-Ismail, H-T Lin. Learning From Data. A short course. AMLbook.com, 2012.

[14] S. Haykin. Neural networks and learning machines. Pearson education, third edition. 2009.

[15] A. Krizhevsky, I. Sutskever, G. Hinton. ImageNet Classification with Deep Convolutional Neural Networks. Neural Information Processing Systems. 25. 2012. 10.1145/3065386.

[16] A. Lindholm, N. Wahlström, F. Lindsten, T. B. Schön. Supervised Machine Learning. Department of Information Technology, Uppsala University. Available at: http://www.it.uu.se/edu/course/homepage/sml/literature/lecture_notes.pdf. March 2019.

[17] D. P. Kingma, J. Ba. Adam: A Method for Stochastic Optimization. Proceedings of the 3rd International Conference on Learning Representations (ICLR). 2014/12/22. arXiv:1412.6980.

[18] M. Sundermeyer, R. Schlüter, H. Ney. LSTM Neural Networks for Language Modeling. Human Language Technology and Pattern Recognition, Computer Science Department, RWTH Aachen University, Aachen, Germany.

[19] FastText's website. Resources. https://fasttext.cc/docs/en/ language-identification.html. Fetched 2020-06-12

[20] A. Joulin, E. Grave, P. Bojanowski, T. Mikolov. Bag of Tricks for Efficient Text Classification. Facebook AI Research. August 9, 2016.

[21] P. Bojanowski, E. Grave, A. Joulin, T. Mikolov. Enriching Word Vectors with Subword Information. Facebook AI Research. July 16, 2016.

[22] I. Mozetic, M. Grcar, J. Smailovic. Multilingual Twitter Sentiment Classification: The Role of Human Annotators. PLOS ONE. May 5, 2016.

[23] Sarah Perez. Techcrunch blog.https://techcrunch.com/2017/11/07/ twitter-officially-expands-its-character-count-to-280-starting-today/. Published November 7, 2017. Fetched 2020-06-12. 4

International Conference on Futuristic Trends in Engineering, Science & Technology

Title: Language Identification, Classification and Translation

Yugal. D. Bihune[1], Lokesh. C. Hiwarkar[2], Ankit. K. Patil, Madhulika Gajbhiye, Minal Parkhad

[1]*Student, Computer Science Engineering, Nagpur Institute of Technology, Maharashtra, India,* Yugalbihune260@gmail.com

[2]*Student, Computer Science Engineering, Nagpur Institute of Technology, Maharashtra, India,*
lokeshchiwarkar7057@gmail.com

[3]*Student, Computer Science Engineering, Nagpur Institute of Technology, Maharashtra, India,*
Patilankit625@gmail.com

[4]*Student, Computer Science Engineering, Nagpur Institute of Technology, Maharashtra, India,*
madhulikagajbhiye2003@gmail.com

[5]*Student, Computer Science Engineering, Nagpur Institute of Technology, Maharashtra, India,*
minalparkhad4937@gmail.com

ABSTRACT:-

Language recognition is the function of automatically detecting the language(s) present in it Document based on document content. In this work, we solve the problem of finding documents containing text from more than one language (multilingual documents). We present a method that enables Detect if a document is multilingual, identify and estimate the languages presentative proportions. We demonstrate the effectiveness of our method on synthetic data, as well as real-world multilingual documents collected from the web.

Keywords: Cross-Language Information Retrieval (CLIR), Part-of-Speech Tagger (POS), Parsing, Morphology, Natural Language Processing (NLP), Stemming.

1. INTRODUCTION:-

The rapid expansion of lesser-known languages on the internet has created a pressing need for advanced language recognition and multilingual data retrieval systems, particularly in applications involving machine translation, spell checks, and more. This task is highly complex due to several factors. Firstly, different languages utilize varying character sets, and even a single language may have multiple character sets. Additionally, some languages that share the same script further complicate the scenario. Automatic processing of these diverse texts is crucial for applications requiring natural language processing, such as web indexing, querying, and providing reading support.

Early identification of the language used is imperative for effective processing. For instance, morphologically based stemming has proven to be vital in enhancing information retrieval. Language-specific algorithms play a significant role, necessitating a deep understanding of the language being processed. Systems requiring access to dictionaries must identify the language for language-specific lemmatization. In the realm of text mining, which focuses on uncovering patterns in textual data, the challenge lies in classifying documents, where each instance represents a document type. Documents are characterized by the words they contain, and the presence or absence of specific words or characters is treated as Boolean features. These features are akin to a 'bag of words,' emphasizing the importance of word occurrences rather than their frequencies. In this paper, the authors employ the concept of document classification for natural language identification, specifically focusing on English, Marathi, and Hindi, showcasing the intricate nature of language processing in the digital age.

Hindi stands as one of the primary languages spoken by 769 million people in India, with 528 million individuals considering it their mother tongue. In the state of Maharashtra, Hindi is spoken by 44 million people, out of which 35 use voiceless sounds, and 9 use voiced sounds. Hindi exhibits a range of sounds, including short vowels, long vowels, and diphthongs. It holds significance as one of the 22 languages recognized in the Eighth Schedule of the Indian Constitution, serving as the official and administrative language of Maharashtra. Hindi belongs to the Indo-Aryan language family, specifically the South Dravidian Group, which distinguishes it from the Dravidian family. Indo-Aryan languages, including Hindi, differ significantly from other language families like Sino-Tibetan and Austro-Asiatic, encompassing variations in phonological, morphological, lexical, syntactic, and semantic structures.

Marathi, a Southern Indo-Aryan Zone, is the most widely spoken language in Maharashtra and serves as the official state language of India. According to the 2001 census, Marathi ranks third in terms of mother tongue speakers in India, with 74 million speakers. It is the 15th most spoken language globally among Indo-Aryan languages. In the context of Indian languages, English uses the ASCII coding system for character specifications. However, Indian languages require Unicode systems like UTF-8, UTF-16, UTF-32, and ISCII for encoding due to their unique characters. Specifically, Kannada and Marathi utilize the UTF-8 text format for coding, ensuring accurate representation of these languages in digital contexts.

1.1 WRITTEN LANGUAGE BASICS:-

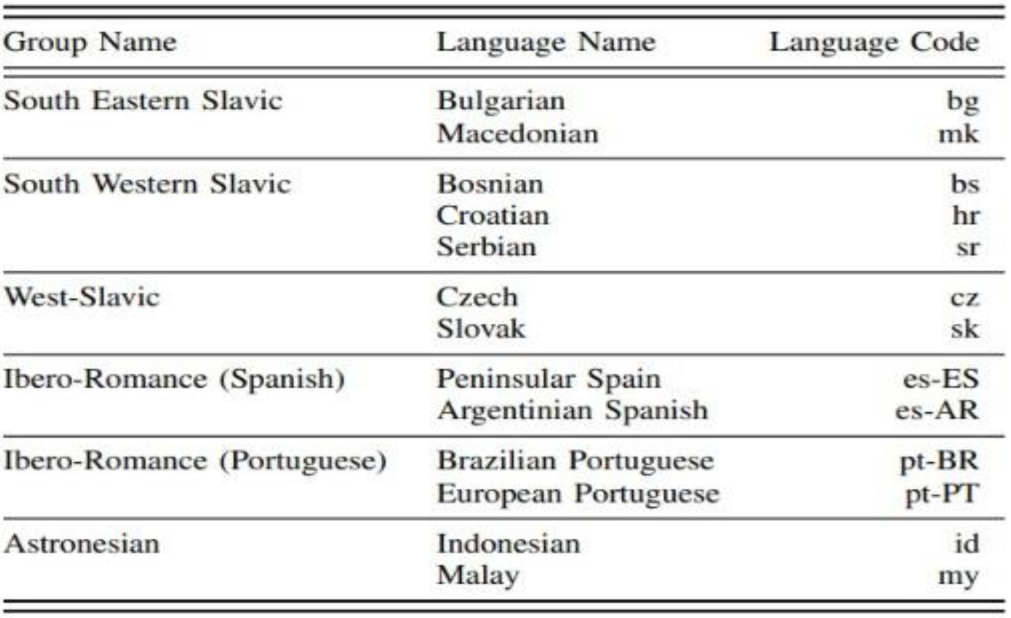
Unicode is a standardized system for encoding text, where each character is uniquely defined by an integer called a Unicode code point. As of January 2018, Unicode 10.0 encompassed

136,755 characters, including Latin, Arabic, Greek, Cyrillic, and Chinese characters. Unicode 10.0 also includes various symbol characters like emojis and currency symbols. It serves as the fundamental building block of characters in different scripts. For instance, the Latin alphabet "ABCabc" consists of 6 letters, while the Cyrillic alphabet has 33 letters, and the Hebrew script "שׁרב" consists of 3 letters, each represented by specific Unicode code points. Unicode code points are bidirectional characters, for example, "\u202b \u05c1 \u05e9 \u05e8 \u05d1 \u202c." The term "character" in this context refers to Unicode code points, and the length of a piece of text is determined by the number of Unicode code points it contains.

Unicode handles characters and their combinations, such as the combination of U+006E (n) and U+0303 (~) to form "ñ," which is equivalent to a single Unicode code point U+00F1 (ñ). To prevent discrepancies, Unicode normalization in form C is employed, where conventional decomposition is followed by conventional synthesis.

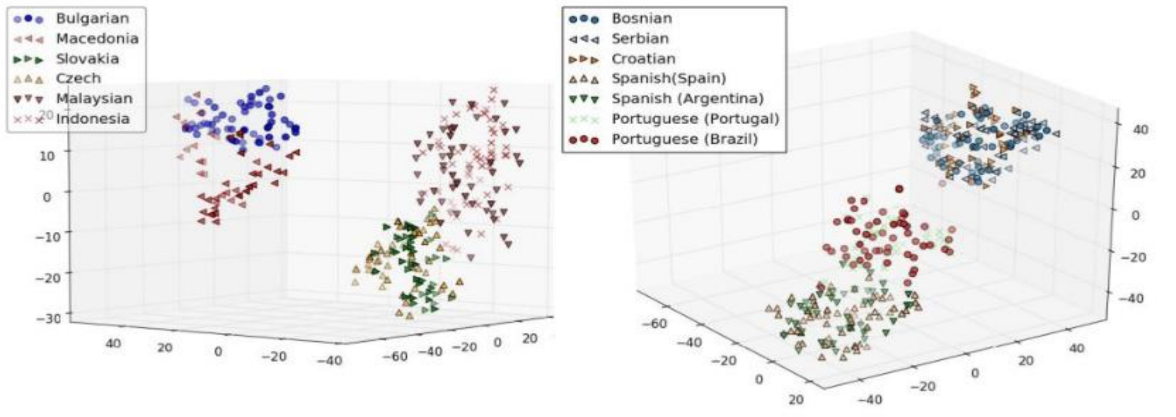
This article focuses on language recognition and discusses various methods to analyze text, including obtaining language indexes based on metadata. Examples of metadata-based language identification include geolocation and website tags, which can be specified in HTML using elements like.

1.2 DATASET DESCRIPTION:-



Group Name	Language Name	Language Code
South Eastern Slavic	Bulgarian	bg
	Macedonian	mk
South Western Slavic	Bosnian	bs
	Croatian	hr
	Serbian	sr
West-Slavic	Czech	cz
	Slovak	sk
Ibero-Romance (Spanish)	Peninsular Spain	es-ES
	Argentinian Spanish	es-AR
Ibero-Romance (Portuguese)	Brazilian Portuguese	pt-BR
	European Portuguese	pt-PT
Astronesian	Indonesian	id
	Malay	my

TABLE I: Benchmark results of available solutions



(a) Easily separable (b) Difficult to separate

In Figure 1, the t-SNE visualization of language groups is presented. Additional plots, including a 3D animated version, can be accessed at http://SeeYourLanguage.info. To create these visualizations, each sentence was vectorized using 1 to 5-grams of tokens delimited by white space characters. The resulting plot (Figure 1) illustrates significant overlap among languages within the same group, while languages from different groups demonstrate clear linear separability. For a comprehensive 3-dimensional visualization of all the languages, a video can be viewed at https://www.youtube.com/watch?v=mhRdfC26q78.

The data utilized for this project was sourced from the "Discriminating between Similar Language (DSL) Shared Task 2015" [19]. Specifically, the dataset included 20,000 instances per language, with 18,000 instances used for training (train.txt) and 2,000 instances for evaluation (test.txt), covering a total of 13 diverse world languages. Additionally, a subset of the training data (devel.txt) was employed for hyper-parameter tuning. The languages were categorized into groups, as outlined in Table I, and these group names were frequently referenced in subsequent sections of the project.

Each entry in the dataset represented a complete sentence extracted from journalistic corpora and composed in one of the specified languages. These entries were tagged with the corresponding language group and the country of origin. To introduce variability into the data, a similar set of mixed-language instances was included, adding noise to the dataset. For the final evaluation, a distinct gold test data (test-gold.txt) was provided.

In our analysis, we applied the t-SNE algorithm to visualize the instances in a 3D Euclidean space [20], [21] for effective representation. Additionally, during the process of feature extraction, measures were taken to address plagiarism concerns and ensure the integrity of the dataset.

2. METHODOLOGY: -

Language identification for multilingual documents poses a challenge as it involves a multi-label classification task. In this context, a document can be associated with any number of labels from a predetermined closed set. Throughout this paper, we represent the complete set of languages as L. A document, denoted as D, containing languages Lx and Ly is expressed as D → {Lx, Ly}, where Lx, Ly ∈ L. If a document does not include a specific language, it is denoted as D → {Lx}, although, for brevity, languages not present in the document are generally omitted. Classifier output is indicated using the notation '.'; for instance, D. {La, Lb} signifies that the document D has been predicted to contain text in languages La and Lb.

2.1 DOCUMENT REPRESENTATION AND FEATURE SELECTION:-

Represent each document D as a frequency distribution over a byte sequence of n g such that as shown in Table 1. Each document is converted into a vector. Here, each entry counts the number of times a particular N-gram byte occurs in the document. This is similar to the bag-of-words model, where a vocabulary of "words" is a set of byte sequences. It is chosen to distinguish between languages.

The exact feature set is selected from the following:

Training data using information gained (IG).

An information theory metric has been developed as a partitioning criterion for decision trees (Quinlan, 1993). Combining IG-based feature selection with Naive Bayes, the classifier has been shown to be particularly effective for language recognition (Louis and Baldwin, 011).

2.2 Generative Mixture Models:-

Generative mixture models are popular for text modeling tasks where a mixture of influences governs the content of a document, such as in multi-label document classification (McCallum, 1999; Ramage et al., 2009), and topic modeling (Blei et al., 2003).

Such models normally assume full exchangeability between tokens (i.e. the bag-of-words assumption), and label each token with a single discrete label.

Multi-label text classification, topic modeling, and our model for language identification in multilingual documents share the same fundamental representation of the latent structure of a document. Each label is modeled with a probability distribution over tokens, and each document is modeled as a probabilistic mixture of labels. As presented in Griffiths and Steyvers (2004), the probability of the ith token (wi) given a set of T labels z1· · ·zT is modeled as:

The set of tokens w is the document itself, which in all cases is observed. In the case of topic modeling, the tokens are words and the labels are topics, and z is latent. Whereas topic modeling is generally unsupervised, multi-label text classification is a supervised text modeling task, where the labels are a set of pre-defined categories (such as RUBBER, IRON-STEEL, TRADE, etc. in the popular Reuters21578 data set (Lewis, 1997)), and the tokens are individual words in documents. z is still latent, but constrained in the training data (i.e. documents are labeled but the individual words are not). Some approaches to labeling unseen documents require that the labels for the training data be inferred, and methods for doing this include an application of the Expectation Maximization (EM) algorithm (McCallum, 1999) and Labeled LDA (Ramage et al.).,

The model that we propose for language identification in multilingual documents is similar to multilabel text classification. In the framework of Equation 1, each per-token label zi is a language, and the vocabulary of tokens is not given by words but rather by specific byte sequences (Section 3.1). The key difference with multi-label text classification is that we use monolingual (i.e. mono-label) training data.

Hence, z is effectively observed for the training data (since all tokens must share the same label). To infer z for unlabeled documents, we utilize a Gibbs sampler, closely related to that proposed by Griffiths and Steyvers (2004) for LDA. The sampling probability for a label zi for token w in a document d is:

(d)j is assumed to have a Dirichlet distribution with hyperparameter α, and the word distribution for each topic φ (w) j is also assumed to have a Dirichlet distribution with hyperparameter β. Griffiths (2002) describes a generative model for LDA where both φ(w) j and θ (d) j are inferred from the output of a Gibbs sampler. In our method, we estimate φ (w) j using maximum likelihood estimation (MLE) from the training data. Estimating φ (w) j through MLE is equivalent to a multinomial Naive Research.

2.3 LANGUAGE IDENTIFICATION IN MULTILINGUAL DOCUMENTS:

The model described in Section 3.2 can be used to compute the most likely distribution to generate an unlabeled document on a given set Languages for which we have monolingual training data, letting terms w be byte n-grams. We select sequences using per-language information gain (Section 3.1), and allowing label z. The set of all languages L. Using the training data, we compute φ^(w) j

(Equation 3), and then we estimate P(L_j|D) for each L_j ∈ L for an unlabeled document, running the Gibbs sampler until. Samples for z_i converge and then tabulating z_i over d and |d|. Normalize by frankly, we can identify the languages present in a document by D. {Lx if ∃(zi = Lx or D. However, closely related languages have the same frequency distribution over byte n-gram features, so it is possible that some tokens will be incorrectly mapped to a language that is similar to a "real" language. We solve this problem by finding the subset of languages from the λ training set L that is maximized

P(λ|D) (a similar approach is adopted in McCallum (1999)). By applying Bayes' theorem, P(λ|D) ∝ P(D|λ)·P(λ), noting that P(D) is The normalizing constant and can be omitted. We assume that P(λ) is stationary (i.e. Any subset of languages is equally likely, a reasonable assumption absence of other evidence), and therefore maximize P(D|λ). For any given D = w1 · · wn and λ, we Estimate P(D|λ) from the output of the Gibbs sampler:

$P(D|\lambda) = \Pi\ N\ i=1\ P(wi\ |\lambda)$ (5) $= \Pi\ N\ i=1\ \Sigma\ j\lambda\ P(wi\ |zi = j)P(zi = j)$ (6)where both $P(wi|zi = j)$ and $P(zi = j)$ are estimated by their maximum likelihood estimation. In practice, a complete evaluation of Powerset the value of L is prohibitively expensive, and so we greedily estimate the optimal λ using Algorithm 1. In essence, we initially rank all candidate languages by computing the maximum likelihood distribution on the complete set of candidate languages. Then, for each of the top-n languages, top-n language."

In the evaluation phase, we aimed to assess the effectiveness of each method in two key aspects: firstly, accurately identifying the language(s) present in each test document, and secondly, for multilingual documents, estimating the relative proportion of the document written in each language. In the context of language identification, the task primarily constitutes a classification problem, adhering to standard metrics such as precision (P), recall (R), and F-n languages. We followed established practices in language recognition research, reporting both document-level micro-averages and language-level macro-averages. Consistent with Baldwin and Lui (2010a), we utilized macro-average F-scores, calculated as the harmonic average of macro-average precision and recall. It's important to note that this approach was chosen to ensure equal emphasis on both precision and recall, with β set to 1, aiming for a balanced evaluation.

To establish the statistical significance of performance differences among systems, we utilized an approximate randomization procedure (Yeh, 2000) with 10,000 repetitions. All reported differences between systems within the results tables (Tables 2, 3, and 4) were statistically significant at the $p < 0.05$ level.

When evaluating the predictions regarding the relative proportion of documents (D) written in each detected language (Li), we compared the predicted subject proportions to the gold-standard ratios determined by our model. This comparison was quantified in terms of a byte ratio, calculated as follows:

gs(Li|D)= Length of D in bytes length of Li part of D in bytes (7).

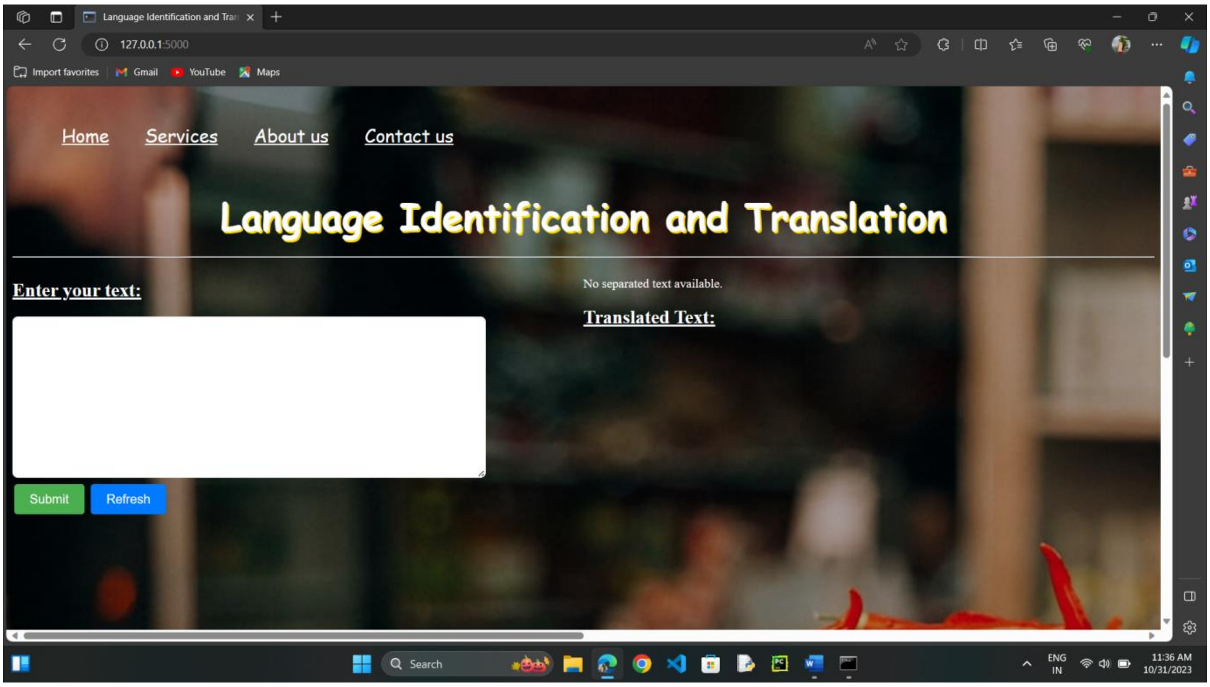
The correlation between predicted and actual proportions was reported using Pearson's R coefficient. Additionally, we calculated the mean absolute error (MAE) over all document-language pairs to provide a comprehensive assessment of the model's performance in estimating language proportions.

2.4 EXPERIMENTAL SETUP, RESULTS, AND PERFORMANCE GRAPH:-

Multiple corpora were assembled to estimate the parameters and assess the performance of the proposed models and approach. One such corpus, BUBShabda Sagara-2011 [17], consisted of a dictionary comprising 9000 words for training in Kannada/Telugu. This corpus included a bilingual word list used for training. In our experiments, the performance of word translation was evaluated based on precision and recall rates in bytes (7). Each experiment considered a single word in the source language and its corresponding translation in the target language. The system's performance was evaluated using a consistent set of 500 unique sentences, which were deliberately excluded from the training corpus.

The results of our experiments, as depicted in Fig. 5, demonstrate a significant improvement in performance. The systems exhibit markedly enhanced accuracy, which remains consistently competitive, especially with the expansion of the corpus size. This observation underscores the positive impact of increasing the dataset size on the effectiveness of our approach.

3. **RESULT:-**

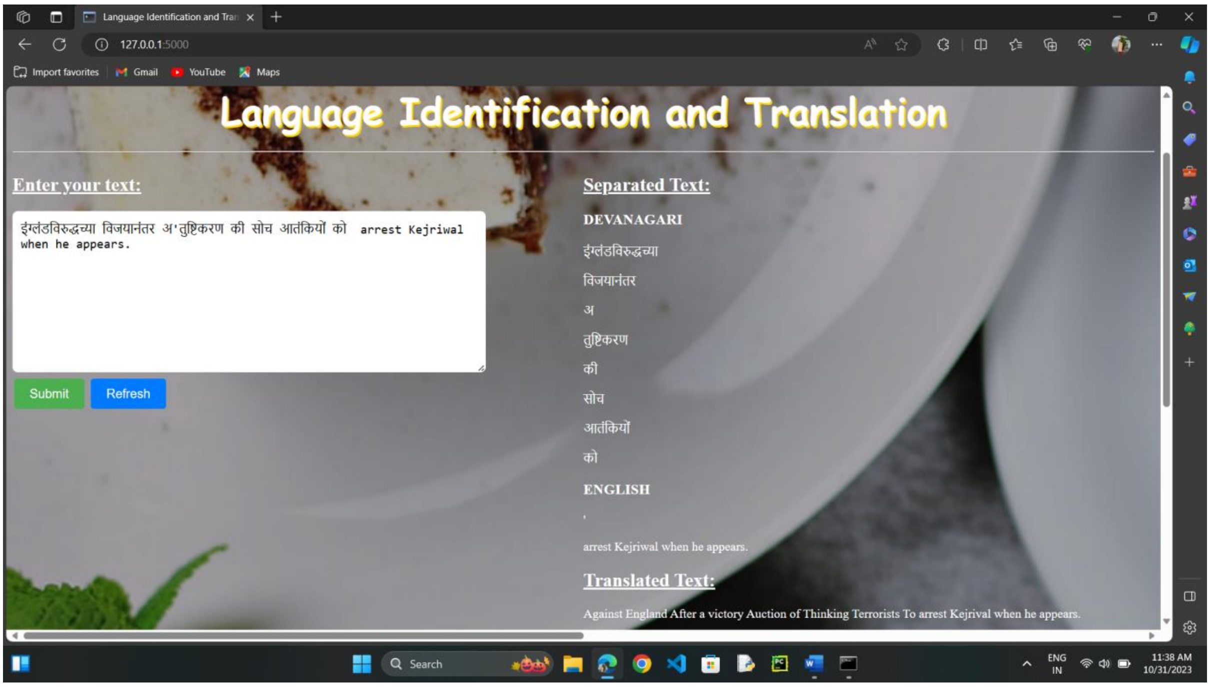


Language Identification and Translation

Enter your text:

No separated text available.

Translated Text:



Language Identification and Translation

Enter your text:

```
इंग्लंडविरुद्धच्या विजयानंतर अ'तुष्टिकरण की सोच आतंकियों को  arrest Kejriwal
when he appears.
```

Separated Text:

DEVANAGARI

इंग्लंडविरुद्धच्या

विजयानंतर

अ

तुष्टिकरण

की

सोच

आतंकियों

को

ENGLISH

'

arrest Kejriwal when he appears.

Translated Text:

Against England After a victory Auction of Thinking Terrorists To arrest Kejrival when he appears.

4. CONCLUSION AND NEXT STEPS:

We have introduced a language identification scheme based on deep neural networks, achieving nearly perfect accuracy in classifying different languages and approximately 90% accuracy in distinguishing highly similar languages. Notably, the languages within the West Slavic group presented considerable challenges. Despite efforts to expand the corpus of these languages using external sources, the improvement was limited. This limitation primarily stemmed from the absence of specific word n-grams unique to certain languages, which were not incorporated into the extended corpus.

To address this challenge, we employed convolutional and recurrent neural network models to identify structures distinctive to each language. However, due to our lack of expertise in these particular languages, we were unable to engineer additional features. Consequently, we believe that further enhancements can only be achieved by designing rule-based features, a task that necessitates consultation with language experts or native speakers from the respective communities.

ACKNOWLEDGEMENT

I owe my sincere feelings of gratitude to Prof. A. Gahankari for his valuable guidance and suggestions, which helped me a lot to write this paper. This paper is in continuation of the major research project entitled “Language identification web application” sanctioned to Prof. A. Gahankari, Department of Computer Science and Applications by the University Grant Commission carried out at the RTMNU Nagpur, India.

REFERENCES

[1] `Aparitosh Gahankari, Dr. Avinash S. Kapse, Dr. V. M. Thakre, " Word Sense Disambiguation - Supervised Approaches: Present Scenario", International Journal of Scientific Research in Science and Technology (IJSRST), Print ISSN: 2395-6011, Online ISSN: 2395-602X, Volume 5, Issue 6, pp.150-154, January-February-2020. Journal URL: https://ijsrst.com/IJSRST208230.

[2] Aparitosh Gahankari, Dr. Avinash S. Kapse, Dr.Mohammad Atique, Dr. V.M. Thakare & Dr. Arvind S. Kapse. (2022). "Word Sense Disambiguation in Marathi Language using Fast-Text model and Indo-Wordnet", Computer Integrated Manufacturing Systems, 28(11), 1607–1617. Retrieved from http://cims-journal.com/index.php/CN/article/view/376.

[3] Aparitosh Gahankari, etal,"A Review on Methods to Solve Polysemy Problem in WSD Occurring while Processing Marathi Text." International Journal of Advanced Research in Science, Communication and Technology (IJARSCT), vol. 2, issue 1, January 2022, ISSN (Online) 2581-9429, https://ijarsct.co.in/jani1.html, DOI: 10.48175/IJARSCT-2463

[4] Aparitosh Gahankari,Vaishnavi Wasankar, Yashika S. Nimje, Kajal Pardhi, Divyani C. Shende, "An Empirical Analysis of Word Sense Disambiguation through Machine Learning Approaches", International Journal of Scientific Research in Computer Science, Engineering

and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 8, Issue 2, pp.104-114, March-April-2022.

[5] ITU Telecommunication Development Bureau. Measuring digital development Facts and figures 2019. ITU Publications. https://www.itu.int/en/ITU-D/Statistics/ Documents/facts/FactsFigures2019.pdf, retrieved 2020-07-07.

[6] M.A. Nejla Qafmolla. Automatic Language Identification. European Journal of Language and Literature Studies Volume 3 Issue 1. 2017. 2411-4103.

[7] D. W. Otter, J. R. Medina, J. K. Kalita. Survey of the Usages of Deep Learning for Natural Language Processing. IEEE Transactions on Neural Networks and Learning Systems. April, 21, 2020.

[8] T. Jauhiainen, M. Lui, M. Zampieri, T. Baldwin, K. Lindén Automatic language identification: A survey. Journal of Artificial Intelligence research, Vol 65. August 25, 2019.

[9] S. Russell, P. Norvig, Artificial intelligence - A modern approach. Pearson education, New Jersey. Third edition, 2010.

[10] A. L. Samuel. Some studies in machine learning using the game of checkers. IBM Journal of research and development, 210-229. 1959.

[11] G. James, D. Witten, T. Hastie, R. Tibshirani. An Introduction to Statistical Learning with Applications in R. 7 ed. New York: Springer Science+Business Media. 2013.

[12] Wikipedia. Bayes' theorem. https://en.wikipedia.org/wiki/Bayes%27_theorem. Fetched 2020-06-05.

[13] A-M. Yaser S., M. Magdon-Ismail, H-T Lin. Learning From Data. A short course. AMLbook.com, 2012.

[14] S. Haykin. Neural networks and learning machines. Pearson education, third edition. 2009.

[15] A. Krizhevsky, I. Sutskever, G. Hinton. ImageNet Classification with Deep Convolutional Neural Networks. Neural Information Processing Systems. 25. 2012. 10.1145/3065386.

[16] A. Lindholm, N. Wahlström, F. Lindsten, T. B. Schön. Supervised Machine Learning. Department of Information Technology, Uppsala University. Available at: http://www.it.uu.se/edu/course/homepage/sml/literature/lecture_notes.pdf. March 2019.

[17] D. P. Kingma, J. Ba. Adam: A Method for Stochastic Optimization. Proceedings of the 3rd International Conference on Learning Representations (ICLR). 2014/12/22. arXiv:1412.6980.

[18] M. Sundermeyer, R. Schlüter, H. Ney. LSTM Neural Networks for Language Modeling. Human Language Technology and Pattern Recognition, Computer Science Department, RWTH Aachen University, Aachen, Germany.

[19] fastText's website. Resources. https://fasttext.cc/docs/en/ language-identification.html. Fetched 2020-06-12.

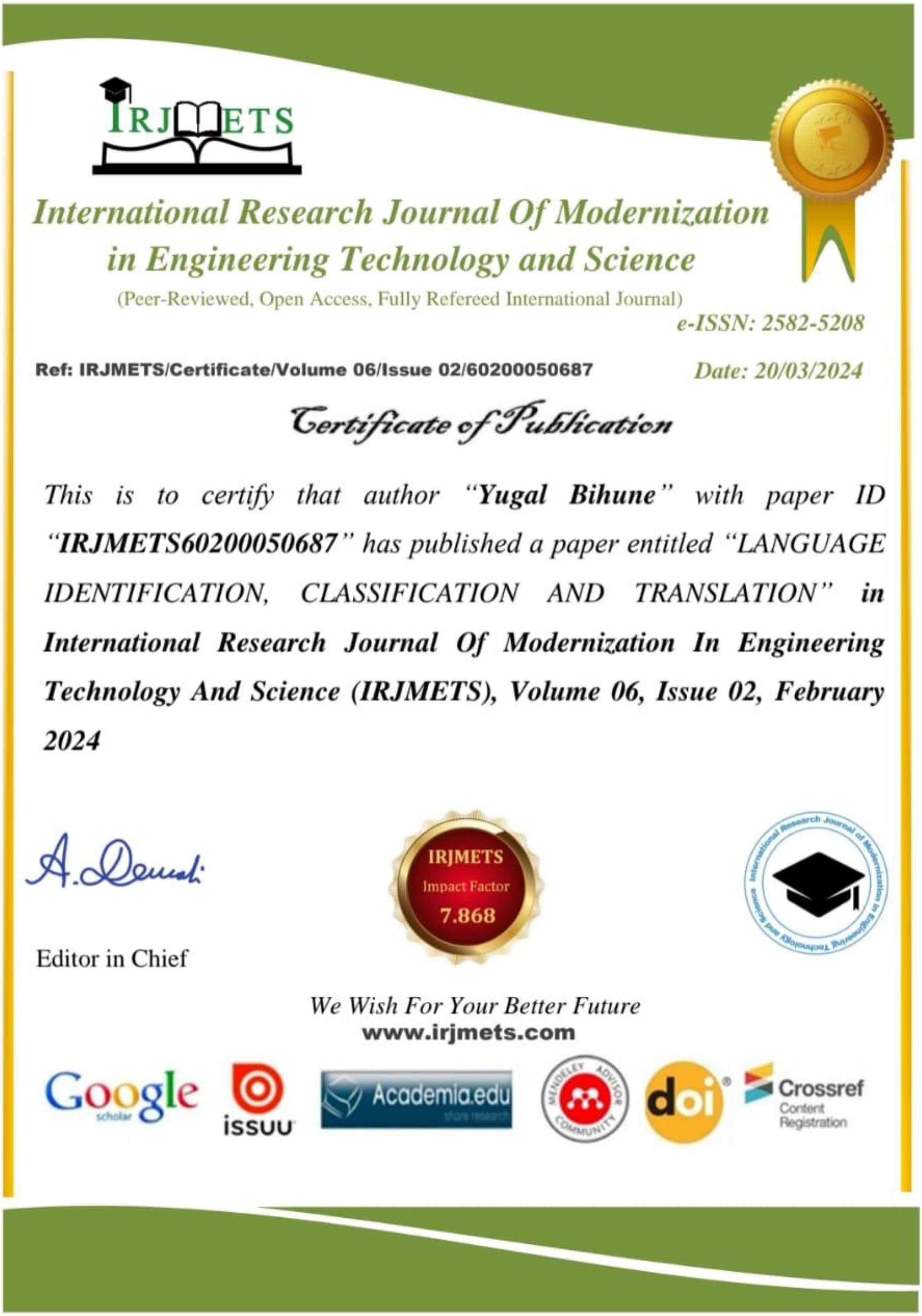
[20] A. Joulin, E. Grave, P. Bojanowski, T. Mikolov. Bag of Tricks for Efficient Text Classification. Facebook AI Research. August 9, 2016.

[21] P. Bojanowski, E. Grave, A. Joulin, T. Mikolov. Enriching Word Vectors with Subword Information. Facebook AI Research. July 16, 2016.

[22] I. Mozetic, M. Grcar, J. Smailovic. Multilingual Twitter Sentiment Classification: The Role of Human Annotators. PLOS ONE. May 5, 2016.

[23] Sarah Perez. Techcrunch blog.https://techcrunch.com/2017/11/07/ twitter-officially-expands-its-character-count-to-280-starting-today/. Published November 7, 2017. Fetched 2020-06-12.

PAPER PUBLISHED CERTIFICATE



IRJMETS

International Research Journal Of Modernization in Engineering Technology and Science

(Peer-Reviewed, Open Access, Fully Refereed International Journal)

e-ISSN: 2582-5208

Ref: IRJMETS/Certificate/Volume 06/Issue 02/60200050687 *Date: 20/03/2024*

Certificate of Publication

*This is to certify that author "**Yugal Bihune**" with paper ID "**IRJMETS60200050687**" has published a paper entitled "LANGUAGE IDENTIFICATION, CLASSIFICATION AND TRANSLATION" **in International Research Journal Of Modernization In Engineering Technology And Science (IRJMETS), Volume 06, Issue 02, February 2024***

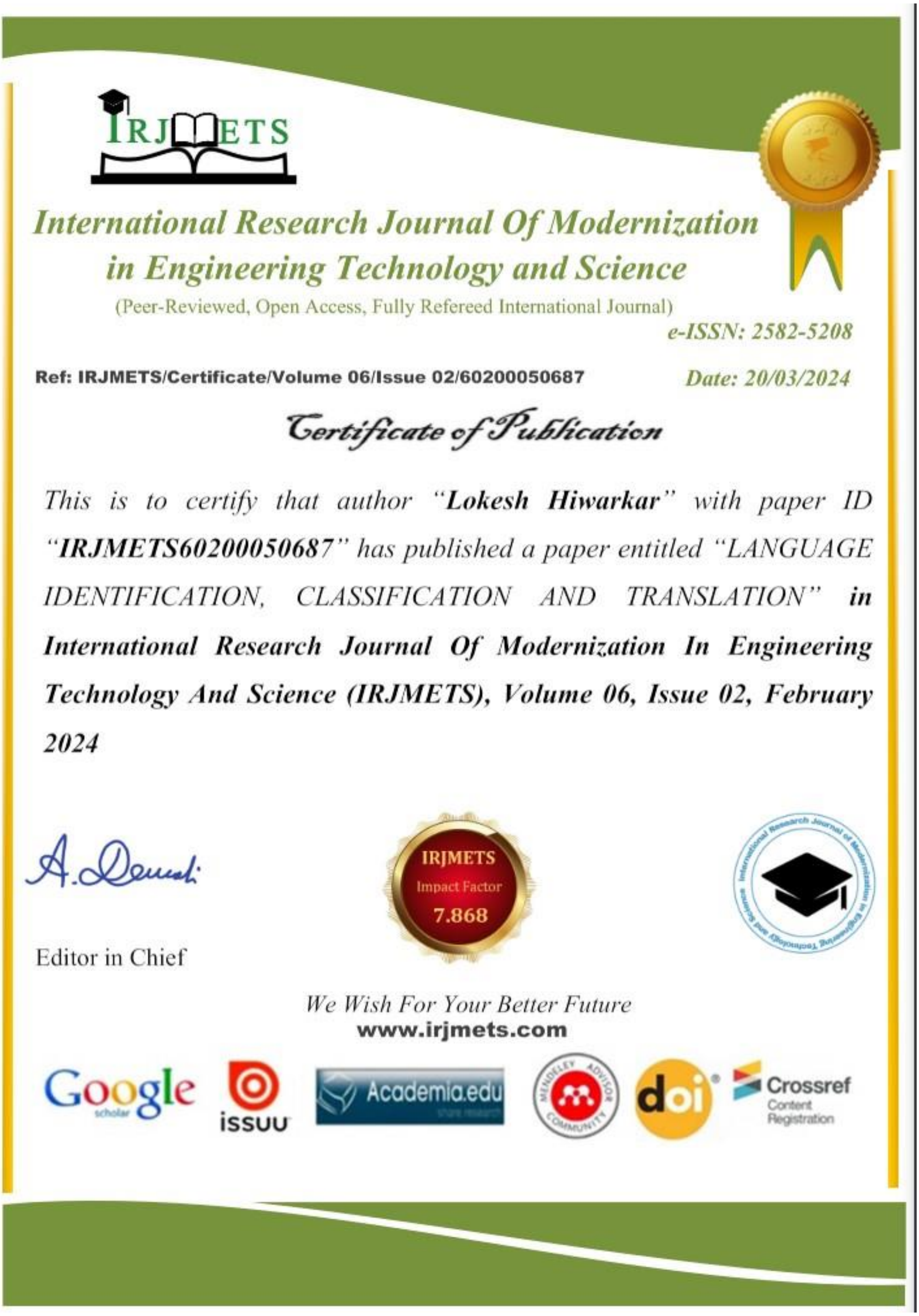
Editor in Chief

IRJMETS
Impact Factor
7.868

International Research Journal of Modernization in Engineering Technology and Science

We Wish For Your Better Future
www.irjmets.com

Google scholar issuu Academia.edu MENDELEY ADVISOR COMMUNITY doi Crossref Content Registration



IRJMETS

International Research Journal Of Modernization in Engineering Technology and Science

(Peer-Reviewed, Open Access, Fully Refereed International Journal)

e-ISSN: 2582-5208

Ref: IRJMETS/Certificate/Volume 06/Issue 02/60200050687 *Date: 20/03/2024*

Certificate of Publication

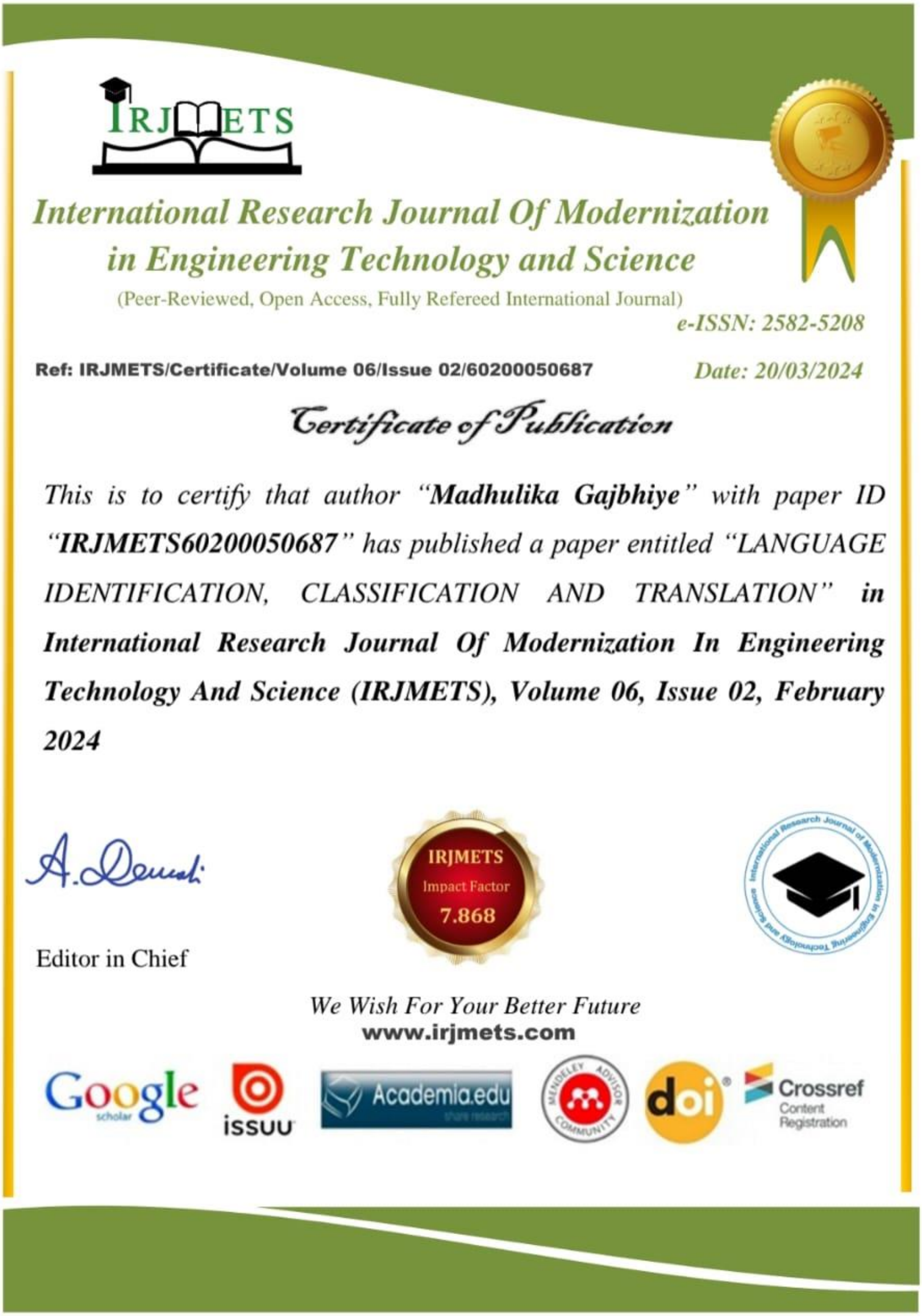
*This is to certify that author "**Lokesh Hiwarkar**" with paper ID "**IRJMETS60200050687**" has published a paper entitled "LANGUAGE IDENTIFICATION, CLASSIFICATION AND TRANSLATION" **in International Research Journal Of Modernization In Engineering Technology And Science (IRJMETS), Volume 06, Issue 02, February 2024***

Editor in Chief

IRJMETS Impact Factor 7.868

We Wish For Your Better Future
www.irjmets.com

Google scholar · issuu · Academia.edu · Mendeley Advisor Community · doi · Crossref Content Registration



IRJMETS

International Research Journal Of Modernization in Engineering Technology and Science

(Peer-Reviewed, Open Access, Fully Refereed International Journal)

e-ISSN: 2582-5208

Ref: IRJMETS/Certificate/Volume 06/Issue 02/60200050687 *Date: 20/03/2024*

Certificate of Publication

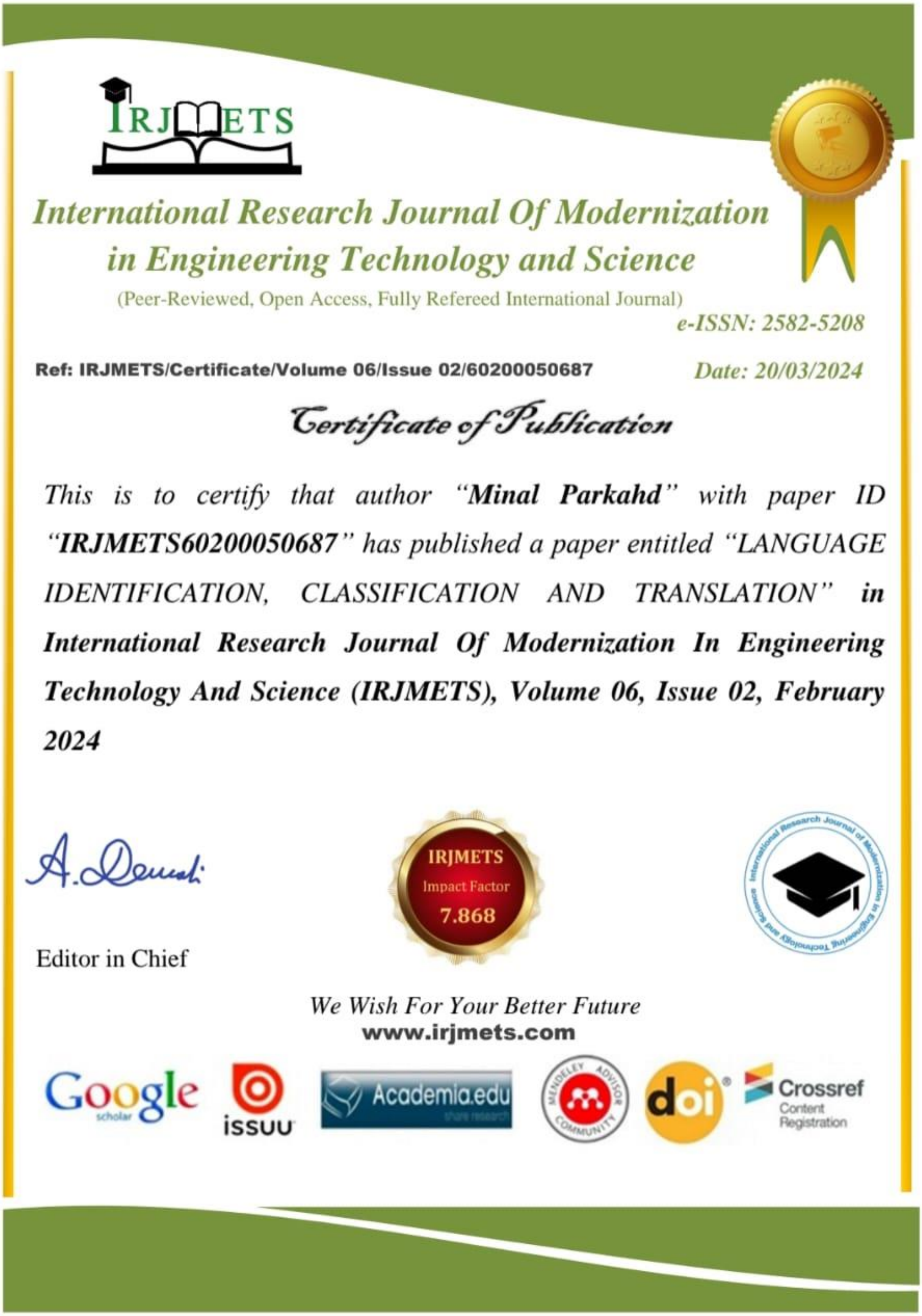
*This is to certify that author "**Madhulika Gajbhiye**" with paper ID "**IRJMETS60200050687**" has published a paper entitled "LANGUAGE IDENTIFICATION, CLASSIFICATION AND TRANSLATION"* ***in International Research Journal Of Modernization In Engineering Technology And Science (IRJMETS), Volume 06, Issue 02, February 2024***

IRJMETS Impact Factor 7.868

Editor in Chief

We Wish For Your Better Future

www.irjmets.com



IRJMETS

International Research Journal Of Modernization in Engineering Technology and Science

(Peer-Reviewed, Open Access, Fully Refereed International Journal)

e-ISSN: 2582-5208

Ref: IRJMETS/Certificate/Volume 06/Issue 02/60200050687 *Date: 20/03/2024*

Certificate of Publication

*This is to certify that author "**Minal Parkahd**" with paper ID "**IRJMETS60200050687**" has published a paper entitled "LANGUAGE IDENTIFICATION, CLASSIFICATION AND TRANSLATION"* ***in International Research Journal Of Modernization In Engineering Technology And Science (IRJMETS), Volume 06, Issue 02, February 2024***

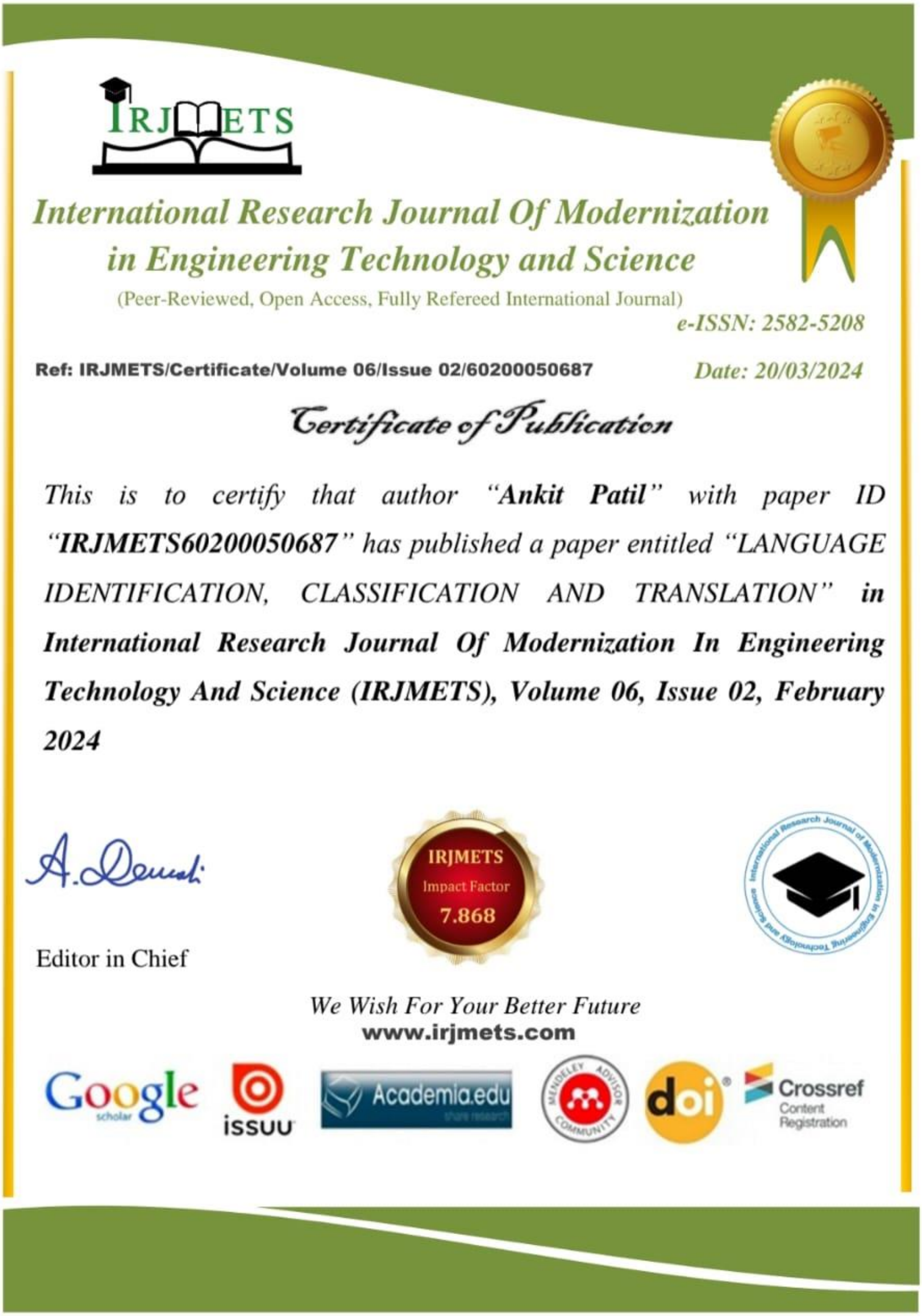
Editor in Chief

IRJMETS
Impact Factor
7.868

International Research Journal of Modernization in Engineering Technology and Science

We Wish For Your Better Future
www.irjmets.com

Google scholar · issuu · Academia.edu share research · Mendeley Advisor Community · doi · Crossref Content Registration



IRJMETS

International Research Journal Of Modernization in Engineering Technology and Science

(Peer-Reviewed, Open Access, Fully Refereed International Journal)

e-ISSN: 2582-5208

Ref: IRJMETS/Certificate/Volume 06/Issue 02/60200050687 *Date: 20/03/2024*

Certificate of Publication

*This is to certify that author "**Ankit Patil**" with paper ID "**IRJMETS60200050687**" has published a paper entitled "LANGUAGE IDENTIFICATION, CLASSIFICATION AND TRANSLATION" **in International Research Journal Of Modernization In Engineering Technology And Science (IRJMETS), Volume 06, Issue 02, February 2024***

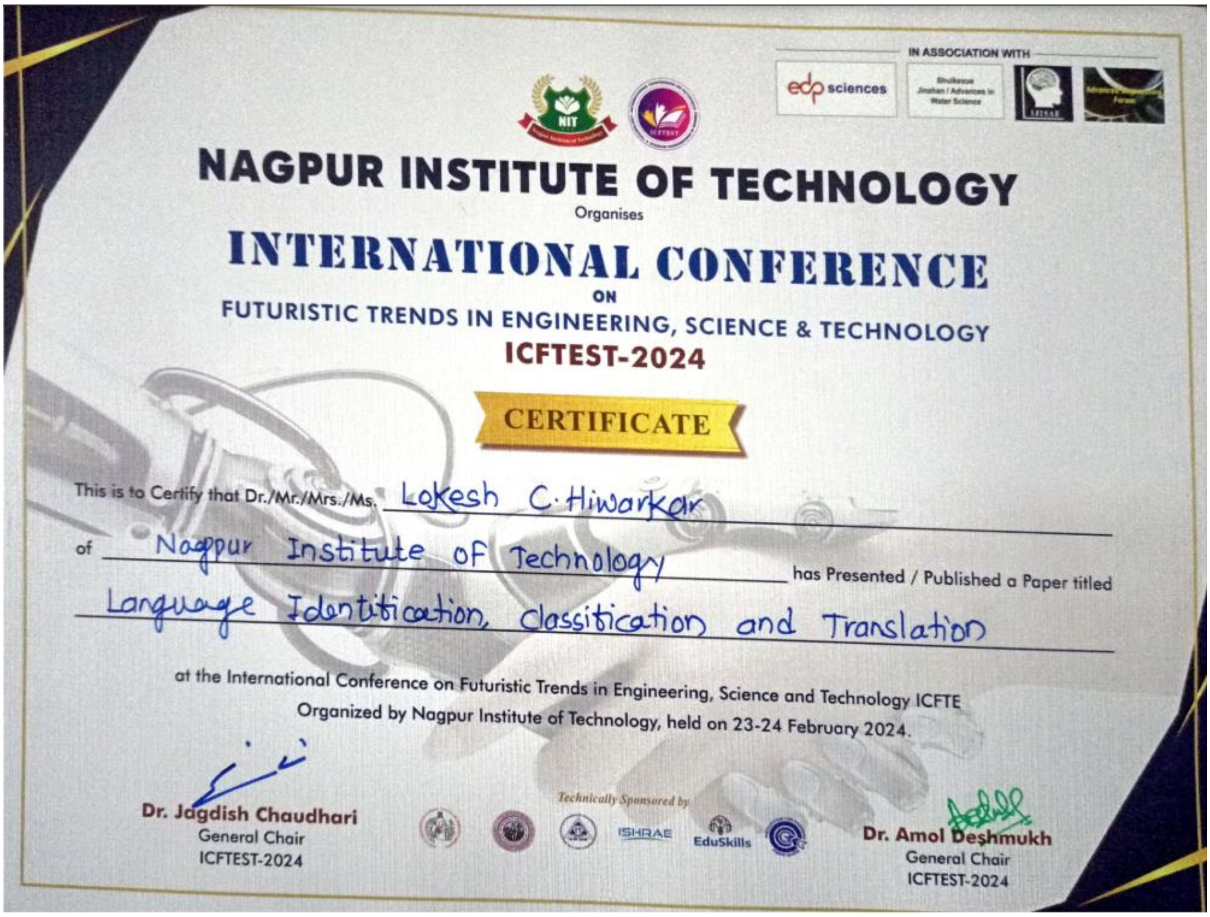
Editor in Chief

IRJMETS
Impact Factor
7.868

International Research Journal of Modernization in Engineering Technology and Science

We Wish For Your Better Future
www.irjmets.com

Google scholar | issuu | Academia.edu share research | Mendeley Advisor Community | doi | Crossref Content Registration



IN ASSOCIATION WITH

edp sciences

Shuikexue Jinzhan / Advances in Water Science

IJISAE

Advanced Engineering Forum

NAGPUR INSTITUTE OF TECHNOLOGY

Organises

INTERNATIONAL CONFERENCE

ON

FUTURISTIC TRENDS IN ENGINEERING, SCIENCE & TECHNOLOGY

ICFTEST-2024

CERTIFICATE

This is to Certify that Dr./Mr./Mrs./Ms. Lokesh C. Hiwarkar

of Nagpur Institute of Technology has Presented / Published a Paper titled

Language Identification, classification and Translation

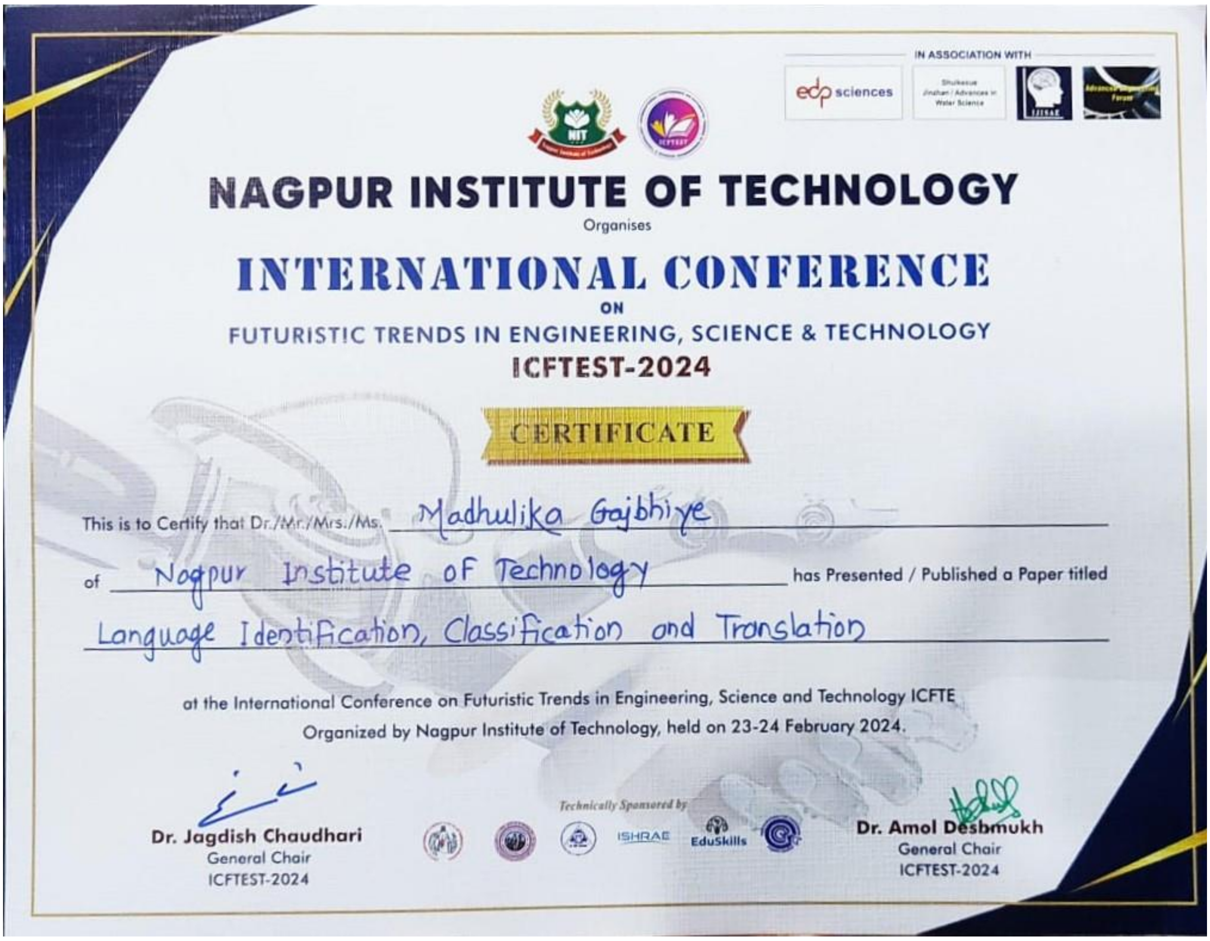
at the International Conference on Futuristic Trends in Engineering, Science and Technology ICFTE Organized by Nagpur Institute of Technology, held on 23-24 February 2024.

Technically Sponsored by

ISHRAE EduSkills

Dr. Jagdish Chaudhari
General Chair
ICFTEST-2024

Dr. Amol Deshmukh
General Chair
ICFTEST-2024



IN ASSOCIATION WITH

edp sciences

Shuikexue Jinzhan / Advances in Water Science

Advanced Engineering Forum

NAGPUR INSTITUTE OF TECHNOLOGY

Organises

INTERNATIONAL CONFERENCE

ON

FUTURISTIC TRENDS IN ENGINEERING, SCIENCE & TECHNOLOGY

ICFTEST-2024

CERTIFICATE

This is to Certify that Dr./~~Mr.~~/Mrs./Ms. Madhulika Gajbhiye

of Nagpur Institute of Technology has Presented / Published a Paper titled

Language Identification, Classification and Translation

at the International Conference on Futuristic Trends in Engineering, Science and Technology ICFTE
Organized by Nagpur Institute of Technology, held on 23-24 February 2024.

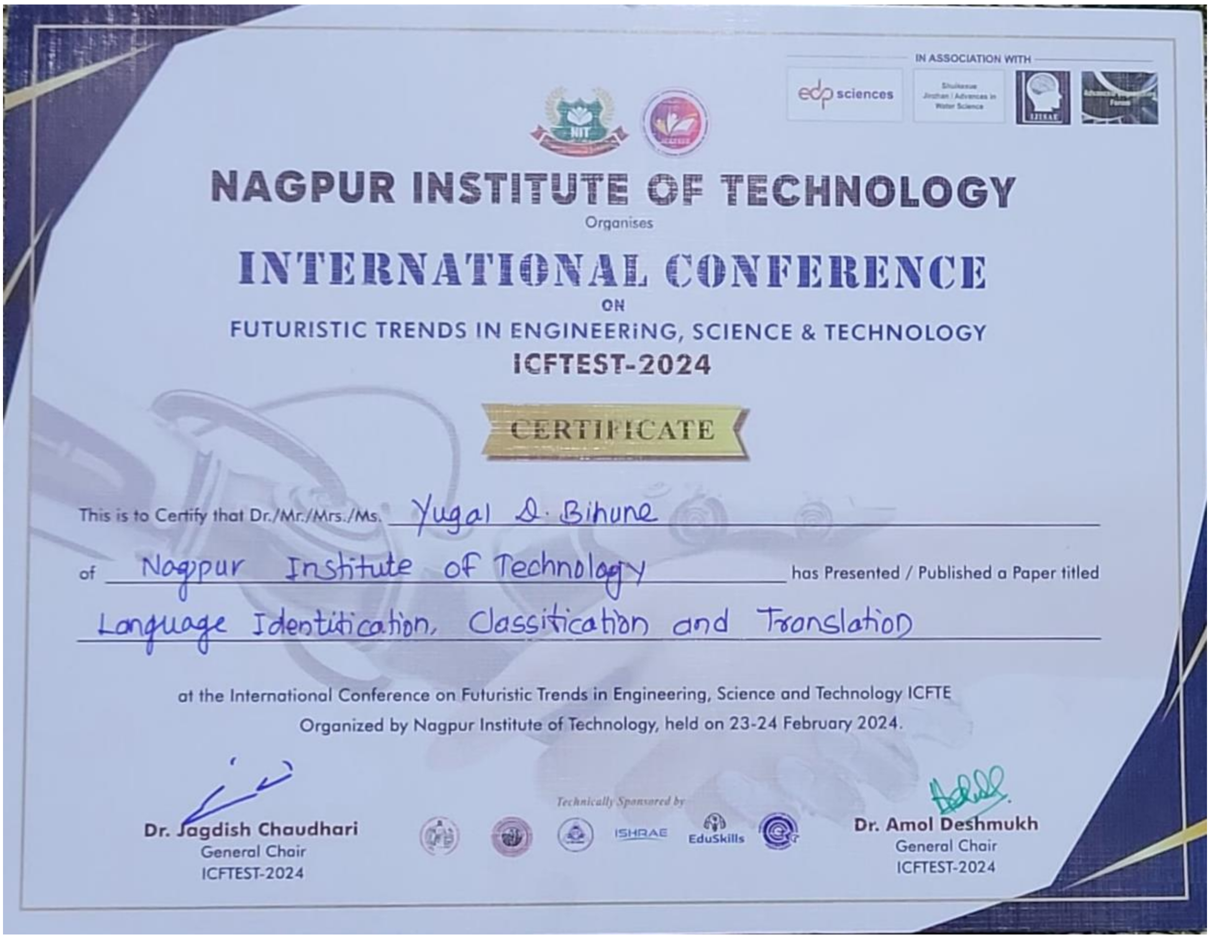
Technically Sponsored by

ISHRAE

EduSkills

Dr. Jagdish Chaudhari
General Chair
ICFTEST-2024

Dr. Amol Desbmukh
General Chair
ICFTEST-2024



IN ASSOCIATION WITH

edp sciences

Shuikexue Jinzhan / Advances in Water Science

NAGPUR INSTITUTE OF TECHNOLOGY

Organises

INTERNATIONAL CONFERENCE

ON

FUTURISTIC TRENDS IN ENGINEERING, SCIENCE & TECHNOLOGY

ICFTEST-2024

CERTIFICATE

This is to Certify that Dr./Mr./Mrs./Ms. Yugal S. Bihune

of Nagpur Institute of Technology has Presented / Published a Paper titled

Language Identification, Classification and Translation

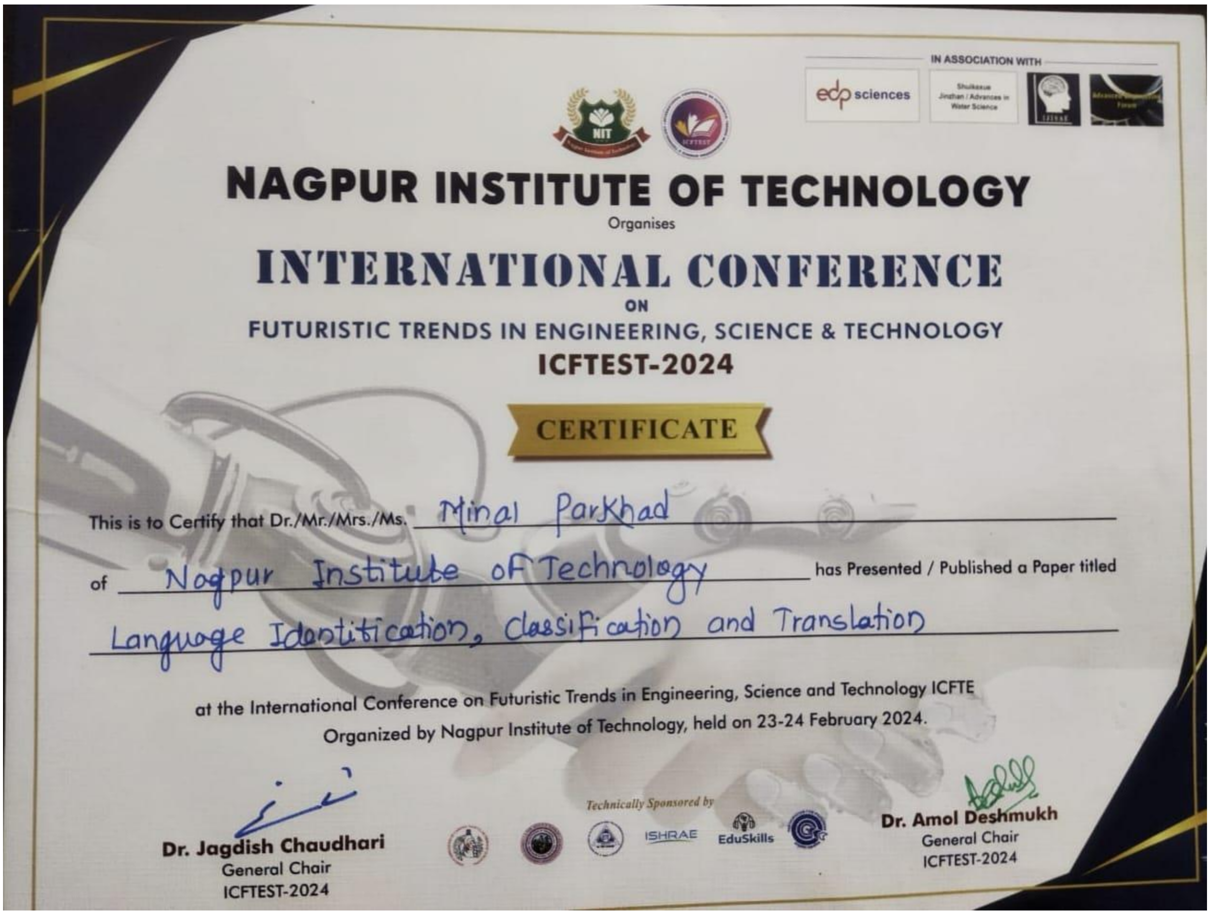
at the International Conference on Futuristic Trends in Engineering, Science and Technology ICFTE Organized by Nagpur Institute of Technology, held on 23-24 February 2024.

Dr. Jagdish Chaudhari
General Chair
ICFTEST-2024

Technically Sponsored by

ISHRAE EduSkills

Dr. Amol Deshmukh
General Chair
ICFTEST-2024



IN ASSOCIATION WITH

edp sciences

Shuikexue Jinzhan / Advances in Water Science

NAGPUR INSTITUTE OF TECHNOLOGY

Organises

INTERNATIONAL CONFERENCE

ON

FUTURISTIC TRENDS IN ENGINEERING, SCIENCE & TECHNOLOGY

ICFTEST-2024

CERTIFICATE

This is to Certify that Dr./Mr./Mrs./Ms. Minal Parkhad

of Nagpur Institute of Technology has Presented / Published a Paper titled

Language Identitication, classification and Translation

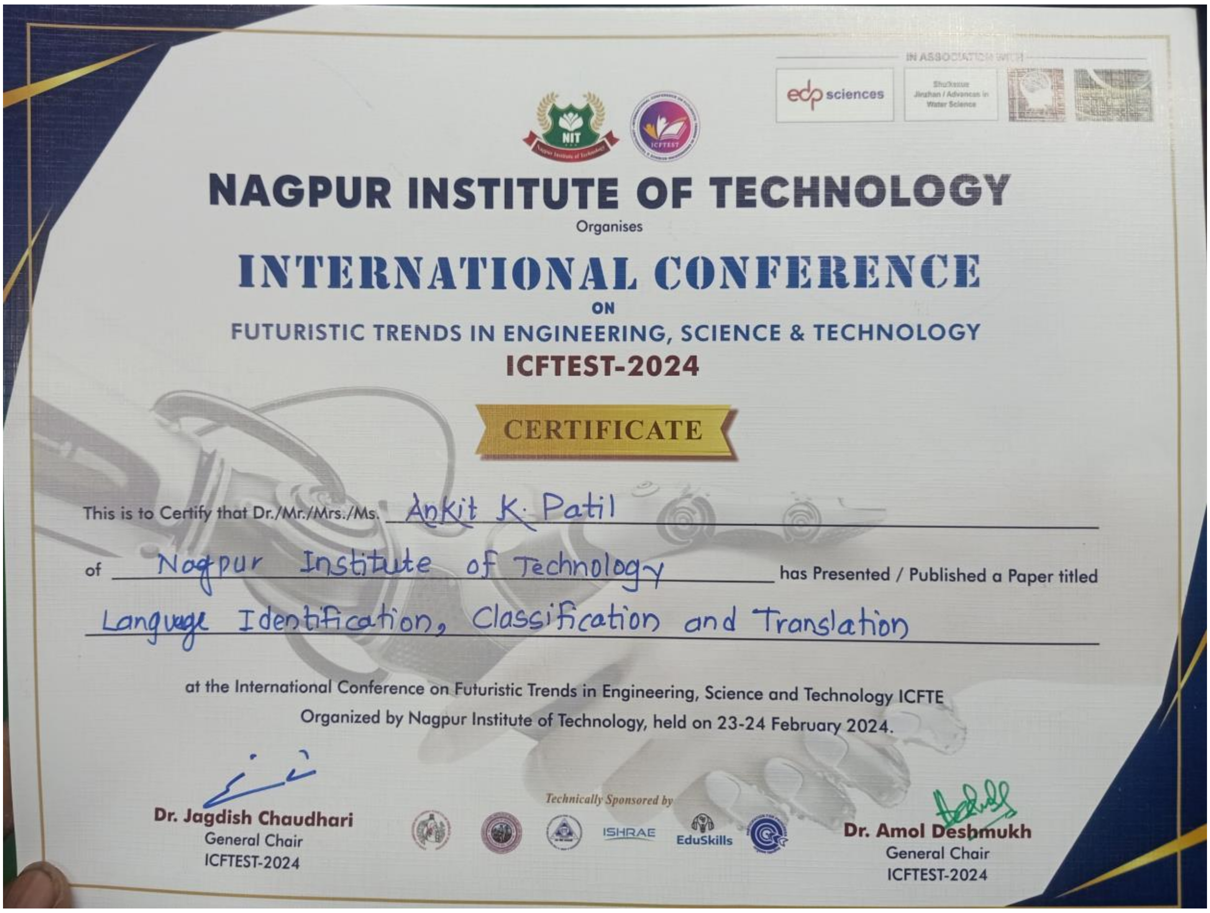
at the International Conference on Futuristic Trends in Engineering, Science and Technology ICFTE Organized by Nagpur Institute of Technology, held on 23-24 February 2024.

Technically Sponsored by

ISHRAE EduSkills

Dr. Jagdish Chaudhari
General Chair
ICFTEST-2024

Dr. Amol Deshmukh
General Chair
ICFTEST-2024



IN ASSOCIATION WITH

edp sciences

Shuikexue Jinzhan / Advances in Water Science

NAGPUR INSTITUTE OF TECHNOLOGY

Organises

INTERNATIONAL CONFERENCE

ON

FUTURISTIC TRENDS IN ENGINEERING, SCIENCE & TECHNOLOGY

ICFTEST-2024

CERTIFICATE

This is to Certify that Dr./Mr./Mrs./Ms. Ankit K. Patil

of Nagpur Institute of Technology has Presented / Published a Paper titled

Language Identification, Classification and Translation

at the International Conference on Futuristic Trends in Engineering, Science and Technology ICFTE Organized by Nagpur Institute of Technology, held on 23-24 February 2024.

Dr. Jagdish Chaudhari
General Chair
ICFTEST-2024

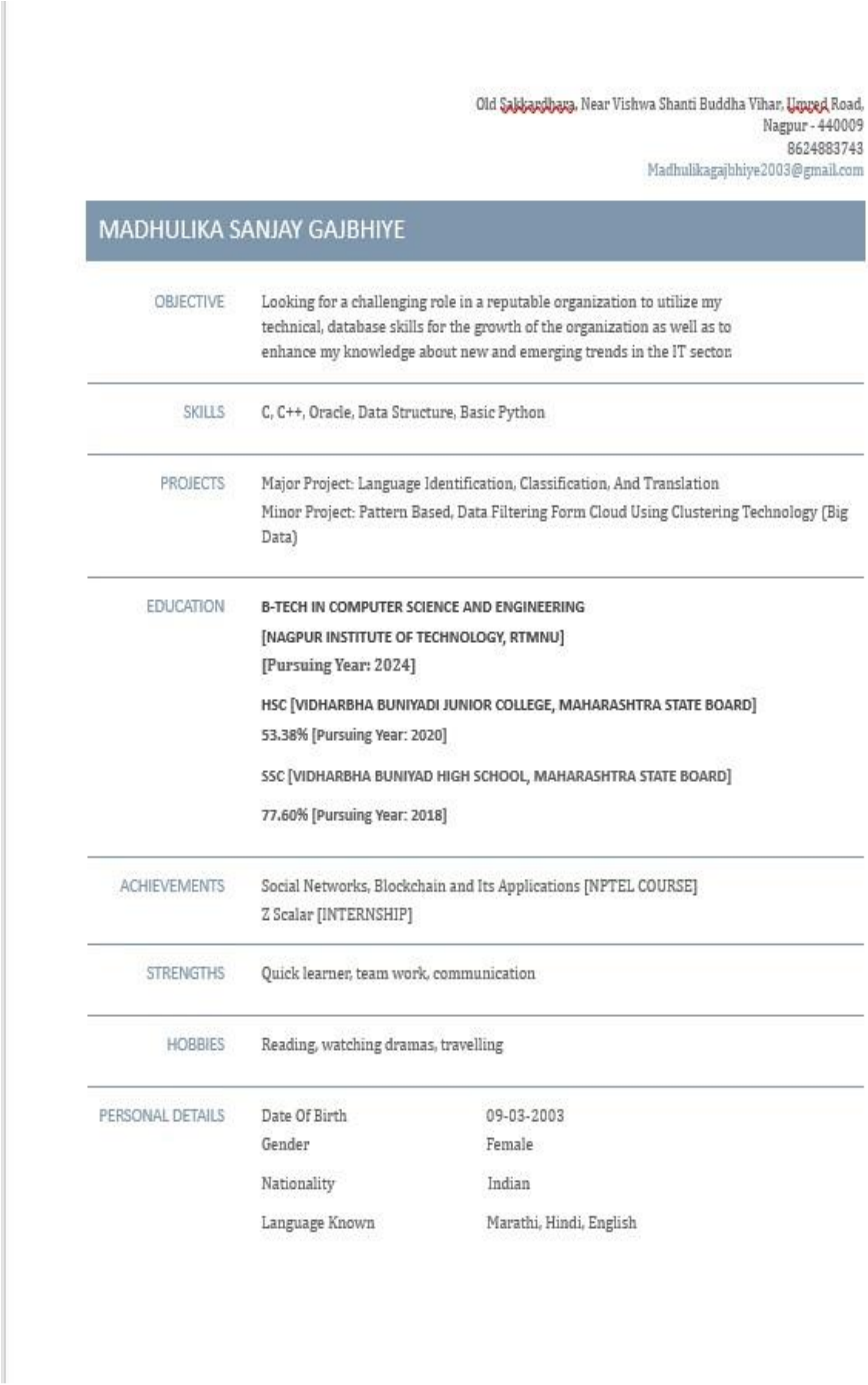
Technically Sponsored by

ISHRAE

EduSkills

Dr. Amol Deshmukh
General Chair
ICFTEST-2024

ALL TEAM MEMBERS RESUME



Old Sakkardhara, Near Vishwa Shanti Buddha Vihar, Umred Road,
Nagpur - 440009
8624883743
Madhulikagajbhiye2003@gmail.com

MADHULIKA SANJAY GAJBHIYE

OBJECTIVE

Looking for a challenging role in a reputable organization to utilize my technical, database skills for the growth of the organization as well as to enhance my knowledge about new and emerging trends in the IT sector.

SKILLS

C, C++, Oracle, Data Structure, Basic Python

PROJECTS

Major Project: Language Identification, Classification, And Translation

Minor Project: Pattern Based, Data Filtering Form Cloud Using Clustering Technology (Big Data)

EDUCATION

B-TECH IN COMPUTER SCIENCE AND ENGINEERING
[NAGPUR INSTITUTE OF TECHNOLOGY, RTMNU]
[Pursuing Year: 2024]

HSC [VIDHARBHA BUNIYADI JUNIOR COLLEGE, MAHARASHTRA STATE BOARD]
53.38% [Pursuing Year: 2020]

SSC [VIDHARBHA BUNIYAD HIGH SCHOOL, MAHARASHTRA STATE BOARD]

77.60% [Pursuing Year: 2018]

ACHIEVEMENTS

Social Networks, Blockchain and Its Applications [NPTEL COURSE]
Z Scalar [INTERNSHIP]

STRENGTHS

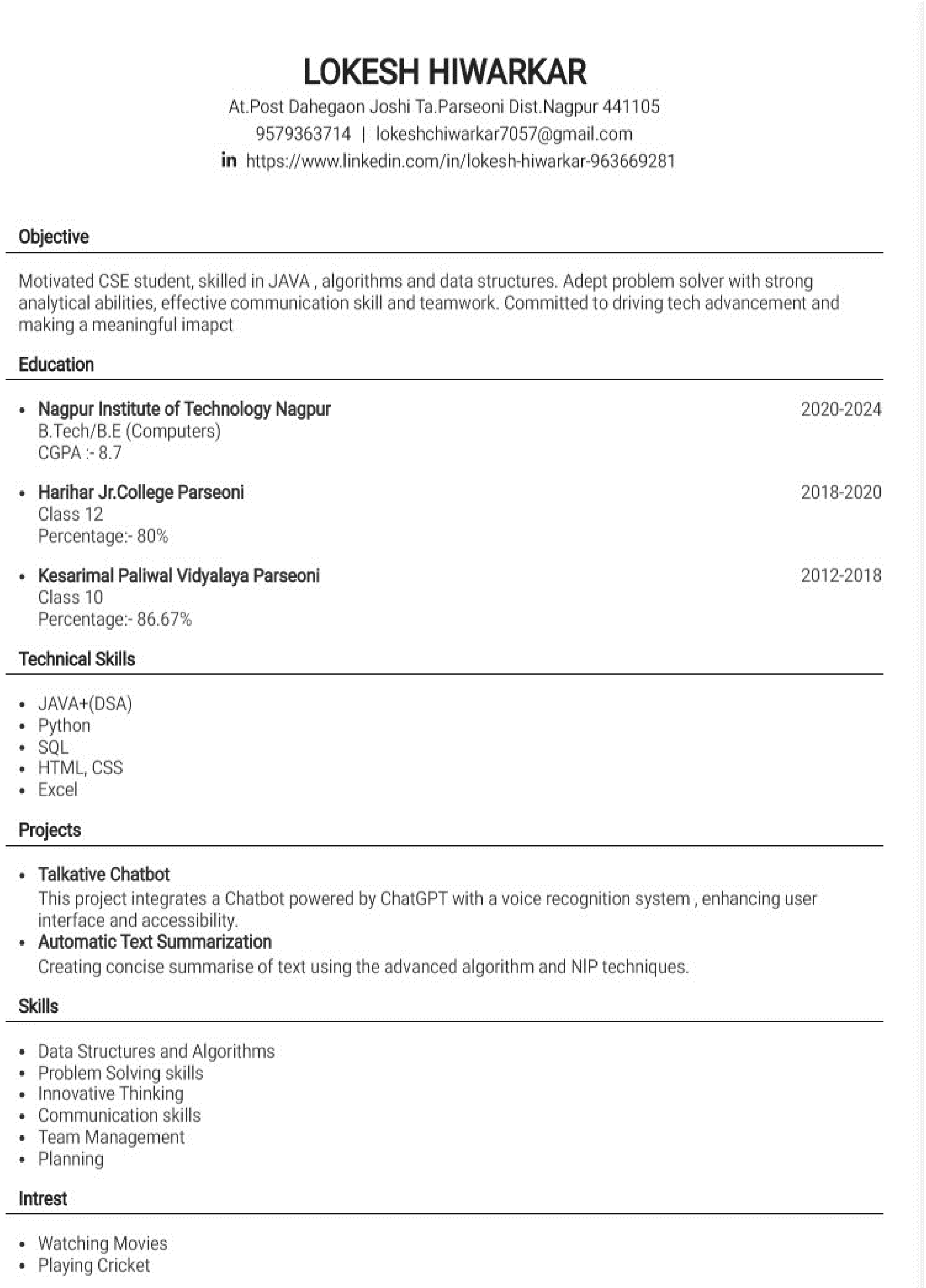
Quick learner, team work, communication

HOBBIES

Reading, watching dramas, travelling

PERSONAL DETAILS

Date Of Birth	09-03-2003
Gender	Female
Nationality	Indian
Language Known	Marathi, Hindi, English



LOKESH HIWARKAR

At.Post Dahegaon Joshi Ta.Parseoni Dist.Nagpur 441105
9579363714 | lokeshchiwarkar7057@gmail.com
in https://www.linkedin.com/in/lokesh-hiwarkar-963669281

Objective

Motivated CSE student, skilled in JAVA , algorithms and data structures. Adept problem solver with strong analytical abilities, effective communication skill and teamwork. Committed to driving tech advancement and making a meaningful imapct

Education

- **Nagpur Institute of Technology Nagpur** 2020-2024
  B.Tech/B.E (Computers)
  CGPA :- 8.7
- **Harihar Jr.College Parseoni** 2018-2020
  Class 12
  Percentage:- 80%
- **Kesarimal Paliwal Vidyalaya Parseoni** 2012-2018
  Class 10
  Percentage:- 86.67%

Technical Skills

- JAVA+(DSA)
- Python
- SQL
- HTML, CSS
- Excel

Projects

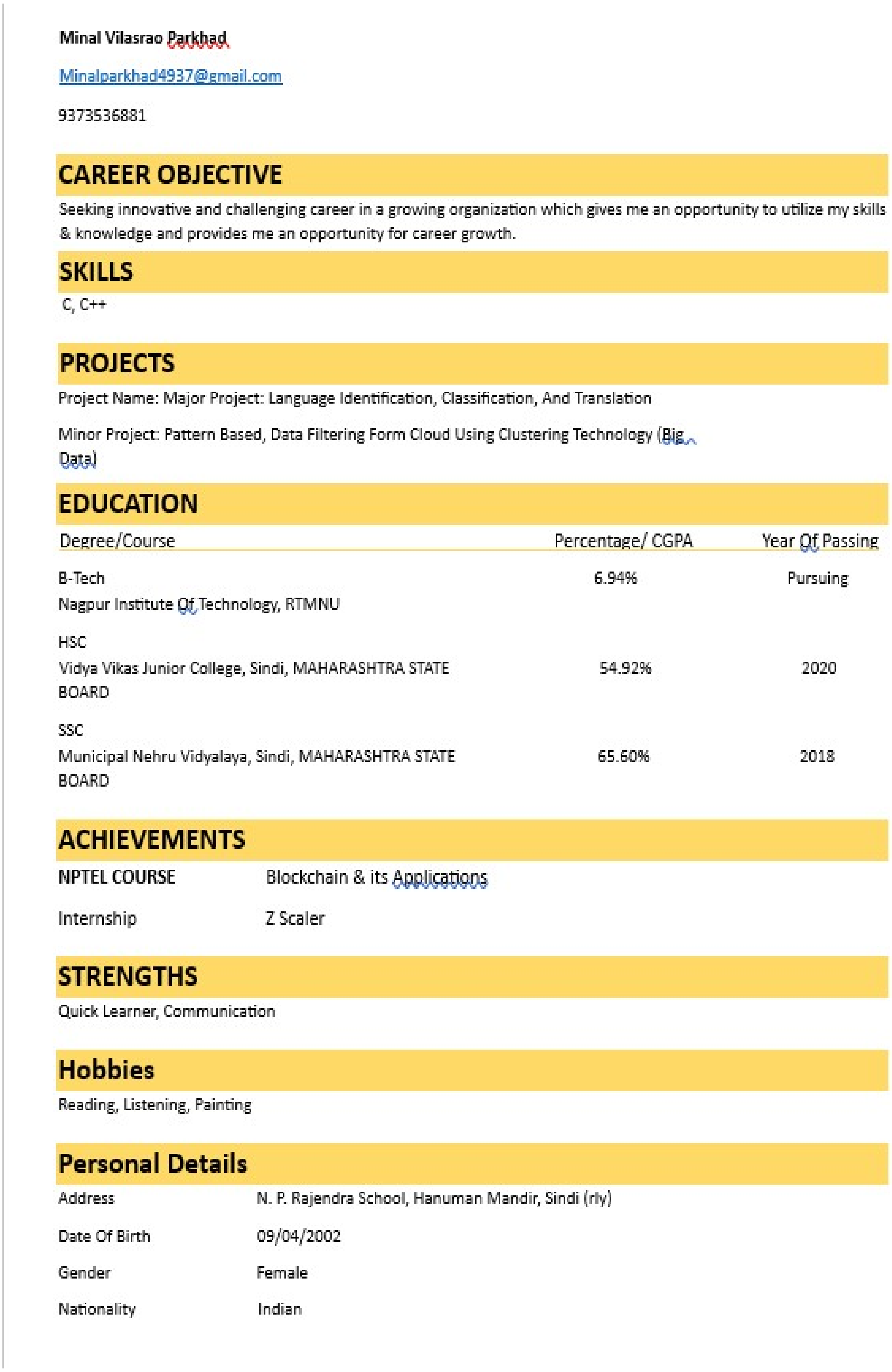
- **Talkative Chatbot**
  This project integrates a Chatbot powered by ChatGPT with a voice recognition system , enhancing user interface and accessibility.
- **Automatic Text Summarization**
  Creating concise summarise of text using the advanced algorithm and NIP techniques.

Skills

- Data Structures and Algorithms
- Problem Solving skills
- Innovative Thinking
- Communication skills
- Team Management
- Planning

Intrest

- Watching Movies
- Playing Cricket



Minal Vilasrao Parkhad

Minalparkhad4937@gmail.com

9373536881

CAREER OBJECTIVE

Seeking innovative and challenging career in a growing organization which gives me an opportunity to utilize my skills & knowledge and provides me an opportunity for career growth.

SKILLS

C, C++

PROJECTS

Project Name: Major Project: Language Identification, Classification, And Translation

Minor Project: Pattern Based, Data Filtering Form Cloud Using Clustering Technology (Big Data)

EDUCATION

Degree/Course	Percentage/ CGPA	Year Of Passing
B-Tech Nagpur Institute Of Technology, RTMNU	6.94%	Pursuing
HSC Vidya Vikas Junior College, Sindi, MAHARASHTRA STATE BOARD	54.92%	2020
SSC Municipal Nehru Vidyalaya, Sindi, MAHARASHTRA STATE BOARD	65.60%	2018

ACHIEVEMENTS

NPTEL COURSE	Blockchain & its Applications
Internship	Z Scaler

STRENGTHS

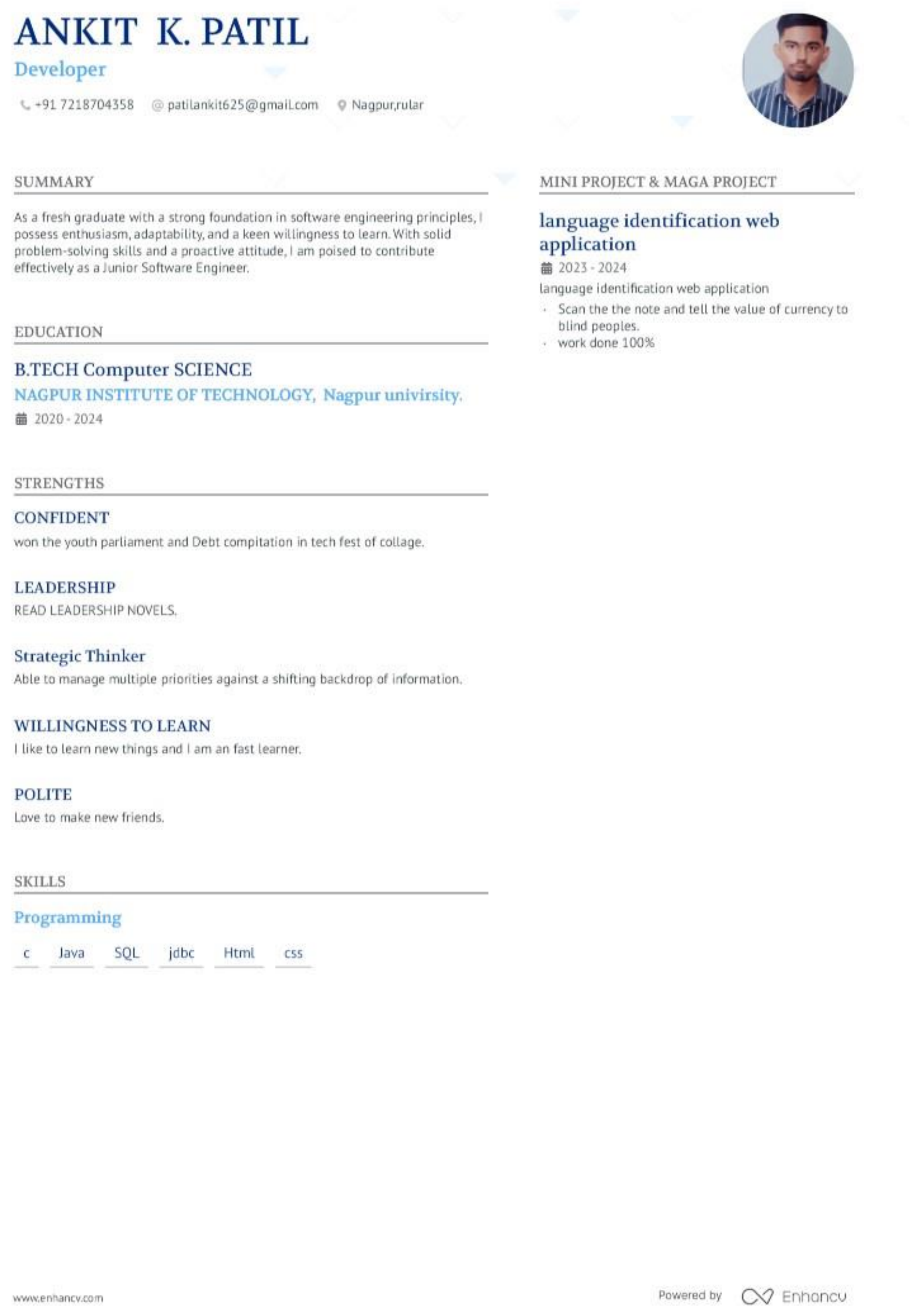
Quick Learner, Communication

Hobbies

Reading, Listening, Painting

Personal Details

Address	N. P. Rajendra School, Hanuman Mandir, Sindi (rly)
Date Of Birth	09/04/2002
Gender	Female
Nationality	Indian



ANKIT K. PATIL

Developer

☎ +91 7218704358 @ patilankit625@gmail.com ⚲ Nagpur,rular

SUMMARY

As a fresh graduate with a strong foundation in software engineering principles, I possess enthusiasm, adaptability, and a keen willingness to learn. With solid problem-solving skills and a proactive attitude, I am poised to contribute effectively as a Junior Software Engineer.

EDUCATION

B.TECH Computer SCIENCE

NAGPUR INSTITUTE OF TECHNOLOGY, Nagpur univirsity.

2020 - 2024

STRENGTHS

CONFIDENT

won the youth parliament and Debt compitation in tech fest of collage.

LEADERSHIP

READ LEADERSHIP NOVELS.

Strategic Thinker

Able to manage multiple priorities against a shifting backdrop of information.

WILLINGNESS TO LEARN

I like to learn new things and I am an fast learner.

POLITE

Love to make new friends.

SKILLS

Programming

c Java SQL jdbc Html css

MINI PROJECT & MAGA PROJECT

language identification web application

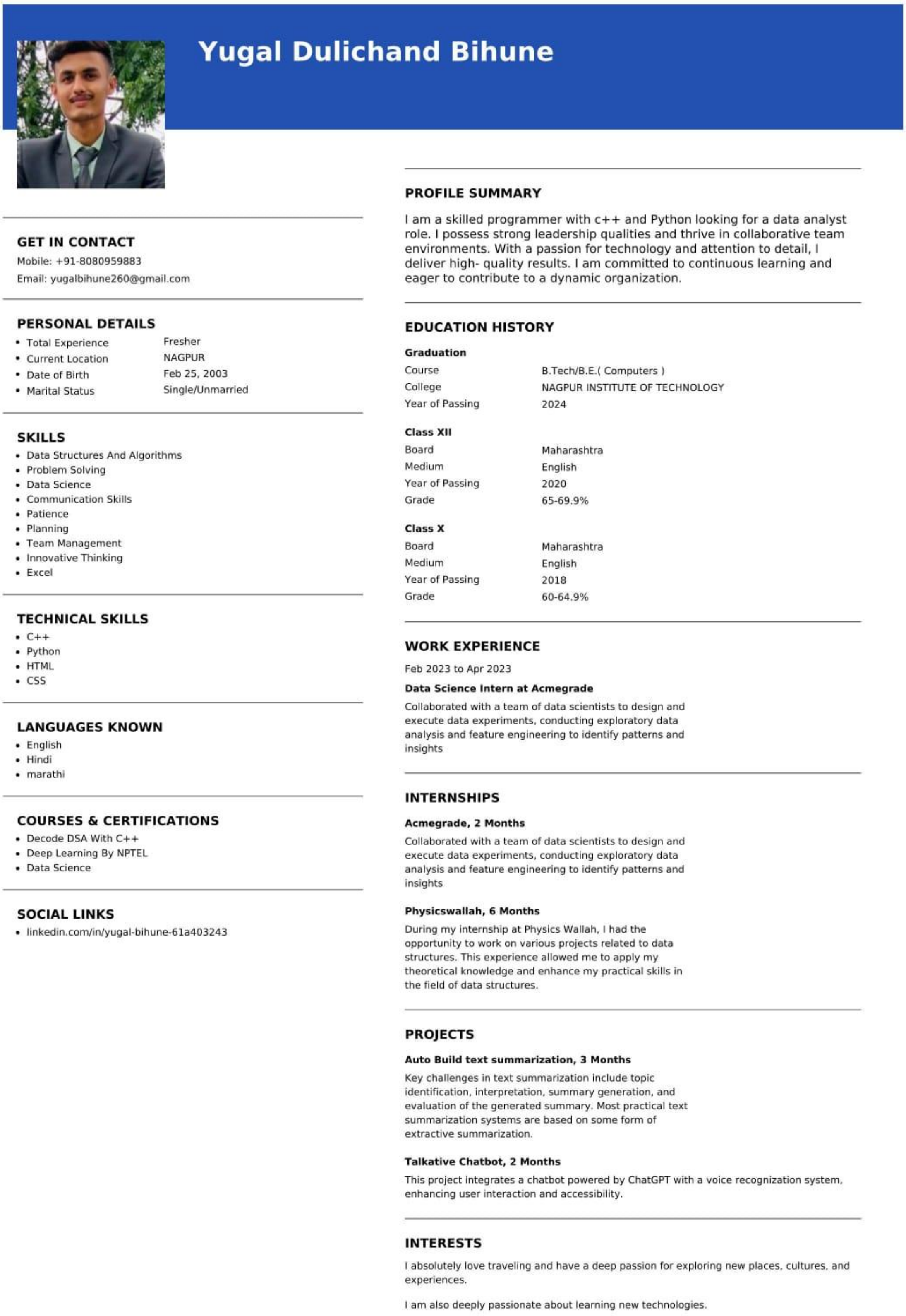
2023 - 2024

language identification web application

- Scan the the note and tell the value of currency to blind peoples.
- work done 100%

www.enhancv.com

Powered by Enhancv



Yugal Dulichand Bihune

PROFILE SUMMARY

I am a skilled programmer with c++ and Python looking for a data analyst role. I possess strong leadership qualities and thrive in collaborative team environments. With a passion for technology and attention to detail, I deliver high- quality results. I am committed to continuous learning and eager to contribute to a dynamic organization.

GET IN CONTACT

Mobile: +91-8080959883

Email: yugalbihune260@gmail.com

PERSONAL DETAILS

• Total Experience	Fresher
• Current Location	NAGPUR
• Date of Birth	Feb 25, 2003
• Marital Status	Single/Unmarried

SKILLS

- Data Structures And Algorithms
- Problem Solving
- Data Science
- Communication Skills
- Patience
- Planning
- Team Management
- Innovative Thinking
- Excel

TECHNICAL SKILLS

- C++
- Python
- HTML
- CSS

LANGUAGES KNOWN

- English
- Hindi
- marathi

COURSES & CERTIFICATIONS

- Decode DSA With C++
- Deep Learning By NPTEL
- Data Science

SOCIAL LINKS

- linkedin.com/in/yugal-bihune-61a403243

EDUCATION HISTORY

Graduation

Course	B.Tech/B.E.(Computers)
College	NAGPUR INSTITUTE OF TECHNOLOGY
Year of Passing	2024

Class XII

Board	Maharashtra
Medium	English
Year of Passing	2020
Grade	65-69.9%

Class X

Board	Maharashtra
Medium	English
Year of Passing	2018
Grade	60-64.9%

WORK EXPERIENCE

Feb 2023 to Apr 2023

Data Science Intern at Acmegrade

Collaborated with a team of data scientists to design and execute data experiments, conducting exploratory data analysis and feature engineering to identify patterns and insights

INTERNSHIPS

Acmegrade, 2 Months

Collaborated with a team of data scientists to design and execute data experiments, conducting exploratory data analysis and feature engineering to identify patterns and insights

Physicswallah, 6 Months

During my internship at Physics Wallah, I had the opportunity to work on various projects related to data structures. This experience allowed me to apply my theoretical knowledge and enhance my practical skills in the field of data structures.

PROJECTS

Auto Build text summarization, 3 Months

Key challenges in text summarization include topic identification, interpretation, summary generation, and evaluation of the generated summary. Most practical text summarization systems are based on some form of extractive summarization.

Talkative Chatbot, 2 Months

This project integrates a chatbot powered by ChatGPT with a voice recognization system, enhancing user interaction and accessibility.

INTERESTS

I absolutely love traveling and have a deep passion for exploring new places, cultures, and experiences.

I am also deeply passionate about learning new technologies.