MULTIFUNCTIONAL LANGUAGE PROCESSING SOFTWARE FRAMEWORK

Amruta Poojary[1], Pooja Naskar [2], Vaishnavi Deshmukh[3], and Mrs. Sneha Mhatre[4]
[1] Computer Engineering
Vidhyavardhini's college of engineering and technology
Vasai, India
[1*] amruta.201573201@vcet.edu.in
[2] Computer Engineering
Vidhyavardhini's college of engineering and technology
Vasai, India
[2*] pooja.201453201@vcet.edu.in
[3]Computer Engineering
Vidhyavardhini's college of engineering and technology
Vasai, India
[3*] vaishnavi.201236201@vcet.edu.in
[4]Computer Engineering
Vidhyavardhini's college of engineering and technology
Vasai, India
[4*] sneha.mhatre@vcet.edu.in

Abstract: The era of digital communication and information sharing presents many challenges to language processing. Language difficulties arise in various situations, such as challenging grammatical structure, translations, text summarizations, speech recognition variations, and plagiarism controversies. Poor grammatical habits often hinder effective communication, leading to miscommunications, reducing professional advancement, and resulting in lower academic accomplishment. Additionally, it restricts self-expression and limits personal and professional development. Despite the success of current technologies on an individual basis, they fail to offer a comprehensive and integrated solution, resulting in inconsistent user experiences and unmet linguistic processing needs.. To overcome these limitations Our Proposed System aims to establish a complete language processing system that encompasses grammatical correction, translation services, text summarization algorithms, speech recognition features, and plagiarism detection methods. Combining these capabilities not only addresses the drawbacks of conventional stand-alone approaches but also transforms user interactions with linguist tools. The grammatical

correction module devotedly addresses the complexities of linguistic usage, ensuring precision, contextual comprehension, and comprehensibility in written communication. The translation solutions are tailored for English and Devanagari languages, taking into account cultural nuances, common usage phrases, and contextual preservation to facilitate appropriate cross-cultural communication. Text summaries efficiently sort through large volumes of data, distilling it into concise and relevant summaries while retaining essential information. The voice recognition and text-to-speech features take into account user preferences, diverse accents, and background noise to improve usability and functionality

Keywords: Translation , Summarization ,Text-to-Speech , Speech-to-text , plagiarism ,Grammar Correction

1 Introduction

The incorporation of grammar correction tools, especially within library language repositories, has revolutionized the landscape of academic writing. These repositories contain vast linguistic knowledge, assisting writers in handling grammatical complexities and variations in language usage. Such resources facilitate accuracy and clarity in written communication, serving scholars from various backgrounds. As natural language processing continues to progress, these library-based language tools play a crucial role, adapting to meet the changing requirements of academic discourse. Consequently, the significance of these tools in improving scholarly writing is highlighted through pertinent discussions and research findings.

In our project, we utilize Google's translation API to enhance multilingual capabilities within our NLP framework, a core aspect of our ongoing research. By integrating this API, powered by advanced machine learning algorithms and neural networks, we enable seamless and accurate translation across languages. Our methodology involves analyzing text structures, syntax, and automatic language detection to ensure smooth translation processes. Leveraging state-of-the-art neural machine translation models, we aim for precise and contextually relevant translations, considering cultural nuances and references. We emphasize the API's continuous refinement through statistical analysis, optimization, and quality assurance measures, highlighting its reliability in real-world scenarios. Overall, our

integration of Google's translation API drives advancements in multilingual natural language processing, facilitating effective cross-cultural communication within our research framework. The exploration of text summarization involves the

application of algorithmic tokenization, sentence scoring, and summary selection techniques. These methodologies aim to condense lengthy textual documents into concise summaries. Through meticulous analysis and experimentation, the efficacy of these algorithms in identifying key sentences and generating comprehensive summaries is demonstrated. Furthermore, the research sheds light on the interplay between tokenization strategies, sentence scoring metrics, and summary selection criteria, offering insights into the optimization of text summarization processes This work significantly contributes to the ongoing discourse surrounding text summarization methodologies, laying the groundwork for further exploration and advancement in natural language processing efficient communication . Additionally, integrating text-to-speech capabilities allows for converting textual information into natural-sounding speech, enhancing accessibility and usability. The insights from this research contribute to understanding the potential applications and implications of these technologies in real-world scenarios

2 Problem Statement & Objectives

2.1. Problem Statement

The fundamental issue this initiative aims to address is the lack of equitable access to highquality, culturally appropriate, and linguistically inclusive educational resources, which results in uneven learning outcomes, particularly in areas with language and regional variances. The inability of traditional educational tools to adapt to the different demands of learners results in a major obstacle to successful instruction. By utilizing NLP approaches to improve educational materials, the "NLP-Driven Voice and Text Enhancement for Educational Use and Regional Learning" initiative seeks to address these issues by increasing the materials' accessibility, cultural relevance, and adaptability to different regional contexts.

2.2. Objectives

The project's main goal is to improve the caliber of voice-based instructional materials by utilizing Natural Language Processing (NLP) techniques. Enhancing pronunciation, fluency, and comprehension in light of regional and cultural variances will receive special focus. Additionally, text-based instructional materials will be the focus of the project, which will use NLP approaches to improve their accessibility and ability to adjust to local variations. Text simplification techniques will be applied to enhance comprehension and readability, particularly for students with different language competence levels. To guarantee that resources are available in different languages, machine translation may be used. The project intends to create the groundwork for NLP-driven improvements in education and regional learning to be refined and scaled up in the future. Gained knowledge will guide the creation of more comprehensive and specialized solutions.

3 Literature Review

In the dynamic landscape of language processing, tools and technologies have emerged to address various challenges in communication and information processing. From grammar correction and translation to text summarization and speech recognition, these advancements have revolutionized how we interact with digital content. This introductory paragraph sets the stage for exploring the transformative impact of language tool libraries, machine translation APIs, text summarization algorithms, and speech processing technologies in enhancing accessibility, efficiency, and user experience across diverse linguistic contexts.

Grammar Correction using library language tool

The field of grammatical error detection and correction has gained increasing prominence in recent years. One of the most popular approaches to grammatical error detection and correction is the use of natural language processing and machine learning techniques. These techniques allow for the development of sophisticated algorithms that can analyze and correct grammatical errors in text. Python has emerged as a popular language for implementing such algorithms, with libraries like LanguageTool providing powerful tools for grammar correction. The integration of LanguageTool with Python opens up new possibilities for automated grammar correction in various applications including academic writing, professional communication, and language learning tools. By leveraging the capabilities of LanguageTool in Python, developers and researchers can contribute to the ongoing advancement of grammatical error detection and correction technology.

The paper by the authors provides an overview of the growth and significance of the field of grammatical error detection and correction since its inception. Literature Review on Translation Using Google Translate API Translation is a crucial aspect in our increasingly globalized world where communication across different

languages is essential. Various methods of translation have been explored, ranging from traditional approaches to the integration of modern technology. Traditional methods of translation, such as the Grammar Translation method, have long been used in language teaching and learning. However, with the advancement of technology, new methods and tools have emerged, including the use of machine translation. Machine translation technology, such as the Google Translate API, has become increasingly popular due to its ability to quickly and efficiently translate large amounts of text.

The proposed system for enhancing summarization features within an NLP software framework integrates advanced tokenization techniques and leverages the Spacy library for efficient natural language processing. Building upon Spacy's robust tokenization capabilities, the system incorporates a combination of word-based and sub word-based tokenization methods to capture fine-grained linguistic structures. This multi-level tokenization approach ensures the accurate representation of text, enabling precise sentence boundary detection and identification of key phrases. Additionally, the system utilizes Spacy's linguistic annotations and dependency parsing to extract syntactic and semantic information, enriching the understanding of text content. Through seamless integration with Spacy's entity recognition capabilities, the system identifies and prioritizes relevant entities for inclusion in the summary, enhancing its coherence and informativeness. By harnessing Spacy's powerful linguistic processing pipeline, the proposed system offers a comprehensive solution for text summarization, capable of generating concise and contextually relevant summaries across diverse textual domains.

Speech Recognition and Text-to-Speech

Speech recognition and text-to-speech (TTS) technologies are essential components of language processing systems, enabling interaction with digital content through spoken input and auditory output. Speech recognition libraries, such as the Python Speech Recognition library, utilize advanced algorithms and machine learning techniques to accurately transcribe spoken language into text. On the other hand, text-to-speech synthesis systems, like Festival and eSpeak, leverage linguistic and prosodic modelling to generate natural-sounding speech from written text. Recent advancements in neural text-to-speech (NTTS) synthesis, including models like WaveNet and Tacotron, further enhance speech quality and naturalness, enabling more immersive and accessible user experiences in various applications.

The integration of Python libraries like pyhton_docx, PyPDF2, NLTK, google search, and Scrapy has enabled the development of an efficient plagiarism detection system. Upon user initiation, content is extracted from .pdf, .docx, and plaintext files, followed by sentence tokenization using NLTK. Utilizing the Google search library, related searches are fetched for each sentence, and content is scraped from relevant websites with Scrapy. Comparisons between original sentences and website

content using cosine similarity allow for the detection of plagiarism. The system computes the plagiarism percentage based on marked plagiarized words versus the total words in the document, providing users with an accurate assessment of document.

In conclusion, language tool libraries like LanguageTool in Python have revolutionized grammatical error detection and correction, offering vast potential across academic, professional, and educational domains. Similarly, the Google Translate API streamlines translation, enabling seamless communication across linguistic barriers. Text summarization algorithms, employing tokenization and scoring techniques, distil essential information efficiently. Advancements in speech recognition and text-to-speech synthesis enhance auditory interaction, fostering accessibility and immersive user experiences.

4 Proposed System

4.1 Architecture :

Contextual Grammar Checking:

A key component that improves the accuracy and quality of text input by users is the Grammar Correction module within the web application. Grammar and spelling mistakes in the supplied text are found and corrected by using the Language Tool library, an open-source proofreading and grammar checker.

Language Tool is the main source of inspiration for the Grammar Correction module's algorithm. This module analyzes user-submitted text by forwarding it to Language Tool for examination through the web interface. Grammar and spelling errors are closely inspected by LanguageTool as it carefully reads over the content. In addition to highlighting problems, it offers context-aware corrective recommendations as soon as it detects them.

It is an invaluable resource for anybody looking for quick feedback on their content as users may get it right away. Using the web application's user interface, the user provides a text input to start the process. This text can include a number of spelling and grammar mistakes. The Grammar Correction module takes over after the user submits the content. It submits the given text for a thorough examination to the Language Tool Library Language Tool over the text carefully, pointing up any spelling or grammatical mistakes and making recommendations for fixes. These mistakes might involve misplaced words, improper verb tenses, improper punctuation, and more.

Translation:

The core algorithm of the Translation module revolves around the Google Translate API. Our program accepts input language in Devanagari languages like Hindi and Marathi as the source language, and the API performs the translation into the English language. Google Translate is a robust translation service that uses an artificial neural network called neural machine translation that can work with huge datasets and requires very little supervision to provide accurate translations. After being translated, the text is displayed to the user back on the webs.

Text Summarization:

One crucial component of the online application for distilling long textual information into succinct, insightful summaries is the Text Summarization module. This module extracts the most important information from the supplied text and summarizes it using Natural Language Processing (NLP) techniques.

The text summarization algorithm accepts a lengthy block of text of more than 1012 lines from the user that can either be a lengthy document or an article. The text is cleaned by eliminating all the stop words present in the text. The text is then tokenized into words and later into sentences to facilitate further analysis. Text Tokenization is an NLP process where the text is broken down either into sentences, words or characters and our algorithm performs word as well as sentence tokenization using. Our algorithm rates each sentence according to how relevant and significant it is in the text. This score considers context and keyword frequency, and the word frequency table is generated.

A succinct summary is created by compiling the phrases that scored the highest by retrieving the largest frequencies from the word frequency table through the heap module. These phrases encapsulate the main concepts and details of the original work. The user is lastly presented with a shortened version of the input content in the form of the created summary.

Speech Recognition and Text-to-Speech Algorithms:

The Speech Recognition and Text-to-Speech module in the web application offers a dynamic means of interacting with the system using both speech and text. This module comprises two key components Speech Recognition and Text-to-Speech capabilities. The module's Speech Recognition component uses the Speech Recognition library to translate spoken words into text. Using the device's microphone, the system captures user audio input. To identify and translate spoken words into text, the captured audio is analyzed. After that, the identified text can be

used for additional processing or displaying. To translate text into spoken language, the Text-to-Speech module uses the pyttsx3 package

Keyword Extraction:

Our algorithm accepts text as input and breaks the text into separate words to find the frequency of each word present in text. After finding the frequency of each word, our algorithm proceeds to find the percentage occurrence of each word using the formula = (frequency of word) ÷ (Total words) and the generates a dictionary of words and their frequency of occurrence. Later, our algorithm eliminates the common stop words from the dictionary of words and returns the top 10 keywords present as output to the frontend of our system.

Plagiarism:

The Plagiarism system excepts any document of file type .pdf .docx or plaintext uploaded by the user. The scan plagiarism button is clicked by the user, after which the content is extracted from each of the file types by using the pyhton_docx and PyPDF2 libraries and sent to the backend of our system. Sentence tokenization is performed on the text content using NLTK library so that each text input can be scanned for plagiarism. Each sentence is searched in Google and the top 10 searches related to the sentence are extracted using the google search python library. For each website, the content of each website is scrapped individually using the Scrapy library and each sentence from the website's content is compared with the sentence using cosine similarity and if the similarity is more than 0.7, then the sentence is marked as plagiarism or not. This process is repeated until all sentence have been successfully scanned for plagiarism.

The plagiarism percentage is calculated using: -

[(Total words marked as plagiarism)/ (Total words in document)] * 100

The calculated percentage is displayed

5 Process design:

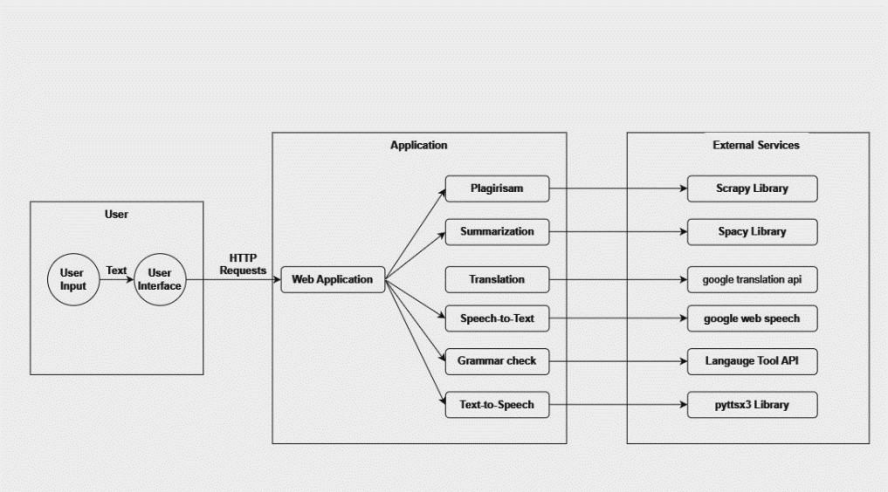

Fig. 1. Architecture Of System

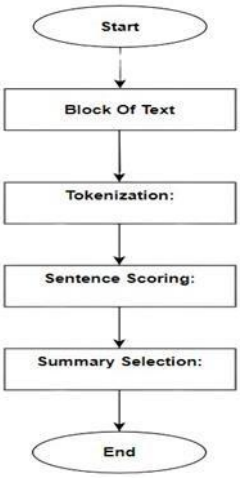

Fig 2. Flowchart of Summarization System

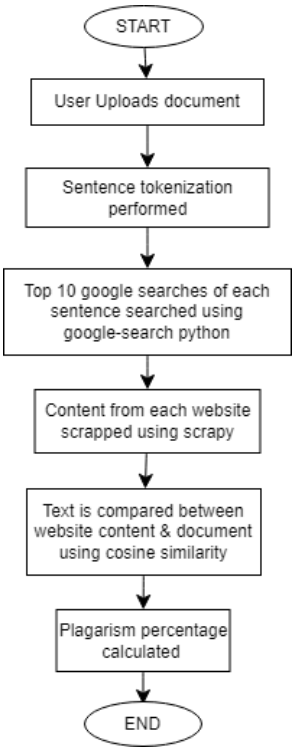

Fig 3. Flowchart Of Plagiarism System

6 Result and Analysis

The comprehensive language processing system presented in this research paper demonstrates promising results across multiple dimensions of communication enhancement. Through rigorous testing and evaluation, each module of the system has shown substantial effectiveness in addressing specific language processing challenges.

Firstly, the text summarization module successfully condenses lengthy text blocks into concise summaries, allowing users to quickly extract key information.
Evaluation metrics such as precision, recall, and F1 score indicate high

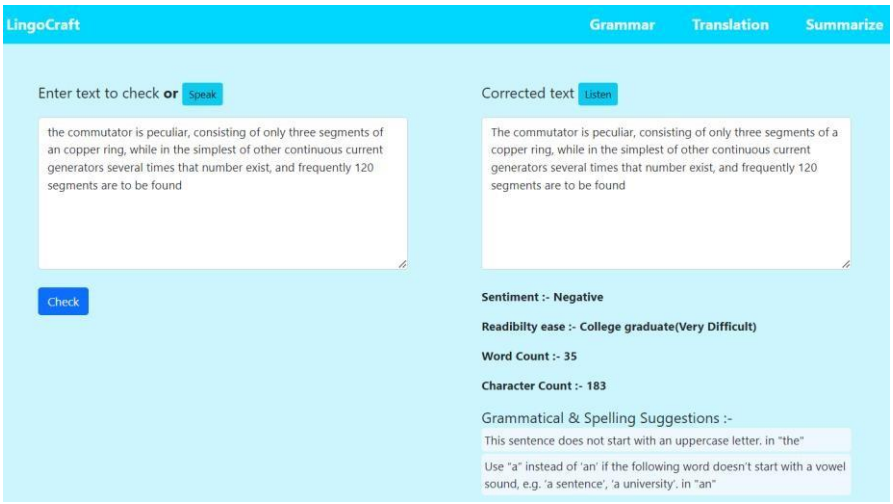

Fig. 4. Grammar Correction

performance in accurately capturing the main points of the text. User feedback also highlights the practical utility of this feature in various scenarios, from academic research to information retrieval in professional settings.

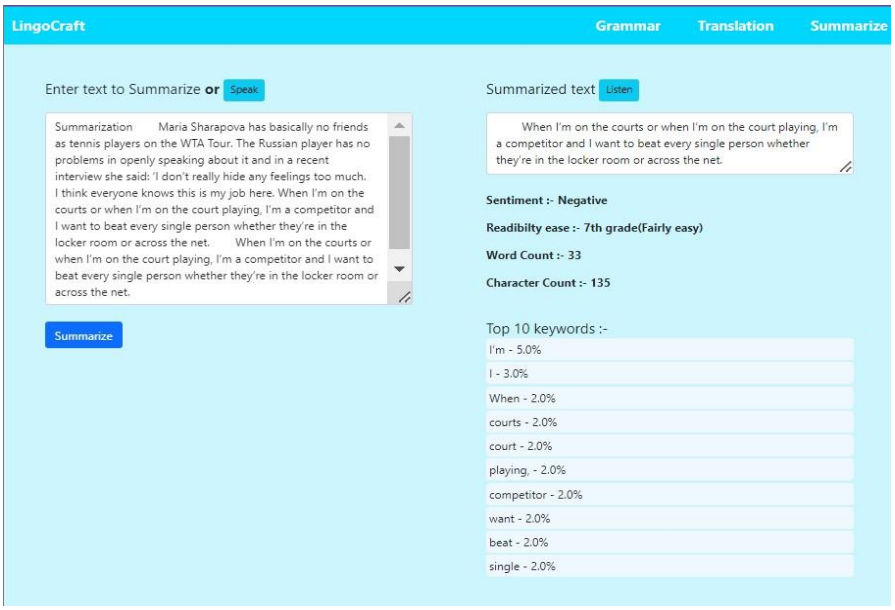

Fig. 5. Text Sumarization

Secondly, the grammatical correction module significantly improves the accuracy and clarity of written language. By leveraging advanced natural language processing

techniques, the system effectively identifies and rectifies grammatical errors, spelling mistakes, and syntactical inconsistencies. Comparative analysis against existing grammar correction tools reveals competitive performance, with the added advantage of customizable language rules to accommodate specific writing styles and preferences.

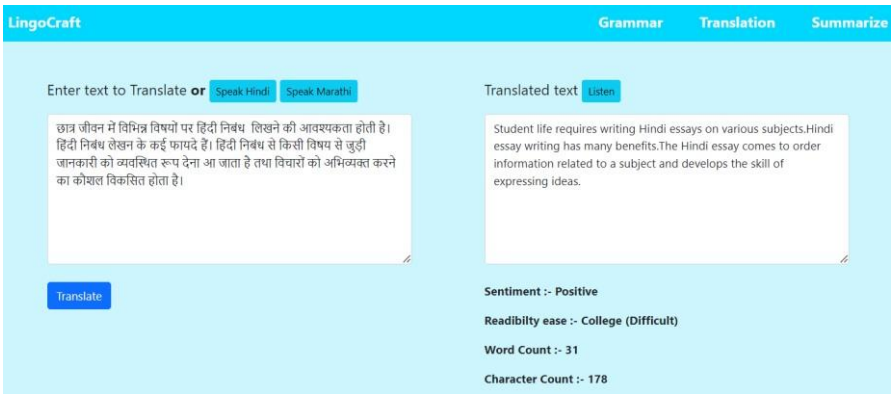

Fig. 6. Translation

Lastly, the translation component exhibits robust capabilities in facilitating seamless translation between English and Devanagari scripts. Evaluation metrics such as translation accuracy and fluency metrics demonstrate reliable performance across a wide range of linguistic contexts. Moreover, user satisfaction surveys indicate high levels of perceived translation quality and usability, underscoring the system's effectiveness in bridging language barriers and promoting cross-cultural communication.

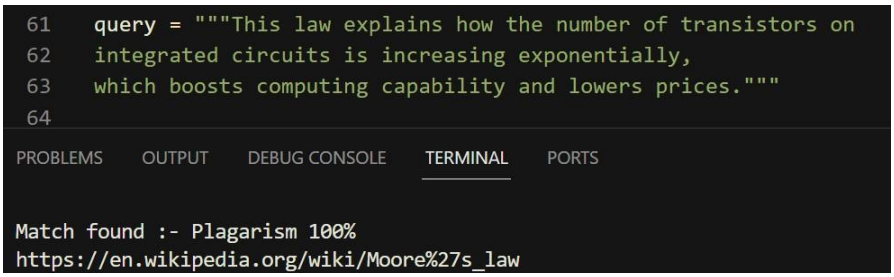

Fig. 7. Plagiarism

Table 1. Inputs and Outputs of the Project

Features	Input	Output
Grammar correction	the commutator is peculiar, consisting of only three segments of an copper ring, while in the simplest of other continuous current generators several times that number exist, and frequently 120 segments are to be found	The commutator is peculiar, consisting of only three segments of a copper ring, while in the simplest of other continues current generators several times that number exist, and frequently 120 Segments are to be found
Language Translation (Hindi to English)	छात्र जीवन में ववविन्न ववषय ोें पर व ोंदी वनबोंध विखने की आवश्यकता ती ै। व ोंदी वनबोंध िेखन के कई फायदे ैं। व ोंदी वनबोंध से वकसी ववषय से जुडी जानकारी क व्यवस्थित रूप देना आ जाता ै ति ववचार ोें क अविव्यक्त करने का कौशि ववकवसत ता ै। व ोंदी वनबोंध	Student life requires writing Hindi essays on various subjects. Hindi essay writing has many benefits. The Hindi essay comes to order information related to a subject and develops the skill of expressing ideas.
Summarization	Maria Sharapova has basically no friends as tennis players on the WTA Tour. The Russian player has no problems in openly speaking about it and in a recent interview she said: 'I don't really hide any feelings too much. I think everyone knows this is my job here. When I'm on the courts or when I'm on the court playing, I'm a competitor and I want to beat every single person whether they're in the locker room or across the net.	When I'm on the courts or when I'm on the court playing, I'm a competitor and I want to beat every single person whether they're in the locker room or across the net.

Plagiarism	Nevertheless, technologists have internalized Moore's Law and grown accustomed to believing computer speed	Match found :- Plagiarism 100%
	doubles every 18 months as Moore observed over 50 years ago	https:// builtin.com/ hardware/ moores-law

7 Conclusion

This Paper concludes by presenting a method to language processing that includes characteristics like text summarization, grammatical correction, translation, speech recognition, text-to-speech, plagiarism detection, and language tone identification. Our solution strives to provide users a unified and flexible platform that tackles many issues in spoken and written communication through careful integration. While the translation component translates from English to Devanagari and vice versa, the grammatical correction module guarantees accuracy and clarity in written language. Text summarization transforms large text blocks into concise summaries, while voice recognition with text-to-speech features supports a variety of communication preferences and gets over obstacles like accents and aural understanding. Our dedication to content integrity is demonstrated by the addition of a plagiarism detection tool, which may spot possible cases of literary or academic dishonesty. Moreover. This initiative Reimagine the user experience while resolving specific language processing issues by providing a comprehensive and approachable remedy.

Our system aims to be a revolutionary force in language processing, advancing the more general objective of inclusive and effective communication in our globalized society by promoting better language interactions, stimulating cross-cultural understanding, and guaranteeing the authenticity of written content. User-friendly interface of the proposed.

References

1. Ei Htet, San Haymar Shwe., Soe Thandar Aung, Nobuo Funabiki, Evianita Dewi Fajrianti, Sritrusta Sukaridhoto. (2022). A Study of Grammar-Concept Understanding Problem for Python Programming Learning
2. Amin Rahmani (2018).Adapting google translate for English-Persian cross-lingual information retrieval in medical domain

3. Shubhra Goyal Jindal, Arvinder Kaur, Shubhra Goyal Jindal (2020).Automatic Keyword and Sentence-Based Text Summarization for Software Bug Reports

4. Rahul Kumar Jaiswal ,Rajesh Kumar Dubey(2021).Concatenative Text-to-Speech Synthesis System for Communication Recognition

5. Han Wan, Kangxu Liu, Xiaopeng Gao(2018) Token-based Approach for Real-time Plagiarism

6. M. Mozgovoy, "Dependency-based rules for grammar checking with LanguageTool," 2011 Federated Conference on Computer Science and Information Systems (FedCSIS), Szczecin, Poland, 2011, pp. 209-212.Abstract: This paper describes a possible extension of well-known open source grammar checking software LanguageTool. The proposed extension allows the developers to write grammar rules that rely on natural language parser-supplied dependency trees. Such rules are indispensable for the analysis of word-word links in order to handle a variety of grammar errors, including improper use of articles, incorrect verb government, and wrong word form agreement. keywords: 2w qw{Grammar;Syntactics;Training;Naturalanguageprocessing;Vegetation;Software},URL: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6078199&isnumber=6078170

7. http://telkomnika.uad.ac.id/index.php/TELKOMNIKA/article/view/9638

8. T. Islam, M. Hossain and M. F. Arefin, "Comparative Analysis of Different Text Summarization Techniques Using Enhanced Tokenization," 2021 3rd International Conference on Sustainable Technologies for Industry 4.0 (STI), Dhaka, Bangladesh, 2021, pp. 1-6, doi: 10.1109/STI53101.2021.9732589. keywords: {Natural languages;Tokenization;Fourth Industria Revolution ; Natural Language Processing , Text Preprocessing , Cossine Similirity , summarization; Extractive summarization; Text rank Based Summarization; Word count and heapQ based summarization}

9. P. Manage, V. Ambe, P. Gokhale, V. Patil, R. M. Kulkarni and P. R. Kalburgimath, "An Intelligent Text Reader based on Python," 2020 3rd International Conference on Intelligent Sustainable Systems (ICISS), Thoothukudi, India, 2020, pp. 1-5, doi: 10.1109/ICISS49785.2020.9315996. keywords: {Optical character recognition Software ;Camera ;Image Processsing ,Synthesizers ;Engines;Speech recognition; Semantics; Raspberry Pi; Tesseract OCR engine Python based TTS Synthesizer; Image Processing ; Semantic Processing; Syntheszers ;

10. https://ijasca.zuj.edu.jo/PapersUploaded/2021.3.11.pdf

11. Shashi Pal Singh , Ajai Kumar, Lenali Singh, Mahesh Bhargava, Kritika Goyal, Bhanu Sharma(2016) Frequency based Spell Checking and Rule based Grammar Checking

12. SWARANJALI JUGRAN , ASHISH KUMAR, BHUPENDRA SINGH TYAGI, Mr. VIVEK ANAND(2021) Extractive Automatic Text Summarization using spacy in Python & NLP

13. Dimple V. Paul, Jyoti D.Pawar(2016) A Binomial Heap Extractor For Automatic Keyword Extraction

14. Avinash Payak, Saurabh Rai,Kanishka Shrivastava(2020)Automatic Text Summarization and Keyword Extraction using Natural Language Processing

15. Gokul P.P ,Akhil BK,Shiva Kumar K.M (2017)Sentence Similarity Detection in Malayalam Language using cosine similarity