International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048
Volume 7, Issue 10, pp 131-138, October-2023
https://doi.org/10.47001/IRJIET/2023.710017

Automated Spelling and Grammar Checker Tool for Sinhala

[1]Navoda M.R., [2]Weerasooriya O.W.R.Y., [3]Siriwardhana A.U.A., [4]Sonali L.D.A., [5]Jenny Krishara, [6]Poorna Panduwawala

[1,2,3,4,5,6]Department of Information Technology, Sri Lanka Institute of Information Technology, Malabe, Sri Lanka
Authors E-mail: [1]it20111656@my.sliit.lk, [2]it20144562@my.sliit.lk, [3]it20034504@my.sliit.lk, [4]it20152246@my.sliit.lk, [5]Jenny.k@sliit.lk, [6]Poorna.p@sliit.lk

***Abstract* - Spelling and grammar are crucial aspects of language proficiency, particularly for primary students who are in the formative stages of their linguistic development. In this research paper, we present a comprehensive tool for the detection and correction of spelling and grammar errors in the Sinhala language, tailored specifically to cater to the needs of primary students. The main features of our tool include spelling error checking and text preprocessing, spell error correction, subject-verb identification, predicting sentence patterns, and analyzing subject characteristics, as well as grammar error detection and correction. Through advanced algorithms and linguistic analysis, our tool identifies spelling errors in Sinhala text, offering suggestions for accurate word choices. Additionally, it employs sophisticated techniques to identify subject-verb agreement and predict sentence patterns, enhancing students' understanding of grammar rules. Moreover, the tool provides grammar error detection capabilities, and corrects them by the tool itself. By integrating these features, our tool aims to enhance the overall language proficiency of primary students in Sinhala. The user-friendly interface and interactive feedback system make it accessible and engaging for young learners, promoting active participation and self-correction.**

Keywords: spelling error checking, spell error correction, subject-verb identification, sentence pattern prediction, subject characteristics analysis, grammar error detection, grammar error correction, primary students, Sinhala language.

I. INTRODUCTION

One of Sri Lanka's official and national languages, Sinhala is an Indo-Aryan language that is spoken by over 16 million people [5]. Sinhala is one of the oldest languages in the world, having existed for more than 2300 years [6]. The Maldivian and Dhivehi languages are the most closely related to it [3]. The Sinhala writing system is a member of the Aramaic script family and is descended from the ancient Indian Brahmi script [7].

16 vowels, 41 consonants, 2 semi-consonants, and 13 consonant modifiers have been identified in the Sinhala language by the Committee on Adaption of National Technology (CANLIT) [8].

Different languages have integrated technical developments as language and technology have advanced together. The English language is a shining example of how technology and language have coevolved in the study of natural language processing (NLP), with creatures thriving across multiple techniques and tools for evaluating grammar and spelling. In order to develop effective methods of processing information and producing precise outcomes, comprehensive study has been performed.

Nevertheless, numerous organizations and entities embraced technological solutions when the pandemic hit the world in 2020, which led to unemployment. The shift to working remotely established the new standard, and English language adaptation was simple. Languages like Sinhala, in contrast, encountered difficulties because technology lagged behind. Teachers and staff members had to put in extra time and effort to complete routine chores that might possibly have been performed quickly by technology. For instance, a straightforward text page needed human cross-referencing and modifications, whereas a well-designed tool might deal with it more effectively and accurately. There is still potential for improvement even though there are some solutions on the market.

The shortage of a standard approach in comparison to English is justifiable given the complicated structure of a language like Sinhala and its rich morphology. There are differences between the spoken and written forms of Sinhala, each featuring a unique grammar and vocabulary. The writing system for Sinhala, an Indo-Aryan language, depends on consonant-vowel sequences transcribed as single units. Due to its historical connections with various languages, the dialect has also taken words from English, Portuguese, Dutch, Tamil.[9]

The natural language processing (NLP) techniques in Sinhala are uncommon [2]. However, a methodical approach

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 7, Issue 10, pp 131-138, October-2023

https://doi.org/10.47001/IRJIET/2023.710017

is needed while analyzing input. The data goes through preprocessing, which removes non-Sinhala characters to make it usable. It then moves on to the module that detects spelling errors before moving on to the module that corrects them. It is ready for the grammatical modules after the data has been cleaned up and the spellings checked. The subject identification module, followed by the subject analyses and tense analyses modules, finds sentence patterns. These two modules are extremely important in deciding how accurate the grammar is.

On the basis of the guidance given via the subject analyses and tense analyses modules, the data is scanned and grammatical decisions are made. Corrections or recommendations are offered if any grammatical irregularities are found.

II. RELATEDWORK

Sinhala is the national dialect of Sri Lanka and widely spoken. Pali and Sanskrit serve as itssources, originating from natural language. Sinhala language has two variants while being used, namely spoken and written. The complex Sinhala language includes requirementsfor grammar and spelling. This creates an opportunity for the Sinhalalanguage to become increasingly challenging. The level of language formality is completely determined bythe established conventions with regards to written context. Applying all of these rules in oneapplication is quite challenging [1].

While there exist automated programs for detecting spelling and grammar mistakes in various languages,such as English, the availability of similar tools for Sinhala remains significantly limited.

The e/akuru sduva/ is an online platform that was studied by Abeyrathne et al. This platform is used for Sinhala language spell and grammar checking. The study focused on two approaches for spell checking: rule based and data driven. Implementing the rule based approach proved to be challenging due to difficulties in defining spelling rules for Sinhala. The study utilized two methodologies: data gathering and spell checking. The researchers identified four key components involved in spell checking: language model, error model, candidate model, and selection process [4]. The grammar checking component incorporates pre-determined computational grammar rules such as subject object verb structure, active-passive voice, and masculine/feminine pronouns. These elements are vital in verifying the grammatical validity of sentences. Furthermore. The module also provides recommendations for improving the output.

Pabasara et al.'s research "Grammar error detection and correction model for Sinhala Language" integrated a machine learning algorithm-based technique with a traditional rule-based approach [3]. Given the scarcity of resources for Sinhala, current tools and systems exhibit reduced accuracy levels. To establish the subject within statements, researchers utilized both a POS tagger and morphological analyzer. Consequently, they stored both the identified subject as well as its respective tag separately to facilitate future usage. Within this project scope, an adapted version of Moratuwa University-developed POS tagger specifically addressed select attributes pertaining to the subject. A "polyglot" morphology analyzer was also used, with higher accuracy due to the limited accuracy of the morphology analyzer [3] The overall accuracy of this system is higher than that of a single method, however, 88.6%. Furthermore, it is suggested that the morphology analyzer should be improved for correct noun classification because it provides too many verb suffixes, making it difficult to choose the correct one [3 This study uses a rule- method based; However, even using a separate dictionary for that purpose is not a practical answer because the third-person system has such a wide range and number of terms [3].

After testing 200 sentences, Liyanage and other Sinhalese dictionaries achieve an accuracy of 60 percent [1]. Feature-Based Context-Free Grammar (CFG) and the free and open-source Natural Language Toolkit (NLTK) were used to distinguish a noun phrase from its main components, including person (first, second, and third), gender (masculine or feminine), numeral (singular or plural), capitalization, and animation. Out of 25, these ten were found in simple sentence structure [1]. Since the accuracy is somewhat poor, it was thought to use more morphology analyzers and word segmentation algorithms to generate more accurate systems [1].

Combining rules-based and stochastic approaches built around HMM (Hidden Markov Model), the hybrid speech tagger component for Sinhalese by UCSC showed significant accuracy on an overall basis, as demonstrated by [10]. Revealed it, with an roughly percentage of unknown words of 20%. To determine which P.O.S. tag set provides the best tagging for Sinhalese, three different P.O.S. tag sets were tested. P.O.S. tagging is the process of identifying words and tagged them based on their context in a sentence. The rule-based method is used whenever the H.M.M. method unable to recognize the symbols of the unrecognized word. Dictionaries were used to compare lexical language rules and a rule-based tagger was needed to store the results in a database. This hybrid approach seems to be better than through the rule-based approach alone because it helps to deal with the drawbacks of the rule-based approach [10][11].

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 7, Issue 10, pp 131-138, October-2023

https://doi.org/10.47001/IRJIET/2023.710017

According to Pabasara et al., input data are either pre-processed data or specifically reviewed data. In a paragraph, sentences are broken up, then words and spaces are separated, and finally other lines are erased. For datasets, more data was used, 1050 active Sinhalese words with no errors in spelling. For their lexical analysis, which they enhanced a previously developed p.o.s. tag system [3], which was improved for better performance. An algorithm-based pattern recognition method was used to identify sentence structures. They claimed to obtain 32 sentence structures.

One of the NLP techniques used in Sinhalese spellchecker using open data [8] is n-gram based data processing method. Before checking for spelling errors, the syntax will be processed first. A list of unique words, which will include homophones and acceptable spelling changes, will be compared with unique words in the text. If a word is found in the exception list, it will be removed from the special vocabulary. The word "ext" can be supplied before the exceptions are compared to the list. If a word is found in the exception list, it will be removed from the special vocabulary. The selected algorithm shows that it can handle only single words. After pre-processing, the variable generation module accepts the processed terms, and all conceivable term variables are used to form a core module, where one gram, two grams, or three grams are multiplied the more a word uses the better recommendation the module is selected. The most frequent term will yield the best recommendation [7].

SinSpell: A Comprehensive Spelling Checker for Sinhala [9] that uses Hunspell engine to provide rule based Spelling Error Detection and suggestion generator system this two important parts are auto-correction module and spelling error detection module. The Hunspell engine receives a dictionary file and a suffix file to distinguish between correct and incorrect spellings. Data were collected and reviewed from multiple sources to generate vocabulary and background files for this study. A similar list of adjectives and adverbs was constructed. Many documents, including letters and reports, were put through an error checking process to find errors.

The research "මහාරාවණා" /mahārāvanā/ [13] done by University of Moratuwa is one of the studies based on grammar testing environment using three-word sentences. This work uses a rule-based approach. Further institutional development is a challenge because of the depth of the Sinhalese language. Proper lexical analysis has yet to be developed for Sinhalese. Colombo, Kelania, University of Moratuwa (UOM) and Sri Lanka Institute of Information Technology (SLIIT) have conducted a lot of research in this field. This study highlights the importance of updating Sinhala language and spelling tests to keep them in line with contemporary changes.

III. METHODOLOGY

The system's goal is to identify and fix spelling mistakes and grammatical errors in text written in the Sinhala language. In other words, it aims to automatically detect and correct mistakes in spelling and grammar when given Sinhala language input. The system is composed into four major components, which are as follows.

1) Text Pre-processing and spelling error analysis
2) Rectification (Correction) of misspelled words
3) Detecting Sentence Pattern with Subject & Verb Prediction and analyze Subject Characteristics
4) Grammar Error Processing (Correction)

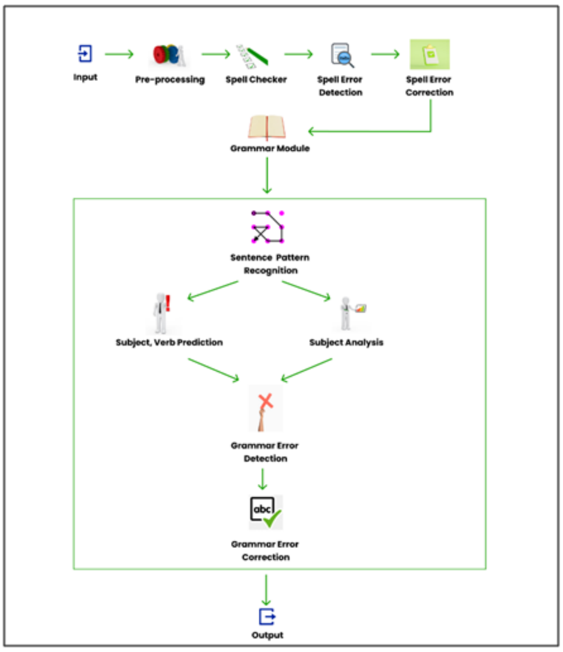


Figure 1: System overview diagram

A) Creating Datasets

The Sinhala language has a substantially lesser number of publicly available datasets than languages such as English and Chinese. To address this issue, we built four datasets separately for each component. (1000 – 7000 words for each component). These datasets are specifically tailored to include the most often used Sinhala words. By creating these customized datasets, we aimed to overcome the limitations posed by the lack of comprehensive existing datasets for Sinhala, enabling them to train the models more effectively and accurately.

	Input	Status
1	ළමයා මල් කඩයි	1
2	ලමයා මල් කඩයි	0
3	සමනලයා පියඹා යයි	1
4	සමනළයා පියඹා යයි	0

Figure 2: Dataset Example of Pre-processing and spell error analysis

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 7, Issue 10, pp 131-138, October-2023

https://doi.org/10.47001/IRJIET/2023.710017

No	Correct	Incorrect	Incorrect
1	ළමයා	ලමයා	
2	සමනලයා	සමනළයා	සමණලයා
3	දණහිස	දනහිස	
4	පහන	පහණ	
5	වළට	වලට	

Figure 3: Dataset Example of Rectification (Correction) of Misspelled Words

1	Sentence	Pattern	Subject	Verb	S1	S2	S3	S4	S5	S
2	මම [illegible] ගියෙමි	0	මම	ගියෙමි	N	I	D	S		A
3	ළමයා මල් කඩයි	0	ළමයා	කඩයි	M	K	D	S		A
4	අපි බුදුන් සරණ යමු	0	අපි	යමු	N	I	D	P		A
5	සමනලයා පියඹා යයි	0	සමනලයා	යයි	N	K	D	S		A

Figure 4: Dataset Example of Detecting Grammatical Errors by Analyzing Subjects & their Characteristics

#	Incorrect	Correct
1	අපි බුදුන් සරණ යමි	අපි බුදුන් සරණ යමු
2	අක්කා පැළ සිටෙව්වෝය	අක්කා පැළ සිටෙව්වාය
3	ඔහු වීරෝදාර සෙබළෙක්	ඔහු වීරෝදාර සෙබළෙකි
4	ඇය කනිෂ්ඨ අංශයේ ඉගෙන ගනී	ඇය කනිෂ්ඨ අංශයේ ඉගෙන ගන්නීය

Figure 5: Dataset Example of Grammar Error Processing (Correction)

B) Text Pre-processing and Spelling error analysis

Whenever a Sinhala sentence is entered into the system, non-Sinhala characters such as numbers as well as punctuation marks may appear. Non-Sinhala letters and punctuation must be deleted before spell-checking the text, and the words in the input sentence must be separated into a list. Before delivering them to the Spelling Error Detection Module, this preprocessing step guarantees that only Sinhala words remain.

The word list, obtained after splitting the input sentence, is then fed into the Spelling Error Detection component. Each word in the list is checked for spelling errors using a deep learning algorithm, Recurrent Neural Network (RNN). This algorithm is trained using the created dataset, allowing it to identify patterns in the data and make predictions about which words are spelled correctly and which are spelled incorrectly. Once the errors are detected, they are then forwarded to the Spell Error Correction component.

Enter Text : [නංගීගේ දනහිස තුවාලවිය]
['නංගීගේ' , 'දනහිස' , 'තුවාල' , 'විය']

Figure 6: Text Pre- Processing

In summary, the system first removes non-Sinhala characters and separates the input sentence into a word list. Then, the Spelling Error Detection module, trained with a dataset, identifies spelling errors in the words. Finally, they will be sent to the Spell Error Correction module for correction.

Rectification (Correction) of misspelled words

Misspelled words refer to the process of identifying and correcting words that are spelled incorrectly. When text is written, it is common for mistakes to occur, such as typographical errors or unintentional misspellings. After errors are detected in the text, they are passed on to the Spelling Error Correction module for correction.

Detected Error : ['නංගීගේ', 'දනහිස', 'තුවාල', 'විය']

Figure 7: Examples of detected errors

So, the logistic regression model is used to correct spelling errors. This module is designed specifically to handle the task of identifying and rectifying spelling errors in the text. Logistic regression is one of the possible approaches that can be used for spelling error correction, although it is not the only method. Logistic regression is a statistical model commonly used for binary classification tasks, where the goal is to predict the probability of an event belonging to one of two classes. Here, a dataset comprising both accurately spelled words and misspelled words has been provided for training a logistic regression model. The model identifies a feature from individual words that makes them different from the misspelled words. These features can include the word's length, specific characters' presence, or character combinations. Then the model is trained to use the features it identified using the dataset. The model learns and understands patterns that indicate whether a word is spelled correctly or not. When a new word is input into the system, the logistic regression model calculates the probability of the word being misspelled. So, the system sets a threshold (a point range) (e.g., 0 = correct, 0.5 = may be correct or wrong, 1 = incorrect) to decide whether to consider the word as misspelled or not.

['නංගීගේ' , 'දණහිස' , 'තුවාල' , 'විය']
[නංගීගේ දණහිස තුවාල විය]

Figure 8: Corrected text using a logistic regression algorithm

In addition to classifying words as either correctly spelled or misspelled, the system can generate suggestions for potential corrections when a word is identified as misspelled. This functionality enhances the system's ability to provide accurate and helpful suggestions to users by offering possible corrections that align with standard spellings.

C) Detecting Sentence Pattern with Subject & Verb Prediction and Analyze Subject Characteristics

Subject and Verb Prediction

Predicting the Subject and Verb of a sentence will be done using a "Random Forest Classifier". We employed the Count Vectorizer class to convert the sentences into numerical

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 7, Issue 10, pp 131-138, October-2023

https://doi.org/10.47001/IRJIET/2023.710017

representations based on token counts. The transformed data were then used to train Random Forest classifiers for subject and verb prediction. We used a manually created dataset for this task. There is one Random Forest classifier for subject prediction and another Random Forest classifier for verb prediction. Each classifier was trained on the generated features and corresponding subject or verb labels. The prediction accuracy of the classifiers was evaluated using a separate validation dataset. The accuracy scores were calculated using the accuracy score function from sci-kit-learn.

```
# predict the subject and verb for a new sentence
new_sentence = "අම්මා පහන දල්වයි"
new_features = vectorizer.transform([new_sentence])
predicted_subject = subject_rf.predict(new_features)[0]
predicted_verb = verb_rf.predict(new_features)[0]
print(f"Predicted subject: {predicted_subject}")
print(f"Predicted verb: {predicted_verb}")

Predicted subject: අම්මා
Predicted verb: දල්වයි
```

Figure 9: Subject Verb Prediction

Sentence Tense Pattern detection:-

We used a manually created dataset for this task and which classifies the given sentences according to basic sentence patterns. We have included 04 basic sentence patterns. They are,

1) Subject-Object-Verb.
2) Subject-Verb-Object.
3) Object-Verb-Subject.
4) Object-Subject-Verb.

The special mechanism used here is an 'Artificial Neural Network'. Used 'Rectified Linear Unit' as the activation parameter. Because it is a popular activation parameter that introduces non-linearity. In the output phase, the activation is set to 'softmax' because it's preferred for multiclass classification problems. Since I have added 4 classes.

```
# Assume the new Sinhala sentence you want to predict is stored in a variable called 'new_sentence'
new_sentence = 'සංගී අම්මා කෑවි'

# Preprocess the new sentence
new_sequence = tokenizer.texts_to_sequences([new_sentence])
new_sequence = keras.preprocessing.sequence.pad_sequences(new_sequence, maxlen=max_sequence_length)

# Make predictions using the trained model
predictions = model.predict(new_sequence)
predicted_label = label_encoder.inverse_transform([predictions.argmax()])

# Print the predicted pattern label
print("Predicted Pattern Label:", predicted_label)

1/1 [==============================] - 0s 113ms/step
Predicted Pattern Label: [0.]
```

Figure 10: Pattern Recognition

Sentence Tense Pattern detection:-

Analyzing the characteristics of subjects constitutes the subsequent phase. This step encompasses six primary characteristics, namely: Gender, Person, Definite/Indefinite, Number, Honorifics, and Animacy.

To discern each characteristic, we employ a Random Forest Classifier, which aids in precise identification. This paragraph contributes to the scholarly discourse in our research paper.

```
# Encode the new word using Unicode representation
new_word = "නාන්නා"
encoded_word = [ord(c) for c in new_word]

Predicted Class for S1: M
Predicted Class for S2: K
Predicted Class for S3: D
Predicted Class for S4: S
Predicted Class for S5: NA
Predicted Class for S6: A
```

Figure 11: Subject Characteristics Identification

D) Grammar Error Processing (Correction)

The Grammar checking component will go through several phases to achieve the objectives, focusing on the sentence itself as well as the subject matter and tense information collected by the Sentence Tense Detection module as inputs.

The Grammar Error Detector will examine the actual sentence for any grammatical errors. This will be done by using deep learning algorithms. So, if any grammatical errors are detected, the grammar correction module will focus on the tense and verb of the sentence, which will form the third phase of the component. The dataset is handled by using SVM (Support Vector Machine) which a popular supervised machine is learning algorithm used for classification and regression tasks. It is primarily used for solving binary classification problems but can also be extended for multi-class classification. Therefore, after determining the appropriate tense, the "Verb" will be modified according to the "Subject". Sinhala sentences that consist of three words are primarily considered to follow a basic pattern.

All the grammatical errors will be detected by considering whether the verb matches the analyzed subject through Sentence pattern recognition. As well as grammatical errors will be rectified by considering the Tense of the sentence through Sentence Tense Detection. The actual sentence is returned as the output if no errors are identified. However, if there are grammatical errors, the sentence is passed to the Grammar Error Correction Module. The result will comprise the corrected sentence. A collection of sentences

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 7, Issue 10, pp 131-138, October-2023

https://doi.org/10.47001/IRJIET/2023.710017

that are free of grammatical errors are input into the grammar correction module in the final stage. As a result, the system will output a sentence that has been properly grammar checked along with the corrected sentence/output.

IV. RESULT AND DISCUSSION

The research paper presents a system designed to automatically identify and rectify spelling mistakes and grammatical errors in Sinhala language text. The system comprises four main components: Pre-processing and spell error analysis, Rectification (Correction) of misspelled words, Detecting Grammatical Errors by analyzing Subjects & their Characteristics, and Grammar Error Processing (Correction).

To address the scarcity of publicly available datasets for the Sinhala language, customized datasets were created for each component. These datasets, specifically tailored to include commonly used Sinhala words and sentences, enabled more effective and accurate training of the system.

The pre-processing and spell error analysis component focuses on removing non-Sinhala characters and punctuation marks from the input sentence and splitting it into a word list. The Spelling Error Detection module, utilizing a Recurrent Neural Network (RNN) trained on the created dataset, identifies spelling errors in the words. The detected errors are then passed to the Spelling Error Correction module for rectification.

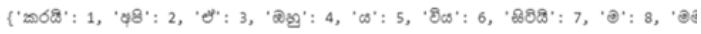


Figure 12: Tokenizing the words

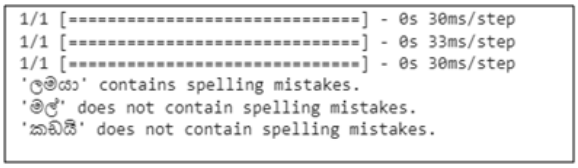
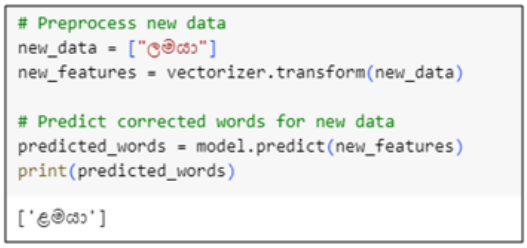


Figure 13: Predicted Output of Analyzing the Spelling mistakes

The Rectification of misspelled words component employs a logistic regression model trained on a dataset comprising correctly spelled and misspelled words. This model identifies features that differentiate correctly spelled words from misspelled words and calculates the probability of a word being misspelled. By setting a threshold, the system determines whether a word is misspelled or not. Additionally, the system provides suggestions for potential corrections, enhancing its ability to offer accurate suggestions to users.



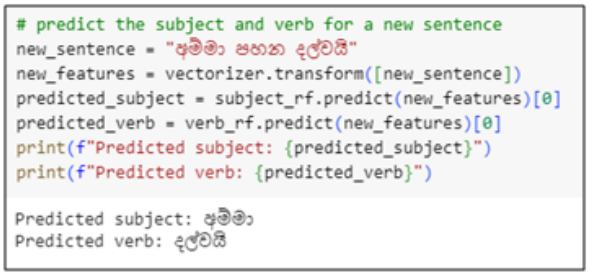
```
# Preprocess new data
new_data = ["ලමයා"]
new_features = vectorizer.transform(new_data)

# Predict corrected words for new data
predicted_words = model.predict(new_features)
print(predicted_words)
```

```
['ළමයා']
```

Figure 14: Correcting of Spelling Mistakes

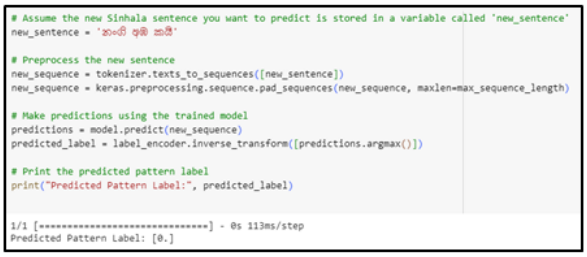
The Detecting Grammatical Errors by analyzing Subjects & their Characteristics component involves subject and verb prediction using Random Forest classifiers. The classifiers are trained on manually created datasets and evaluate the accuracy of predictions using a separate validation dataset. Sentence tense pattern detection is performed using an Artificial Neural Network, which classifies sentences based on basic sentence patterns.



```
# predict the subject and verb for a new sentence
new_sentence = "අම්මා පහන දැල්වයි"
new_features = vectorizer.transform([new_sentence])
predicted_subject = subject_rf.predict(new_features)[0]
predicted_verb = verb_rf.predict(new_features)[0]
print(f"Predicted subject: {predicted_subject}")
print(f"Predicted verb: {predicted_verb}")
```

```
Predicted subject: අම්මා
Predicted verb: දැල්වයි
```

Figure 15: Subject Verb Analysis



```
# Assume the new Sinhala sentence you want to predict is stored in a variable called 'new_sentence'
new_sentence = 'සංජී අපි සමී'

# Preprocess the new sentence
new_sequence = tokenizer.texts_to_sequences([new_sentence])
new_sequence = keras.preprocessing.sequence.pad_sequences(new_sequence, maxlen=max_sequence_length)

# Make predictions using the trained model
predictions = model.predict(new_sequence)
predicted_label = label_encoder.inverse_transform([predictions.argmax()])

# Print the predicted pattern label
print("Predicted Pattern Label:", predicted_label)
```

```
1/1 [==============================] - 0s 113ms/step
Predicted Pattern Label: [0.]
```

Figure 16: Sentence Pattern Recognition

The Grammar Error Processing (Correction) component utilizes deep learning algorithms to examine the actual sentence for grammatical errors. If errors are detected, the system focuses on the tense and verb of the sentence. Support Vector Machine (SVM) is employed to handle the dataset and perform classification. The system determines the appropriate tense and modifies the verb according to the subject. Sentence pattern recognition and Sentence Tense Detection contribute to the accurate identification and correction of grammatical errors.

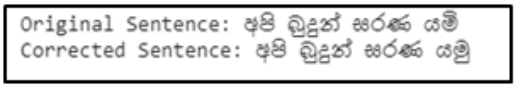


Figure 17: Grammar Error Correction

ISSN (online): 2581-3048
Volume 7, Issue 10, pp 131-138, October-2023
https://doi.org/10.47001/IRJIET/2023.710017

V. CONCLUSION AND FUTURE WORK

In this Research, we developed a system for automatically identifying and correcting spelling mistakes and grammatical errors in the Sinhala language. The system consists of four main components: pre-processing and spell error analysis, the rectification of misspelled words, detecting grammatical errors by analyzing subjects and their characteristics, and grammar error processing.

Through the creation of customized datasets, we addressed the scarcity of publicly available datasets for Sinhala, enabling us to train the models more effectively and accurately. The pre-processing and spell error analysis component successfully identified spelling errors in Sinhala text using a deep learning algorithm. The rectification of misspelled words component, employing a logistic regression model, corrected the identified spelling errors and provided suggestions for potential corrections.

The detection of grammatical errors by analyzing subjects and their characteristics was achieved using Random Forest Classifiers and an Artificial Neural Network. These components demonstrated promising accuracy in predicting subjects, verbs, and sentence tense patterns, allowing for the detection and correction of grammatical errors.

Overall, the developed system shows the potential to significantly improve the quality of written Sinhala text by automatically detecting and correcting spelling mistakes and grammatical errors. While this research presents a comprehensive system for spelling and grammar correction in Sinhala, there are several avenues for future work and improvement such as Integration of Language-specific Rules, Extension to other Languages, User Interface and Feedback and etc.

ACKNOWLEDGMENT

We would like to express our sincere gratitude to all those who have contributed to the successful completion of this research project. Your support, encouragement, and expertise have been invaluable, and we are truly grateful for the opportunity to undertake this work.

The authors express their sincere gratitude and appreciation to the language experts and Sinhala teachers for their invaluable support and encouragement during the system's development phase. Their assistance in analyzing grammar rules has been instrumental in the progress of this project.

REFERENCES

[1] C.Liyanage, R.Pushpananda, D.L.Herath, R.Weerasinghe, "A computational grammar of Sinhala, "International Conference on Intelligent Text Processing and Computational Linguistics, 2012.

[2] L.Abeyrathne, S.Edirisinghe, R.Premachandra, A.Warsha, N. De Silva, S. Thelijjagoda, "Spell and grammar checking tool for Sinhalese-අකුරුඝෙෝදුව-සයබඇසුබඇක්කරනුරිසඇසයනි,"2018.

[3] H.M.U.Pabasara, S.Jayalal, "Grammatical error detection and correction model for Sinhala language sentences, "2020 International Research Conference on Smart Computing and Systems Engineering (SCSE), 2020.

[4] A.Wasala, R.Weerasinghe, C.Liyanage, R.Pushpananda, E.Jayalatharchchi, "An Open-Source Data Driven Spell Checker for Sinhala," 2009.

[5] "Statistics.gov.lk," 2017. [Online]. Available: http://www.statistics.gov.lk/PopHouSat/CPH2011/in-ex.php?fileName=pop42gp=Activitiestpl=3.[Accessed 6 June 2019].

[6] R.C.Widyalankara, "A cause-effect analysis of phonology of Sri Lanka," 2011.

[7] R. Salomon, Guide to the Study of Inscriptions in Sanskrit, Prakrit, and the other Indo-Aryan Languages, 2000.

[8] V.Samaranayake, J.Dissanayake, A.Weerasinghe, H.Wijayawardhana, "An Introduction to UNICODE for Sinhala Characters," 2003.

[9] H.P. Ray, The archaeology of seafaring in ancient South Asia, 2003.

[10] U.Liyanapathirana, K.Gunasinghe, G.Dias, "SinSpell: A Comprehensive Spelling Checker for Sinhala," 2021.

[11] D.Gunasekara, W.V.Welgama, A.R.Weerasinghe, "Hybrid Part of Speech tagger for Sinhala Language, "in Sixteenth International Conference on Advances in ICT for Emerging Regions (ICTer), 2016.

[12] Waltl, Bernhard, Georg Bonczek, Florian Matthes, "Rule-based Information Extraction: Advantages, Limitations, and Perspectives,"2018.

[13] "Spelling and grammar," [Online]. Available: https://www.findbestopensource.com/product/maharavana. [Accessed 21 May 2019].

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 7, Issue 10, pp 131-138, October-2023

https://doi.org/10.47001/IRJIET/2023.710017

Citation of this Article:

Navoda M.R., Weerasooriya O.W.R.Y., Siriwardhana A.U.A., Sonali L.D.A., Jenny Krishara, Poorna Panduwawala, "Automated Spelling and Grammar Checker Tool for Sinhala" Published in *International Research Journal of Innovations in Engineering and Technology - IRJIET*, Volume 7, Issue 10, pp 131-138, October 2023. Article DOI https://doi.org/10.47001/IRJIET/2023.710017

****

