International Journal Scientific and Profesional (IJ-ChiProf)
Vol 4 Issue 2 March-May 2025, pp:278-286
ISSN: 2829-2618 (Online)
DOI: https://doi.org/10.56988/chiprof.v4i2.81

Implementing Conditional Random Fields on English Text Grammar Analysis

Fadhil Ahmad[1], Tata Sutabri[2]
[1,2]Department of Information Technology, Universitas Binadarma Palembang, Indonesia
Email: 242420005@student.binadarma.ac.id, tata.sutabri@gmail.com

Article Info
Article history:
Received April 21, 2025
Revised April 22, 2025
Accepted April 22, 2025

Keywords:
Analysis
Conditional Random fields
English Language
Grammar
Part of speech tagging

ABSTRACT
This study explores the implementaion of the Conditional Random Fields (CRF) algorithm in the grammatical analysis of English texts, specifically in the task of Part of Speech (POS) tagging. CRF is a statistical model effective in classifying words into grammatical categories such as nouns, verbs, adjectives, and others. The research methodology includes a literature review and experimental implementation using labeled datasets, integrated into a web-based application. The implementation results demonstrate that the CRF model provides accurate tagging results and can be utilized for sentence structure analysis in English texts. The application is developed using the Python programming language, supported by the NLTK and sklearn-crfsuite libraries, and uses the Flask framework for the user interface. This research is expected to contribute to the development of technology-based tools for English language learning.

Corresponding Author:
Fadhil Ahmad
Department of Information Technology
Universitas Binadarma Palembang,
Jl. Jenderal Ahmad Yani No.3, 9/10 Ulu, Kecamatan Seberang Ulu I, Kota Palembang, Sumatera Selatan 30111
Email: 242420005@student.binadarma.ac.id

1. INTRODUCTION

Communication is a crucial aspect of social and societal life. To establish effective communication, a sufficient understanding of the language used is essential. One common and fundamental concept is the use of language itself. Language is a familiar element for many people. Each region and country possesses one or more languages used in daily life. Mastering more than one language in the digital era offers numerous advantages and facilitates access to a broader range of information, compared to relying solely on information from a single domestic source. English serves as the dominant lingua franca—a language commonly used by individuals whose native language is not English, such as the people of Indonesia. Indonesia holds a proficiency rating of 38.45–54.06 in English language usage among ASEAN countries [1] Rahadi argues that English language skills can be improved through activities such as language courses, self-directed learning, and formal education in schools [2]. Common challenges frequently encountered in mastering English are found in the areas of reading, writing, speaking, and listening. An individual is considered proficient when they demonstrate competence across these aspects [3].

This study aims to support the English learning process through the use of technology or algorithms. Prior to that, it is important to understand the concept of Part of Speech (POS) tagging. POS tagging is the process of assigning labels to a text based on grammatical categories. This process enables a deeper understanding of the grammatical structure of a text and plays a significant role in various natural language processing (NLP) tasks such as syntactic parsing, semantic analysis, text processing, information retrieval, machine translation, and named entity recognition [4]. In recent years, Conditional Random Fields (CRFs) have become a highly valuable tool in the field of natural language processing, particularly in handling sequential data such as text analysis. A study by Wang (2024) demonstrated the application of CRFs in conjunction with Rasch Measurement Models to analyze English paragraphs. This research confirmed the effectiveness of the model in capturing linguistic patterns and enhancing the identification of grammatical components [5]. CRFs function by assigning labels such as nouns, adjectives, and verbs [6]. The model

International Journal Scientific and Profesional (IJ-ChiProf)
Vol 4 Issue 2 2025, pp: 278-286
ISSN: 2829-2618 (Online)

is capable of recognizing the relationships between words and their contextual meanings, thus enabling accurate labeling even when the available data is limited or ambiguous [7]. At a glance, the CRF algorithm appears well-suited for English grammar text analysis aimed at improving the language learning process [8].

2. RESEARCH METHOD

This study was conducted using a literature review and experimental method by applying the Conditional Random Fields (CRF) algorithm, with the aim of assisting users in the analysis and classification of grammatical components and elements in English, such as verbs, nouns, adjectives, prepositions, determiners, and others. Once the CRF-based part-of-speech tagging model is trained, it will be utilized to help users identify the grammatical components of the input text. The Conditional Random Fields (CRF) algorithm is a commonly used model for part-of-speech tagging. This natural language processing technique is employed to classify labels or tags for individual words. Part-of-speech tagging itself consists of various approaches, such as rule-based and stochastic methods. Conditional Random Fields fall under the stochastic category, which relies on statistical measures, probabilities, and frequency distributions [9]. The process of part-of-speech tagging typically involves several stages, namely annotation, the placement of grammatical components, and parsing. Annotation refers to the process of assigning POS tags to words within a sentence. Before this step begins, the set of tags or labels must first be determined. Once the annotation is complete, it is necessary to ensure the accurate placement of grammatical components. The final step is parsing, which functions to determine whether the processed symbols or sentences conform to grammatical rules [10].

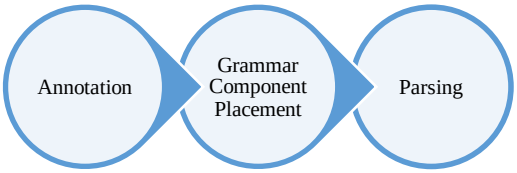

Image 1. Part of speech tagger main elements

In the context of language learning, understanding the components of grammar is of great importance. A part-of-speech tagger plays a significant role in facilitating the identification of common grammatical components. Conditional Random Fields fall under the category of discriminative models. The following are several grammatical components in the English language that will serve as references during the model training process using CRF [11].

Table 1. Open class POS Tags

Tag	Decription	Example
ADJ	Adjective	Happy, clever
ADV	Adverb	So, very
NOUN	Noun	University
VERB	Verb	Work
PROPN	Proper Noun	Ahmad, Indonesia

Table 2. Closed class POS Tag

Tag	Description	Example
ADP	Preposition	In, on, at
AUX	Auxiliary verb	Can, may, will
CCONJ	Conjuction	But, and
DET	Determiner	This, that, the
NUM	Numeral	Two, 21, 23:00
PART	Particle	Is, not, to
PRON	Pronoun	She, he, I
SCONJ	Subordinating conjuction	Whether, because

International Journal Scientific and Profesional (IJ-ChiProf)
Vol 4 Issue 2 March-May 2025, pp:278-286
ISSN: 2829-2618 (Online)
DOI: https://doi.org/10.56988/chiprof.v4i2.81

Table 3. Other POS Tags category

Tag	Description	Example
PUNCT	Punctuation	, . ; ()
SYM	Symbol	@#$%
X	Other	Asdf, gwf

This research was conducted through several systematic steps. First, the researcher collected a dataset consisting of English sentences that had been labeled with part-of-speech (POS) tags, such as data from Universal Dependencies or nltk.corpus.treebank. The next step involved preprocessing, which included tokenizing the sentences into individual words, normalizing the text to lowercase, and removing irrelevant symbols. Subsequently, key features were extracted from each word, including the word form itself, whether it begins with a capital letter, word prefixes and suffixes, and the word's position in the sentence, such as at the beginning (BOS) or end (EOS). Once the features were extracted, the Conditional Random Fields model was trained using Python libraries such as sklearn-crfsuite with the labeled data to produce accurate POS tagging. The trained model was then evaluated using a test dataset by measuring accuracy, precision, recall, and F1-score. As a final step, the POS tagging results were used to analyze sentence grammar structures, including identifying subject, predicate, object, and other syntactic elements. Previous studies on the application of Conditional Random Fields in natural language processing have demonstrated significant results in various linguistic tasks, including morphological analysis and named entity recognition.

CRF-based frameworks are capable of addressing the label bias problem and offer flexibility in selecting interdependent features. This approach enables more accurate analysis of word sequences, which is crucial in understanding the grammatical structure of a sentence [12]. The article 'White Blood Cell Classification Using SMOTE-SVM Method with Hybrid Feature Extraction and Image Segmentation Using Gaussian Mixture Model' (ECTI-CIT, 2021) illustrates how machine learning approaches can be optimized through a combination of feature extraction techniques and statistical models, although it is not specifically within the context of language processing [13]. The following image will show the steps of research.

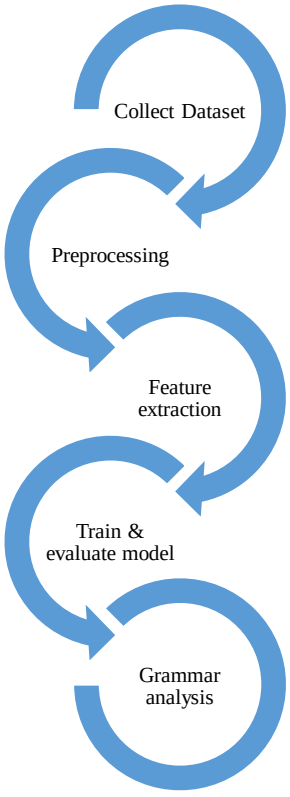

Image 2. Research steps

Vol 4 Issue 2 2025, pp: 278-286
ISSN: 2829-2618 (Online)

3. RESULTS AND ANALYSIS

In his book *Information System Analysis* (2012), Tata Sutabri defines a system model as consisting of input, process, and output. This study views the process of grammar analysis in English texts using the Conditional Random Fields (CRF) algorithm as part of an information system. Within this system, English text data serves as the input, the CRF algorithm performs the processing to identify grammatical structures, and the output of this analysis becomes valuable for various applications, such as natural language processing and the development of linguistic information systems [14]. In studies related to English language learning, such as Task-Based Language Learning (TBLL), the CRF algorithm automatically identifies grammatical elements—such as subject, predicate, and object—through the POS tagging process, which serves as the foundation for sentence structure analysis in English [15]. To enable users to more easily visualize the practical output of CRF algorithm implementation, it is necessary to develop a simple interface. This can be systematically achieved by applying software development frameworks such as RAD, Prototype, Waterfall, Agile, and others [16]. To narrow the scope and ensure timely delivery, a basic prototype will be developed. The CRF implementation utilizes a Python library called *sklearn-crfsuite*. This library provides an intuitive interface that is compatible with the *scikit-learn* ecosystem, thereby simplifying the sequence labeling process such as POS tagging. With features such as support for the L-BFGS optimization algorithm, regularization settings, and automated evaluation reporting, *sklearn-crfsuite* serves as an efficient and flexible tool for computational analysis of English grammar structures [17]. The main objective of this study is to develop a simple application integrated with the Conditional Random Fields machine learning algorithm. The POS tagger is supported by the *Natural Language Toolkit* (NLTK), a Python library that facilitates the implementation of POS taggers and CRF in the developed software.

Table 4. Required Tools and Libraries

Number	Tool name	Purpose
1	Python 3.11	Scripting langauge used to code and program the application.
2	NLTK library	Library that allows POS tagger implementation
3	Sklearn-crfsuite library	Library used to implement CRF into the application.
4	Flask framework	Python framework that runs the application.
5	HTML/CSS/Javascript	Used to design and handle the frontend

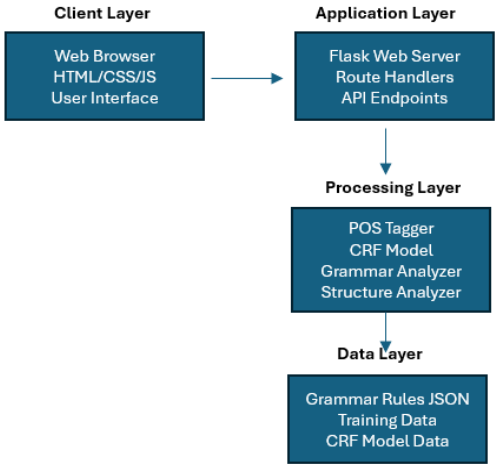

Image 3. System Architecture

This application is structured into four layers to ensure an effective grammar analysis process. The client layer handles user interactions through an interface developed using HTML, CSS, and JavaScript. The application layer, powered by Flask, manages HTTP requests and route handling. The processing layer contains the core components of NLTK, including the POS tagger and the CRF model, which are responsible for performing the grammar analysis. The data layer stores and manages grammar rules, training data, and configuration models in JSON format.This layered

International Journal Scientific and Profesional (IJ-ChiProf)
Vol 4 Issue 2 March-May 2025, pp:278-286
ISSN: 2829-2618 (Online)
DOI: https://doi.org/10.56988/chiprof.v4i2.81

architecture allows the software to be more scalable, enabling the addition of new features or options that enhance user experience and facilitate the use of the application.

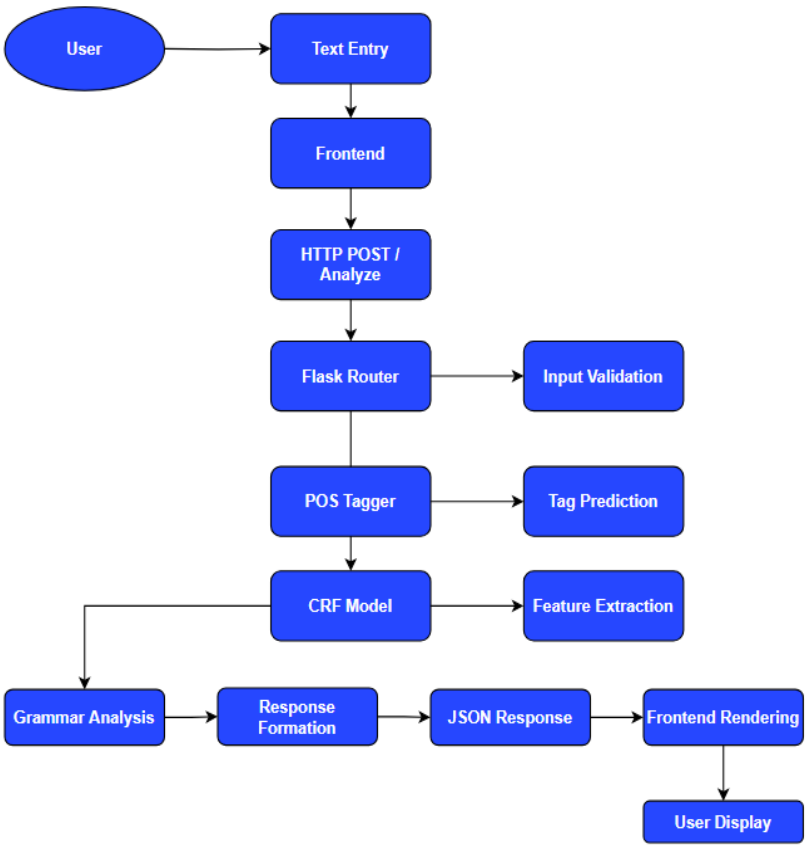

Image 4. Data flow diagram

The data flow begins when the user inputs a sentence or paragraph through the web-based user interface. This input is captured by the JavaScript frontend, which is responsible for managing dynamic behavior on the client side. Once the input is validated, it is transmitted via an HTTP POST request to the Flask backend server. Flask acts as the application controller, receiving the request and passing the input to the processing pipeline. Within the processing layer, the input text undergoes several stages of natural language processing. Initially, the text is tokenized and annotated using a POS (Part-of-Speech) tagger, which extracts linguistic features such as word shape, capitalization, affixes, and positional context (e.g., beginning-of-sentence or end-of-sentence). These features are then passed to the Conditional Random Fields (CRF) model, which has been trained on a labeled corpus. The CRF model performs probabilistic sequence labeling, predicting the most likely grammatical tags for each token based on learned contextual dependencies. Following POS tagging, a grammar analysis component conducts two key types of analysis: structural analysis, which evaluates sentence construction, complexity, and grammatical form (e.g., identifying clause boundaries or passive voice structures); and token-level analysis, which provides descriptive explanations for each POS tag assigned. This dual-layered analysis supports deeper understanding and interpretability for learners or downstream NLP applications.

The annotated and analyzed output is then formatted into a structured JSON response. This format allows for efficient rendering and integration on the frontend. The response includes detailed grammatical annotations, sentence-level summaries, and optionally, color-coded tags or tooltips for enhanced readability. The frontend parses this JSON data and dynamically displays the results using interactive visual elements, ensuring the information is both accessible and pedagogically useful to the user. This modular and layered approach not only ensures a robust and efficient processing workflow but also enables easy maintenance and scalability. Additional functionalities—such as grammar suggestions, error highlighting, or export options—can be integrated with minimal changes to the core architecture.

International Journal Scientific and Profesional (IJ-ChiProf)
Vol 4 Issue 2 2025, pp: 278-286
ISSN: 2829-2618 (Online)

Furthermore, the clear separation between the data, logic, and presentation layers helps preserve data integrity and maintain consistency throughout the grammar analysis process.

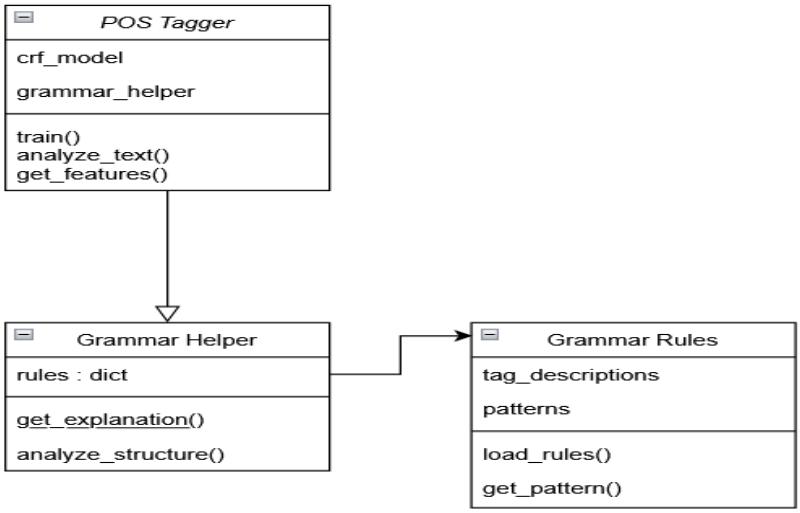

Image 5. Entity relations diagram

The application's data structure is centered around three main classes: POSTagger, GrammarHelper, and GrammarRules. The POSTagger class manages the CRF model and the core text analysis functions. The GrammarHelper class is responsible for interpreting grammatical rules and performing structural analysis, while the GrammarRules class stores tag descriptions and pattern definitions. Key data is stored in JSON format, including grammar rules (which contain tag definitions and examples), analysis responses (which hold token-level and structural information), and training data (which maintains the relationship between sentences and tokens). This architecture enables efficient data management while supporting comprehensive grammatical analysis capabilities. It is important to note that the application does not require a traditional Entity-Relationship Diagram (ERD), as it does not utilize a relational database. Instead, it relies on file-based storage for grammar rules, model configurations, and associated data, which simplifies deployment and enhances portability.

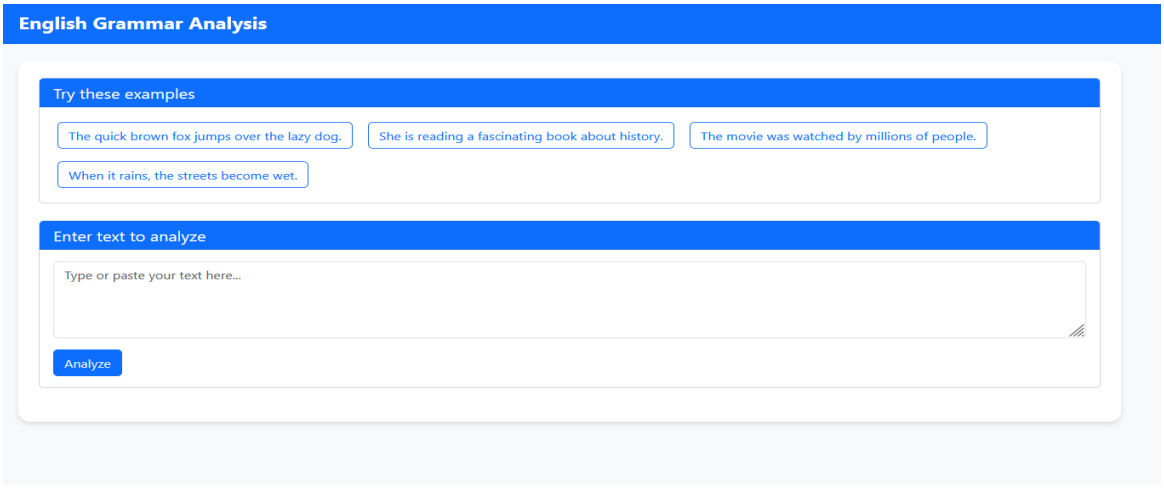

Image 6. Initial interface of the homepage

On the initial interface displayed when users access the application via the website, several primary components are presented. These include a text input field where users can enter the text they wish to analyze. For users who are unsure of what to input, several sample texts are provided to demonstrate how the grammar analysis works. Once the

International Journal Scientific and Profesional (IJ-ChiProf)
Vol 4 Issue 2 March-May 2025, pp:278-286
ISSN: 2829-2618 (Online)
DOI: https://doi.org/10.56988/chiprof.v4i2.81

text is entered into the input field, the user can click the 'Analyze' button to initiate the analysis process. The results will then be displayed, consisting of both structural and grammatical component analysis in the form of POS (Part-of-Speech) tagging.

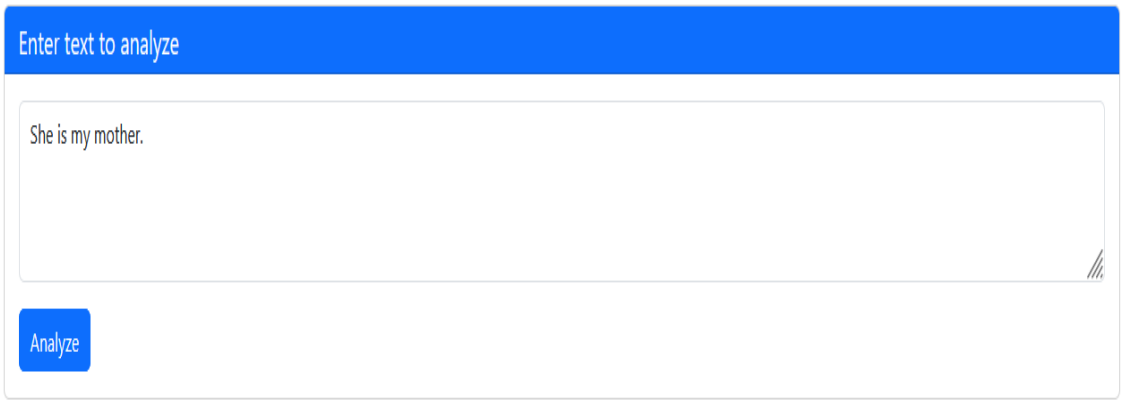

Image 7. Input field and column for text analysis

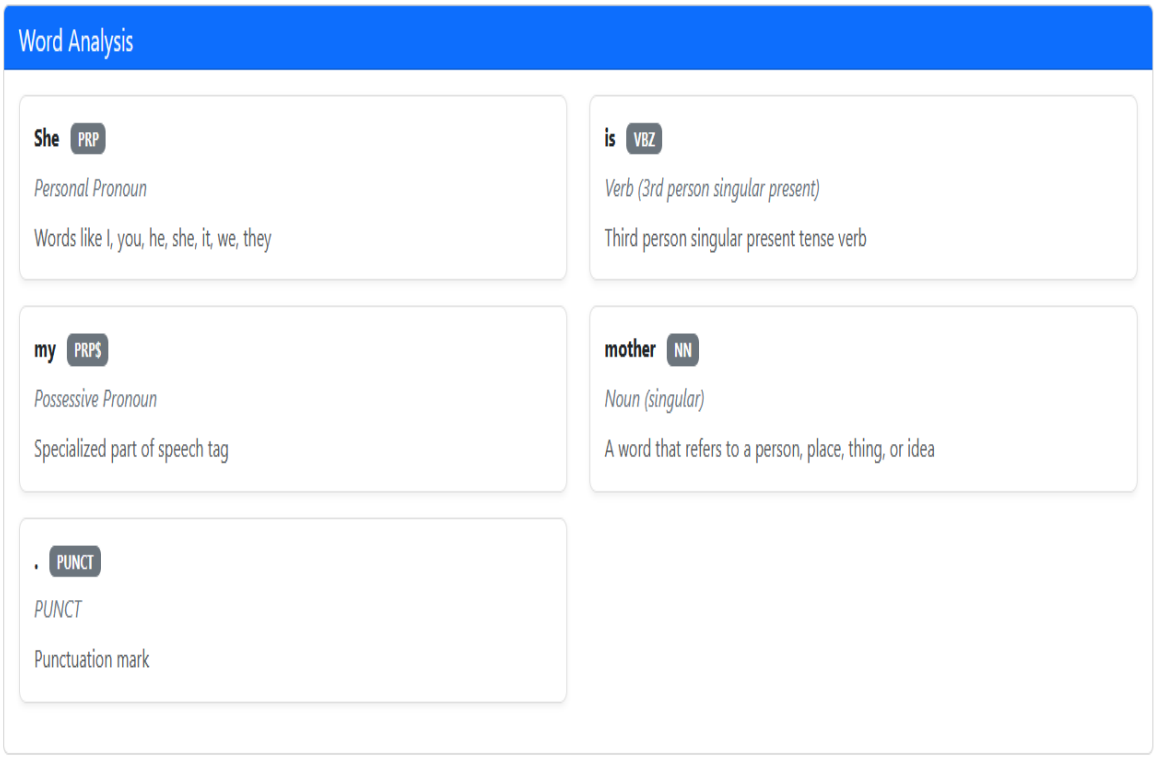

Image 8. Word analysis column

After the grammar analysis is successfully completed, users are presented with a word analysis table. In this section, users can view the POS (Part-of-Speech) tags corresponding to each word entered in the initial input field. This feature allows users to clearly observe the grammatical classification of each token in the sentence, supporting a deeper understanding of the syntactic structure.

International Journal Scientific and Profesional (IJ-ChiProf)
Vol 4 Issue 2 2025, pp: 278-286
ISSN: 2829-2618 (Online)

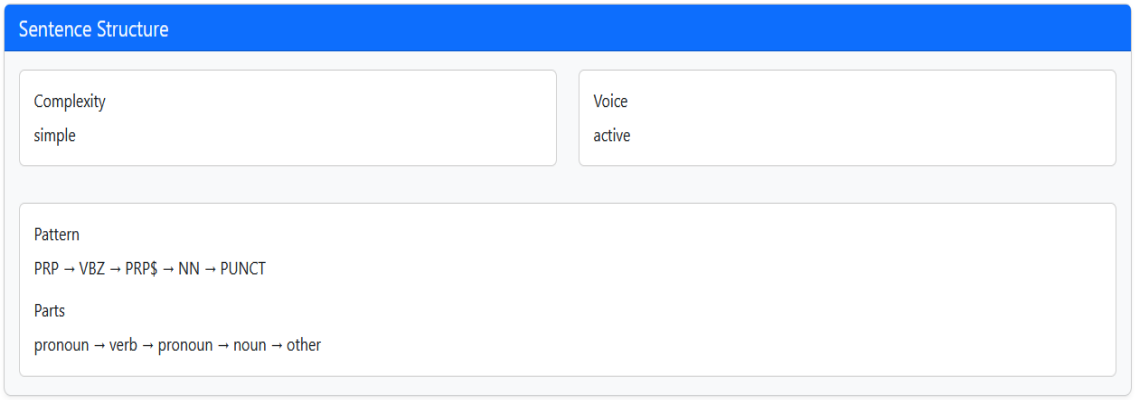

Image 9. Sentence structure anaylsis

Users can also view the structural analysis of the input sentence to gain a deeper understanding of the core grammatical components within the English text.

4. CONCLUSION

Based on the results of the conducted study, it can be concluded that the Conditional Random Fields (CRF) algorithm demonstrates high effectiveness in the grammatical analysis of English texts, particularly in Part-of-Speech (POS) tagging tasks. Through a statistical and machine learning-based approach, the CRF model is capable of accurately identifying and classifying grammatical elements, even in situations involving limited data or ambiguous sentence contexts. The system development process was carried out through a series of stages, including labeled data collection, linguistic feature extraction, model training, and performance evaluation using metrics such as accuracy, precision, recall, and F1-score. The implementation of a web-based application, developed using Python libraries such as NLTK and sklearn-crfsuite, alongside the Flask framework, successfully provided an interactive user interface that supports sentence structure analysis and word classification based on grammatical categories. The findings of this research demonstrate that the integration of Natural Language Processing (NLP) technologies with the CRF algorithm offers an innovative solution to support English language learning. Moreover, it opens up opportunities for the development of more comprehensive and adaptive linguistic information systems in the future.

5. ACKNOWLEDGEMENTS

First and foremost, the authors would like to express their deepest gratitude to Allah SWT for the strength, guidance, and blessings that made this research possible. We extend our sincere appreciation to Universitas Binadarma Palembang, especially the Department of Information Technology, for the support and resources provided throughout this study. Special thanks are due to Mr. Tata Sutabri, whose mentorship and academic guidance played a key role in the development of this research. We are also grateful to the contributors of the open-source libraries used in this work, including NLTK, sklearn-crfsuite, and the Flask framework. Finally, heartfelt thanks go to the first author's parents for their constant encouragement and support, and to all peers and fellow researchers who provided valuable feedback and motivation during this research journey.

REFERENCES

[1] N. Syafitri, F. Annisa, E. Purnomo, M. Lutfi, S. Suhairi, and J. Manajemen, "PENGGUNAAN BAHASA INGGRIS SEBAGAI STRATEGI KOMUNIKASI GLOBAL DALAM INDUSTRI PARIWISATA," *Jurnal Ilmiah PGSD FKIP Universitas Mandiri*, vol. 9, pp. 1–12, Dec. 2023.

[2] A. Murti and A. D. A. Kusuma, "Kecakapan Berbahasa Inggris Serta Keterlibatan Masyarakat Dalam Pengembangan Desa Ekowisata Pancoh," *Khasanah Ilmu - Jurnal Pariwisata Dan Budaya*, vol. 14, no. 1, pp. 21–29, Apr. 2023, doi: 10.31294/khi.v14i1.15305.

[3] I. Gusti, A. Agung, and D. Susanthi, "KENDALA DALAM BELAJAR BAHASA INGGRIS DAN CARA MENGATASINYA," *Linguistic Community Service Journal* |, vol. 1, no. 2, pp. 1–7, 2021, doi: 10.22225/licosjournal.v1i2.2658.

[4] M. Ali, M. Khan, and Y. Alharbi, "A conditional random field based approach for high-accuracy part-of-speech tagging using language-independent features," *PeerJ Comput Sci*, vol. 10, 2024, doi: 10.7717/peerj-cs.2577.

[5] Y. Wang, "The application effect of the Rasch measurement model combined with the CRF model: An analysis based on English discourse," *PLoS One*, vol. 19, no. 8, Aug. 2024, doi: 10.1371/journal.pone.0309001.

[6] C. Intelligence and Neuroscience, "Retracted: English Grammar Detection Based on LSTM-CRF Machine Learning Model," *Comput Intell Neurosci*, vol. 2023, no. 1, Jan. 2023, doi: 10.1155/2023/9818431.

[7] R. Wang, "An Automatic Error Detection Method for Engineering English Translation Based on the Deep Learning Model," *Math Probl Eng*, vol. 2022, 2022, doi: 10.1155/2022/9918654.

[8] R. Takhanov and V. Kolmogorov, "Combining pattern-based CRFs and weighted context-free grammars," *Intelligent Data Analysis*, vol. 26, no. 1, pp. 257–272, 2022, doi: 10.3233/IDA-205623.

[9] J. Awwalu, S. E.-Y. Abdullahi, and A. E. Evwiekpaefe, "PARTS OF SPEECH TAGGING: A REVIEW OF TECHNIQUES," *FUDMA JOURNAL OF SCIENCES*, vol. 4, no. 2, pp. 712–721, Oct. 2020, doi: 10.33003/fjs-2020-0402-325.

[10] K. S. M. Anbananthen, J. K. Krishnan, Mohd. S. Sayeed, and P. Muniapan, "Comparison of Stochastic and Rule-Based POS Tagging on Malay Online Text," *Am J Appl Sci*, vol. 14, no. 9, pp. 843–851, Sep. 2017, doi: 10.3844/ajassp.2017.843.851.

[11] D. Jurafsky and J. H. Martin, "Speech and Language Processing An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models Third Edition draft Summary of Contents," Colorado, Jan. 2024.

[12] T. Sutabri, A. Suryatno, D. Setiadi, and E. S. Negara, "Improving Naïve Bayes in Sentiment Analysis For Hotel Industry in Indonesia," in *2018 Third International Conference on Informatics and Computing (ICIC)*, 2018, pp. 1–6. doi: 10.1109/IAC.2018.8780444.

[13] T. Sutabri and C. Adi Putra, "_White+Blood+Cell+Classification+Using+SMOTE-SVM+Method+with+," Jan. 2025.

[14] Tata Sutabri, *Analisis Sistem Informasi*, 1st ed., vol. 1. Yogyakarta: ANDI OFFSET, 2012. Accessed: Dec. 03, 2024. [Online]. Available: https://books.google.co.id/books?id=ro5eDwAAQBAJ&printsec=copyright#v=onepage&q&f=false

[15] F. Ahmad and T. Sutabri, "Implementasi Metode Task Based Language Learning untuk Meningkatkan Kompetensi Bahasa Inggris Berbasis Android," Palembang, 2024.

[16] F. Ahmad, N. Sari, and T. Sutabri, "Pengembangan Sistem Informasi E-Permit Menggunakan Metode Rapid Application Development Pada Polsek Semendawai Suku III," 2024.

[17] D. Chopra, N. Joshi, and I. Mathur, *Mastering Natural Language Processing With Python*. 2016.