Journal of Informatics, Computer Science, Data Science And Arificial Intelligence
Volume 4 No. 1 June 2026, ISSN : 3031-9145

IMPLEMENTATION OF HIDDEN MARKOV MODEL WITH VITERBI ALGORITHM FOR DETECTING MEDAN SLANG

Gilang Ramadhan[1], Khairuddin Nasution[2], Darjat saripurna[3], Zainuddin Nasution[4]
[1]Universitas Islam Sumatera Utara, gilangramadhannn498@gmail.com.
[2]Universitas Islam Sumatera Utara, khairuddin_nst@uisu.ac.id.
[3]STMIK Triguna Dharma, darjatsaripurna@gmail.com.
[4]Universitas Islam Sumatera Utara, zainuddin.nasution@fisip.uisu.ac.id.

ARTICLE INFO

ARTICLE HISTORY:
Received : 10 June 2026
Revised : 22 June 2026
Accepted : 30 June 2026

Keywords:
Hidden Markov Model, Viterbi Algorithm, Natural Language Processing, Medan Slang,Text Classification.

ABSTRACT

The increasing use of slang language in digital communication has created challenges in automatic text analysis, particularly for regional slang variations such as Medan slang. This study aims to implement the Hidden Markov Model (HMM) with the Viterbi Algorithm to detect Medan slang words in fictional texts. HMM is used to model probabilistic relationships between words through two hidden states, namely standard language and slang language, while the Viterbi Algorithm determines the most probable sequence of states based on observed words. The dataset consists of training and testing data collected from fictional stories containing Medan slang expressions. The proposed method was implemented in a web-based application that provides slang detection, history management, and dataset administration features. Experimental results show that the HMM-Viterbi approach can successfully identify slang words and classify them according to their linguistic categories. The evaluation indicates that the proposed model performs effectively in detecting Medan slang patterns within textual data. This research contributes to the development of Natural Language Processing applications for regional language analysis and demonstrates the potential of probabilistic sequence modeling for local slang detection

INTRODUCTION

Language is one of the most important elements of human communication and cultural identity. Along with the rapid growth of digital technology and social media platforms, language usage has evolved significantly, particularly among teenagers and young adults. One of the most notable phenomena is the increasing use of slang language in online and offline communication. Slang language is often characterized by informal expressions, shortened words, modified spellings, and region-specific vocabulary that differ from standard language conventions.

In Medan, slang language has developed into a distinctive communication style influenced by Indonesian, local dialects, and cultural interactions. Expressions such as *awak*, *klen*, *cemana*, *kedan*, and *lek* are commonly used by young people in daily conversations. Although these expressions enrich local linguistic identity, they create challenges for automatic text processing systems because many of them do not follow standard Indonesian language structures.

The emergence of social media platforms such as Instagram, Twitter, TikTok, and online messaging applications has accelerated the spread of slang language. As a result, the need for automated systems capable of identifying and classifying slang expressions has increased. Such systems may support linguistic studies, educational applications, sentiment analysis, and regional language preservation efforts.

Several previous studies have investigated slang language from sociolinguistic and communication perspectives. However, only a limited number of studies have applied computational approaches for automatically detecting regional slang languages. Existing studies primarily focus on language usage patterns and social impacts, while computational detection methods remain relatively unexplored.

To address this research gap, this study proposes the implementation of Hidden Markov Model (HMM) combined with the Viterbi Algorithm for Medan slang language detection. HMM is capable of modeling probabilistic relationships among sequential words, while the Viterbi Algorithm identifies the most probable sequence of hidden states based on observed words. Through this approach, words can be automatically classified into standard language and slang language categories. The objective of this study is to develop and evaluate a web-based Medan slang detection system using Hidden Markov Model and the Viterbi Algorithm and analyze its effectiveness in detecting regional slang expressions.

LITERATURE REVIEW

Medan Slang Language

Medan slang language is an informal language variety commonly used by young people and local communities in Medan. It is formed through a combination of Indonesian language, local dialects, and cultural influences. Medan slang is characterized by shortened expressions, phonological modifications, and unique vocabulary that differs from standard Indonesian language. Examples include *awak* (I), *klen* (you all), *cemana* (how), and *kedan* (close friend).

Natural Language Processing

Natural Language Processing (NLP) is a field of artificial intelligence that focuses on enabling computers to understand, interpret, and process human language. NLP techniques are widely applied in machine translation, sentiment analysis, speech recognition, and text classification. In this study, NLP techniques are employed to analyze textual data and identify slang language patterns automatically.

Hidden Markov Model

Hidden Markov Model (HMM) is a probabilistic model used to represent systems with hidden states and observable outputs. In Natural Language Processing, HMM has been extensively applied to tasks such as part-of-speech tagging, speech recognition, and sequence prediction. The model consists of initial probabilities, transition probabilities, and emission probabilities that collectively describe the relationship between hidden states and observed words.

Viterbi Algorithm

The Viterbi Algorithm is a dynamic programming algorithm used to determine the most probable sequence of hidden states in a Hidden Markov Model. The algorithm performs probability calculations through initialization, recursion, termination, and backtracking processes. Due to its computational efficiency, the Viterbi Algorithm is widely used in sequence classification and speech recognition applications.

Related Studies

Previous studies have investigated slang language and probabilistic language processing from various perspectives. However, limited research has focused specifically on Medan slang detection using Hidden Markov Models and the Viterbi Algorithm. Therefore, this study contributes by applying probabilistic sequence modeling techniques to detect regional slang language automatically.

RESEARCH METHODS

Research Design

This study employs a quantitative approach with a computational experimental method to implement and evaluate the Hidden Markov Model and the Viterbi Algorithm for detecting Medan slang language in textual data. The research process consists of data collection, data preparation, model training, Viterbi decoding, system implementation, and performance evaluation.

Data Collection

Data collection was conducted through two methods:

1. Literature Study: Literature sources were obtained from scientific journals, conference proceedings, repositories, and academic databases such as Google Scholar, SINTA, and GARUDA.
2. Observation: The observational stage involved collecting Medan slang expressions from social media platforms such as Instagram, Twitter, and TikTok, as well as daily conversations among young people in Medan.

The collected texts consisted of fictional stories, social media comments, captions, and conversational expressions containing Medan slang vocabulary. These data were compiled into a dataset for training and testing purposes.

Data Preparation and Labeling

After the data collection process, all textual data were transformed into a structured dataset suitable for training the Hidden Markov Model. Each sentence was tokenized into individual words and manually labeled according to its linguistic category. Two labels were defined in this study, namely *Standard Language (B)* and *Slang Language (G)*. The labeling process was conducted to provide training data for estimating transition and emission probabilities within the Hidden Markov Model.
A sample of the labeled training dataset is presented in

Table 1. Slang word dataset

Kalimat	**Observasi**	**Label**
Mengkek kali dia bujang	mengkek, kali, dia, bujang	G,B,B,G
Saya otw ke kampus	saya, mau, ke, kampus	B,G,B,B
Longor kali dia	longor, kali, dia	G,B,B

Asik becewek aja dia bre	asik, becewek, aja, dia, bre	B,G,B,G

Hidden Markov Model Training

This study applies the Hidden Markov Model (HMM) combined with the Viterbi Algorithm to detect Medan slang words in textual data. The dataset was collected from fictional stories and daily conversational expressions containing Medan slang vocabulary. Each word in the dataset was manually labeled into two hidden states: Standard Language (B) and Slang Language (G).

The research process consists of data collection, preprocessing, model training, Viterbi decoding, and performance evaluation. During preprocessing, each sentence is tokenized into individual words and assigned corresponding labels. The labeled dataset is then used to estimate the parameters of the Hidden Markov Model.

The Hidden Markov Model is represented as:

$$\lambda = (A, B, \pi)$$

where A denotes the transition probability matrix, B denotes the emission probability matrix, and π represents the initial state probability.

Based on the training dataset, the initial state probabilities are calculated as shown in Table 1.

Table 1. Initial State Probabilities

Matriks π (initial probabilities)

First label G, B, G, B, G

Total number of sentences = 5

- G = 3
- B = 2

State	Number Of Sentences Starting With State	Total Sentencest	Probabilitas Fir (π)
Gaul (G)	3	5	0,6

Baku ((B)	2	5	0.4

The transition probability matrix is shown in Table 2.

Table 2. Transition Probability Matrix

Matriks A (transition probabilities)

All transition

Sentences 1: G → B → B → G

- G→B
- B→B
- B→G

Sentences 2: B → G → B → B

- B→G
- G→B
- B→B

Sentences 3: G → B → B

- G→B
- B→B

Sentences 4: B → G → B → G

- B→G
- G→B
- B→G

Sencentes 5: G → B → G → B

- G→B
- B→G
- G→B

From G:

- G→B = 6
- G→G = 0

  Total = 6
- P(G→B) = 6/6 = **1.00**
- P(G→G) = 0/6 = **0.00**

From B:

- B→G = 5
- B→B = 3

  Total = 8
- P(B→G) = 5/8 = **0.625**
- P(B→B) = 3/8 = **0.375**

From → To	**G**	**B**
Gaul (G)	0.00	1.00
Baku (B)	0.625	0.375

The emission probabilities are estimated from word frequencies observed in each state. Examples are presented in Table 3.

Table 3. Emission Probability Examples

Matriks B (emission probabilities)

State G (total sentences G = 8)

All slang words appear once.

- All words G = 1/8 = **0.125**

State B (total sentences B = 12)

Time (B)

- Appears 3 times
- 3/12 = **0.25**

He (B)

- Appears 4 times
- 4/12 = **0.33**

Tabel emisi state G

All words in G appear 1 time, so

Word	**Calculation**	**Prob**
mengkek	1/8	0.125
bujang	1/8	0.125
otw	1/8	0.125
longor	1/8	0.125
becewek	1/8	0.125
bre	1/8	0.125
bengak	1/8	0.125

cuy	1/8	0.125

Tabel emisi state B

Word	**Amount**	**Calculation**	**Prob**
kali	3	3/12	0.25
dia	4	4/12	0.33
saya	1	1/12	0.083
ke	1	1/12	0.083
kampus	1	1/12	0.083
asik	1	1/12	0.083
aja	1	1/12	0.083

After the HMM parameters are obtained, the Viterbi Algorithm is applied to determine the most probable hidden state sequence. The recursive equation used in the Viterbi Algorithm is:

$$\boldsymbol{\delta_t(j) = max_i\left[\delta_{(t-1)}(i) \times a_i j\right] \times b_j(O_t)}$$

where δt(j) is the maximum probability of state j at time t, aij is the transition probability from state i to state j, and bj(Ot) is the emission probability of observation Ot.

The output of the algorithm is the most probable sequence of labels representing whether each observed word belongs to standard language or Medan slang language.

Example of Viterbi Algorithm Implementation

The Viterbi Algorithm was applied to determine the most probable sequence of hidden states for each observed word in the input sentence. The algorithm utilizes the initial

probability, transition probability, and emission probability obtained during the Hidden Markov Model training process.

To illustrate the implementation, the following sentence is used as an example:
Input sentences:

"Saya otw ke kampus"

The observation sequence is defined as:
O = {saya, otw, ke, kampus}

The hidden states used in this study consist of:
S = {B, G}

Where:

- **B** = Standard Language (Bahasa Baku)
- **G** = Medan Slang Language (Bahasa Gaul)

During the initialization stage, the probability of each state is calculated using the initial probability (π) and the emission probability of the first observed word. Next, the recursion stage computes the highest probability for every subsequent observation by combining the transition probability matrix (A) and the emission probability matrix (B). Finally, the backtracking stage determines the most probable hidden state sequence.
The classification result produced by the Viterbi Algorithm is presented in Table 4.

Table 4. Example of Viterbi Detection Result

Word	**Predicated State**	**Clasification**
Saya	B	Standard Language
Otw	G	Medan Slang
Ke	B	Standard Language
Kampus r	B	Standard Language

Based on the probability calculation, the Viterbi Algorithm generates the hidden state sequence:

B → G → B → B

The result indicates that the word **"otw"** is classified as Medan slang, while the remaining words are classified as standard Indonesian. This demonstrates that the proposed Hidden Markov Model combined with the Viterbi Algorithm is capable of identifying slang words by considering the probability of transitions between hidden states and the probability of each observed word appearing in a particular state.
The implementation confirms that probabilistic sequence modeling enables more reliable classification than simple dictionary matching because the prediction is determined based on contextual state transitions rather than solely on individual word lookup.

To evaluate the performance of the proposed model, a testing dataset consisting of manually labeled words was used. The classification results are summarized using a confusion matrix.

Table 5. Confusion Matrix

Aktual \ Prediksi	**Slang**	**Not slang**
Bahasa Gaul	35 (TP)	5 (FN)
Bukan Bahasa Gaul	6 (FP)	54 (TN)

Based on the confusion matrix, the evaluation metrics are calculated as follows:

Accuracy = (TP + TN) / (TP + TN + FP + FN)

Precision = TP / (TP + FP)

Recall = TP / (TP + FN)

F1-Score = 2 × (Precision × Recall) / (Precision + Recall)

The performance evaluation results are shown in Table 6.

Table 6. Performance Evaluation Results

Metrik	**Mark**
Akurasi	89%
Precision	0.85
Recall	0.88
F1-Score	0.86

The experimental results show that the Hidden Markov Model combined with the Viterbi Algorithm can effectively detect Medan slang words within textual data. The model is capable of recognizing sequential word patterns and distinguishing slang expressions from standard Indonesian words with satisfactory performance.

UML Modeling

Unified Modeling Language (UML) was used to model the proposed system before implementation. The system was designed using Use Case Diagrams, Activity Diagrams, and Class Diagrams. These diagrams describe user interactions, system workflows, and relationships among system entities.

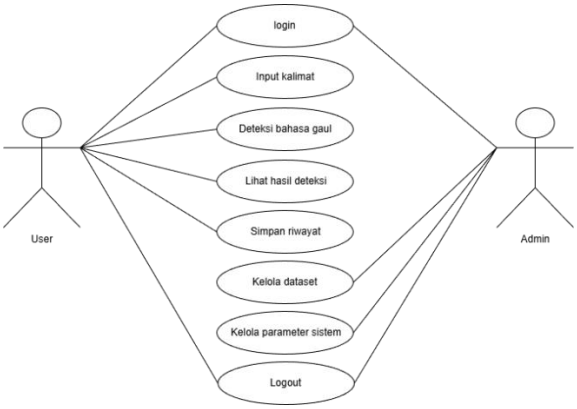

Figure 2 System Use Case Diagram

RESULTS AND DISCUSSION

System Implementation

The proposed system was implemented as a web-based application using PHP, MySQL, HTML, CSS, and JavaScript. The implementation integrates the Hidden Markov Model and Viterbi Algorithm to classify input words into standard language and Medan slang categories.

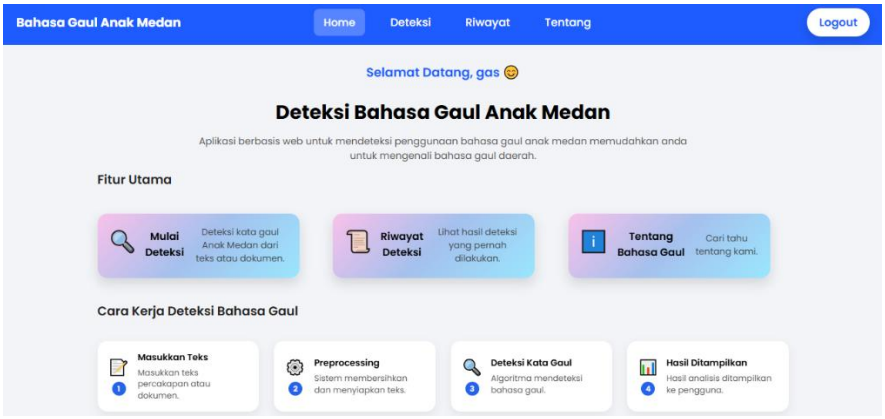

Figure 3 Application dashboard page

The dashboard serves as the main interface of the Medan slang detection system. This page provides users with access to the primary functionalities of the application, including slang detection, detection history, and information related to the system. The dashboard is designed to offer a simple and user-friendly interface, allowing users to navigate the application efficiently. In addition, the dashboard provides an overview of the system and facilitates interaction between users and the slang detection features implemented using the Hidden Markov Model and Viterbi Algorithm.

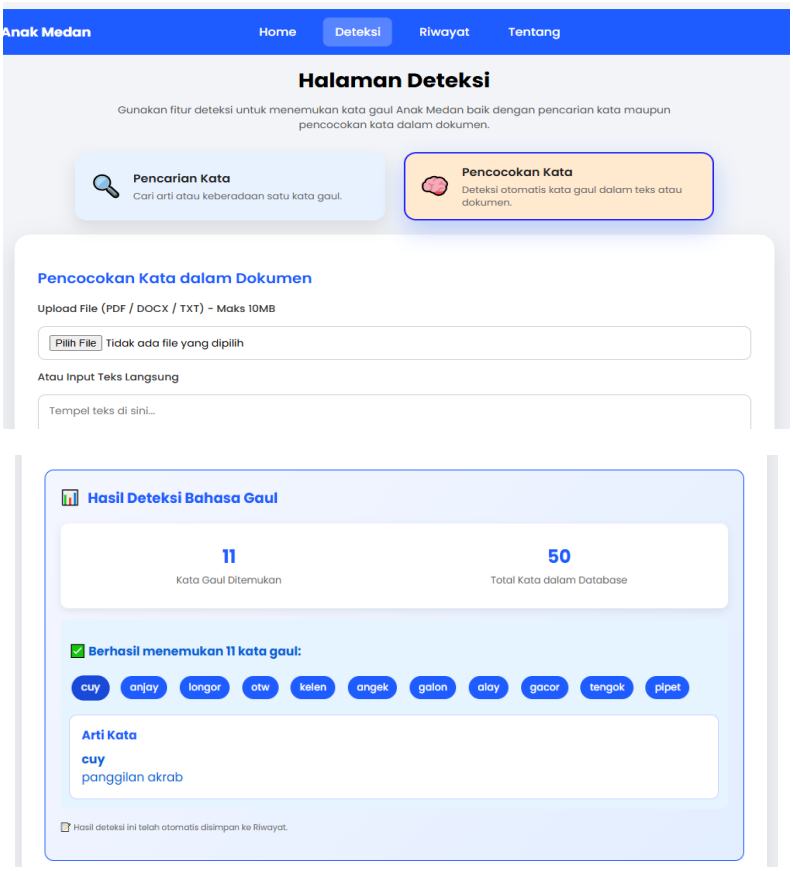

Figure 4 Slang word detection and meaning page

The slang detection page is the core component of the proposed system. On this page, users can either enter text manually or upload a text file containing sentences to be analyzed. The system processes the input text using the Hidden Markov Model and Viterbi Algorithm to determine the most probable linguistic category of each word. The detection results are displayed in tabular form, showing the identified slang words along with their corresponding classifications. This feature enables users to automatically recognize Medan slang expressions within textual data and demonstrates the practical implementation of probabilistic sequence modeling in Natural Language Processing applications.

CONCLUSION

This study successfully implemented a Hidden Markov Model (HMM) combined with the Viterbi Algorithm for detecting Medan slang language in textual data. The proposed approach was integrated into a web-based application capable of processing input text and automatically classifying words into standard language and slang language categories.

The implementation process involved data collection, dataset labeling, probability estimation using Hidden Markov Model parameters, and sequence classification using the Viterbi Algorithm. The developed system was able to identify Medan slang expressions based on contextual word sequences and provide classification results automatically.

Performance evaluation demonstrated that the proposed method effectively distinguishes slang words from standard Indonesian words. The evaluation results indicate that the HMM-Viterbi approach is suitable for regional slang language detection and can be applied as a Natural Language Processing solution for local language analysis.

Future research may focus on expanding the dataset size, incorporating additional regional slang variations, and comparing the proposed approach with modern machine learning or deep learning techniques to improve classification performance and scalability.

REFERENCES

Adelina Sitanggang, Bertania Permata Sari, Fahrysyah Aulya, Mira Cahya Hendra, & Kurnia Pulungan. (2024). Pengaruh Penggunaan Bahasa Gaul Dikalangan Siswa Terhadap Penggunaan Bahasa Indonesia Di SMK Negeri 10 Medan. Lencana: Jurnal Inovasi Ilmu Pendidikan, 2(3), 238–244. https://doi.org/10.55606/lencana.v2i3.3787

Afriansyah, A., & Syaripudin, A. (2022). Perancangan Sistem Informasi Absensi Dewan Guru Tenaga Harian Lepas Berbasis Web Pada Sekolah Dasar Negeri Kunciran 6 Kota Tangerang. Biner: Jurnal Ilmiah Informatika Dan Komputer,1(1),17–25. https://pdfs.semanticscholar.org/82a5/e7ac47f769e004c752ca456dac973984e751.pdf

Amin, R. M., Fakhriza, M., & Sinaga, I. A. (2024). SISTEM PENJADWALAN RAPAT BERBASIS SITUS WEB DI SEKRETARIAT DPRD KOTA BINJAI. JTIK (Jurnal Teknik Informatika Kaputama), 8(1), 39–48.

https://scholar.archive.org/work/vkd5we6bgrasbnqlf6bbjdqvfm/access/wayback/https://jurnal.kaputama.ac.id/index.php/JTIK/article/download/492/409

Anggini, N., Afifah, N. Y., & Syaputra, E. (2022). Pengaruh Bahasa Gaul (SLANG) Terhadap Bahasa Indonesia Pada Generasi Muda. Jurnal Multidisiplin Dehasen (MUDE), 1(3), 143–148. https://jurnal.unived.ac.id/index.php/mude/article/view/2477

Arianti, T., Fa'izi, A., Adam, S., & Wulandari, M. (2022). Perancangan sistem informasi perpustakaan menggunakan diagram UML (Unified Modelling Language). Jurnal Ilmiah Komputer Terapan Dan Informasi, 1(1), 19–25. https://download.garuda.kemdikbud.go.id/article.php?article=3053203&val=27786&title=PERANCANGAN

Dewi, N., Prasetya, A., & Sitanggang, A. R. U. (2023). Analisis Faktor Yang Mempengaruhi Remaja Indonesia Memakai Bahasa Gaul. Indonesian Journal of Language, Literature, and Education Language, Literature, 1(1), 1–6. https://jurnal.academiacenter.org/index.php/IJOL/article/view/26

Elza L. Saragih, Samuel Banjarnahor, Kartina Siagian, & Lasro Jumanta Sihite. (2025). Bahasa Gaul di Kalangan Remaja Kota Medan, Tantangan dalam Pembinaan Bahasa Baku. Jurnal Yudistira : Publikasi Riset Ilmu Pendidikan dan Bahasa, 3(4), 41–49. https://doi.org/10.61132/yudistira.v3i4.2309

Emira Hayatina Ramadhan, Selvina Maharani, & Dona Aji Karunia Putra. (2024). Analisis Pola Bunyi dalam Bahasa Gaul di Media Sosial. Tabasa: Jurnal Bahasa, Sastra Indonesia, dan Pengajarannya, 4(01), 13–25. https://doi.org/10.22515/tabasa.v4i01.7236

Fauzi, J. R. (2020). Algoritma dan flowchart dalam menyelesaikan suatu masalah. J. Tek. Inform, 3(2), 12. https://www.academia.edu/download/65240641/Makalah_20330044_Johan_Reza_Fauzi.pdf

Firmansyah, I., Adikara, P. P., & Adinugroho, S. (2021). Klasifikasi kelas kata (part-of-speech tagging) untuk bahasa madura menggunakan algoritme viterbi. Jurnal Teknologi Informasi Dan Ilmu Komputer, 8(5), 1039–1048. https://jtiik.ub.ac.id/index.php/jtiik/article/view/4483

Hendra, H., & Riti, Y. F. (2023). Perancangan dan implementasi website dengan konsep UI/UX untuk mengoptimalkan marketing perusahaan. Jurnal Informatika Dan Teknik Elektro Terapan, 11(3s1). https://journal.eng.unila.ac.id/index.php/jitet/article/view/3430

Irfani, A., Amelia, R., & Saptanti, D. (2006). Algoritma Viterbi dalam Metode Hidden Markov Models pada Teknologi Speech Recognition. Laboratorium Ilmu Dan Rekayasa Komputasi. Departemen Teknik Informatika, Institut Teknologi Bandung.

Iswatiningsih, D., Fauzan, & Pangesti, F. (2021). Ekspresi remaja milenial melalui penggunaan bahasa gaul di media sosial. KEMBARA Journal of Scientific Language Literature and Teaching, 7(2), 476–489. https://doi.org/10.22219/kembara.v7i2.18301

Jollyta, D., Oktarina, D., & Johan, J. (2020). Tinjauan Kasus Model Speech Recognition: Hidden Markov Model. JEPIN (Jurnal Edukasi Dan Penelitian Informatika), 6(2), 202–209. https://jurnal.untan.ac.id/index.php/jepin/article/download/39231/75676586693

Juanda, R., & Yadi, I. Z. (2020). Penerapan Rule Based Dengan Algoritma Viterbi Untuk Deteksi Kesalahan Huruf Kapital Pada Karya Ilmiah. Journal of Computer and Information Systems Ampera, 1(1), 53–62. https://doi.org/10.51519/journalcisa.v1i1.5

Manurian, W., Mubarok, I., Agustin, A. S., & Sania, N. (2020). Perancangan Sistem Informasi Pencatatan Poin Pelanggaran Tata Tertib Siswa Berbasis Website Pada SMK YP Karya 1 Tangerang. Jurnal Gerbang STMIK Bani Saleh, 10(1).

http://download.garuda.kemdikbud.go.id/article.php?article=1735070&val=15069&title=Perancangan%20Sistem%20Informasi%20Pencatatan%20Poin%20Pelanggaran%20Tata%20Tertib%20Siswa%20Berbasis%20Website%20Pada%20Smk%20YP%20Karya%201%20Tangerang

Meiyanti, R. (2020). PENDETEKSIAN KEBOHONGAN MENGGUNAKAN ALGORITMA VITERBI DAN TRANSFORMASI HANKEL PADA SUARA SECARA REAL TIME. 2.