Fake News Detection Using Textual Features: A Performance Comparison of Feature Selection Approaches and Machine Learning Models

Jhune Eleazar P. Camatura
Department of Computer Science, College of Computing and Information Sciences, Caraga State University

Shane D. Cale
Department of Computer Science, College of Computing and Information Sciences, Caraga State University

Takeyasu V. Nakazato
Department of Computer Science, College of Computing and Information Sciences, Caraga State University : Associate Professor

Regien B. Nakazato
Department of Computer Science, College of Computing and Information Sciences, Caraga State University : Assistant Professor I

https://doi.org/10.5109/7395551

出版情報：Proceedings of International Exchange and Innovation Conference on Engineering & Sciences (IEICES). 11, pp.421-426, 2025-10-30. International Exchange and Innovation Conference on Engineering & Sciences
バージョン：

Fake News Detection Using Textual Features: A Performance Comparison of Feature Selection Approaches and Machine Learning Models

Jhune Eleazar P. Camatura[1], Shane D. Cale[1], Takeyasu V. Nakazato[2], Regien B. Nakazato[3]
[1]Student, Department of Computer Science, College of Computing and Information Sciences, Caraga State University, Butuan City, Philippines
[2]Associate Professor, Department of Computer Science, College of Computing and Information Sciences, Caraga State University, Butuan City, Philippines
[3]Assistant Professor I, Department of Computer Science, College of Computing and Information Sciences, Caraga State University, Butuan City, Philippines

Corresponding author email: rbnakazato@carsu.edu.ph

Abstract: *The proliferation of fake news on digital platforms necessitates robust detection systems. This study investigates the effectiveness of three feature selection methods Manual selection, Genetic Algorithm (GA), and SelectKBest in enhancing fake news detection models using purely textual features. A comprehensive literature review identified key linguistic features, while automated methods optimized feature subsets for improved performance. The dataset comprised 6,635 news articles, preprocessed and evaluated using Random Forest, SVM, Logistic Regression, and Neural Networks. Results demonstrated that automated feature selection, particularly GA, significantly outperformed manual methods, achieving 80.4% accuracy with Neural Networks compared to 72.1% with manual selection. Entity count emerged as a pivotal feature across all selection methods, providing strong classification signals with up to 68.4% accuracy even as a standalone predictor. The findings highlight the superiority of data-driven feature selection approaches in improving detection accuracy, offering practical insights for developing more effective systems to combat misinformation.*

Keywords: Fake news detection; Feature selection; Genetic Algorithm; SelectKBest; Textual features; Machine learning

1. INTRODUCTION

The proliferation of digital media has profoundly transformed the way information is produced, shared, and consumed. Social media platforms have become central to news dissemination, offering instant access to vast amounts of content globally. While this accessibility offers benefits, it has also facilitated the rapid spread of "fake news" false or misleading information masquerading as factual news. The spread of such misinformation poses serious threats to public trust, political stability, and the integrity of information systems [1]. Misinformation can manipulate public opinion, fuel social unrest, and undermine confidence in credible news sources. In computer science, developing automated fake news detection systems has become an important research area, particularly through machine learning (ML) and natural language processing (NLP) techniques. A critical component in building these systems is feature engineering: the process of extracting and selecting relevant data attributes to improve model performance. This involves deriving meaningful variables from
raw text (feature extraction) and identifying the most predictive subset of these features (feature selection) [2]. Researchers have explored various linguistic features, including syntactic, grammatical, sentiment, and readability metrics, demonstrating their potential in identifying fake news [3]. A significant challenge in fake news detection is developing effective methodologies for selecting text-based features that accurately distinguish between real and fake news articles. This research focuses solely on linguistic and textual features extracted directly from news content, excluding external contextual information. It is often unclear which specific textual features contribute most significantly to this distinction, and there is a need to better understand the effectiveness of different feature selection methods both manual and automated in improving detection accuracy. This study aims to address these challenges by investigating and comparing feature engineering techniques for fake news detection. The primary objective is to identify the most impactful textual features and the most effective selection method to enhance overall model performance. Specifically, this research will: 1) identify commonly used linguistic and textual features from existing literature for manual selection; 2) implement automated feature selection methods, namely SelectKBest and Genetic Algorithm (GA), to identify a highly predictive subset of features; 3) perform a comparative analysis of these three feature selection methods (manual, SelectKBest, GA) using four machine learning models: Support Vector Machine (SVM), Logistic Regression Algorithm (LRA), Random Forest (RF), and a Neural Network (NN).

It is also important to highlight that machine learning models play a critical role in evaluating the effectiveness of selected features. These models, particularly Neural Network architectures, are capable of capturing complex patterns by leveraging multiple processing layers that learn hierarchical representations of data. This capability enhances the robustness and generalizability of the feature selection process [8].

The goal is to identify a minimal yet effective set of text-based features and the most reliable selection approach to improve the accuracy of fake news detection based on text content alone.

2. MATERIALS AND METHODS

The methodology employed in this study focuses on evaluating and comparing two distinct feature selection approaches for detecting fake news from text data. The first approach is manual feature selection, which involves conducting an extensive literature review to identify the most effective features reported in previous research studies. These features are selected based on their proven impact on classification accuracy and their relevance to fake news detection tasks. This method relies on human expertise and domain knowledge to inform the selection process. The second approach is automated feature selection, which uses computational techniques to identify the most significant features from the dataset. Specifically, we utilize algorithms such as SelectKBest and Genetic Algorithm (GA) to score and select features based on statistical relevance and optimization techniques. Following feature selection, machine learning models are developed and trained separately using the features derived from each method. The models are then evaluated and compared to determine which feature selection approach yields better performance in terms of accuracy. This comparative analysis aims to provide insights into the strengths and limitations of both manual and automated feature selection in the context of fake news detection.

2.1 Dataset

The dataset utilized in this study is the "Fake News Dataset" from Kaggle. It comprises 6,635 news articles, equally distributed between "Real" and "Fake" labels, with columns providing the number, title, text, and label for each article. This dataset was split into training and testing sets for model development and evaluation.

2.2 Data Preprocessing

The "Fake News Dataset" undergoes a two-phase transformation to prepare it for machine learning: feature extraction followed by numerical transformation. This process converts raw textual content into structured input while preserving key linguistic patterns.In the feature extraction phase, 20 linguistic and textual features are derived directly from the original news articles. These features aim to capture quantifiable patterns that help distinguish between real and fake news. Minimal text cleaning is applied to maintain the integrity of the content. The selection of features is based on their relevance to fake news detection as identified in prior studies, as well as their ability to ensure computational efficiency. Next, during numerical transformation, each extracted feature is systematically converted into a machine-readable format. Lexical features such as word count are stored as integers, sentiment-related ratios are represented as floating-point values, and categorical attributes like emotion or linguistic formality are encoded numerically. This enables machine learning algorithms to interpret and learn from the data effectively for classification tasks.

2.3 Feature Selection

Feature selection is a crucial phase in constructing an effective fake news detection model. It involves identifying the most relevant attributes in the dataset that significantly contribute to distinguishing between real and fake news. This study adopts a hybrid approach that integrates both manual and automated methods to ensure that the selected features are grounded in research and optimized for performance. Manual feature selection was performed through a thorough review of recent literature. The process was guided by clear criteria to avoid arbitrary decisions: recurrence across multiple studies, documented impact on model performance, publication recency (within the last seven years), and versatility across different models and datasets. This ensured that only features with strong empirical support were shortlisted for further evaluation. To refine and validate these manually selected features, and to potentially discover additional useful ones, automated feature selection techniques were applied. The first method, SelectKBest [4], uses statistical tests to evaluate the individual relevance of features. In our case, the ANOVA F-test was employed, which is well-suited for continuous language-based variables. Based on previous findings [5], ANOVA offered a good balance between interpretability and performance, achieving better results with Random Forest classifiers than chi-square or mutual information on similar tasks. The algorithm for SelectKBest is presented on Fig.1

```
FUNCTION SelectKBestFeatureSelection:
1. FOR each feature in feature_set:
        a. Compute ANOVA F-score (feature vs. target)
        b. Store normalized score
2. RANK features by normalized scores
3. SELECT features with score > threshold
4. VALIDATE via cross-validation
5. RETURN optimal feature subset
END FUNCTION
```

Fig. 1. Pseudocode for SelectKBest Feature Selection.

Another method we used is the Genetic Algorithm (GA). GA was selected as one of the algorithms for feature selection due to its proven reliability in estimating various property types and its successful application not only in textual domains but also in the estimation of diverse variables, such as the compression index [6]. It is a population-based evolutionary search technique for feature selection. In this approach, features are encoded as binary chromosomes indicating whether each feature is included (1) or excluded (0). The algorithm starts with a randomly initialized population of feature subsets. Each individual's fitness is evaluated through k-fold cross-validation using a Random Forest classifier, with fitness scores based on classification accuracy. Parents are selected via tournament selection, favoring higher fitness individuals. Offspring is generated by crossover, combining features from two parents, and mutation, which randomly flips bits to maintain genetic diversity and prevent premature convergence. Elitism preserves the best-performing individuals across generations [6]. This iterative process balances exploration and exploitation, enabling the algorithm to evolve feature subsets that maximize model performance while avoiding local optima. The algorithm for the GA is presented in Fig. 2.

```
FUNCTION GeneticAlgorithmFeatureSelection:

    Initialize population with random feature subsets

    WHILE termination criteria not met:
        Evaluate fitness via cross-validation
        Select parents via tournament selection
        Create offspring through crossover + mutation
        Repair offspring to maintain 5 features
```

Fig. 2 Pseudocode for Genetic Algorithm-based Feature Selection

2.4 Model Training and Evaluation

Four machine learning models were selected and trained using the feature subsets identified by each selection method: Support Vector Machine (SVM), Logistic Regression Algorithm (LRA), Random Forest (RF), and Neural Network (NN). These models were chosen for their diverse and complementary strengths in handling engineered features. To ensure robust and reproducible results, a consistent training and evaluation strategy was applied across all models. The dataset was split into 80% training and 20% test sets with stratification to preserve class distribution. Feature scaling was applied where appropriate (for LR, SVM, and NN), and balanced class weights were used to address potential class imbalance.Model performance was evaluated on the unseen test set using a combination of standard metrics suitable for binary classification tasks: accuracy, precision, recall, F1-score, and Area Under the Receiver Operating Characteristic Curve (AUC).

Accuracy: Measures the overall proportion of correctly classified news articles:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Precision: Indicates the proportion of true positive predictions out of all positive predictions, reflecting the model's ability to avoid false positives:

$$Precision = \frac{TP}{TP + FP}$$

Recall: Represents the proportion of true positive predictions out of all actual positive instances, measuring the model's ability to capture all positives:

$$Recall = \frac{TP}{TP + FN}$$

F1-score: The harmonic mean of precision and recall, balancing both metrics into a single performance measure:

$$F1 - Score = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

Area Under the ROC Curve (AUC) assesses the model's ability to discriminate between classes across different threshold settings, where 1.0 indicates perfect discrimination and 0.5 indicates random guessing.

2.5 Comparative Analysis

A comparative analysis was conducted to assess the performance of the four machine learning models when trained with features selected by the manual, Genetic Algorithm, and SelectKBest methods. This analysis aimed to identify the most effective feature selection strategy and model combination for fake news detection.

3. EVALUATION OF FEATURE SELECTION METHODS

This section presents the performance results of the machine learning models when trained with features selected by each of the three methods, as well as a baseline performance using the unfiltered feature set. To establish a baseline for comparison, all 20 extracted features were used to train each of the four machine learning models without applying any feature selection. The performance metrics for this baseline are summarized in Table 1.

Table 1: Categorization of Extracted Features for Fake News

Category	Features
Lexical	word_count, punctuation_count, caps_ratio, entity_count, verb_count, adj_count, exclamation_ratio, passive_voice_ratio, sentence_count.
Readability & Style	readability, formality_score, subjectivity_score, flesch_kincaid_grade
Emotion & Sentiment	dominant_emotion, emotional_intensity, sentiment_score, pos_sentiment, neg_sentiment, neu_sentiment, sentiment_numeric

Table 2: Performance of All Models Using All 20 Features

Model	Accuracy	Precision	Recall	F1-Score	AUC
RF	81.8	82.0	81.8	82.0	90.1
SVM	81.5	81.0	83.0	81.8	90.1
LRA	76.3	74.4	81.7	77.7	84.4
NN	84.1	84.0	84.8	84.4	91.1.

3.1 Evaluation of Feature Selection Methods

This section presents the performance results of the machine learning models when trained with features selected by each of the three methods, as well as a baseline performance using the unfiltered feature set. To establish a baseline for comparison, all 20 extracted features were used to train each of the four machine learning models without applying any feature selection. The performance metrics for this baseline are summarized in Table 2.

The baseline results indicate the inherent discriminative power of the full set of extracted textual features before any optimization.

The manual feature selection process, guided by a literature review, identified five features: readability,

entity count, word count, dominant emotion, and sentiment numeric. The performance of the models when trained on this manually selected feature set is presented in Table 3.

Table 3: Performance Results of Manual Feature Selection

Model	Accuracy	Precision	Recall	F1-Score	AUC
RF	0.74	0.73	0.77	0.75	0.82
SVM	0.72	0.73	070	0.71	0.78
LRA	0.64	0.64	0.65	0.65	0.70
NN	0.72	0.71	0.76	0.73	0.78

The results show that models trained on the manually selected features achieved varying levels of performance, providing a point of comparison for the automated methods.

The Genetic Algorithm (GA) identified a subset of five features: entity_count, word_count, caps_ratio, passive_voice_ratio, and flesch_kincaid_grade. The performance of the models using this GA-selected feature set is shown in Fig. 3 and Table 4.

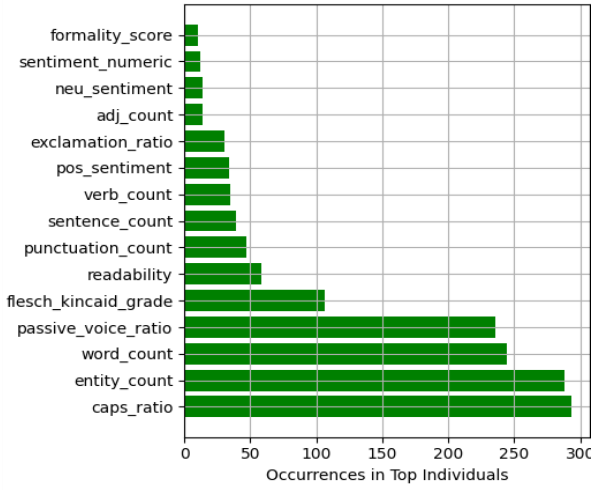

Table 4: Performance Results of Genetic Algorithm Feature Selection

Model	Accuracy	Precision	Recall	F1-Score	AUC
RF	80.4	80.5	80.7	80.6	88.7
SVM	80.0	78.8	82.0	80.4	87.6
LRA	73.6	70.5	82.0	75.8	81.0
NN	80.4	80.2	84.2	81.3	87.7

Fig. 3. GA Top 15 Most Frequent Feature

The GA-selected features generally led to improved performance compared to manual selection, highlighting the algorithm's ability to identify effective feature combinations.

The SelectKBest method, using statistical tests (ANOVA F-test), identified the following five features: dominant_emotion, passive_voice_ratio, exclamation_ratio, caps_ratio, and entity_count. The performance of the models with these features is summarized in Fig. 4 and Table 5.

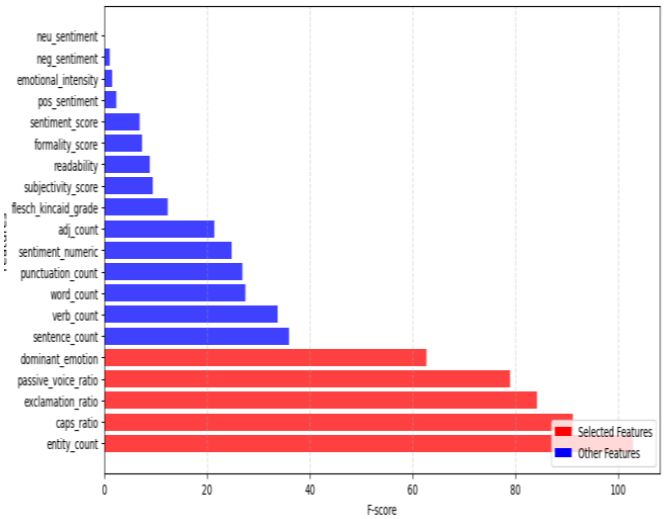

Fig. 4. Feature Scores (F-values)

Table 5: Performance Results of SelectKBest Feature Selection

Model	Accuracy	Precision	Recall	F1-Score	AUC
RF	76.1	74.6	79.8	77.1	85.0
SVM	75.7	74.1	80.0	76.8	84.1
LRA	70.1	67.3	81.0	74.0	78.6
NN	76.2	74.1	78.7	76.3	81.8

SelectKBest also demonstrated an improvement over manual selection, indicating the value of a data-driven statistical approach to feature ranking.

3.2 Venn Diagram of Selected Features

Venn diagram was used to visualize the common and unique features selected by the three feature selection methods. Features found in the intersection of methods are considered strong indicators for classification. To evaluate their effectiveness, a model was trained using only these shared features. This helps determine whether features commonly selected by all methods can still produce good performance with fewer inputs.

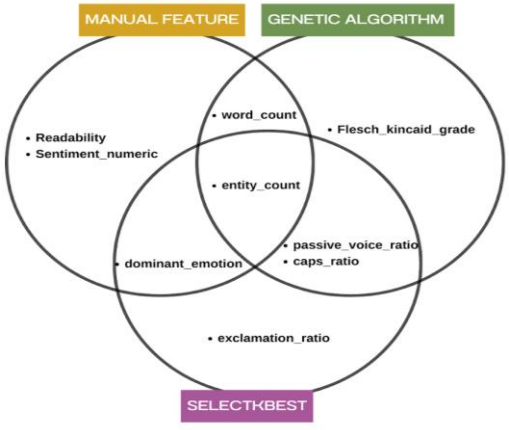

Fig. 5. Venn diagram of selected features across methods

The Venn diagram shown in Fig. 5 revealed that entity_count was the only feature consistently selected by all three methods. This suggests its strong relevance for fake news detection across different selection strategies. Given its consistent selection, the performance of the models was evaluated using entity_count as the sole feature.

Table 6: Performance Comparison of Models Using Entity Count Feature

Model	Accuracy	Precision	Recall	F1-Score	AUC
RF	68.4	64.0	70.6	67.1	71.6
SVM	67.3	66.3	71.7	69.0	71.2
LRA	65.2	68.7	57.1	62.4	69.2
NN	67.6	65.7	75.0	70.0	70.6

Table 6 shows that using entity count as the sole feature achieves accuracies between 65.2% and 68.4%, with corresponding F1-scores from 62.4% to 70.0%. These results, well above the 50% random-guessing baseline, demonstrate that the number of named entities provides a robust signal for distinguishing fake from real news. As a standalone predictor, entity count delivers solid performance, with Neural Networks achieving the highest F1-score (70.0%) and Random Forest attaining peak accuracy (68.4%). This underscores its value as a pivotal feature in fake news detection.

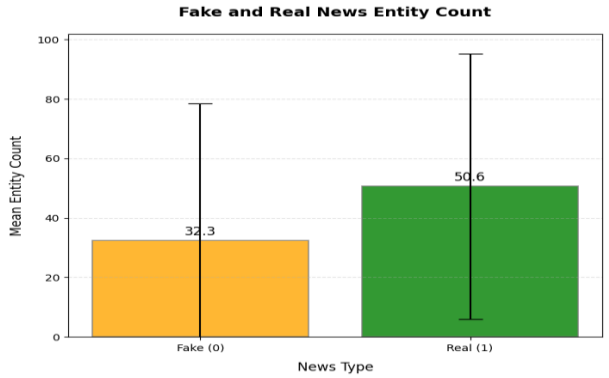

Fig. 6. Mean Entity Count in Fake vs. Real News Articles

These results shown in Fig. 6 indicate that entity count is a reliable feature for separating real from fake news with classification accuracy ranging from 65–68%. This aligns with findings by Shahi et al. [7], who observed that real news stories tend to contain more named entities, reflecting greater use of references and factual information. In contrast, fake news stories often lack such references, consistent with the structural differences identified in this study. The 18.3-entity gap observed across the dataset further underscores the robustness of entity count as a distinguishing feature, demonstrating its effectiveness in identifying fake news.

3.3 Comparison Analysis of Feature Selection Methods

Table 7: Average Model Performance Across Manual and Automated Feature Selection Methods

Model	Accuracy	Precision	Recall	F1-Score	AUC
Manual method	70.5	70.2	72.2	71.1	77.7
Genetic Algorithm	78.6	77.5	82.2	79.5	86.3
SelecttKBest	74.5	72.5	80	76.1	82.4

Among the feature selection methods evaluated shown in Table 7, the Genetic Algorithm (GA) clearly demonstrated superior performance, achieving an average accuracy of 78.6%, which surpasses manual selection by 8.1 percentage points. GA's strong recall (82.2%) and AUC (86.3%) underscore its ability to effectively detect fake news cases while maintaining overall classification quality. SelectKBest emerged as a balanced middle ground, delivering 74.5% accuracy and 80.0% recall. Though less computationally demanding than GA, it focuses on identifying individual high-impact features but lacks the capacity to capture complex feature interactions that GA's evolutionary process exploits. While manual selection showed lower accuracy (70.5%) and AUC (77.7%), it retains value for its transparency and grounding in domain expertise. Notably, features such as entity count consistently ranked highly across all approaches, highlighting their foundational importance in fake news detection. GA's dynamic optimization via evolutionary operations enables it to uncover synergistic feature combinations, leading to superior model performance and particularly strong recall critical for reducing false negatives in fake news classification. From an efficiency perspective, GA substantially reduces feature dimensionality with minimal accuracy loss. For example, a Neural Network trained on only five GA-selected features retained 81.2% accuracy, preserving 96.5% of the full 20-feature model's performance while reducing the feature space by 75%. This, alongside stable holdout test results, confirms GA's practical advantage in delivering robust, efficient fake news detection models.

3.4 Runtime Analysis

To evaluate the computational efficiency of the feature selection methods used in this study, we measured the runtime performance of both the Genetic Algorithm and SelectKBest approaches. These evaluations were conducted on a standard mid-range system equipped with an AMD Ryzen 5 5500U processor and 16GB of RAM. The goal was to assess not only how long each method takes to execute but also how efficiently they handle operations such as generations (in GA) or feature scoring (in SelectKBest). Table 8 summarizes the total execution time and average processing time per unit operation for both methods, highlighting the trade-offs between speed and complexity in their respective algorithms.

Table 8: Feature Selection Runtime Comparison

Method	**Total Runtime**	**Average Time per Generation/Operation**

Genetic Algorithm	13.01 mins	45.91 seconds average time per generation
SelectKbest	12.45 sec	0.62 seconds average tme per feature

4. CONCLUSION

This study conclusively demonstrates that automated feature selection methods significantly enhance fake news detection performance compared to manual selection, achieving higher accuracy, recall, and AUC with a reduced feature set. The Genetic Algorithm, in particular, proved most effective in identifying optimal feature combinations, resulting in superior model performance, especially for Neural Networks and Random Forests. The consistent significance of the entity count feature across all selection methods highlights its robustness as a key indicator for distinguishing fake from real news, even providing a notable signal as a standalone predictor. These findings underscore the critical role of data-driven feature optimization in developing accurate and efficient misinformation detection systems.

5. REFERENCES

[1] Shu, K., Sliva, A.L., Wang, S., Tang, J., & Liu, H. (2017). Fake News Detection on Social Media: A Data Mining Perspective. ArXiv, abs/1708.01967. [2] Patel, H. (2024, April 29). Feature engineering explained. Built-in.https://builtin.com/articles/feature-engineerig

[2] Choudhary, A., & Arora, A. (2021). Linguistic feature-based learning model for fake news detection and classification. Journal of Computational Science, 43,101104.https://www.sciencedirect.com/science/article/abs/pii/S095741742030909X

[3] Kavya, D. (2023). Optimizing performance: SelectKBest for efficient feature selection in machine learning.Medium.https://medium.com/@Kavya2099/optimizing-performance-selectkbest-for-efficient-feature-selection-in-machine-learning-3b635905ed48

[4] Tariq, M. A. (2024). A study on comparative analysis of feature selection algorithms for students grades prediction. Journal of Intelligent & Fuzzy Systems, 48(1). [5] Gómez, F. (2023). Genetic algorithms for feature selection in machine learning. NeuralDesigner.https://www.neuraldesigner.com/blog/genetic_algorithms_for_feature_selection/

[5] Shahi, G. K., Seneviratne, O., & Spaniol, M. (2025). SemCAFE: When Named Entities make the Difference – Assessing Web Source Reliability through Entity-level Analytics. Proceedings of the 17th ACM Web Science Conference 2025. https://doi.org/10.1145/3717867.3717880

[6] S. Co, J. Du, H. Ong, M. Uy Ching, Estimation of Compression Index Using Genetic Algorithm Approach, Proceedings of International Exchange and Innovation Conference on Engineering & Sciences (IEICES), vol. 10, pp. 379-384, 2024, Interdisciplinary Graduate School of Engineering Sciences, Kyushu University.

[7] S. Muhammad, Detection of Slums from Very High-Resolution Satellite Images Using Machine Learning Algorithms: A Case Study of Fustat Area in Cairo, Egypt, Proceedings of International Exchange and Innovation Conference on Engineering & Sciences (IEICES), vol. 6, pp. 219-224, 2022, Interdisciplinary Graduate School of Engineering Sciences, Kyushu University.