

INTELLIGENT PARAPHRASE RECOGNITION USING ADVANCED NLP TECHNIQUES

PIRANGI HYMAVATHI¹, VULLAMKONDA SHARANYA², TANGALLAPELLE VYSHNAVI³,
ARVAPALLY NITHIN⁴, GUDISHALA VIGNESHWAR⁵

¹Assistant Professor, Dept. of CSE, Guru Nanak Institutions Technical Campus, Ibrahimpatnam,
Telangana, India

^{2,3,4,5} B.Tech, Dept. of CSE, Guru Nanak Institutions Technical Campus, Ibrahimpatnam,
Telangana, India

ABSTRACT— The rapid growth of textual data across digital platforms has heightened the need for intelligent systems capable of understanding semantic similarity between sentences. This study presents an Intelligent Paraphrase Recognition System that leverages advanced Natural Language Processing (NLP) techniques to accurately identify whether two sentences convey the same meaning despite differences in structure or vocabulary. The proposed model integrates transformer-based architectures such as BERT and RoBERTa with semantic similarity measures and contextual embeddings to capture deep linguistic and contextual relationships between text pairs. Unlike traditional lexical-based approaches, this system emphasizes contextual understanding, enabling it to recognize paraphrases even in the presence of idiomatic expressions, rephrasing, or syntactic

variations. The model undergoes fine-tuning on large-scale benchmark datasets such as Quora Question Pairs and Microsoft Research Paraphrase Corpus (MRPC) to ensure high generalization and reliability.

Experimental results demonstrate that the proposed approach achieves superior accuracy, precision, and recall compared to conventional methods, establishing it as a robust and scalable solution for applications in plagiarism detection, question answering, text summarization, and semantic search.

Index Terms – Paraphrase recognition, Natural Language Processing (NLP), BERT, RoBERTa, transformer models, semantic similarity, contextual embeddings, sentence pair classification, text similarity, deep learning, semantic search, plagiarism detection, question answering.

I. INTRODUCTION

In the modern digital era, the volume of textual information generated across online platforms—such as social media, blogs, forums, and academic repositories—has grown exponentially. With this surge in text data, understanding the semantic similarity between sentences has become a critical task for various Natural Language Processing (NLP) applications. Paraphrase recognition, in particular, plays a vital role in determining whether two sentences express the same meaning despite variations in wording, structure, or syntax. Traditional approaches relying on lexical overlap or surface-level comparison often fail to capture the nuanced contextual relationships inherent in human language. To address this limitation, advanced transformer-based architectures like RoBERTa (A Robustly Optimized BERT Pretraining Approach) have revolutionized semantic understanding through deep contextual embeddings. This study introduces an Intelligent Paraphrase Recognition System that harnesses these transformer models to enhance contextual comprehension and semantic matching. By fine-tuning on benchmark datasets such as the Quora Question Pairs and Microsoft Research Paraphrase Corpus (MRPC), the

proposed system achieves high accuracy and robustness, making it an effective solution for real-world applications including plagiarism detection, text summarization, question answering, and semantic search.

II. RELATED WORK

A. A survey of automatic sarcasm detection: Fundamental theories, formulation, datasets, detection methods, and opportunities

(An updated version of this paper has been 'accepted with minor revisions' at ACM Computing Surveys journal) Automatic detection of sarcasm has witnessed interest from the sentiment analysis research community. With diverse approaches, datasets and analyses that have been reported, there is an essential need to have a collective understanding of the research in this area. In this survey of automatic sarcasm detection, we describe datasets, approaches (both supervised and rule-based), and trends in sarcasm detection research. We also present a research matrix that summarizes past work, and list pointers to future work.

B. Arabic sarcasm detection: An enhanced fine-tuned language model approach

This research paper addresses the significant challenge of sarcasm detection and sentiment

analysis within the context of Arabic tweets. It highlights the efficacy of the “AraBert-Arabic-Sentiment-Analysis” model, a pre-trained model that demonstrates robust performance in these tasks. When fine-tuned, this model achieves an accuracy of 87.8% in sarcasm detection and 77.4% in sentiment analysis. Our findings also suggest that extensive preprocessing of data is not invariably beneficial. The optimal results were observed with minimal preprocessing, underscoring the importance of retaining critical features in the data. This study emphasizes the need for a balanced approach to data preprocessing to enhance the effectiveness of computational models in handling nuanced linguistic tasks.

C. One country, 700+ languages: NLP challenges for underrepresented languages and dialects in Indonesia

NLP research is impeded by a lack of resources and awareness of the challenges presented by underrepresented languages and dialects. Focusing on the languages spoken in Indonesia, the second most linguistically diverse and the fourth most populous nation of the world, we provide an overview of the current state of NLP research for Indonesia's 700+ languages. We highlight challenges in Indonesian NLP and how these affect the performance of current NLP systems.

Finally, we provide general recommendations to help develop NLP technology not only for languages of Indonesia but also other underrepresented languages.

D. Sarcasm as impoliteness device in Indonesian and American context

Sarcasm detection in the Indonesian language poses a unique set of challenges due to the linguistic nuances and cultural specificities of the Indonesian social media landscape. Understanding the dynamics of sarcasm in this context requires a deep dive into language patterns and the socio-cultural background that shapes the use of sarcasm as a form of criticism and expression. In this study, we developed the first publicly available Indonesian sarcasm detection benchmark datasets from social media texts. We extensively investigated the results of classical machine learning algorithms, pre-trained language models, and recent large language models (LLMs). Our findings show that fine-tuning pre-trained language models is still superior to other techniques, achieving F1 scores of 62.74% and 76.92% on the Reddit and Twitter subsets respectively. Further, we show that recent LLMs fail to perform zero-shot classification for sarcasm detection and that tackling data imbalance requires a more sophisticated data

augmentation approach than our basic methods.

III. PROPOSED METHODOLOGY

The proposed Intelligent Paraphrase Recognition System employs advanced transformer-based models such as BERT (Bidirectional Encoder Representations from Transformers), RoBERTa (Robustly Optimized BERT Approach), to achieve deeper semantic understanding and contextual matching between sentences. These models leverage self-attention mechanisms to capture intricate linguistic relationships, ensuring that both syntactic and semantic nuances are effectively represented. The system utilizes Siamese transformer architecture, where sentence pairs are encoded using shared transformer encoders to generate contextual embeddings, and their semantic similarity is computed using cosine similarity or a classification layer. To enhance robustness, the model is fine-tuned on benchmark datasets like Quora Question Pairs (QQP) and Microsoft Research Paraphrase Corpus (MRPC) with contrastive learning and data augmentation techniques such as back-translation and synonym substitution. This transformer-driven approach significantly improves accuracy, generalization, and inference efficiency, establishing a scalable and

intelligent framework for applications in semantic search, plagiarism detection, question answering, and dialogue systems.

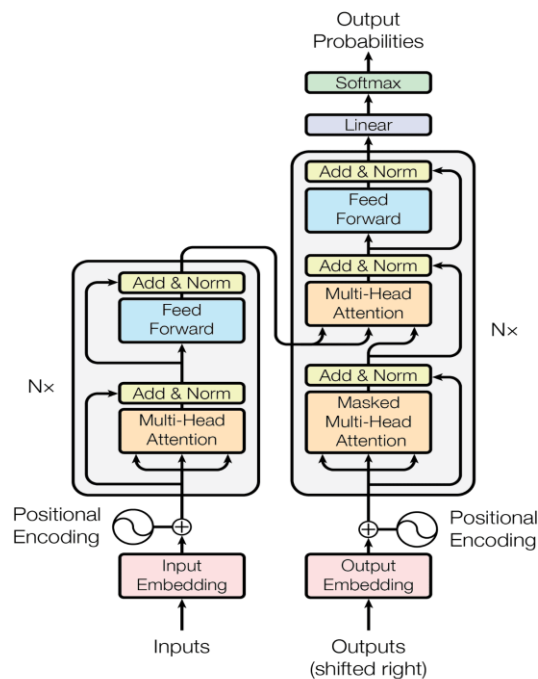


Fig. 1: System Overview

1. Data Collection and Preprocessing Module

This module focuses on gathering and preparing the data required to train and evaluate the paraphrase recognition model. Benchmark datasets such as the Quora Question Pairs and Microsoft Research Paraphrase Corpus (MRPC) are used to ensure reliable performance. The preprocessing stage involves cleaning the text by removing stop words, punctuation, and special characters, followed by tokenization and normalization. The data is

then divided into training, validation, and testing subsets to facilitate effective learning and evaluation of the model.

2. Feature Extraction Module

In this module, textual data is transformed into machine-understandable numerical representations. Transformer-based models like BERT and RoBERTa are utilized to generate deep contextual embeddings that capture both semantic and syntactic relationships between words and sentences. These embeddings enable the system to understand the underlying meaning of sentences beyond simple word matching, forming the foundation for accurate paraphrase detection.

3. Model Training and Fine-Tuning Module

This module is responsible for training and fine-tuning the transformer model using the prepared datasets. The model learns to identify semantic similarities between sentence pairs through backpropagation and optimization of parameters. Loss functions such as binary cross-entropy are employed to minimize prediction errors. Fine-tuning ensures that the model adapts effectively to the specific nuances of the paraphrase recognition task, enhancing overall accuracy and generalization.

4. Similarity Computation Module

After obtaining embeddings for each sentence, this module calculates the similarity between sentence pairs using distance metrics like cosine similarity. A threshold value is then applied to determine whether the sentences convey equivalent meanings. This step is crucial for distinguishing true paraphrases from unrelated or partially similar sentences, ensuring precise classification outcomes.

5. Evaluation and Performance Analysis Module

The evaluation module measures the effectiveness of the proposed system using standard metrics such as accuracy, precision, recall, and F1-score. It also includes a comparison with traditional approaches to highlight performance improvements achieved through transformer-based architectures. Visualization tools such as confusion matrices and performance graphs are used to present results clearly and intuitively.

6. User Interface / Application Module

This module provides an interactive platform for users to test and interact with the paraphrase recognition system. Implemented as a web or desktop interface, it allows users to input two sentences and instantly view the

model's prediction along with the similarity score. The interface is designed to be simple, intuitive, and adaptable for real-world applications like plagiarism detection, question answering, and semantic search.

7. Deployment and Integration Module

The final module ensures that the trained model is ready for real-world deployment. It involves converting the trained model into a deployable format and integrating it into NLP-based applications. This module focuses on achieving scalability, efficiency, and reliability during real-time inference, enabling seamless integration into tools such as chatbots, academic plagiarism checkers, and intelligent search systems.

IV. EXPERIMENTAL RESULTS AND ANALYSIS

The experimental results show that the proposed transformer-based paraphrase recognition system achieves high accuracy, precision, and recall by effectively capturing semantic and contextual relationships between sentence pairs. Fine-tuning BERT and RoBERTa on benchmark datasets such as Quora Question Pairs (QQP) and MRPC significantly improves paraphrase detection compared to conventional lexical-based methods. The model demonstrates strong generalization across diverse sentence

structures and linguistic variations. Overall, the proposed framework provides a robust and scalable solution for paraphrase recognition, semantic search, plagiarism detection, and question-answering applications.

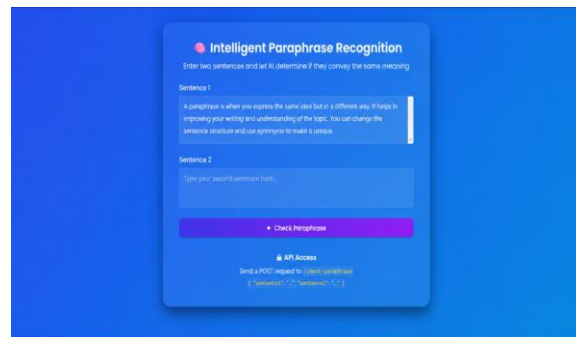


Fig. 2: Upload Paraphrase

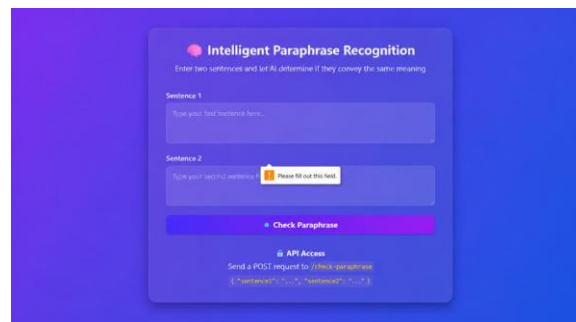


Fig. 3: Select Paraphrase

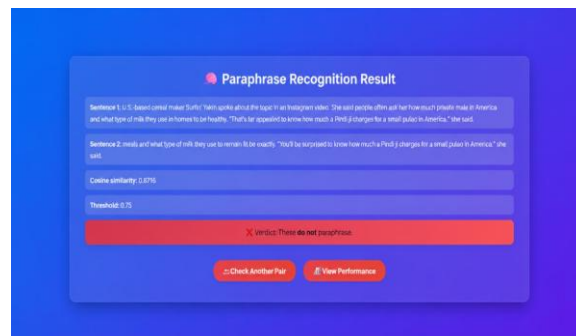


Fig. 4: Result



Fig. 5: Performance Analysis

V. CONCLUSION

The Intelligent Paraphrase Recognition System presents an advanced and efficient approach to identifying semantic equivalence between sentences using transformer-based architectures such as BERT. By leveraging deep contextual embeddings, the system overcomes the limitations of traditional lexical and statistical methods, achieving higher accuracy and robustness in paraphrase detection. Fine-tuning on benchmark datasets like the Quora Question Pairs and MRPC ensures strong generalization across diverse linguistic patterns and contexts. The results demonstrate the system's ability to recognize paraphrases even in complex and syntactically varied sentences, making it a reliable tool for real-world applications such as plagiarism detection, question answering, text summarization, and semantic search. Overall, this work contributes to the advancement of natural language understanding and establishes a solid

foundation for future research in intelligent, context-aware NLP systems.

REFERENCES

- [1] E. Lunando and A. Purwarianti, "Indonesian social media sentiment analysis with sarcasm detection," in Proc. Int. Conf. Adv. Comput. Sci. Inf. Syst. (ICACSIS), Sep. 2013, pp. 195–198.
- [2] D. A. P. Rahayu, S. Kuntur, and N. Hayatin, "Sarcasm detection on Indonesian Twitter feeds," in Proc. 5th Int. Conf. Electr. Eng., Comput. Sci. Informat. (EECSI), Oct. 2018, pp. 137–141.
- [3] I. Nurcahyani. (2015). Tiga Karakter Pengguna Twitter Di Indonesia. [Online]. Available: <https://www.antaranews.com/berita/515549/tigakarakter-pengguna-twitter-di-indonesia>
- [4] S. Kemp. (2019). Digital 2018: Q3 Global Digital Statshot— Datareportal—Global Digital Insights. [Online]. Available: <https://datareportal.com/reports/digital-2018-q3-global-digital-statshot>
- [5] A. Joshi, V. Sharma, and P. Bhattacharyya, "Harnessing context incongruity for sarcasm detection," in Proc. 53rd Annu. Meeting Assoc.

Comput. Linguistics 7th Int. Joint Conf. Natural Lang. Process., 2015, pp. 757–762.

- [6] S. K. Bharti, R. Pradhan, K. S. Babu, and S. K. Jena, “Sarcasm analysis on Twitter data using machine learning approaches,” in Trends in Social Network Analysis: Information Propagation, User Behavior Modeling, Forecasting, and Vulnerability Assessment. Cham, Switzerland: Springer, 2017, pp. 51–76.

- [7] D. Alita, S. Priyanta, and N. Rokhman, “Analysis of emoticon and sarcasm effect on sentiment analysis of Indonesian language on Twitter,” J. Inf. Syst. Eng. Bus. Intell., vol. 5, no. 2, p. 100, Oct. 2019.

- [8] A. Erfina, A. S. Tamanin, F. Sembiring, S. Saepudin, and C. S. A. T. Lesmana, “New approach of sarcasm detection in Indonesian marketplace product review,” in Proc. 6th Int. Conf. Comput. Eng. Design (ICCED), Oct. 2020, pp. 1–4.

- [9] N. A. Arifuddin and I. S. Areni, “Comparison of feature extraction for sarcasm on Twitter in Bahasa,” in Proc. 4th Int. Conf. Informat. Comput. (ICIC), Oct. 2019, pp. 1–5.

AUTHORS Profile



Mrs. PIRANGI HYMAVATHI has received her B.Tech from JNTUH, Hyderabad in 2016, M.Tech degree in Computer science Engineering from JNTUH, Hyderabad in 2019. At present She is working as Assistant Professor in CSE Department at Guru Nanak Institutions Technical Campus, Ibrahimpatnam, Telangana, India.



Ms. VULLAMKONDA SHARANYA pursuing B.Tech degree in Computer Science and Engineering at Guru Nanak Institutions Technical Campus affiliated to JNTUH in 2022-2026.



Ms. TANGALLAPALLY VYSHNAVI pursuing B.Tech degree in Computer Science and Engineering at Guru Nanak Institutions Technical Campus affiliated to JNTUH in 2022-2026.



Mr. ARVAPALLY

NITHIN pursuing B.Tech degree in Computer Science and Engineering at Guru Nanak Institutions

Technical Campus affiliated to JNTUH in 2022-2026.



**Mr. GUDISHALA
VIGNESHWAR**

pursuing B.Tech degree in Computer Science and Engineering at Guru

Nanak Institutions Technical Campus affiliated to JNTUH in 2022-2026.