NATURAL LANGUAGE PROCESSING AND CNN MODEL FOR INDONESIAN SARCASM DETECTION

M. RAMAN KUMAR[1], POTHARAJU RAHUL[2], GURRAM HEMANTH [3], PICHARA VIYATH RAO[4], DAYYALA SAMPATH [5]

[1]Assistant 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— Sarcasm detection presents unique challenges due to the complex linguistic and contextual nature of sarcastic expressions. Understanding sarcasm in text requires advanced methods capable of capturing nuanced patterns that are not easily detectable by traditional approaches. In this study, we propose the use of Convolutional Neural Networks (CNN) for sarcasm detection. The CNN model is designed to identify and classify sarcastic content by learning intricate patterns from the text. Our approach demonstrates that CNNs, with their ability to effectively capture spatial hierarchies and context in textual data, offer superior performance in detecting sarcasm compared to simpler methods. The study also emphasizes the importance of sophisticated data augmentation techniques to address issues like data imbalance, further enhancing the model's effectiveness. This work contributes to advancing sarcasm detection, providing valuable insights for applications in sentiment analysis and natural language understanding.

Index Terms – Sarcasm Detection, Indonesian NLP, CNN, Deep Learning, Text Classification, Sentiment Analysis, Social Media Analytics, Machine Learning, Semantic Analysis, Artificial Intelligence.

I. INTRODUCTION

Malaria Sarcasm detection in text is a complex challenge in Natural Language Processing (NLP) because it involves understanding subtle linguistic patterns, contextual meanings, and variations in human communication. Unlike direct expressions, sarcastic statements often convey meanings that differ from their literal interpretation, making them difficult for traditional sentiment analysis techniques to

classify accurately. Factors such as word ambiguity, cultural context, emotional tone, and informal language usage further increase the difficulty of identifying sarcastic expressions.

Traditional machine learning approaches for sarcasm detection generally rely on manually engineered features such as sentiment polarity, keyword frequency, punctuation patterns, and linguistic characteristics. However, these methods may have limitations in capturing deeper contextual relationships and complex semantic representations within text. As a result, there is a need for advanced deep learning approaches capable of automatically learning meaningful features from large-scale textual data.

In this study, we propose the use of a Convolutional Neural Network (CNN) model for sarcasm detection in Indonesian text. CNNs, which are widely used in image processing, have also demonstrated strong performance in various NLP applications due to their ability to extract important local patterns and hierarchical features from sequential data. Through convolution operations, CNN models can identify significant word combinations, phrases, and contextual structures that contribute to sarcastic expressions.

II. RELATED WORK

A. 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 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.

B. 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.

C. Large language models only pass primary school exams in indonesia: A comprehensive test on IndoMMLU

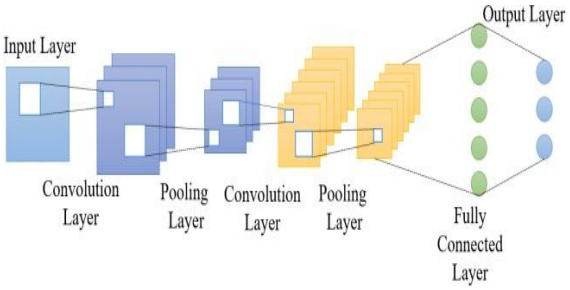
Although large language models (LLMs) are often pre-trained on large-scale multilingual texts, their reasoning abilities and real-world knowledge are mainly evaluated based on English datasets. Assessing LLM capabilities beyond English is increasingly vital but hindered due to the lack of suitable datasets. In this work, we introduce IndoMMLU, the first multi-task language understanding benchmark for Indonesian culture from primary school to university entrance exams in Indonesia. By employing professional teachers, we obtain 14,981 questions across 64 tasks and education levels, with 46% of the questions focusing on assessing proficiency in the Indonesian language and knowledge of nine local languages and cultures in Indonesia.

III. PROPOSED METHODOLOGY

The A Convolutional Neural Network (CNN) is a type of deep learning architecture designed to process structured data by automatically learning important patterns and features from input data. Originally developed for image processing tasks, CNNs have become one of the most successful approaches in computer vision applications such as image classification, object detection, facial recognition, and image segmentation. Inspired by the biological visual system, CNNs imitate the way humans recognize objects by extracting meaningful features through multiple layers of interconnected neurons.

CNNs are highly effective because they can automatically identify hierarchical patterns within data without requiring manual feature extraction. The network consists of several specialized layers, including convolutional layers, activation functions, pooling layers, and fully connected layers. Convolutional layers apply filters to the input data to detect important local patterns, while activation functions introduce non-linearity, allowing the model to learn complex relationships.

In the convolution operation, small filters or kernels move across the input data to capture important features such as edges, shapes, and patterns. In Natural Language Processing (NLP), CNNs can be applied to text data by treating words or sentences as structured sequences. Through convolution operations, CNNs can detect meaningful word combinations, phrases, and contextual patterns that contribute to text classification



tasks such as sentiment analysis and sarcasm detection.

Fig. 1: System Overview

1. Data Gathering

The dataset is the collection of data that you use to train, test, and validate your machine learning model. It can be any structured or unstructured data relevant to the problem you're solving. In image classification, for example, the dataset would consist of labeled images of different classes. For a text-based task, it could include text samples with corresponding labels.

2. Data cleaning:

Data cleaning refers to the process of identifying and rectifying errors or inconsistencies in the dataset. It ensures the data is accurate, complete, and ready for analysis. This step is essential to improve model performance since poor-quality data can lead to misleading results.

3. Preprocessing the data:

Data preprocessing is the transformation of raw data into a format that is more suitable for the model. It helps to convert raw data into numerical or structured format and scales it as needed to improve the accuracy of the machine learning algorithms.

4. Splitting the data:

Splitting the data involves dividing the dataset into different subsets for training, validation, and testing. This ensures that the model can be trained on one set of data and tested on another to evaluate its generalization capabilities.

5. Model apply:

Applying a machine learning model means selecting a suitable algorithm or architecture to learn from the data. Depending on the task (classification, regression, etc.), you would choose an appropriate model to train on the dataset.

6. Model evaluation:

After the model is trained, it's essential to evaluate its performance on a separate set of data (the validation or test set). Model evaluation helps you understand how well the model generalizes to unseen data and whether any adjustments are needed.

7. Prediction:

After training and evaluating the model, the final step is making predictions on new, unseen data. The model uses the patterns learned during training to predict outcomes based on the input features of the new data.

IV. EXPERIMENTAL RESULTS AND ANALYSIS

Experimental results demonstrated that the model effectively identified sarcastic and non-sarcastic text with high classification accuracy by capturing contextual and semantic patterns in Indonesian language expressions.

The analysis showed that the CNN-based feature extraction mechanism successfully learned local linguistic features and implicit sentiment cues that are difficult to identify using traditional machine learning methods.



Fig. 2: Homepage

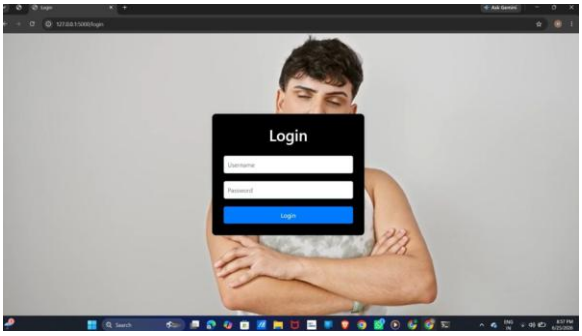


Fig. 3: Login



Fig. 4: Predicted Results

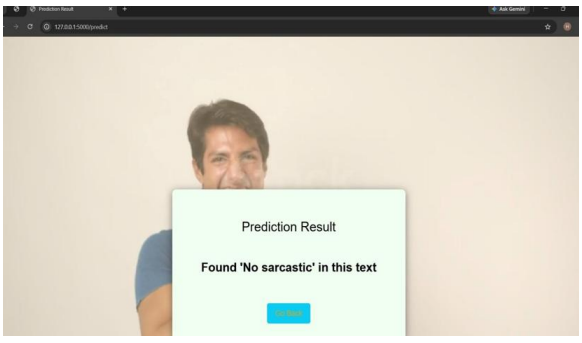


Fig. 5: Predicted Results

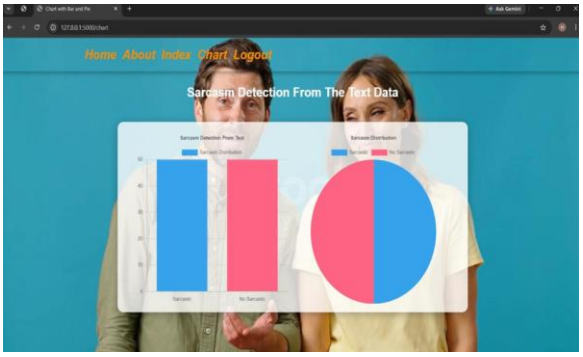


Fig. 6: Predicted Results in Charts

V. CONCLUSION

This study introduced the first publicly available Indonesian sarcasm detection benchmark datasets, contributing valuable resources for research in this area. We compared classical machine learning approaches with modern pre-trained language models, as well as multilingual large language models with zero-shot capabilities. Our findings indicate that pre-trained language models outperformed traditional machine learning methods in sarcasm detection tasks, though classical methods remain more effective in low-resource environments. Additionally, we observed that large language models still struggle with accurate zero-shot sarcasm detection, and synthetic data generation through machine translation and weighted cross-entropy loss did not sufficiently address data imbalance issues. Despite the limitations of our semi-automatic data creation method, this work lays a strong foundation for future advancements in sarcasm detection in low-resource languages like Indonesian.

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.

AUTHORS Profile



Mr. M. RAMAN KUMAR received his B.Tech degree in Computer Science and Engineering from Ranchi University in 2010 and his M.Tech degree in Computer Science and Engineering from Bhagwant University, Ajmer, in 2020. He is presently pursuing his Ph.D. from Atal Bihari Vajpayee Indian Institute of Information Technology and Management (ABV-IIITM), Gwalior. He has been dedicated to the teaching profession for the past 2 years. At present, he is working as an Assistant Professor in the Department of Computer Science and Engineering at Guru Nanak Institutions Technical Campus, Ibrahimpatnam, Telangana, India.



Mr. POTHARAJU RAHUL pursuing B. Tech degree in Computer Science and Engineering at Guru Nanak Institutions Technical Campus affiliated to JNTUH in 2022-2026.



Mr. GURRAM HEMANTH pursuing B. Tech degree in Computer Science and Engineering at Guru Nanak Institutions Technical Campus affiliated to JNTUH in 2022-2026.



Mr. PICHARA VIYATH RAO pursuing B. Tech degree in Computer Science and Engineering at Guru Nanak Institutions Technical Campus affiliated to JNTUH in 2022-2026.



Mr. DAYYALA SAMPATH pursuing B. Tech degree in Computer Science and Engineering at Guru Nanak Institutions Technical Campus affiliated to JNTUH in 2022-2026.