IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 1 May 2024, accepted 12 June 2024, date of publication 20 June 2024, date of current version 28 June 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3416955

RESEARCH ARTICLE

IdSarcasm: Benchmarking and Evaluating Language Models for Indonesian Sarcasm Detection

DERWIN SUHARTONO, (Member, IEEE), WILSON WONGSO, AND ALIF TRI HANDOYO
Computer Science Department, School of Computer Science, Bina Nusantara University, Jakarta 11480, Indonesia

Corresponding author: Derwin Suhartono (dsuhartono@binus.edu)

This work was supported in part by Bina Nusantara University for the open access publication.

ABSTRACT 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.

INDEX TERMS Low-resource data, low-resource languages, Indonesian sarcasm detection, natural language processing, sarcasm detection, sentiment analysis.

I. INTRODUCTION

Sarcasm, also known as irony, has been extensively studied in the fields of linguistics and psychology. In the field of natural language processing, detecting sarcasm within a sentence or message remains a significant challenge because the lexical features extracted from the sentence do not provide sufficient information to detect sarcasm [1]. Sarcasm is a growing research area in English natural language processing. However, there is still a lack of research on detecting sarcasm in Indonesian text.

While some research has been done [1], [2], many of them are still using outdated machine learning techniques. In their research, Twitter is the most common dataset that has been used. This is typical because Twitter is one of the social media platforms that generate millions of data points every day and Indonesia has the fifth-highest number of Twitter users [3]. Thus, Indonesian Twitter data is abundant and worth analyzing.

Because Twitter only allows messages to contain 280 characters, users are forced to write their messages creatively. Most Indonesian Twitter users are active and expressive; they can creatively express themselves on trending topics in a limited number of characters [3]. As part of their creativity, some of them frequently use sarcasm, or positive words to express negative opinions, in their Twitter posts.

Another social media that is worth mentioning is Reddit. Although Reddit is legally banned in Indonesia, it is still one of the most popular social media in Indonesia [4], where users can express themselves in various subreddits. In Reddit, the discussion, conversation, and sarcastic remarks are less constrained. One of those discussions that contain sarcasm is sometimes marked by `/s` in the post. Thus, this dataset is worth considering, especially according to [5] and [6] the difference in text length can impact the sarcasm detection result, meaning that we can analyze and investigate the different results of using the Twitter or Reddit datasets.

The associate editor coordinating the review of this manuscript and approving it for publication was Jolanta Mizera-Pietraszko.

Even though numerous studies have been done to identify sarcasm in English, there is still a lack of research that conducts sarcasm detection in the Indonesian language [2]. Previous studies that work on sarcasm in the Indonesian language also still use very basic and outdated, classical machine learning techniques [2], [7], [8], [9], [10]. For example, research conducted in [10] only uses random forest machine learning to classify sarcastic tweets. Another study [9] used only SVM [11] to classify sarcastic texts. The most recent one by [12] combines various word embedding models, and their experimental result shows that the combination of fastText embeddings and BiGRU classifier produced the best performance in an Indonesian Twitter dataset. It shows that various past research that studied sarcasm detection in the Indonesian language still lack their method of detection, and even the most recent one still uses considerably outdated deep learning models.

Nevertheless, conventional machine learning techniques have shown limitations when faced with implicitly message-carrying sarcastic statements because they are unable to contextualize the entire sentence. This made the switch to deep learning techniques necessary. Consequently, various studies have already implemented sarcasm detection using deep learning techniques [13], [14]. Those studies have adopted a deep learning paradigm for classifying sarcasm, combining methods like hybrid neural networks – which fuse convolutional neural networks (CNN) and bidirectional long short-term memory (LSTM) architectures – and multi-layer perceptrons. However, all those studies focused on English data. Thus, this research is focused on leveraging various pre-trained monolingual and multilingual language models to classify Indonesian sarcastic texts in social media, primarily on Twitter and Reddit. This research also aims to investigate synthetic data and training techniques methods that may be suitable for sarcastic data in the Indonesian language.

In summary, the contributions of this research include:

- Development of Indonesian sarcasm detection benchmark datasets.
- Baseline models and results, covering classical machine learning methods, fine-tuning pre-trained language models, and zero-shot inference via multilingual large language models.
- Attempts to alleviate data imbalance through synthetic data and weighted loss.
- Potential research direction for future Indonesian sarcasm detection datasets.

The remaining section of this paper is structured as follows: In Section II, prior research on sarcasm detection and augmentation methods in sarcasm datasets is reviewed. Section III discusses the datasets, pre-processing methods, suggested models, and experimental procedures. Section IV discusses the experimental results. Finally, in Section V the research's conclusions are covered.

II. RELATED WORKS

Research on sarcasm detection in Indonesia is still limited, despite the increasing use of sarcasm in social media. Several studies have explored sarcasm detection in various languages such as English, Indian, and Indonesian. Despite the limited research on sarcasm detection in Indonesia, there is a growing need to develop effective methods for detecting sarcasm in Indonesian social media. In this session, previous research that focused on sarcasm detection in Indonesian text will be reviewed.

In recent years, there has been a growing interest in sarcasm detection in various languages, including Indonesian. The use of sarcasm in digital communication has become increasingly prevalent, making it important to develop effective sarcasm detection systems for languages beyond English. Previous research has focused primarily on English, and there is a need to explore the unique linguistic and cultural aspects of the Indonesian language to develop accurate sarcasm detection models.

Research on sarcasm detection in the Indonesian language is still in its infancy compared to English and other widely studied languages. However, there have been some notable efforts in this direction. For instance, [7] analyzed sarcastic patterns in Indonesian social media posts using machine learning techniques. Their findings revealed the significance of contextual information and cultural references in detecting sarcasm, indicating the need for culturally sensitive models for accurate detection in the Indonesian language.

Additionally, [1] proposed a technique that combines interjection and punctuation as feature extraction methods for sarcasm detection in Indonesian Twitter feeds. They further applied different weighting and classification algorithms, such as TF-IDF and k-Nearest Neighbor, to achieve better performance in detecting sarcasm.

Furthermore, [11] focused on sarcasm detection using machine learning algorithms, including SVM. Their study demonstrated the effectiveness of CNN in capturing contextual and semantic information for improved sarcasm detection. Another study [9] aimed to detect sarcasm in the Indonesian language using various linguistic features and machine learning techniques. The researchers collected a dataset of 480 train data and 120 test data from Twitter by crawling. After pre-processing and feature extraction, the data were classified using the Support Vector Machine algorithm. The researchers compared the accuracy of different features, including N-gram, POS (part-of-speech) Tags, Punctuation, and Pragmatic, as well as combining all the features. Their proposed approach achieved the highest accuracy of 91.6% with a precision of 92% when all features were combined. However, it can be seen that the research still uses an outdated machine learning technique, and uses a very small amount of data.

Another study [2] further explored the use of deep learning models, specifically a bidirectional long short-term memory neural network, for sarcasm detection in the Indonesian

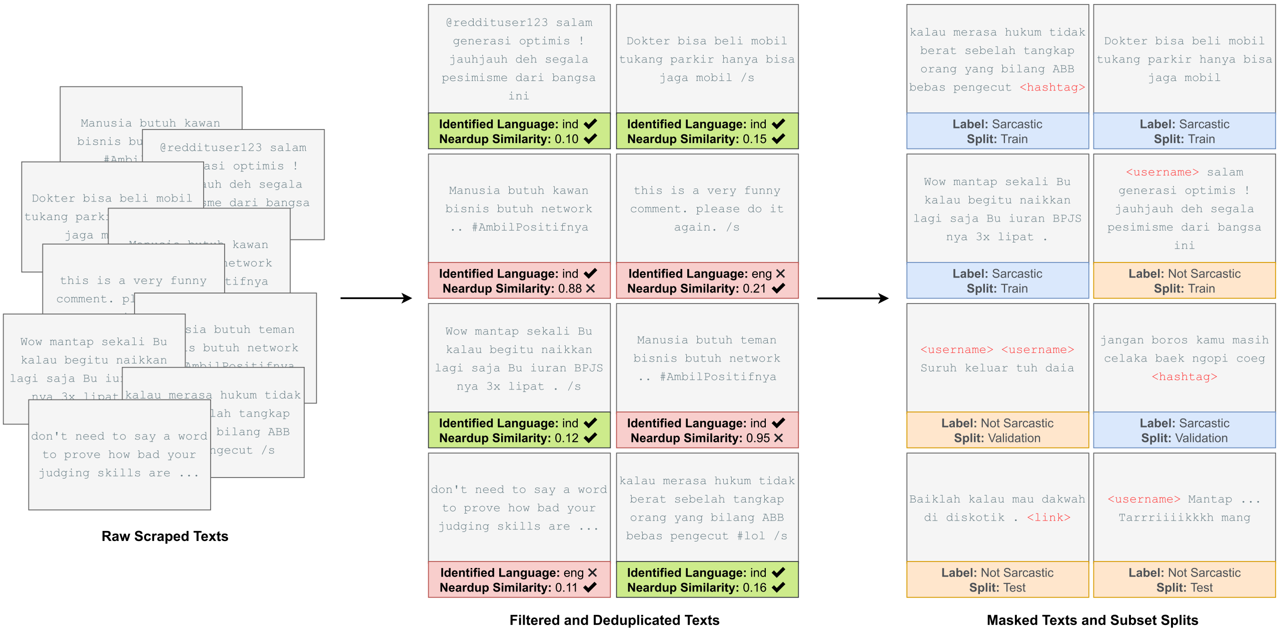

Raw Scraped Texts
@reddituser123 salam generasi optimis ! jauhjauh deh segala pesimisme dari bangsa ini
Identified Language: ind ✓
Neardup Similarity: 0.10 ✓
Dokter bisa beli mobil tukang parkir hanya bisa jaga mobil /s
Identified Language: ind ✓
Neardup Similarity: 0.15 ✓
Manusia butuh kawan bisnis butuh network .. #AmbilPositifnya
Identified Language: ind ✓
Neardup Similarity: 0.88 ✗
this is a very funny comment. please do it again. /s
Identified Language: eng ✗
Neardup Similarity: 0.21 ✓
Wow mantap sekali Bu kalau begitu naikkan lagi saja Bu iuran BPJS nya 3x lipat . /s
Identified Language: ind ✓
Neardup Similarity: 0.12 ✓
Manusia butuh teman bisnis butuh network .. #AmbilPositifnya
Identified Language: ind ✓
Neardup Similarity: 0.95 ✗
don't need to say a word to prove how bad your judging skills are ...
Identified Language: eng ✗
Neardup Similarity: 0.11 ✓
kalau merasa hukum tidak berat sebelah tangkap orang yang bilang ABB bebas pengecut #lol /s
Identified Language: ind ✓
Neardup Similarity: 0.16 ✓
Filtered and Deduplicated Texts
kalau merasa hukum tidak berat sebelah tangkap orang yang bilang ABB bebas pengecut <hashtag>
Label: Sarcastic
Split: Train
Dokter bisa beli mobil tukang parkir hanya bisa jaga mobil
Label: Sarcastic
Split: Train
Wow mantap sekali Bu kalau begitu naikkan lagi saja Bu iuran BPJS nya 3x lipat .
Label: Sarcastic
Split: Train
<username> salam generasi optimis ! jauhjauh deh segala pesimisme dari bangsa ini
Label: Not Sarcastic
Split: Train
<username> <username> Suruh keluar tuh daia
Label: Not Sarcastic
Split: Validation
jangan boros kamu masih celaka baek ngopi coeg <hashtag>
Label: Sarcastic
Split: Validation
Baiklah kalau mau dakwah di diskotik . <link>
Label: Not Sarcastic
Split: Test
<username> Mantap ... Tarrriiiikkkh mang
Label: Not Sarcastic
Split: Test
Masked Texts and Subset Splits

FIGURE 1. Data and text processing pipeline for the creation of the Reddit Indonesia sarcasm dataset.

language. Their findings showed that the deep learning model outperformed traditional machine learning models in detecting sarcasm, achieving a higher accuracy rate. This indicates that incorporating deep learning models may be a more effective approach for sarcasm detection in the Indonesian language. One of the major challenges in sarcasm detection in the Indonesian language is the lack of labeled datasets. While there are ample resources for English sarcasm detection, there is a scarcity of annotated datasets for the Indonesian language, hindering the training and evaluation of detection models.

Overall, previous work in sarcasm detection in the Indonesian language has shown that incorporating specialized models, cultural and contextual factors, interjections, punctuation, TF-IDF, and k-Nearest Neighbor algorithms can significantly improve the accuracy of sarcasm detection in the Indonesian language. As a result, in this study, we investigate several novel models to improve the identification of sarcasm in Indonesian text, and then use data augmentation techniques to address the problem of insufficient data in sarcasm detection.

III. METHODOLOGY

In this section, we first introduce two new datasets for Indonesian sarcasm detection, which we call `id_sarcasm`. Afterward, we show the different baseline models that have been fine-tuned for and evaluated on the two datasets.

A. REDDIT INDONESIA SARCASM DATASET

Reference [15] proposed a dataset creation method to collect sarcastic Indonesian comments from Reddit. However, since the dataset is not available to the public, we decided to reproduce their data creation methodology and develop our own dataset, extending the proposed method to the latest Reddit archive.

We leveraged a previously collected Reddit torrent [16] from January 2020 to September 2023 and filtered for comments in the Indonesian subreddit `r/indonesia`. As conducted in [15], comments which end with a `/s` tag are considered and thereby labeled as sarcastic. Adding a `/s` at the end of a comment implies sarcasm, a common practice done on Reddit. All other comments which do not contain the tag are considered non-sarcastic. As the presence of this suffix tag would reduce this problem to a trivial substring condition, it was removed during pre-processing. Furthermore, near-deduplication using MinHash locality-sensitive hashing (LSH) [17] was performed to remove duplicate comments.

Likewise, since Indonesian is not the only language used as a medium of communication in Reddit, we filtered them using fastText [18], [19] which identifies the language(s) of a given comment. Only comments that were classified as Indonesian (`id`), Javanese (`jv`), Minangkabau (`min`), Malay (`ms`), or Sundanese (`su`) are included after this filtering process. We deliberately chose this set of languages after qualitative checking.

Moreover, as these comments may contain sensitive data such as usernames, hashtags, emails, and URLs, we applied a masking technique to maintain privacy and reduce noise in our data [20]. Namely, these components are masked as `<username>`, `<hashtag>`, `<email>`, and `<link>` respectively.

Finally, since there is an overwhelming amount of non-sarcastic comments than sarcastic ones, we randomly sampled non-sarcastic comments to meet a 1:3 ratio of sarcastic to non-sarcastic comments, following the iSarcasm task found in SemEval 2022 [21]. The re-sampled dataset was then randomly split into train (70%), validation (10%), and test (20%) subsets. Our final dataset consists of 3,529 sarcastic and 10,587 non-sarcastic comments, totaling 14,116 overall comments. Fig. 1 summarizes our overall data and text processing pipeline. Examples of Reddit comments contained in our dataset are shown in Table 3.

B. TWITTER INDONESIA SARCASM DATASET

Reference [22] similarly introduced an Indonesian sarcasm detection dataset based on Indonesian tweets, whose pre-processed version of the dataset is publicly available. The dataset consists of 17,718 tweets that have been previously labeled as either sarcastic or non-sarcastic by an expert. These tweets were collected from March 2013 to February 2020.

They have conducted an extensive pre-processing pipeline, which consisted of normalization from colloquial to standard Indonesian spelling, lowercasing, stop-word removal, and reverse word order. While the first three procedures are expected for a typical text processing step, the latter is rather unusual and has been reverted in our dataset release to maintain its original word order.

Then, a pre-processing pipeline similar to the one performed on the Reddit dataset discussed in Section III-A and shown in Fig. 1 was applied, except for language identification. That is, near-duplicate tweets were removed using the same MinHash LSH algorithm [17]. Surprisingly, this step significantly reduced the initial number of sarcastic tweets from 4,350 to only 671. This steep decrease in the number of sarcastic tweets was verified after qualitatively inspecting the sarcastic tweets. Many of these spam tweets are repeats of an initial tweet, with only one or two words being replaced. While they are not exact duplicates, MinHash LSH was able to cluster them due to their high levels of similarity.

Accordingly, the same masking technique introduced in [20] was applied to this collection of tweets. Usernames, hashtags, emails, and URLs were masked as `<username>`, `<hashtag>`, `<email>`, and `<link>` respectively.

Furthermore, while the number of sarcastic tweets was significantly reduced after the de-duplication step, the number of non-sarcastic tweets was still relatively high as they only contained about 1,000 near duplicates. Applying the same re-sampling procedure, the dataset was re-sampled to a 1:3 ratio of sarcastic to non-sarcastic tweets, which were then randomly split into train (70%), validation (10%), and test (20%) subsets. The resultant dataset consists of 671 sarcastic and 2,013 non-sarcastic tweets, summing to 2,684 tweets in total. Examples of tweets found in our dataset are shown in Table 4.

C. BASELINE MODELS

1) CLASSICAL MACHINE LEARNING

While pre-trained language models generally perform better than classical machine learning approaches on most natural language understanding tasks [20], in some extreme cases such as low-resource languages and limited computational resources, classical approaches are preferred and occasionally outperform more modern techniques [23], [24]. Indeed, classical machine learning algorithms paired with hand-engineered features remain a popular approach to sarcasm detection [21].

Following [23], [25], we leveraged three classical text classification algorithms as our baselines: Logistic Regression, Naive Bayes, and SVM (Support Vector Machine). In short, Logistic Regression [26] predicts binary outcomes by modeling the probabilities as a logistic function of word frequencies, adjusting weights to minimize errors. Naive Bayes [27] calculates the likelihood of each category based on the independence of word occurrences, using Bayes' Theorem to predict text classification. SVMs [11] classify texts by finding a hyperplane in high-dimensional space that best separates different classes with the largest margin, using linear and nonlinear mappings through the kernel trick.

We used Scikit-Learn [28] as our machine learning framework and performed a grid search across a set of hyperparameter ranges shown in Table 5, whose values are from [25]. For feature vectorization of input texts, both Bag of Words and TF-IDF were tested and compared. The best parameters were chosen based on the model's performance on the dataset's validation split.

2) PRE-TRAINED LANGUAGE MODELS

As mentioned previously, pre-trained language models are superior to conventional methods on most downstream tasks, displaying their adaptability and versatility across multiple language understanding tasks [29]. There are two prominent monolingual BERT-based [29] encoder language models available for the Indonesian language, namely IndoNLU's IndoBERT [20] and IndoLEM's IndoBERT [30]. Due to their similar performances on numerous downstream tasks like emotion classification, sentiment analysis, topic modeling, and rhetorical mode classification [23], both models were included in our investigation.

Moreover, as multilingual language models like multilingual BERT (mBERT) [29] and XLM-RoBERTa (XLM-R) [31] continue to provide a strong baseline performance similar to their monolingual counterparts, they were similarly investigated in our experiments.

Following the best hyperparameter settings conducted in the experiments of NusaX [25] and NusaWrites [23], we fine-tuned our language models on the datasets proposed in Section III-A and III-B. These hyperparameters are shown in Table 6.

In our experiments, we replaced the language modeling head of each model with a linear classifier head. To facilitate a binary classification task like sarcasm detection, the classifiers were optimized using the cross-entropy loss function. All of our experiments are conducted using the HuggingFace framework [32] on PyTorch [33].

Additionally, we also experimented by adding synthetic data using machine-translated sarcastic texts and weighted cross-entropy loss, in an attempt to mitigate the effects of data imbalance. Namely, we translated the sarcastic texts found in the English train split of the iSarcasmEval dataset [21] into Indonesian. The translation utilized the Google Translate API and gave 867 additional Indonesian sarcastic texts, which were concatenated to the respective train sets of either Reddit or Twitter datasets.

Furthermore, due to the large imbalance ratio of 1:3 sarcastic to non-sarcastic texts of our datasets, an additional rescaling weight factor to the cross-entropy loss was applied. This factor was automatically calculated using the balanced class weights, assigning a higher weight to the minority class, which in our case is the sarcastic class. The equation for a weighted binary cross-entropy loss is shown in Equation 1.

$$\begin{aligned}\mathcal{L} = -\mathbb{E}[M \cdot w_p \cdot y_{\text{true}} \cdot \log\left(y_{\text{pred}}\right) \\ + M \cdot w_n \cdot (1 - y_{\text{true}}) \cdot \log\left(1 - y_{\text{pred}}\right)]\end{aligned} \tag{1}$$

where w_p and w_n are class weights of the positive and negative classes respectively, and y_{true} and y_{pred} are the true and predicted class labels (0 or 1), respectively.

Since our datasets share a similar class imbalance, the class weights of both datasets are about $w_n = 0.666$ for the non-sarcastic class and $w_p = 2.0$ for the sarcastic class. These values were then further rescaled by an additional weight factor $M = 2$.

Due to the nature of the class imbalance of our datasets, the F1 score was chosen as the main evaluation metric of our models. The equations to calculate the accuracy, precision, recall, and F1 score of a binary classification task are as follows:

$$\begin{aligned}\text{Accuracy} &= \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \\ \text{Precision} &= \frac{\text{TP}}{\text{TP} + \text{FP}} \\ \text{Recall} &= \frac{\text{TP}}{\text{TP} + \text{FN}} \\ \text{F1} &= \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}\end{aligned} \tag{2}$$

where TP, TN, FP, and FN represent the number of counts of true positives, true negatives, false positives, and false negatives, respectively.

We report the evaluation scores of our models on the test subsets of the respective datasets in Tables 1 and 2.

3) ZERO-SHOT INFERENCE VIA LARGE LANGUAGE MODELS

Recently, large language models (LLMs) have become increasingly useful and effective for a diverse set of tasks, including those that they have not seen during training time [34]. Moreover, newer LLMs are multilingual and were trained on a vast range of languages including Indonesian. Because of this, their zero-shot performances are often benchmarked on natural language understanding tasks such as NusaWrites [23] and the BHASA [35] benchmark, both of which include Indonesian and/or regional languages of Indonesia.

Specifically, NusaWrites [23] attempted to conduct a zero-shot evaluation of these LLMs on their language understanding tasks. Interestingly, although these models have been fine-tuned to follow instructions via human-like prompts [34], their performance remains immensely underwhelming compared to both classical machine learning approaches and fine-tuning. Following this setup, we wanted to investigate whether these LLMs will continue to be lackluster in Indonesian sarcasm detection.

Like NusaWrites [23], variants of BLOOMZ and mT0 [36] were selected as zero-shot inference baselines on our datasets. However, due to the lack of computational resources available, we limit the model size to 3.7B parameters only. To obtain the zero-shot predicted labels of these models, the most probable label (sarcastic or non-sarcastic) was selected based on the sum of the log probabilities of the input sequence upon inference. This follows the same zero-shot inference procedure as NusaWrites [23].

Since these models were fine-tuned to follow instructions, human-like prompts shown in Table 7 were thereby used during inference. To de-bias the results and ensure fairness irrespective of the prompts, five different prompts were given and the mean evaluation metrics were reported in Tables 1 and 2. Since these zero-shot evaluation results are mean values of the different prompts' respective metric scores, we note that they do not conform to Equation 2.

IV. RESULTS AND DISCUSSION

A. FINE-TUNING RESULTS

The results of sarcasm detection on our newly curated Indonesian datasets are presented in Tables 1 and 2. As the classical machine learning methods do not involve the addition of contextual embeddings, it is unsurprising that the results are inferior compared to fine-tuning of pre-trained language models. The much simpler Bag of Words and TF-IDF feature vectorizer poorly encodes semantic representation of sarcastic comments and relies more on the text's syntactic feature. However, despite its lackluster performance on the Reddit dataset, classical machine learning approaches remain competitive with fine-tuning pre-trained language models. This is in line with the results of NusaX [25] and NusaWrites [23] whereby classical machine learning algorithms could very well outperform pre-trained language models when it comes to low-resource data settings. Otherwise, when the amount of data is sufficiently large enough, pre-trained language models continue to be superior in performance.

Moreover, there is a steep decrease in the F1 score of the Twitter dataset from the one reported by its original authors [22], and when tested on the original dataset. Notably, the authors were able to achieve remarkably high F1 scores of 94.74% and 98.04% on the imbalance and balanced dataset. We attribute this suspiciously high accuracy on a

TABLE 1. Evaluation scores on the test subset of reddit indonesia sarcasm dataset. Results include all methods examined in this study. The maximum score per metric per method has been bolded for clarity. The overall highest score per metric has been further underlined.

	Reddit			
Model	**Accuracy**	**F1-score**	**Precision**	**Recall**
Classical				
Logistic Regression	0.7829	**0.4887**	0.5943	**0.4150**
Naive Bayes	**0.7889**	0.4591	0.6388	0.3583
SVM	0.7885	0.4467	**0.6461**	0.3413
Fine-tuning				
IndoNLU IndoBERT$_{\text{BASE}}$	0.7921	0.6100	0.5745	0.6501
IndoNLU IndoBERT$_{\text{LARGE}}$	0.7911	0.6184	0.5690	**0.6771**
IndoLEM IndoBERT$_{\text{BASE}}$	0.7670	0.5671	0.5295	0.6105
mBERT	0.7829	0.5338	0.5764	0.4972
XLM-R$_{\text{BASE}}$	0.8031	0.5690	**0.6284**	0.5198
XLM-R$_{\text{LARGE}}$	**0.8120**	**0.6274**	0.6217	0.6331
Fine-tuning with Synthetic Data				
IndoNLU IndoBERT$_{\text{BASE}}$	0.7851	**0.6213**	0.5552	**0.7054**
IndoNLU IndoBERT$_{\text{LARGE}}$	**0.7909**	0.5816	**0.5824**	0.5807
IndoLEM IndoBERT$_{\text{BASE}}$	0.7652	0.5703	0.5257	0.6232
mBERT	0.7688	0.5346	0.5380	0.5312
XLM-R$_{\text{BASE}}$	0.7819	0.5734	0.5610	0.5864
XLM-R$_{\text{LARGE}}$	0.7907	0.5944	0.5766	0.6133
Fine-tuning with Synthetic Data + Weighted Loss				
IndoNLU IndoBERT$_{\text{BASE}}$	0.7755	**0.6253**	0.5365	0.7493
IndoNLU IndoBERT$_{\text{LARGE}}$	0.7794	0.6245	0.5435	0.7337
IndoLEM IndoBERT$_{\text{BASE}}$	0.7153	0.5742	0.4585	**0.7677**
mBERT	0.7316	0.5509	0.4735	0.6586
XLM-R$_{\text{BASE}}$	0.7553	0.5994	0.5074	0.7323
XLM-R$_{\text{LARGE}}$	**0.7890**	0.5940	**0.5722**	0.6176
Zero-shot				
BLOOMZ-560M	0.3046	0.3870	0.2500	0.8895
BLOOMZ-1.1B	0.2737	0.3944	0.2492	0.9490
BLOOMZ-1.7B	**0.3365**	0.3758	0.2500	0.8320
BLOOMZ-3B	0.3055	0.4000	0.2560	0.9260
BLOOMZ-7.1B	0.3058	**0.4036**	**0.2575**	0.9388
mT0$_{\text{SMALL}}$	0.2500	0.4000	0.2500	**1.0000**
mT0$_{\text{BASE}}$	0.2515	0.3990	0.2496	0.9940
mT0$_{\text{LARGE}}$	0.2500	0.3998	0.2499	0.9991
mT0$_{\text{XL}}$	0.2503	0.4001	0.2500	**1.0000**

TABLE 2. Evaluation scores on the test subset of twitter indonesia sarcasm dataset. Results include all methods examined in this study. The maximum score per metric per method has been bolded for clarity. The overall highest score per metric has been further underlined.

	Twitter			
Model	**Accuracy**	**F1-score**	**Precision**	**Recall**
Classical				
Logistic Regression	**0.8661**	**0.7142**	0.7627	**0.6716**
Naive Bayes	0.8531	0.6721	0.7570	0.6044
SVM	0.8624	0.6782	**0.8125**	0.5820
Fine-tuning				
IndoNLU IndoBERT$_{\text{BASE}}$	0.8662	0.7273	0.7385	0.7164
IndoNLU IndoBERT$_{\text{LARGE}}$	0.8643	0.7160	0.7480	0.6866
IndoLEM IndoBERT$_{\text{BASE}}$	0.8290	0.6462	0.6667	0.6269
mBERT	0.8030	0.6467	0.5843	0.7239
XLM-R$_{\text{BASE}}$	0.8513	0.7386	0.6570	**0.8433**
XLM-R$_{\text{LARGE}}$	**0.8885**	**0.7692**	**0.7937**	0.7463
Fine-tuning with Synthetic Data				
IndoNLU IndoBERT$_{\text{BASE}}$	**0.8513**	0.7015	0.7015	0.7015
IndoNLU IndoBERT$_{\text{LARGE}}$	0.8420	0.6444	**0.7333**	0.5746
IndoLEM IndoBERT$_{\text{BASE}}$	0.8123	0.6731	0.5943	**0.7761**
mBERT	0.7900	0.6319	0.5607	0.7239
XLM-R$_{\text{BASE}}$	0.8420	**0.7038**	0.6601	0.7537
XLM-R$_{\text{LARGE}}$	0.8383	0.6904	0.6599	0.7239
Fine-tuning with Synthetic Data + Weighted Loss				
IndoNLU IndoBERT$_{\text{BASE}}$	**0.8532**	0.6749	**0.7523**	0.6119
IndoNLU IndoBERT$_{\text{LARGE}}$	0.8476	0.6985	0.6884	0.7090
IndoLEM IndoBERT$_{\text{BASE}}$	0.8178	0.6573	0.6184	0.7015
mBERT	0.7974	0.6280	0.5786	0.6866
XLM-R$_{\text{BASE}}$	0.8476	0.6555	0.7500	0.5821
XLM-R$_{\text{LARGE}}$	0.8271	**0.7103**	0.6096	**0.8507**
Zero-shot				
BLOOMZ-560M	**0.2672**	0.3916	0.2469	0.9507
BLOOMZ-1.1B	0.2591	0.3987	**0.2499**	0.9865
BLOOMZ-1.7B	0.2620	0.3885	0.2448	0.9417
BLOOMZ-3B	0.2527	0.3847	0.2419	0.9402
BLOOMZ-7.1B	0.2531	0.3968	0.2483	0.9865
mT0$_{\text{SMALL}}$	0.2490	0.3988	0.2490	**1.0000**
mT0$_{\text{BASE}}$	0.2494	0.3985	0.2489	0.9985
mT0$_{\text{LARGE}}$	0.2494	**0.3989**	0.2491	**1.0000**
mT0$_{\text{XL}}$	0.2490	0.3988	0.2490	**1.0000**

semantically challenging task like sarcasm detection to the high percentage of duplicates found in the original dataset. As discussed in Section III-B, the original number of sarcastic tweets dropped from 4,350 to only 671 (−84.5%) with near-duplicate removal. We hypothesize that their model might have learned the repeated sarcastic tweet formats and, due to data leakage, reduced the classification problem to a more trivial substring matching task.

In both Reddit and Twitter datasets, the cross-lingual XLM-R$_{\text{LARGE}}$ model achieved the highest F1 scores, which can be attributed to the model being the largest out of the other pre-trained language models, with 561M parameters. In contrast, the smaller, monolingual IndoNLU IndoBERT$_{\text{BASE}}$ with only 127M parameters provided the second-best results despite its size. This is parallel with the results of many other Indonesian language understanding benchmarks like NusaX [25], NusaWrites [23], and IndoNLU [20], where this monolingual model remains competitive with cross-lingual variants due to the former's specialization in the language and the latter's curse of multilinguality [31]. On the contrary, the multilingual mBERT underperformed even against the monolingual models. Previous studies [37], [38] show that mBERT usually only outperforms XLM-R on sequence tagging tasks like Named Entity Recognition (NER) or Part-of-Speech (POS) tagging.

B. ZERO-SHOT RESULTS

Despite being fine-tuned on a variety of downstream tasks through human-like instructions, the zero-shot sarcasm detection results via LLMs severely underperform compared to the fine-tuning methods. Although the recall scores are very high, the precision scores are very low, indicating that the models are merely guessing that the texts are sarcastic and unable to distinguish them from non-sarcastic ones. Again, these results are similar to the findings of NusaWrites [23] whereby there is a large gap between zero-shot LLMs like BLOOMZ and mT0 [36] and fine-tuning approaches on downstream Indonesian language understanding tasks.

In another study, a new benchmark dataset was developed to measure the capabilities of LLMs on local Indonesian examination questions, ranging from primary school questions to university entrance exams [39]. These questions cover a wide range of subjects and evaluate how well current LLMs truly understand the Indonesian language and regional languages of Indonesia. Indeed, BLOOMZ, mT0, and many other multilingual LLMs merely pass primary school examinations and fail at higher educational levels – indicating that current LLMs need better contextualization and understanding of Indonesian languages, let alone for a challenging task like sarcasm detection.

C. SYNTHETIC DATA AND WEIGHTED LOSS IMPACT

The techniques attempted to alleviate the effects of data imbalance, namely synthetic data and weighted cross-entropy loss, seem to worsen the training results instead of improving them. Although both methods are quite popular for handling class imbalance, they are still not satisfactory for sarcasm detection on our datasets. Nevertheless, while some results in Tables 1 and 2 show that these techniques could slightly increase the evaluation results, they are only beneficial to certain models and datasets (e.g. IndoBERT, mBERT, XLM-R$_{\text{BASE}}$ on Reddit) and harmful to others (e.g. Twitter).

Several reasons are plausible for the fluctuating results, such as the small synthetic dataset size used in this study (only 867), the translation quality (machine-translated, not translated by expert humans), and the mismatch between sarcasm domains. Indonesian sarcasm is often impolite, insulting, or inappropriate [40], which greatly differs from American English sarcasm which involves more satire, irony, and humor elements [41], for instance. Likewise, translating sarcasm from one language to another involves very sophisticated and nuanced expertise [42] which is hard to mimic and propagate, especially through machine translation.

Therefore, further experimentation is needed to examine the feasibility of adding synthetic datasets, in potential combination with weighted loss. Also, instead of translating sarcastic texts, generating sarcastic texts from scratch using LLMs is a plausible future research direction. However, as discussed in Section IV-B, current multilingual LLMs still very much struggle to understand the Indonesian language, much less generate sarcastic texts that adhere closely to the culture.

Sarcastic comments and tweets shown in Table 3 and Table 4, for example, require a prior understanding of Indonesia's religious upbringing, everyday interactions with traffic laws, and local food prices. These illustrate how culturally and locally informed a language model must be to correctly classify a text's sentiment. Based on this and the analysis done by [23], we similarly show the frequency and proportion of common Indonesian local words analyzed in the culturally-nuanced NusaWrites corpus [23] in both Reddit and Twitter sarcasm detection datasets in Table 8. As shown, the more challenging Reddit dataset contains more local words than the Twitter dataset.

D. CHALLENGES WITH INDONESIAN SARCASM DETECTION DATASETS

Despite basing our Reddit sarcasm detection dataset creation procedure on the methodology proposed in [15], there are potential improvements to the dataset. For instance, both the Reddit and Twitter datasets depend on special tags like `/s` and `#sarcasm` to be considered automatically as sarcastic texts. However, this doesn't guarantee that the content is actually sarcastic and requires further verification by an expert. This data curation step will not only be timely to develop but could also increase the expenses needed to curate such a high-quality dataset for a large amount of text.

Similarly, several of these texts found in the dataset often lack conversational context. While certain texts are independently sarcastic on their own, others require a more complete conversational history/context to confidently determine whether or not it is sarcastic [43]. The addition of a conversational history that a Reddit comment or a Twitter tweet is replying to could improve the robustness of the classification verdict.

Further, like many other studies related to Indonesian natural language processing (NLP), researchers who worked on Indonesian sarcasm detection often do not release the results of their data collection [44]. This greatly hinders the future progress of the field, which is precisely why we decided to release our sarcasm detection datasets as open-source for other researchers to leverage.

Finally, the datasets which we have released are limited to social media domains only. The usage of sarcasm in the Indonesian language goes beyond social media and includes other domains such as literary works (films, novels, songs), mass media (newspapers, television programs), and public environments (audible speech) [40]. A diverse range of text domains is necessary when designing a language understanding benchmark dataset to thoroughly examine a language model's capability [20].

In light of these findings and the complex nuance of sarcasm, we strongly encourage the future development of an expert human-annotated Indonesian sarcasm detection dataset, much like iSarcasm [21], [45]. Instead of relying on limited synthetic datasets and/or data augmentation techniques that are inadequate, linguistic experts in the Indonesian language can help write culturally nuanced and sentiment-enriched sarcastic texts across diverse domains [46] to combat the imminent data imbalance and increase the quality of the sarcastic corpus. Likewise, multimodal sarcasm detection can be further explored as multimodal language models are becoming more widespread [47].

TABLE 3. Data samples of reddit indonesia sarcasm dataset.

Comment
Sarcastic
Beda, ini inovasi terbaru, Apple kan king of innovation
(It's different, this is the latest innovation, Apple is king of innovation)
taat perintah tuhan : "kalau ada razia mendingan kabur"
(obey God's command: "if there is a raid, it's better to run away")
Not Sarcastic
Tpi yg ini beneran custom handcraft ya soalny?
(But this one is really a custom handcraft, right?)
Hubby: Ngga say, ini mas 10 menit lagi berangkat.
(Hubby: No darling. I'm leaving in 10 minutes.)

TABLE 4. Data samples of twitter indonesia sarcasm dataset.

Tweet
Sarcastic
Beli gorengan 80juta Kenyaaang :v <hashtag>
(Buy fried food for 80 million Fulll :v <hashtag>)
Pintar banget sih mbak ngehancurin mood seseorang
(You're really clever at ruining someone's mood)
Not Sarcastic
Bandel tuh diluar jgn dikandang!<hashtag>
(Stay outside if you're undisciplined, not caged!<hashtag>)
Biarkan dia bekerja Kita tunggu hasilnya aja . <hashtag>
(Let him work. Let's just wait for the results. <hashtag>)

V. CONCLUSION

In this study, we proposed the first publicly available Indonesian sarcasm detection benchmark datasets, one of which was originally curated by scraping Indonesian Reddit comments and the other by leveraging and carefully processing an existing Indonesian Twitter sarcasm dataset. Our study compared classical machine learning methods with pre-trained language models and novel multilingual large language models with zero-shot capabilities.

Our experiments show that pre-trained language models remain superior on sarcasm detection tasks and classical methods are preferred in low-resource settings. On the contrary, large language models are still unable to satisfactorily perform accurate zero-shot sarcasm detection. Further, synthetic data obtained through machine translation and weighted cross-entropy loss during fine-tuning are provably ineffective at combatting data imbalance. Whilst we acknowledge the limitations of the proposed semi-automatic data creation method, we hope to have faithfully developed a robust initial baseline to gauge the semantic capabilities of recent advances in language models.

APPENDIX A
DATA SAMPLES

We show several data samples of our Reddit Indonesia sarcasm dataset in Table 3 and Twitter Indonesia sarcasm dataset in Table 4.

TABLE 5. Hyperparameters for classical machine learning methods, to be further leveraged in a grid-based hyperparameter search. Values based on Table 9 of [25].

Hyperparameter	Naive Bayes	SVM	Logistic Regression
feature vectorizer	{Bag of Words, TF-IDF}		
alpha	(0.001 - 1)	-	-
C	-	(0.01 - 100)	(0.001 - 100)
kernel	-	{rbf, linear}	-

TABLE 6. Hyperparameters for fine-tuning pre-trained language models. Values based on and modified from Table 11 of [23].

Hyperparameter	Value
learning rate	1e-5
batch size	32
#epochs	100
early stop	3
optimizer	AdamW
AdamW β	(0.9, 0.999)
AdamW ϵ	1e-8
scheduler	cosine
weight decay	0.03
sequence length	128

TABLE 7. Zero-shot prompt templates for multilingual large language models.

Prompt Templates
{`text`} => Sarcasm: {`label`}
Text: {`text`} => Sarcasm: {`label`}
{`text`}\nIs this text above sarcastic or not? {`label`}
Is the following text sarcastic?\nText: {`text`}\nAnswer: {`label`}
Text: {`text`}\nPlease classify the text above for sarcasm. {`label`}

APPENDIX B
HYPERPARAMETERS

We provide the hyperparameters used in our experiments. Namely, the set of hyperparameter ranges used in classical machine learning methods and the subsequent grid search is provided in Table 5, while the pre-trained language models are fine-tuned with hyperparameters shown in Table 6.

APPENDIX C
ZERO-SHOT PROMPTS

We provide five zero-shot prompt templates used in our zero-shot inference experiments in Table 7, whose mean results are thus evaluated and compared against other methods used in this study. `text` is a placeholder variable for the current text being inferred, and `label` can either be "sarcastic" or "not sarcastic".

APPENDIX D
COMMON INDONESIAN LOCAL WORDS

Inspired by NusaWrites [23], we analyzed the frequency and proportion of common Indonesian local words in both Reddit and Twitter sarcasm detection datasets in Table 8. The same

TABLE 8. Common indonesian local words and their frequency and proportion in both reddit and twitter sarcasm detection datasets.

Topic	Word	Reddit		Twitter	
		freq	prop (%)	freq	prop (%)
food	indomie	37	0.0176	1	0.0020
	rendang	11	0.0052	0	0.0000
	tempe	10	0.0048	4	0.0078
	gule	0	0.0000	0	0.0000
	sate	9	0.0043	1	0.0020
transportation	angkot	11	0.0052	0	0.0000
	ojol	9	0.0043	0	0.0000
	gojek	16	0.0076	7	0.0137
religion	doa	17	0.0081	12	0.0235
	gaib	5	0.0024	0	0.0000
	alhamdulilah	1	0.0005	5	0.0098
	insyaallah	2	0.0010	1	0.0020
adjective	bule	38	0.0181	0	0.0000
	santun	4	0.0019	11	0.0216
	alay	4	0.0019	1	0.0020
Total		**174**	**0.0829**	**43**	**0.0844**

list of words is taken from the corresponding NusaWrites corpus analysis.

ACKNOWLEDGMENT

The authors would like to thank Bina Nusantara University for facilitating and supporting this entire research process.

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/tiga-karakter-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.

[10] Y. Yunitasari, A. Musdholifah, and A. K. Sari, "Sarcasm detection for sentiment analysis in Indonesian tweets," *Indonesian J. Comput. Cybern. Systems*, vol. 13, no. 1, pp. 53–62, Jan. 2019.

[11] P. Schiilkop, C. Burgest, and V. Vapnik, "Extracting support data for a given task," in *Proc. 1st Int. Conf. Knowl. Discovery Data Mining*, 1995, pp. 252–257.

[12] M. A. Rosid, D. Siahaan, and A. Saikhu, "Pre-trained word embeddings for sarcasm detection in Indonesian tweets: A comparative study," in *Proc. 9th Int. Conf. Inf. Technol., Comput., Electr. Eng.*, Aug. 2022, pp. 281–286.

[13] J. Lemmens, B. Burtenshaw, E. Lotfi, I. Markov, and W. Daelemans, "Sarcasm detection using an ensemble approach," in *Proc. 2nd Workshop Figurative Lang. Process.*, 2020, pp. 264–269.

[14] R. Misra and P. Arora, "Sarcasm detection using hybrid neural network," 2019, *arXiv:1908.07414*.

[15] K. S. Ranti and A. S. Girsang, "Indonesian sarcasm detection using convolutional neural network," *Int. J. Emerg. Trends Eng. Res.*, vol. 8, no. 9, pp. 4952–4955, 2020.

[16] (2023). *Reddit Comments/Submissions 2005-06 to 2023-09*. [Online]. Available: https://academictorrents.com/details/89d24ff9d5fbc1efcdaf9d7689d72b7548f699fc

[17] A. Z. Broder, "On the resemblance and containment of documents," in *Proc. Compress. Complex. Sequences*, 1997, pp. 21–29.

[18] A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov, "Bag of tricks for efficient text classification," 2016, *arXiv:1607.01759*.

[19] A. Joulin, E. Grave, P. Bojanowski, M. Douze, H. Jégou, and T. Mikolov, "FastText.Zip: Compressing text classification models," 2016, *arXiv:1612.03651*.

[20] S. Cahyawijaya, G. I. Winata, B. Wilie, K. Vincentio, X. Li, A. Kuncoro, S. Ruder, Z. Y. Lim, S. Bahar, M. Khodra, A. Purwarianti, and P. Fung, "IndoNLG: Benchmark and resources for evaluating Indonesian natural language generation," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2021, pp. 1–10.

[21] I. Abu Farha, S. V. Oprea, S. Wilson, and W. Magdy, "SemEval-2022 task 6: ISarcasmEval, intended sarcasm detection in English and Arabic," in *Proc. 16th Int. Workshop Semantic Eval.*, 2022, pp. 802–814.

[22] S. Khotijah, J. Tirtawangsa, and A. A. Suryani, "Using LSTM for context based approach of sarcasm detection in Twitter," in *Proc. 11th Int. Conf. Adv. Inf. Technol.*, Jul. 2020, doi: 10.1145/3406601.3406624.

[23] S. Cahyawijaya, H. Lovenia, F. Koto, D. Adhista, E. Dave, S. Oktavianti, S. M. Akbar, J. Lee, N. Shadieq, T. W. Cenggoro, H. W. Linuwih, B. Wilie, G. P. Muridan, G. I. Winata, D. Moeljadi, A. F. Aji, A. Purwarianti, and P. Fung, "NusaWrites: Constructing high-quality corpora for underrepresented and extremely low-resource languages," 2023, *arXiv:2309.10661*.

[24] M. N. Nityasya, H. A. Wibowo, R. E. Prasojo, and A. F. Aji, "Costs to consider in adopting NLP for your business," 2020, *arXiv:2012.08958*.

[25] G. I. Winata, A. F. Aji, S. Cahyawijaya, R. Mahendra, F. Koto, A. Romadhony, K. Kurniawan, D. Moeljadi, R. E. Prasojo, P. Fung, T. Baldwin, J. H. Lau, R. Sennrich, and S. Ruder, "NusaX: Multilingual parallel sentiment dataset for 10 Indonesian local languages," in *Proc. 17th Conf. Eur. Chapter Assoc. Comput. Linguistics*, 2023, pp. 815–834.

[26] J. S. Cramer, "The origins of logistic regression," *SSRN Electron. J.*, vol. 119, pp. 167–178, 2002.

[27] H. Zhang, "The optimality of naive Bayes," *Aa*, vol. 1, no. 2, p. 3, 2004.

[28] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and É. Duchesnay, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, Nov. 2011.

[29] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, vol. 1, Jun. 2019, pp. 4171–4186.

[30] F. Koto, A. Rahimi, J. Han Lau, and T. Baldwin, "IndoLEM and IndoBERT: A benchmark dataset and pre-trained language model for Indonesian NLP," 2020, *arXiv:2011.00677*.

[31] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, "Unsupervised cross-lingual representation learning at scale," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 8440–8451.

[32] T. Wolf et al., "Transformers: State-of-the-art natural language processing," in *Proc. Conf. Empirical Methods Natural Lang. Process. Syst. Demonstrations*, 2020, pp. 38–45.

[33] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Köpf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, *PyTorch: An Imperative Style, High-Performance Deep Learning Library*. Red Hook, NY, USA: Curran Associates Inc., 2019.

[34] J. Wei, M. Bosma, V. Y. Zhao, K. Guu, A. Wei Yu, B. Lester, N. Du, A. M. Dai, and Q. V. Le, "Finetuned language models are zero-shot learners," 2021, *arXiv:2109.01652*.

[35] W. Qi Leong, J. Gang Ngui, Y. Susanto, H. Rengarajan, K. Sarveswaran, and W. Chandra Tjhi, "BHASA: A holistic Southeast Asian linguistic and cultural evaluation suite for large language models," 2023, *arXiv:2309.06085*.

[36] N. Muennighoff, T. Wang, L. Sutawika, A. Roberts, S. Biderman, T. Le Scao, M. S. Bari, S. Shen, Z.-X. Yong, H. Schoelkopf, X. Tang, D. Radev, A. F. Aji, K. Almubarak, S. Albanie, Z. Alyafeai, A. Webson, E. Raff, and C. Raffel, "Crosslingual generalization through multitask finetuning," 2022, *arXiv:2211.01786*.

[37] A. Ebrahimi and K. Kann, "How to adapt your pretrained multilingual model to 1600 languages," in *Proc. 59th Annu. Meeting Assoc. Comput. Linguistics 11th Int. Joint Conf. Natural Lang. Process.*, Aug. 2021, pp. 4555–4567.

[38] X. Wang, S. Ruder, and G. Neubig, "Expanding pretrained models to thousands more languages via lexicon-based adaptation," 2022, *arXiv:2203.09435*.

[39] F. Koto, N. Aisyah, H. Li, and T. Baldwin, "Large language models only pass primary school exams in indonesia: A comprehensive test on IndoMMLU," 2023, *arXiv:2310.04928*.

[40] S. Syafruddin, A. Thaba, A. R. Rahim, M. Munirah, and S. Syahruddin, "Indonesian people's sarcasm culture: An ethnolinguistic research," *Linguistics Culture Rev.*, vol. 5, no. 1, pp. 160–179, Jun. 2021.

[41] H. Pratama, "Sarcasm as impoliteness device in Indonesian and American context," *Proc. English Lang. Teach., Literature, Transl.*, vol. 10, no. 1, pp. 38–42, Feb. 2022.

[42] R. Sukmaningrum, "The analysis of translation techniques of irony and sarcasm in novel entitled the sign of the four," *Eternal English Teach. J.*, vol. 7, no. 1, pp. 1–20, Oct. 2018.

[43] D. Ghosh, A. R. Fabbri, and S. Muresan, "Sarcasm analysis using conversation context," *Comput. Linguistics*, vol. 44, no. 4, pp. 755–792, Dec. 2018.

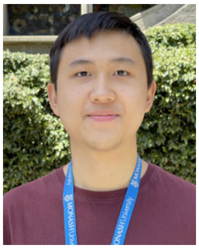
[44] A. F. Aji, G. I. Winata, F. Koto, S. Cahyawijaya, A. Romadhony, R. Mahendra, K. Kurniawan, D. Moeljadi, R. E. Prasojo, T. Baldwin, J. H. Lau, and S. Ruder, "One country, 700+ languages: NLP challenges for underrepresented languages and dialects in Indonesia," in *Proc. 60th Annu. Meeting Assoc. Comput. Linguistics*, 2022, pp. 7226–7249.

[45] M. A. Galal, A. Hassan Yousef, H. H. Zayed, and W. Medhat, "Arabic sarcasm detection: An enhanced fine-tuned language model approach," *Ain Shams Eng. J.*, vol. 15, no. 6, Jun. 2024, Art. no. 102736.

[46] W. Chen, F. Lin, G. Li, and B. Liu, "A survey of automatic sarcasm detection: Fundamental theories, formulation, datasets, detection methods, and opportunities," *Neurocomputing*, vol. 578, Apr. 2024, Art. no. 127428.

[47] H. Liu, R. Wei, G. Tu, J. Lin, C. Liu, and D. Jiang, "Sarcasm driven by sentiment: A sentiment-aware hierarchical fusion network for multimodal sarcasm detection," *Inf. Fusion*, vol. 108, Aug. 2024, Art. no. 102353.

DERWIN SUHARTONO (Member, IEEE) received the Ph.D. degree in computer science from Universitas Indonesia, in 2018. He is currently a Faculty Member of Bina Nusantara University, Indonesia. His research interest includes natural language processing. Recently, he is continually doing research in argumentation mining and personality recognition. He actively involves in Indonesia Association of Computational Linguistics (INACL), a national scientific association in Indonesia, IndoCEISS, and Aptikom. He has his professional memberships in IEEE, ACM, INSTICC, and IACT. He also takes role as reviewer in several international conferences and journals.



WILSON WONGSO received the Bachelor of Computer Science degree from Bina Nusantara University, Indonesia. He is currently a Machine Learning Engineer, specializing in natural and speech language processing. His research interests include natural language processing for low-resource languages and Indonesia-related languages.

ALIF TRI HANDOYO received the master's degree from Bina Nusantara University, in 2023. His research interests include natural language processing, image processing, and the Internet of Things.