Bulletin of Electrical Engineering and Informatics
Vol. 15, No. 2, April 2026, pp. 1379~1390
ISSN: 2302-9285, DOI: 10.11591/eei.v15i2.10935

Multitask deep learning for sentiment analysis with sarcasm detection in bilingual code-mixed social media content

Mohd Suhairi Md Suhaimin[1,2], Adi Wibowo[3], Ervin Gubin Moung[1], Patricia Anthony[4], Mohd Hanafi Ahmad Hijazi[1,5]

[1]Data Technology and Applications Research Group, Faculty of Computing and Informatics, Universiti Malaysia Sabah, Kota Kinabalu, Malaysia
[2]Politeknik Kota Bharu, Department of Polytechnic and Community College Education, Kota Bharu Kelantan, Malaysia
[3]Department of Computer Science, Universitas Diponegoro, Semarang, Indonesia
[4]Centre for Geospatial and Computing Technologies, Lincoln University, Lincoln, New Zealand
[5]Creative Advanced Machine Intelligence Research Centre, Faculty of Computing and Informatics, Universiti Malaysia Sabah, Kota Kinabalu, Malaysia

Article Info

Article history:

Received Jun 28, 2025
Revised Feb 4, 2026
Accepted Mar 5, 2026

Keywords:

Bilingual code-mixed
Deep learning
Hybrid feature engineering
Language model
Multitasking

ABSTRACT

Sentiment analysis in social media often hindered by sarcasm, which can reverse text meaning, and bilingual code-mixing, which adds complexity in non-English primary context. Existing approaches extract separate features for each language and translate them into a single language, resulting in the loss of contextual meaning and omission of crucial features. This paper proposes a multitask learning model for sentiment analysis with sarcasm detection tailored to bilingual code-mixed social media content. A hybrid feature engineering technique is integrated into a multitask deep learning architecture designed to capture the nuances of sentiment and sarcasm while addressing the complexities of processing bilingual code-mixed content. The hybrid technique combines domain-knowledge-based natural language processing (NLP) with a deep learning-based embedding approach. It includes rule-based preprocessing, normalization, spellchecking, feature extraction and selection, and feature representation. The engineered features are integrated into a multitask deep learning network using bidirectional long short-term memory (Bi-LSTM) combined with gated recurrent units (GRU). Using a public dataset that contains bilingual code-mixed social media content related to public security, our proposed model achieved a higher F1-score compared to two baseline models that employ single task and multitask approaches.

Corresponding Author:

Mohd Hanafi Ahmad Hijazi
Data Technology and Applications Research Group, Faculty of Computing and Informatics
Universiti Malaysia Sabah
Kota Kinabalu, Malaysia
Email: hanafi@ums.edu.my

1. INTRODUCTION

In the digital era, social media platforms have become the primary medium for expressing opinions, sharing experiences, and disseminating information. The domain has expanded beyond product review applications to areas such as stock markets, elections, disasters, healthcare, and software engineering. Sentiment analysis computationally classifies opinions expressed in text as either positive or negative [1]. Sentiment analysis and opinion mining are often used interchangeably, encompassing various tasks such as sentiment mining, sentiment classification, opinion extraction, subjectivity analysis, affect analysis, review

mining, entity extraction, and emotion analysis [1], [2]. It is particularly useful in social media for identifying the polarity of opinions expressed toward a particular topic, product, service, or event.

Despite its importance and widespread use, sentiment analysis faces significant challenges. A major issue is the use of sarcasm, which reverses the intended meaning of words, often for humor or irony [3]. For example, the social media expression "thanks to you, coronavirus, all things are messed up" may be misclassified as positive due to the presence of the word "thanks", despite the overall negative sentiment. This complexity is further compounded by the prevalence of bilingual code-mixed content in social media. In regions where English is not the primary language, users often blend their native language with English, creating complex linguistic structures that conventional analysis tools struggle to interpret accurately [4], [5]. For example, the expression: "Doesn't feel like Merdeka w all the chaos that's happening in the country, but nevertheless, Malaysia *Tanah Airku & jangan takut, kita #Lawan*!". It expresses frustration and disappointment about the current chaos in the country, which dampens the usual celebratory feeling of Merdeka (Malaysia's Independence Day). However, this is contradicted by a strong sense of patriotism and resilience, as shown by the phrases "Malaysia *Tanah Airku*" (Malaysia, my homeland) and "*jangan takut, kita #Lawan!*" (don't be afraid, we will fight/resist!). Detecting sarcasm within code-mixed text is critical, as its misinterpretation can lead to errors in sentiment analysis. Overcoming this challenge is important in critical domains such as public security, where accurate classification of sentiment is essential [6].

To address these challenges, this paper proposes a multitask learning approach designed to resolve ambiguities associated with sarcasm and improve the overall sentiment analysis. Using a two-task approach, the model can better understand the true sentiment behind the text by simultaneously detecting sarcasm and reducing misclassification caused by sarcastic expressions, thereby improving overall sentiment. Specifically, the approach focuses on bilingual code-mixed social media content in Roman (Latin) script. Hybrid feature engineering techniques and a multitask deep learning model were utilized to improve the accuracy of sentiment classification, while effective sarcasm detection was combined as a supportive task. Although this study experimented with the English-Malay language pair, the method and approach can be generalized to other bilingual language pairs with appropriate linguistic adjustments and preprocessing steps. The contributions of this work are: i) the development of a multitask deep learning model that performs both sentiment analysis and sarcasm detection, specifically designed for bilingual code-mixed social media content and ii) the introduction of a novel hybrid feature engineering technique which combines domain knowledge-based natural language processing (NLP) features with deep learning-based embeddings, to better capture the linguistic nuances in bilingual code-mixed text.

The remainder of the paper is organized as follows. Section 2 reviews related work. Section 3 describes the proposed approach. Section 4 outlines the experimental setup. Section 5 presents the result and analysis. Section 6 discusses the limitations of this work, and outlines directions for future work.

2. RELATED WORK

2.1. Features engineering for bilingual code-mixed content

Code-mixing, also referred as code-switching in inter-sentential, is a widespread linguistic phenomenon on social media, which occurs when bilingual or multilingual speakers alternate between two or more languages in conversation or expression [7]. This phenomenon poses unique challenges for sentiment analysis because traditional tools designed primarily for monolingual text often struggle with code-mixed content, leading to poor accuracy and a loss of contextual meaning [8].

Research in this area has predominantly explored various strategies for handling bilingual and code-mixed content. Some works have attempted to translate code-mixed text into a single dominant language before analysis [9]-[11]. However, this often results in the loss of contextual expressions and culturally specific content. To overcome this limitation, researchers have applied direct analysis of code-mixed text without translation, leveraging on attention mechanism using a deep-learning approach such as convolutional neural network (CNN) and bidirectional long short-term memory (Bi-LSTM) [8]. Researchers are increasingly advocating for advanced NLP tools that can understand and process multiple languages simultaneously. These tools leverage lexical features including tokens and syntactic features [12], such as parts-of-speech (POS) tags and specific patterns [9] that are unique to code-mixed language usage. Additionally, contextual features in form of named entity help to understand and utilize the broader context in which words appear [4]. All these features contribute to identifying sentiments and detecting sarcasm more effectively.

Feature engineering technique encompasses feature extraction, selection and transformation in creating higher order or new features in the dataset using domain knowledge [13]. One important technique is word embedding, where words are converted into dense vector representations, capturing semantic meanings across different languages. These embeddings often incorporate not just the words but also their context within the sentence, enhancing the model's ability to understand and interpret contextual meanings in code-

mixed contents [4], [12]. Attention mechanisms are another vital technique used to weigh the importance of different words in the text. By focusing on the most relevant parts of the input, the mechanisms improve the model's performance [14]. Complementing this, positional embeddings are employed to provide word position context, helping the model maintain the sequence and syntactic relationships between different languages within code-mixed sentences [15].

2.2. Multitask sentiment analysis with sarcasm detection model

Multitask learning involves the simultaneous training of a model on multiple related tasks, which can improve the performance and generalizability over single-task models [11], [16]. In sentiment analysis, this approach is particularly effective for related tasks like sentiment classification and sarcasm detection, which often share underlying linguistic features. Deep learning architectures typically implement this using shared layers to capture common knowledge and task-specific layers to handle individual nuances, resulting in improved overall performance.

Multitask network architectures include a variety of components that leverage the strengths of different neural network architecture. Typically, these architectures use bidirectional recurrent neural networks (RNN), such as gated recurrent units (GRU) or long short-term memory (LSTM). These structures are crucial for modeling the sequential nature of text content and capturing contextual information bidirectionally. Attention mechanisms are frequently integrated within these models to enhance the focus on relevant segments of the text. By weighing the importance of different words or phrases during the learning process, these mechanisms allow the model to develop a more subtle understanding of the text, which is particularly beneficial for distinguishing between literal and sarcastic tones [17].

The utilization of pre-trained models like bidirectional encoder representations from transformers (BERT) [18] and its extension within multitask frameworks has become increasingly popular, Leveraging the deep contextual embeddings generated by these transformer-based can enhance the model's ability to decode subtle and context-dependent meanings in text [19]. Furthermore, large language models (LLMs) such as generative pre-trained transformers (GPTs) have recently been tested within multitask pipelines for sentiment and sarcasm detection, showing moderate performance gains in low-resource settings [20].

As multitask models become increasingly complex, the integration of sophisticated fusion techniques such as neural tensor networks or simple concatenation followed by dense layers becomes essential. These techniques ensure that features from various network components are effectively synthesized to enhance classification and detection tasks [17]. The optimization of these models involves advanced strategies such as the Adam or RMSprop optimizers [21], which adaptively adjust learning rates for different parameters. Building on established multitask learning techniques, our approach integrates these key optimization strategies to effectively handle the complexities of sentiment and sarcasm detection in bilingual code-mixed data.

Multitask learning has also been leveraged in other domains such as biometrics. Xu *et al*. [22] pre-train a shared-weight network on soft palmprint attributes (gender and chirality) and transfer that knowledge to identity recognition to reduce false matches, showing that an auxiliary task can regularize the primary task and improve accuracy. Conceptually, this parallels our use of sarcasm detection to reduce "false sentiment matches". More recently, [23] propose an enhanced multitask learning framework with a classification branch (identity, gender, chirality) and a hashing branch, augmented by channel attention, a customized gate control module to balance shared vs. task-specific experts, and automatic loss-weight adjustment. Although the domain differs, these mechanisms of attention, gating, and dynamic task weighting are general multitask learning techniques and inform our architecture choices.

3. PROPOSED APPROACH

In addressing the challenges of sentiment classification and sarcasm detection in bilingual code-mixed text, we developed an approach that integrates hybrid feature engineering with a multitask deep learning architecture. This section details the two core components of the proposed method.

3.1. Hybrid feature engineering technique based on natural language processing

The proposed hybrid feature engineering technique combines domain-knowledge-based NLP feature technique with a deep learning-based approach. The domain-knowledge-based NLP processes involve tokenization, part-of-speech (POS) tagging, stopword removal, and additional preprocessing steps required for inferences and feature extractions. The deep learning-based approach restructures feature representation and embedding into an interconnected preprocessing step that directly feeds into a neural network architecture. This approach comprises dense embeddings and multiple hidden layers. Figure 1 shows the hybrid feature engineering technique for the bilingual code-mixed content. The technique involves three stages: i) rule-based preprocessing and normalization, ii) feature extraction and selection, and iii) feature representation.

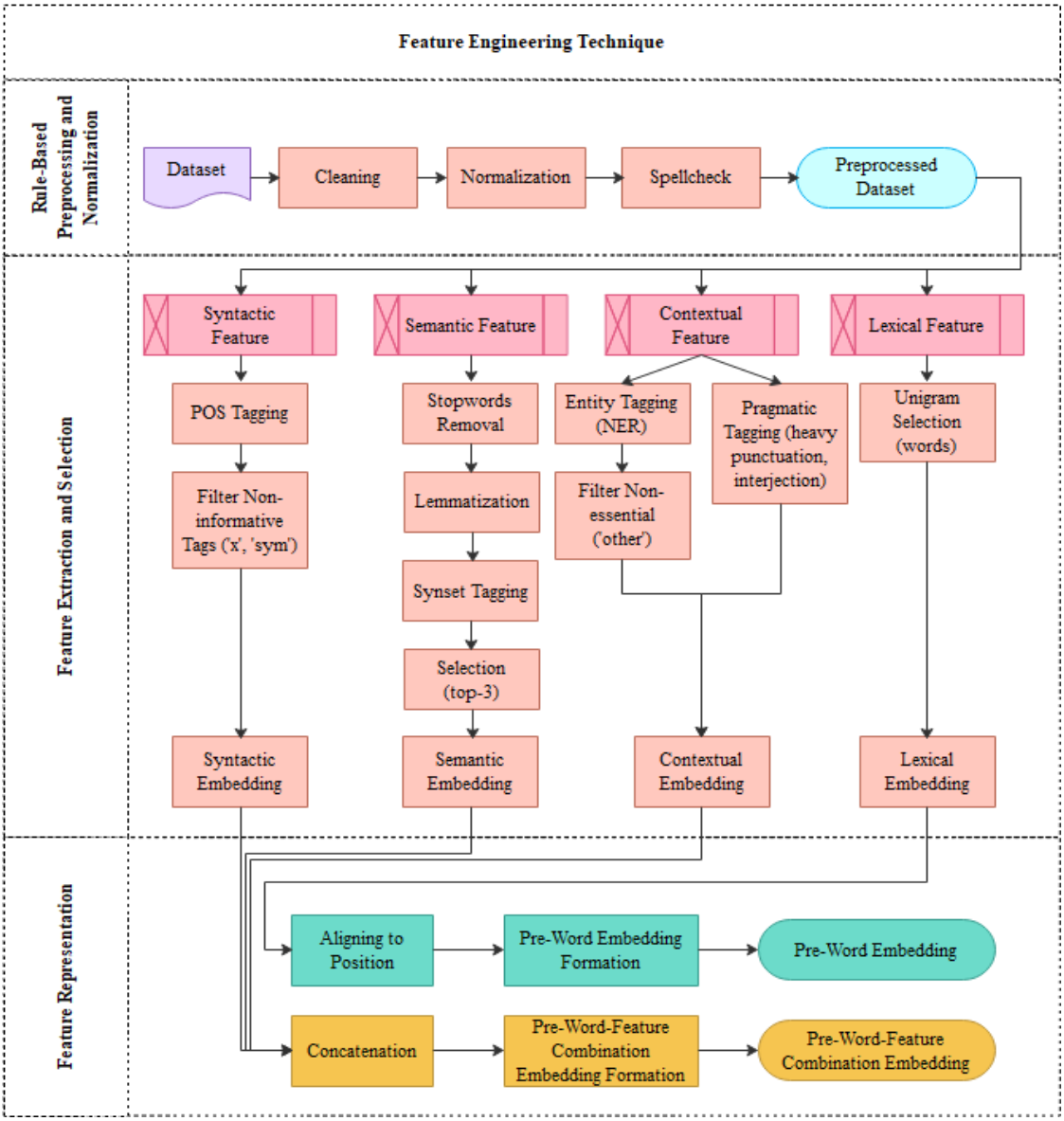

Figure 1. Hybrid feature engineering technique for bilingual code-mixed content

3.1.1. Rule-based preprocessing and normalization

The initial stage involves cleaning, normalizing, and spellchecking the raw dataset to reduce noise and improve consistency. The cleaning process removes punctuation, spaces, and line breaks, but preserves the context of hashtags by only removing the '#' symbol. Normalization then standardizes the text by converting it to lowercase and replacing URLs, usernames, emails, emojis, numbers, and other entities with generic tags (e.g., <url> and <user>). Finally, a bilingual transformer model corrects spelling errors and informal words, ensuring a cleaner dataset for subsequent stages.

3.1.2. Feature extraction and selection

In the second stage, features that differentiate sentiment and sarcasm are extracted. Four types of NLP-based features are used, each uniquely contributing to the nuanced understanding of sentiment and sarcasm in bilingual code-mixed text:

a. Lexical

Lexical features are used to capture the surface-level meaning and sentiment of texts by analyzing word choice [19], [21]. To achieve this, we extract unigram words from the preprocessed dataset. These features provide the foundational word-by-word insights necessary for sentiment classification in complex, bilingual code-mixed content.

b. Syntactic

Syntactic features are used to capture how sentence structure affects meaning, as certain patterns can signal the emphasis or irony crucial for sarcasm detection [24]. To achieve this, POS was used to tag each word with its grammatical role. This process helps identify sentence structures commonly associated with sarcastic expressions. A filter method is then applied to remove non-informative tags (e.g., 'x' for unknown and 'sym' for symbols) for a cleaner, more efficient feature set [13].

c. Semantic

Semantic features are used to understand the meaning and context of words, which is crucial for discerning sentiment nuances and disambiguating complex language that syntax alone cannot resolve [25]. To capture this meaning, we use synonym sets ('synsets'). The process involves removing stopwords from the preprocessed text, lemmatizing each word, and then tagging it with its corresponding synsets from bilingual language databases. A filter method is then applied to select the top three synsets for each word, balancing semantic richness with practical feature sparsity [13].

d. Contextual

Contextual features are used to capture the pragmatic and entity-related aspects of a sentence, such as tone and cultural references. This is essential for identifying sarcasm and emotional undertones that are not explicitly stated in the text [4]. To generate these features, two techniques were used. First, named entity recognition (NER) is applied to identify and classify entities, while non-essential tags like 'other' are removed. Second, pragmatic features like heavy punctuation (e.g., "!!", "??") and interjections are extracted using predefined mappings and bilingual vocabularies. All selected features are then used as input for the subsequent stage.

3.1.3. Feature representation

The features selected in the previous stage are transformed into two non-sparse embedding formats: i) pre-word embedding and ii) pre-word-feature combination embedding. This hybrid technique ensures that the combined lexical, syntactic, semantic, and contextual features effectively preserve the nuances required for analyzing sentiment and sarcasm.

a. Pre-word embedding

This format creates a foundational dense representation by projecting words into a continuous vector space where semantically similar words are positioned closely. To form this embedding, unigram tokens from the preprocessed content are aligned with their original positions. This alignment ensures the subtleties captured by individual NLP features are retained during the vector projection.

b. Pre-word-feature combination embedding

This format extends the embedding by concatenating syntactic, semantic and contextual, to include word-feature combinations, providing a more nuanced input for sentiment and sarcasm analysis. Each NLP feature type is concatenated with the word and preserved in a specific format:

- Syntactic: "word_POS"
- Semantic: "synonym1 synonym2 synonym3"
- Contextual (Entity): "word_NER"
- Contextual (Pragmatic): "punct(n)" for punctuation and "intj(n)" for interjections, where 'n' is the category number.

This word-feature combination ensures the model receives a rich, context-aware input from our hybrid feature engineering technique. Table 1 shows examples of both embedding formations. All representations are aligned with their original content and prepared for input into the multitask deep learning model.

Table 1. Example of feature representation

Original content (raw)	Pre-embedding formation	Feature representation
"*Ceih,... Dkt sabah nie relax jer tiada air ... Aq rasa klu org Penang mai dkt sini, confirm hari2 ada demonstrasi* https://t.co/LPujI9GCU4"	Word embedding	Lexical *ceh dekat sabah ini relax sahaja tiada air aku rasa kalau orang penang datang dekat sini, sah hari-hari ada demonstrasi* <url>
	Word-feature combination embedding	Syntactic ceh_propn dekat_adj sabah_noun ini_det relax_noun sahaja_adv tiada_part air_noun aku_pron rasa_noun kalau_sconj orang_det penang_propn datang_verb dekat_adp sini_noun, punct sah_adj hari-hari_noun ada_verb demonstrasi_noun url_noun Semantic Sabah borneo utara british borneo utara slack up unbend loosen up keseorangan hanya satu sifar kosong gentle wind melody broadcast perasaan sensasi tiba di sini nilai harga kualiti tunjuk perasaan Contextual intj86 sabah_location penang_location

3.2. Hybrid feature engineering technique based on natural language processing

The proposed multitask deep learning model uses a parallel feature fusion architecture to handle the complexity of sentiment classification and sarcasm detection in bilingual, code-mixed contexts. As shown in Figure 2, the structure comprises four main layers: i) input embedding layer, ii) shared multilayer network, iii) transformation layer, and iv) task-specific layer.

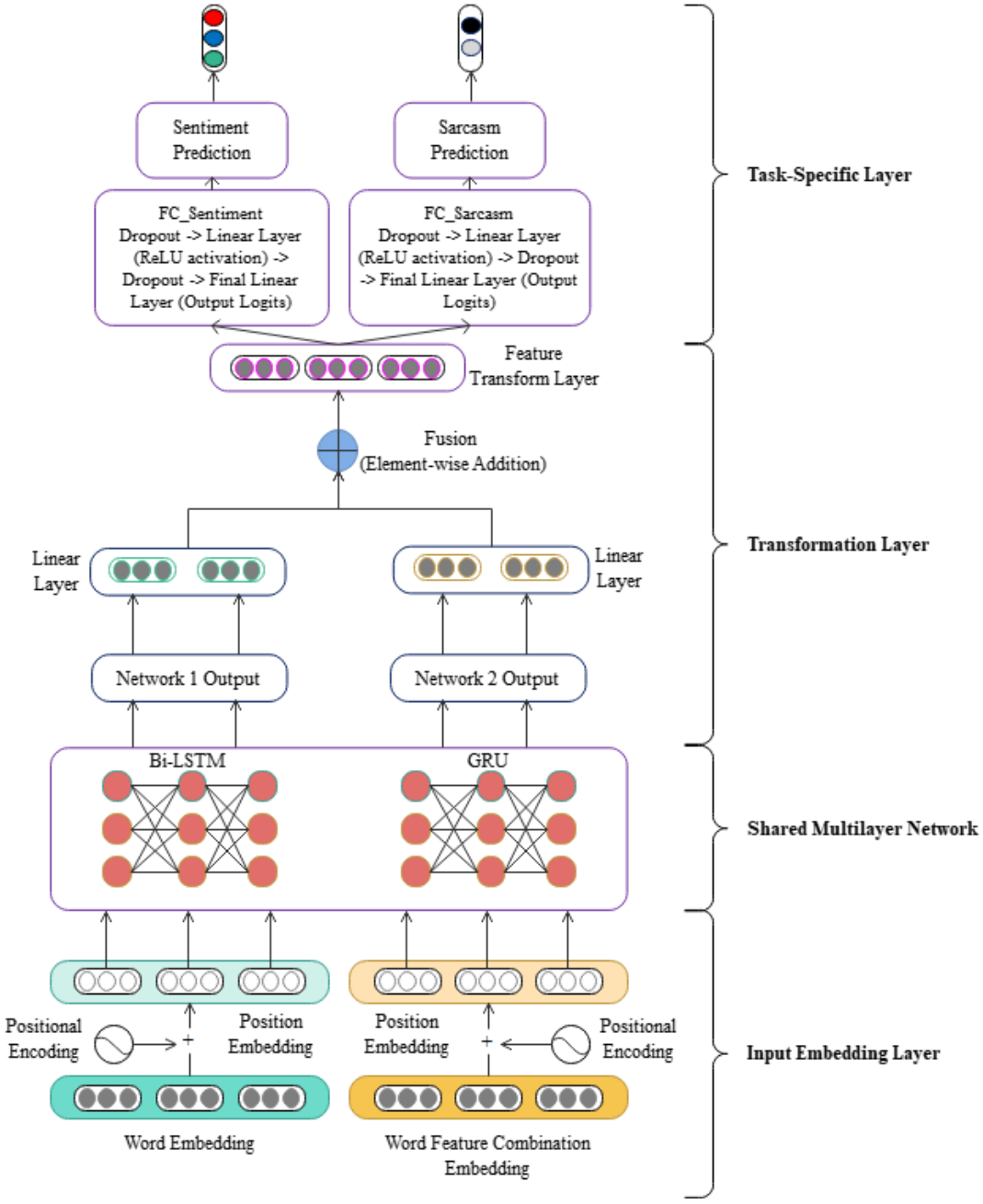

Figure 2. Multitask deep learning for sentiment classification with sarcasm detection model

3.2.1. Input embedding layer

This layer transforms input data into a structured format. It includes two parallel embedding types: word embeddings (green) and word-feature combination embeddings (yellow). They were designed to encode rich linguistic features. Positional encoding is applied to both stream types to ensure the model captures not just linguistic attributes but also sequence order. The resulting sequences are processed in parallel: lexical word embeddings are fed to a Bi-LSTM pathway and word-feature combination embeddings

are fed to a GRU pathway (from Figure 1). These enriched features are then fed in parallel into the shared multilayer network.

3.2.2. Shared multilayer network

As the core of the architecture, the shared network comprises two encoders that learn complementary information: a Bi-LSTM over the lexical stream and a GRU over the word-feature combination stream. Each pathway produces a fixed-length vector via temporal pooling over the encoder states. These vectors are then passed to the transformation layer for dimensional alignment and fusion, allowing the model to learn from multiple aspects of the input and capture patterns vital for sentiment and sarcasm.

3.2.3. Transformation layer

Following the shared network, the transformation layer aligns the two representations and performs fusion ('Network 1 Output' and 'Network 2 Output'). First, the Bi-LSTM vector is passed through a linear projection to match the dimensionality of the GRU output so that fusion is well-defined. The two vectors are then fused by element-wise addition (labeled "Element-wise Addition"). The fused vector passes through a feature transform layer (Linear+ReLU) before being fed to the task-specific heads. This step unifies the diverse features into a single representation for downstream prediction.

3.2.4. Task-specific layer

The final stage consists of two task-specific fully connected heads: FC_Sentiment and FC_Sarcasm. Each head is implemented as an nn.Sequential with the same structure: Dropout → Linear layer (ReLU activation) → Dropout → Final linear (Logits). Separating the tasks at this level enables specialized learning while leveraging the shared representation produced by the preceding layers.

4. EXPERIMENTAL SETUP

The objective of the experiment is to evaluate the performance of various features when integrated into the multitask deep learning model.

4.1. Experiment setting

All experiments were carried out on a system running Ubuntu Linux 22.04, with an Intel i7 CPU clocked at 3.60 GHz and 64 GB of DDR4 RAM. The programming environment was set up using Python Anaconda 2023 distribution. Model training utilized a Nvidia Geforce RTX 4090 GPU with 24 GB of DDR6X memory, facilitated by the PyTorch 2.0 deep learning framework.

4.2. Baseline

Two baseline model categories were used: single-task (for sentiment or sarcasm) and multitask. Both used a simple Bi-LSTM network with tokenized raw content as features. We used a public dataset of bilingual (English-Malay) code-mixed comments from X and TikTok. The data was annotated by three experts, achieving acceptable Fleiss's kappa scores of 0.46 (sentiment) and 0.42 (sarcasm) [26]. To address class imbalance in the original 10,000 comments, we applied selective oversampling [27]. The data distribution before and after this process is as follows:

Sentiment:

- Original: 3,072 positive, 4,197 negative, 2,569 neutral
- Oversampled: 4,965 positive, 5,888 negative, 5,138 neutral

Sarcasm:

- Original: 2,355 sarcastic, 7,645 non-sarcastic
- Oversampled: 5,906 sarcastic, 10,085 non-sarcastic

4.3. Proposed model configuration

The proposed multitask model is designed to process a rich set of features generated through a hybrid feature engineering technique. The key libraries and tools employed for preprocessing and feature extraction were: for preprocessing and spellchecking, we employed several pretrained Malaya language models, including the Wiki-News model, Dump Combine, and LLM GPT-2. For feature extraction, we engineered four distinct categories. Syntactic features (POS tags and NER) were generated using the Malaya (xlnet-base) library. Semantic features were created using synonym sets from English WordNet and Bahasa WordNet, while contextual features were compiled from public source.

Training was conducted for 30 epochs with a batch size of 64. The model used a balanced loss function of categorical cross-entropy for the sentiment task and binary cross-entropy for the sarcasm task. The network was optimized using the RMSprop algorithm with a learning rate scheduler (step size=5 and gamma=0.7) to ensure stable convergence. The rectified linear unit (ReLU) was used as the activation function, and dropout was applied in the fully connected layers to prevent overfitting. The optimal hyperparameters for embedding dimensions, dropout rate, and learning rate were determined through a targeted grid search.

4.4. Evaluation metric

To evaluate the performance of the proposed model, the F1-score (*F1*) is used as the evaluation metric. This aligns with the prevalent evaluation approach in recent works on multitask deep learning for sentiment analysis with sarcasm detection.

5. RESULT AND ANALYSIS

5.1. Proposed features and model result

Table 2 compares the performance of our proposed model against the single-task and multitask baselines. From Table 2, the analysis highlights several key findings:

- Superior performance of proposed model: our model achieved the highest F1-scores, with 0.8629 in sentiment and 0.9226 in sarcasm detection. This success is attributed to the rich, hybrid feature set (lexical, syntactic, semantic, and contextual), which provides deeper linguistic insight than the simple word embeddings used by the baselines.
- Effectiveness of multitask learning: the multitask baseline consistently outperformed the single-task baseline on both tasks, demonstrating the benefit of jointly learning sentiment and sarcasm, which is consistent with prior work [11], [16].
- Task difficulty observation: in all experiments, sarcasm detection achieved higher F1-scores than the three-class sentiment classification, indicating the inherent challenge of nuanced sentiment analysis in text containing sarcasm [3].

Table 2. Proposed model result

Model	Task	F1-score (F1)
Single-task baseline	Sentiment	0.8239
	Sarcasm	0.8732
Multitask baseline	Sentiment	0.8438
	Sarcasm	0.9153
Our proposed model	Sentiment	**0.8629**
	Sarcasm	**0.9226**

5.2. Ablation study

To understand the contribution of different components within the hybrid feature engineering technique, an ablation study was conducted using the proposed multitask model. Table 3 includes results from different settings, showing how various transformer-based language models (Wiki News, Dump Combine, and GPT2) and feature combinations (Lexical+Syntactic, Lexical+Syntactic+Semantic, and Lexical+Syntactic+Semantic+Contextual) impact the model's performance.

Table 3. Features in model comparison results

	Sentiment	Sarcasm	Sentiment	Sarcasm	Sentiment	Sarcasm
Rule-based+normalization+spellcheck (Wiki News)	Lexical+Syntactic		Lexical+Syntactic+Semantic		Lexical+Syntactic +Semantic+Contextual	
	F1: 0.8566	F1: 0.9090	F1: 0.8577	F1: 0.9119	F1: 0.8629	F1: 0.9226
Rule-based+normalization+spellcheck (Dump Combine)	Lexical+Syntactic		Lexical+Syntactic+Semantic		Lexical+Syntactic +Semantic+Contextual	
	F1: 0.8547	F1: 0.9109	F1: 0.8389	F1: 0.8987	F1: 0.8541	F1: 0.9175
Rule-based+normalization + spellcheck (GPT2)	Lexical+Syntactic		Lexical+Syntactic+Semantic		Lexical+Syntactic +Semantic+Contextual	
	F1: 0.8484	F1: 0.9154	F1: 0.8572	F1: 0.9151	F1: 0.8540	F1: 0.9156

The top performance was achieved by the model that employed all features (Lexical+Syntactic+Semantic+Contextual) and utilized the Wiki News library for spellchecking, reaching peak F1-scores of 0.8629 for sentiment and 0.9226 for sarcasm. This demonstrates that the integration of rich, multi-faceted features is critical for capturing the complexities of sentiment and sarcasm. To provide a granular understanding of the factors contributing to this strong performance, the following points detail the key findings.

5.2.1. Library (transformer-based language model) impact in preprocessing stage

The ablation study reveals that the Wiki News language model consistently outperforms the Dump Combine and GPT2 models across all feature combinations, particularly in sarcasm detection. This suggests that the pretraining domain of the transformer-based model (news-related data in Wiki News) is more aligned with the sentiment and sarcasm detection tasks in the bilingual, code-mixed text, which might be more formal or structured similarly to news content.

5.2.2. Feature combination impact in representation stage and model structure

The study highlights that the addition of contextual features to the lexical, syntactic, and semantic features significantly boosts performance, especially in sarcasm detection. The F1-scores improved when contextual features are included, underscoring their importance in capturing the nuanced and often implicit meanings in sarcastic statements.

5.2.3. Overall contribution

The ablation study clearly demonstrates the importance of both the choice of language model and the comprehensive feature set in achieving high performance. The combination of the Wiki News model and the full set of features (Lexical+Syntactic+Semantic+Contextual) achieve the highest F1-scores for both sentiment and sarcasm tasks, validating the efficacy of proposed hybrid approach and multitask model.

6. CONCLUSION

This paper proposes a multitask deep learning model for sentiment analysis with sarcasm detection for bilingual code-mixed social media content. Utilizing hybrid features engineering technique integrated into a multitask deep learning architecture combining Bi-LSTM and GRU networks, our approach demonstrates superior results compared to the single-task baseline. The combination of all proposed NLP-based features (lexical, syntactic, semantic, and contextual) in the developed multitask model also outperforms the raw input features in the multitask baseline. The study focuses on bilingual code-mixed content in the Roman (Latin) script, with specific experiments conducted on the English-Malay language pair. On comparisons with broader state-of-the-art, our model addresses a novel configuration; joint sentiment and sarcasm modeling on bilingual code-mixed English–Malay social media using a new, public dataset—for which, to our knowledge, no directly comparable published model exists. Applying monolingual or single-task models "as-is" would require substantial, non-trivial adaptation (code-mix handling, dual heads, and task-loss balancing) and would not constitute a like-for-like benchmark. Accordingly, our results establish a first strong, publicly reported baseline for this specific task.

While our approach shows promising results, the model's generalizability to other bilingual language pairs has not been fully explored. The model may require further adjustments and linguistic expertise to handle different language pairs effectively. This limitation highlights the need for broader testing across diverse languages and code-mixed content types. For future work, we plan to extend the model's capabilities by experimenting with other bilingual language pairs (including non-Roman scripts), collaborating with language experts to fine-tune the pre-processing and feature extraction for various languages.

FUNDING INFORMATION

This research was funded by the Ministry of Higher Education Malaysia, Fundamental Research Grant Scheme (FRGS), grant number FRGS/1/2022/ICT02/UMS/02/3 and SDPA1.0. This research was supported in 2025 by the World Class University (WCU) Program, Faculty of Science and Mathematics, Universitas Diponegoro, Indonesia.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Mohd Suhairi Md Suhaimin	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓		✓	
Adi Wibowo					✓					✓				
Ervin Gubin Moung	✓		✓	✓			✓			✓		✓	✓	
Patricia Anthony				✓	✓					✓				
Mohd Hanafi Ahmad Hijazi	✓	✓		✓	✓	✓		✓	✓	✓	✓	✓	✓	✓

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

DATA AVAILABILITY

The data that support the findings of this study are openly available in Zenodo at https://doi.org/10.5281/zenodo.11642494.

REFERENCES

[1] B. Liu, *Sentiment analysis: Mining opinions, sentiments, and emotions*. Cambridge University Press, 2020.
[2] N. Gupta and R. Agrawal, "Application and techniques of opinion mining," in *Hybrid Computational Intelligence*: Elsevier, 2020, pp. 1-23.
[3] M. Lydiri, Y. El Mourabit, Y. El Habouz, and M. Fakir, "A performant deep learning model for sentiment analysis of climate change," *Social Network Analysis and Mining*, vol. 13, no. 1, 2023, doi: 10.1007/s13278-022-01014-3.
[4] D. Bansal, R. Grover, N. Saini, and S. Saha, "GenSumm: A Joint Framework for Multi-task Tweet Classification and Summarization using Sentiment Analysis and Generative Modelling," *IEEE Transactions on Affective Computing*, vol. 15, no. 4, pp. 1838-1855, Oct.-Dec. 2024, doi: 10.1109/TAFFC.2021.3131516.
[5] H. Y. Lin and T. S. Moh, "Sentiment analysis on COVID tweets using COVID-Twitter-BERT with auxiliary sentence approach," in *Proceedings of the 2021 ACMSE Conference - ACMSE 2021: The Annual ACM Southeast Conference*, 2021, pp. 234-238, doi: 10.1145/3409334.3452074.
[6] M. S. M. Suhaimin, M. H. A. Hijazi, E. G. Moung, P. N. E. Nohuddin, S. Chua, and F. Coenen, "Social media sentiment analysis and opinion mining in public security: Taxonomy, trend analysis, issues and future directions," *Journal of King Saud University-Computer and Information Sciences*, vol. 35, no. 9, 2023, doi: 10.1016/j.jksuci.2023.101776.
[7] A. Pratapa, G. Bhat, M. Choudhury, S. Sitaram, S. Dandapat, and K. Bali, "Language modeling for code-mixing: The role of linguistic theory based synthetic data," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2018, pp. 1543-1553.
[8] R. R. Frias, R. P. Medina, and A. S. Sison, "Attention-based Bilateral LSTM-CNN for the Sentiment Analysis of Code-mixed Filipino-English Social Media Texts," in *2023 International Conference on Digital Applications, Transformation & Economy (ICDATE)*, Miri, Sarawak, Malaysia, 2023, pp. 1-5, doi: 10.1109/ICDATE58146.2023.1024892.
[9] P. A. R. Azmi, A. W. Z. Abidin, S. Mutalib, I. S. M. Zawawi, and S. A. Halim, "Sentiment Analysis on MySejahtera Application during COVID-19 Pandemic," in *2022 3rd International Conference on Artificial Intelligence and Data Sciences (AiDAS)*, 7-8 Sept. 2022, pp. 215-220, doi: 10.1109/AiDAS56890.2022.9918748.
[10] L. M. M. Silva, C. R. Valêncio, G. F. D. Zafalon, and A. C. Columbini, "Feature Selection with Hybrid Bio-inspired Approach for Classifying Multi-idiom Social Media Sentiment Analysis," in *International Conference on Enterprise Information Systems, ICEIS - Proceedings*, 2022, vol. 1, pp. 297-307, doi: 10.5220/0010972800003179.
[11] S. Chen, Y. Zhang, and Q. Yang, "Multi-task learning in natural language processing: An overview," *ACM Computing Surveys*, vol. 56, no. 12, pp. 1-32, 2024, doi: 10.1145/3663363.
[12] H. R. Utomo and A. Romadhony, "Sentiment Analysis on Indonesia-English Code-Mixed Data," in *2023 IEEE 8th International Conference for Convergence in Technology (I2CT)*, Lonavla, India, 2023, pp. 1-6, doi: 10.1109/I2CT57861.2023.10126234.
[13] M. A. Haque, "Feature Engineering & Selection for Explainable Models A Second Course for Data Scientists," *LULU Internacional*, 2022.
[14] N. Alturayeif, H. Luqman, and M. Ahmed, "Enhancing stance detection through sequential weighted multi-task learning," *Social Network Analysis and Mining*, vol. 14, no. 1, 2023, doi: 10.1007/s13278-023-01169-7.
[15] Mamta and A. Ekbal, "Transformer based multilingual joint learning framework for code-mixed and english sentiment analysis," *Journal of Intelligent Information Systems*, vol. 62, no. 1, pp. 231-253, 2024, doi: 10.1007/s10844-023-00808-x.
[16] S. Ruder, "An overview of multi-task learning in deep neural networks," *arXiv preprint*, 2017, doi: 10.48550/arXiv.1706.05098.
[17] N. Majumder, S. Poria, H. Peng, N. Chhaya, E. Cambria, and A. Gelbukh, "Sentiment and Sarcasm Classification with Multitask Learning," *IEEE Intelligent Systems*, vol. 34, no. 3, pp. 38-43, 2019, doi: 10.1109/MIS.2019.2904691.

[18] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint*, 2018, doi: 10.48550/arXiv.1810.04805.
[19] F. Molavi and J. B. Mohasefi, "BERT Transformers Multitask Learning Sarcasm and Sentiment Classification (BMSS)," in *2023 13th International Conference on Computer and Knowledge Engineering (ICCKE)*, 1-2 Nov. 2023 2023, pp. 515-520, doi: 10.1109/ICCKE60553.2023.10326244.
[20] S. R. Friðriksdóttir, A. Simonsen, A. S. Ásmundsson, G. L. Friðjónsdóttir, A. K. Ingason, V. Snæbjarnarson, and H. Einarsson, "Ice and Fire: Dataset on Sentiment, Emotions, Toxicity, Sarcasm, Hate speech, Sympathy and More in Icelandic Blog Comments," in *Proceedings of the Fourth Workshop on Threat, Aggression & Cyberbullying@ LREC-COLING-2024*, 2024, pp. 73-84.
[21] Y. Y. Tan, C. O. Chow, J. Kanesan, J. H. Chuah, and Y. L. Lim, "Sentiment Analysis and Sarcasm Detection using Deep Multi-Task Learning," *Wireless Personal Communications*, vol. 129, no. 3, pp. 2213-2237, 2023, doi: 10.1007/s11277-023-10235-4.
[22] H. Xu, L. Leng, Z. Yang, A. B. J. Teoh, and Z. Jin, "Multi-task pre-training with soft biometrics for transfer-learning palmprint recognition," *Neural Processing Letters*, vol. 55, no. 3, pp. 2341-2358, 2023, doi: 10.1007/s11063-022-10822-9.
[23] L. Chen, L. Leng, Z. Yang, and A. B. J. Teoh, "Enhanced multitask learning for hash code generation of palmprint biometrics," *International Journal of Neural Systems*, vol. 34, no. 4, 2024, doi: 10.1142/S0129065724500205.
[24] M. M. Bala, M. S. Rao, and M. R. Babu, "Sentiment trends on natural disasters using location based twitter opinion mining," *International Journal of Civil Engineering and Technology*, vol. 8, no. 8, pp. 9-19, 2017.
[25] J. K. Chandra, E. Cambria, and A. Nanetti, "One Belt, One Road, One Sentiment? A Hybrid Approach to Gauging Public Opinions on the New Silk Road Initiative," in *IEEE International Conference on Data Mining Workshops, ICDMW*, 2020, vol. 2020-November, pp. 7-14, doi: 10.1109/ICDMW51313.2020.00011.
[26] T. Ptáček, I. Habernal, and J. Hong, "Sarcasm detection on czech and english twitter," in *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, 2014, pp. 213-223.
[27] S. Behl, A. Rao, S. Aggarwal, S. Chadha, and H. S. Pannu, "Twitter for disaster relief through sentiment analysis for COVID-19 and natural hazard crises," *International Journal of Disaster Risk Reduction*, vol. 55, 2021, doi: 10.1016/j.ijdrr.2021.102101.

BIOGRAPHIES OF AUTHORS

Mohd Suhairi Md Suhaimin received his Master's and Ph.D. degrees in computer science from the Universiti Malaysia Sabah (UMS). He is a lecturer from Polytechnic and Community College, Ministry of Higher Education Malaysia. He is currently a co-researcher at the Data Technology and Applications Research Group, UMS. He has published several articles and has served as a reviewer for reputable journals. His research interests include multitask deep learning, natural language processing, large language models, and face recognition, specifically their intersection, and contribution to artificial intelligence applications. He can be contacted at email: suhairisuhaimin@pkb.edu.my or mohd_suhairi_di21@iluv.ums.edu.my.

Adi Wibowo received the Bachelor's degree in mathematics in 2005 from Universitas Diponegoro, and Master of Computer Science in 2011 from Universitas Indonesia. He received his Ph.D. degree in Engineering from Nagoya University in 2016. He is currently an Associate Professor in Department of Informatics, Universitas Diponegoro, Semarang. His research interests include artificial intelligence, computer vision, and data science. He can be contacted at email: bowo.adi@live.undip.ac.id.

Ervin Gubin Moung received his Bachelor's degree in computer engineering in 2008, Master of Computer Engineering in 2013, and Ph.D. in Computer Engineering from Universiti Malaysia Sabah (UMS) in 2018. He is a Senior Lecturer at the Faculty of Computing and Informatics at UMS and has served as the Deputy Director of the Research Management Centre at UMS since 2022. His research interests include public health, smart health, agriculture, food security, biodiversity, and environmental sustainability. He can be contacted at email: ervin@ums.edu.my.

Patricia Anthony received the Ph.D. degree from University of Southampton, in 2003. She is an Associate Professor in the Faculty of Environment, Society and Design at Lincoln University, New Zealand. Her research interests are in agents and multi-agent systems, machine learning for natural language processing and text analytics, and the applications of artificial intelligence in agriculture and environmental systems. She has published over 100 papers in peer-reviewed journals and international conferences. She can be contacted at email: patricia.anthony@lincoln.ac.nz.

Mohd Hanafi Ahmad Hijazi received his B.Sc. and M.Sc. degrees in computer science from Universiti Teknologi Malaysia, in 2001 and 2005 and the Ph.D. degree in computer science from the University of Liverpool, United Kingdom, in 2012. From 2012 to 2018, he was a Senior Lecturer with the Faculty of Computing and Informatics, Universiti Malaysia Sabah. Since 2018, he has been an Associate Professor at the same faculty, where he currently serves as the Dean. His research interests include data mining and artificial intelligence, with applications in healthcare and biometrics. He can be contacted at email: hanafi@ums.edu.my.