Unveiling Gender in Malay-English Short Text: A Comparative Study of ML, DL and Sequential Models with XAI Misclassification Analysis

Norazlina Khamis[1], Nur Shaheera Shastera Nulizairos[2], Haslizatul Mohamed Hanum[3],
Amirah Ahmad[4], Nor Hapiza Mohd Ariffin[5], Ruhaila Maskat[6]*
Faculty of Computing & Informatics, Universiti Malaysia Sabah, Sabah, Malaysia[1]
Faculty of Computer & Mathematical Sciences, Universiti Teknologi MARA Shah Alam, Selangor, Malaysia[2, 3, 6]
Academy of Language Studies, Universiti Teknologi MARA Selangor, Malaysia[4]
MIS Department-Faculty of Business, Sohar University, Oman[5]

***Abstract*—Gender identification through written text analysis leverages writer-specific characteristics including linguistic patterns and stylistic behaviors, yet research on gender identification in Malay-English (Manglish) using Traditional Machine Learning (ML), Shallow Deep Learning (DL), and Deep Sequential techniques remains limited compared to English-focused studies. This study addresses this gap by investigating gender identification in Manglish across traditional ML, Shallow DL, and Sequential Deep Model approaches using a self-collected dataset of Manglish tweets from 50 anonymized Malaysian public figures. Following preprocessing, feature extraction employed Word2Vec embeddings and TF-IDF methods, revealing that Word2Vec embeddings delivered superior performance across Shallow DL and Deep Sequential models, with Bi-CNN achieving optimal results of accuracy (0.722), precision (0.727), recall (0.722), and F1-score (0.720), while TF-IDF vectorization yielded substandard performance except for Logistic Regression, which achieved consistent metrics of 0.728 across all evaluation criteria. To enhance model interpretability, eXplainable Artificial Intelligence (XAI) tools including SHAP and LIME were applied to analyze misclassifications, identifying key issues such as frequent shortform usage and word misassignment affecting prediction accuracy, and incorporating these XAI insights through iterative refinements yielded modest improvements from 72.4% to 72.8%, demonstrating XAI's value in model optimization despite limitations in capturing dataset biases and complex linguistic patterns. This study contributes the first gender classification dataset for Malay short text and demonstrates that Shallow DL and Deep Sequential models, enhanced by XAI-driven analysis, show significant promise for mixed-language contexts, highlighting the unique challenges of code-switched languages in NLP tasks while suggesting future research should explore large language models to advance classification performance isn multilingual social media environments.**

Keywords*—Gender identification; Manglish; machine learning; shallow deep learning; deep sequential model

I. Introduction

The increasing reliance on digital communication platforms underscores the importance of gender identification in various domains, including cybersecurity, online community management, and fraud detection [1]. Online anonymity, particularly on platforms such as Twitter, complicates the distinction between male and female identities, intensifying the need for robust gender classification methods. Character limitations inherent to social media platforms, coupled with the prevalent use of informal, abbreviated, and slang-filled language, present considerable challenges for conventional natural language processing (NLP) techniques seeking to extract accurate gender cues [2]. The dynamic nature of online vocabulary, with new terms emerging frequently, further complicates the process of interpreting limited and unstructured content.

Male writers are inclined towards unisex references, insults, or profanities, directive tone, strong assertion, and rhetorical questions, and most of the time, they keep it short, direct, and crude [3]. In contrast, female writers' inclinations are strong for using hedges, polite and emotionally expressive words, and they also love to show concern for others [3].

Gender identification within the Malay-English (Manglish) language presents a particularly acute challenge due to the scarcity of established corpora and dedicated resources. Research into gender identification for Indonesian language, for example, is notably sparse compared to studies in English and other languages [4]. A specific gap exists in the comparative analysis of traditional machine learning (ML), shallow deep learning (DL), and deep sequential algorithms for author gender identification in Malay short texts [5]. Furthermore, the lack of transparency in complex models necessitates methods to understand their decision-making, particularly concerning misclassifications.

This work addresses the research gap by investigating gender identification in Manglish using traditional ML, shallow DL, and sequential deep model approaches. A primary objective involves curating a novel dataset of anonymized linguistic data from 50 influential Malaysian public figures to facilitate further inquiry in this domain. Beyond prediction, this study explores how Explainable Artificial Intelligence (XAI) can enhance model comprehension and transparency, particularly by analyzing causes of misclassification. The scope encompasses a comparative evaluation of diverse algorithmic paradigms and an in-depth XAI-driven analysis of model performance.

*Corresponding author.

The ability to accurately classify gender from multilingual short texts holds considerable significance for several reasons. User-specific characteristics, such as gender and personality, often manifest in writing style. Understanding these stylistic differences allows for the customization of digital content and marketing strategies, including recommendations and advertisements, to better suit diverse user groups. For low-resource languages like Malay, research in this area provides foundational knowledge for future NLP applications. Furthermore, the application of XAI in this context offers a methodology for identifying and mitigating potential biases within gender identification models, thereby contributing to more equitable technological systems.

The paper is structured as follows: Section II reviews previous work. Section III covers the methodology for comparison and explanation. Here, we describe the construction and preparation of the dataset, the model chosen based on performance measurements, and the XAI techniques in identifying the misclassifications. Section IV presents the results of all the models used for the experiment. Lastly, in Section V, we conclude the results, address the limitations, and suggest the possible future work in exploring the gender identification of Malay short text authors.

II. Related Works

A. Foundations of Automated Gender Classification

Automated gender classification from text relies on identifying linguistic patterns correlated with author gender. This often involves analyzing lexical choices, syntactic structures, and discourse markers. Research in this field frequently leverages stylistic features that differentiate writing patterns between genders. However, the effectiveness of these methods varies significantly across languages and text types [5]. The absence of comprehensive, labeled datasets for specific language varieties, such as Manglish, impedes the development and validation of robust classification systems.

Gender expression in Malay-English (Manglish) social media texts is characterized by code-mixing, informal abbreviations, and localized slang, posing unique challenges for automated analysis. The specific stylistic nuances of Manglish, including frequently used short forms and the dynamic emergence of new vocabulary, make it difficult for conventional NLP techniques to discern gender-specific linguistic cues. For instance, expressions like "syok gila" (crazy fun) or "makan already" (already eaten) exemplify the local slang patterns that may convey subtle gender information. The lack of established corpora further complicates the identification of consistent gendered patterns, necessitating specialized data collection and processing methods for effective classification.

Multilingual data representation presents substantial challenges for gender identification models. The linguistic complexity of mixed languages, such as Manglish, involves diverse sociolinguistic features that influence how gender is perceived and interpreted within speech processing systems. A significant obstacle lies in the scarcity of established text corpora for Malay-language short texts. This contrasts with English, where extensive resources support such research. Previous efforts, such as developing phoneme distribution datasets for Malay, highlight the critical need for linguistic resources to support speech analysis and automatic speech recognition [6]. Without adequate and representative datasets, models struggle to generalize effectively across the intricate variations present in multilingual social media discourse.

B. Deep Learning Paradigms for Short Text Classification

Deep learning (DL) models offer advanced capabilities for text classification by automatically learning intricate patterns from raw data, often surpassing traditional machine learning in complex linguistic tasks. These paradigms are particularly suited for capturing subtle stylistic cues in short, informal texts prevalent in social media. The models' ability to learn hierarchical representations reduces the need for manual feature engineering, making them adaptable to diverse linguistic contexts. However, their performance can be sensitive to data characteristics and model complexity, requiring careful selection and tuning.

Shallow neural networks, such as basic Artificial Neural Networks (ANN), are capable of capturing local features within text, including short sequences of words or characters that correspond to local slang patterns. These models learn important slang terms directly from the data without explicit manual feature engineering. An example is the identification of gender cues from expressions like "syok gila" or "makan already". Despite their utility in capturing localized patterns, shallow networks may encounter limitations when confronted with highly complex datasets involving intricate feature interactions or when attempting to model long-range contextual dependencies. In such scenarios, their performance can be surpassed by more complex architectures or robust traditional models.

Deep sequential models are specifically designed to process ordered data, where the sequence of words is central to understanding meaning. These models operate on the premise that the meaning of a word is heavily influenced by the words preceding and succeeding it. Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM) networks, and Gated Recurrent Units (GRUs) are prominent examples within this category. GRUs, in particular, are structured to anticipate future words and maintain contextual awareness for accurate predictions. Extensions such as bidirectional and convolutional bidirectional mechanisms are often integrated into these architectures to enhance performance by capturing richer contextual information, thereby facilitating more robust feature extraction. Hybrid neural architectures, such as Convolutional Bidirectional Gated Recurrent Units (C-Bi-GRU), effectively model complex linguistic structures in informal social media content by simultaneously capturing local textual features and long-range contextual relationships, as demonstrated in studies identifying author gender in Egyptian Arabic tweets [7].

C. Explainable Artificial Intelligence in Gender Prediction

Explainable Artificial Intelligence (XAI) addresses the critical need for transparency and interpretability in complex AI systems, particularly in sensitive applications such as gender prediction. As AI models become increasingly sophisticated, understanding their internal logic and decision-making processes becomes essential for ensuring trust,

identifying biases, and enabling human oversight [8]. XAI provides tools and frameworks to bridge the gap between algorithmic complexity and human comprehension [9].

Prominent XAI frameworks and methods, including SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations), are widely utilized for visualizing feature interactions and quantifying feature importance. These post-hoc techniques render complex machine learning models more interpretable and comprehensive. For instance, LIME explanations have been shown to be reasonable for models like XGBoost, aiding researchers in selecting optimal models for deployment and providing end-users with justifications for predictions. XAI tools enable deeper understanding of model behavior while maintaining high predictive performance. For misclassification analysis, an adversarial technique involving SHAP and LIME determines minimal input feature changes required to correct misclassified examples, thereby identifying primary causes of error.

XAI tools play a central role in understanding and diagnosing model misclassifications, particularly in applications like gender identification based on writing styles. By investigating how models arrive at their predictions, XAI approaches help identify potential errors or flaws in the modeling process, such as misinterpretations or biases stemming from training data. This capability is crucial for gender identification, which demands precise observation of user writing styles [10]. Insights derived from XAI, such as those revealing frequent shortform usage and word misassignment, can guide iterative refinements, leading to improved model optimization. Furthermore, XAI enhances human decision-making, as evidenced by increased agreement among moderators using XAI tools for tasks like hate speech classification.

III. Methods and Materials

A. Dataset

The collected dataset comprises a total of 898,884 instances, derived from 25 male and 25 female authors, with 78 attributes, including user information, tweet text, geolocation, and metadata. This dataset is publicly accessible at Maskat et al. [11]. Fig. 1 shows the percentage distribution. Total tweets by female authors are 499,751, while those by male authors are 399,133. Fig. 1 shows the gender-based distribution of tweets, which is sufficiently balanced, ensuring that no imbalance exists in the dataset to avoid any bias issues during the training or evaluation process later on.

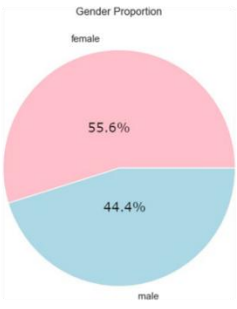

Fig. 1. Proportion of tweets based on gender.

B. Data Preprocessing

The raw Manglish tweet data underwent a meticulous cleaning and preprocessing pipeline. This process included standard NLP steps such as tokenization, lowercasing, removal of punctuation, special characters, and numerical data. Due to the nature of social media text, specific handling of emoticons, URLs, and user mentions was also implemented. Crucially, the pipeline incorporated a custom approach for handling Manglish-specific elements, including frequent short-form words, informal spellings, and code-switching instances. The dynamic and informal nature of Manglish, characterized by slang, newly coined terms, and frequent code-switching between Malay and English, presents unique challenges for traditional preprocessing methods. For instance, a term like "mcm" (Malay for "macam," meaning "like") or "sbb" (Malay for "sebab," meaning "because") requires specific normalization not typically found in standard English or Malay preprocessing tools. This customized cleaning ensures that the linguistic nuances of Manglish are retained or appropriately transformed for feature extraction.

Two distinct feature extraction methods were employed to convert the preprocessed text into numerical vectors: Word2Vec embeddings and Term Frequency-Inverse Document Frequency (TF-IDF).

1) TF-IDF: This method assigns weights to words based on their frequency within a document and their rarity across the entire corpus. It effectively reduces the influence of common words (stopwords) while emphasizing distinctive keywords that are more likely to convey stylistic or linguistic cues related to gender. TF-IDF generates sparse, high-dimensional vectors, which can be effective for traditional machine learning models.

2) Word2Vec: As a word embedding technique, Word2Vec maps words into dense, low-dimensional vector spaces, capturing semantic relationships between words based on their context. This approach allows models to understand the meaning and context of words, which is particularly beneficial for complex linguistic structures found in code-mixed languages. Pre-trained Word2Vec models, or models trained on the specific Manglish corpus, enable the capture of semantically rich embeddings that are less susceptible to the sparsity issues of TF-IDF.

C. Model Selection

A total of 11 models are used in this study, which include the variations of the main models: Logistic Regression (LR), Random Forest (RF), Convolutional Neural Network (CNN), and Gated Recurrent Unit (GRU). Both LR and RF are further enhanced with a powerful gradient boosting algorithm, XGBoost, as these traditional models can sometimes fall short when dealing with highly complex datasets that involve intricate feature interactions. This hybrid ensemble approach ensures that the weaknesses of traditional models are compensated by the strengths of XGBoost, resulting in more robust and efficient predictions overall. Additionally, to enhance efficiency, a stacked ensemble model of LR and RF with XGBoost is constructed. Meanwhile, for CNN and GRU,

two extensions, which are bidirectional and convolutional bidirectional mechanisms, are incorporated as an effort to further improve the performance of the deep learning models, as well as to test the capability of the complex model in accurately predicting the gender of authors. By integrating these bidirectional extensions, both models are better equipped to capture richer contextual information, leading to more robust feature extraction and improved overall performance.

IV. Results and Discussion

A. Experiment 1: Evaluating Machine Learning and Deep Learning Models with TF-IDF Vectorizer

From Table I, when using the TF-IDF vectorizer, it is evident that the traditional model, LR, achieves the highest accuracy, precision, recall, and F1-score of 0.728 across all models. These balanced and consistently strong results demonstrate that LR performs reasonably well in accurately identifying the gender based on the writing styles, capturing discriminative word-level patterns present in the text. Meanwhile, the performances of the deep learning models (CNN, GRY, Bi-CNN, Bi-GRU, C-Bi-CNN, and C-Bi-GRU) are noticeably lower compared to the traditional models, suggesting that the model struggles to correctly identify the gender when using TF-IDF features. In particular, the GRU model appears less suited for the gender identification task in this study, likely due to its limitations in capturing the irregular patterns present in Manglish language. Despite the integration of advanced features, such as bidirectional and concatenated convolutional-bidirectional layers, in CNN and GRU algorithms, no performance improvements were observed. This finding indicates that such enhancements are not practical when paired with a TF-IDF vectorizer for gender classification. Similarly, the ensemble models for traditional machine learning models (XGB-LR, XGB-RF) performed worse than the standalone models, suggesting that the additional complexity introduced by XGBoost did not contribute to higher predictive accuracy and, in fact, may have hindered performance.

TABLE I. Model Performance with TF-IDF Vectorizer

Models	Accuracy	Precision	Recall	F1-Score
LR	**0.728**	**0.728**	**0.728**	**0.728**
RF	0.689	0.689	0.689	0.688
XGB-LR	0.667	0.674	0.667	0.663
XGB-RF	0.535	0.712	0.535	0.419
LR-RF-XGBoost	0.610	0.728	0.610	0.553
CNN	0.540	0.541	0.540	0.538
Bi-CNN	0.516	0.537	0.516	0.440
C-Bi-CNN	0.536	0.547	0.536	0.507
GRU	0.500	0.250	0.500	0.333
Bi-GRU	0.487	0.235	0.487	0.312
C-Bi-GRU	0.467	0.212	0.467	0.294

The findings further highlight that model complexity does not guarantee performance improvements in this case. Instead, the traditional algorithms, despite their simplicity, may be more effective and computationally efficient in identifying the gender of mixed-language users when TF-IDF is used for feature representation. The comparatively lower performance observed in CNN, GRU, Bi-CNN, Bi-GRU, C-Bi-CNN, and C-Bi-GRU models may be attributed to several factors, including the complexity of the task, limitations in the architectural capacity of the models, or challenges in capturing significant trends and patterns from the data. As deep learning models are commonly designed to leverage dense, low-dimensional embeddings (e.g., Word2Vec, GloVe) that capture semantic relationships between words, forcing the models to operate on sparse TF-IDF vectors may hinder their ability to fully exploit their representational strengths. The ineffectiveness may also be caused by the overparameterization of deep learning models. As TF-IDF vectors lack sequential and semantic information due to the insignificant amount of data, it automatically limits the learning capacity of deep learning models, causing the models to overfit to noise in the sparse TF-IDF features rather than learning robust gender-discriminative patterns in the text. Therefore, it can be assumed that the poorer performance of deep learning and ensemble models is not necessarily due to their inherent weaknesses, but rather the incompatibility between the feature extraction and the models' design.

B. Experiment 2: Evaluating Machine Learning and Deep Learning Models with Word2Vec Vectorizer

From Table II, when using the Word2Vec embedding method, the traditional models, particularly LR and RF models, demonstrated relatively strong performance. Among these, the ensemble model XGB-LR slightly outperformed both LR and RF, indicating that combining XGBoost with Logistic Regression can offer incremental improvements in predictive accuracy. On the other hand, XGB-RF and stacked LR-RF+XGBoost models did not perform as well as expected, providing no significant improvement over their base models in predicting gender. These findings indicate that the XGBoost algorithm works more effectively with LR than with RF for gender identification tasks. However, it is worth noting that the differences in performance across all models are relatively small, indicating that Word2Vec embeddings provide a robust foundation that enables each model to achieve a reasonably reliable predictive capability.

TABLE II. Model Performance with Word2Vec Embeddings

Models	Accuracy	Precision	Recall	F1-Score
LR	0.713	0.714	0.713	0.713
RF	0.716	0.719	0.716	0.715
XGB-LR	0.717	0.718	0.717	0.716
XGB-RF	0.693	0.704	0.693	0.689
LR-RF-XGBoost	0.693	0.702	0.693	0.689
CNN	0.718	0.722	0.718	0.717
Bi-CNN	**0.722**	**0.727**	**0.722**	**0.720**
C-Bi-CNN	0.708	0.714	0.708	0.706
GRU	0.710	0.720	0.710	0.708
Bi-GRU	0.707	0.711	0.707	0.706
C-Bi-GRU	0.711	0.718	0.711	0.709

In the case of the deep learning models, the results reveal a contrasting trend compared to TF-IDF. Specifically, Bi-CNN proved to be the most effective model, outperforming even the traditional machine learning models. This demonstrates that Word2Vec embeddings provide dense, semantically rich input representations that deep models, such as CNNs, can effectively exploit, enabling them to capture nuanced patterns in writing styles. However, these findings reveal some notable observations, as the results show inconsistencies in the performances of architectural enhancements. For the GRU algorithm, the C-Bi-GRU model demonstrated good performance, indicating that combining bidirectional and concatenated architectures can enhance the model's effectiveness. It also enhances the GRU algorithm's ability to identify key patterns and dependencies in the gender identification task. However, this was not observed with CNN models, as the performance of the enhanced models appeared to decline. This suggests that the advanced structure of the models may not always contribute to improvements and additional value, and in some cases, may even hinder the models' effectiveness. In conclusion, the efficacy of ensemble models may vary depending on the model type and the nature of the task itself. In this case, Word2Vec embeddings align better with deep learning models because they offer dense, low-dimensional semantic vectors, unlike TF-IDF, which creates sparse and context-independent features. Consequently, deep models, especially Bi-CNN and C-Bi-GRU, benefit more from Word2Vec embeddings. Therefore, to achieve the best results, it is essential for studies to explore additional architectural structures or models to determine the most effective and compatible approach for the task, thereby improving the accuracy of gender identification despite the complexity of the language.

C. Experiment 3: Evaluating Machine Learning and Deep Learning Models with Word2Vec Vectorizer

This approach was only done to maintain consistency and relevance when evaluating the performance of the two techniques. From Table III, which presents the results of experiments using both TF-IDF and Word2Vec vectorizers on the same 10,000-instance dataset, several important patterns emerge.

The traditional machine learning algorithms, including ensemble methods such as LR, RF, and XGB-LR ensemble, consistently achieve stable performance across all evaluation metrics compared to deep learning models, regardless of the vectorization method used. This demonstrates that traditional models are both robust and adaptable in leveraging different feature representations for the task of gender identification. This also highlights that both vectorization techniques are effective in capturing relevant and significant textual features for gender identification, with each vectorizer offering its own benefits. Word2Vec provides semantically rich embeddings, while TF-IDF emphasizes the importance of words within the dataset. The fact that traditional models perform well with either approach indicates that both vectorization methods can capture meaningful textual signals, even in Manglish, a highly irregular and code-mixed language.

TABLE III. MODEL PERFORMANCE WITH BOTH TF-IDF VECTORIZER AND WORD2VEC EMBEDDINGS

Methods	Models	Accuracy	Precision	Recall	F1-Score
TF-IDF	**LR**	**0.728**	**0.728**	**0.728**	**0.728**
	RF	0.689	0.689	0.689	0.688
	XGB-LR	0.667	0.674	0.667	0.663
	XGB-RF	0.535	0.712	0.535	0.419
	LR-RF-XGBoost	0.61	0.728	0.61	0.553
	CNN	0.54	0.541	0.54	0.538
	Bi-CNN	0.516	0.537	0.516	0.44
	C-Bi-CNN	0.536	0.547	0.536	0.507
	GRU	0.5	0.25	0.5	0.333
	Bi-GRU	0.487	0.235	0.487	0.312
	C-Bi-GRU	0.467	0.212	0.467	0.294
Word2Vec	LR	0.568	0.568	0.568	0.565
	RF	**0.596**	**0.599**	**0.596**	**0.594**
	XGB-LR	0.587	0.587	0.587	0.587
	XGB-RF	0.571	0.57	0.571	0.57
	LR-RF-XGBoost	0.571	0.571	0.571	0.568
	CNN	0.558	0.56	0.558	0.556
	Bi-CNN	0.581	0.584	0.581	0.574
	C-Bi-CNN	0.569	0.568	0.569	0.568
	GRU	0.493	0.75	0.493	0.325
	Bi-GRU	0.55	0.551	0.55	0.549
	C-Bi-GRU	0.558	0.558	0.558	0.557

In contrast, deep learning models such as CNN, GRU, and their improved variants like Bi-CNN, Bi-GRU, and C-Bi-GRU struggle to outperform simpler models, especially when combined with TF-IDF. These lower results indicate that deep neural networks are not able to consistently extract the relevant cues for gender classification within this dataset. This limitation may be due to several reasons, such as the unnecessary complexity of the models, which, when combined with limited training data, prevented the models from generalizing effectively, the lack of significant features in the dataset itself, or even the incompatibility of the device used, which somehow restricted the optimization of complex models. The unpredictable grammatical structures, informal spellings, and code-switching common in Manglish may also hinder the ability of sequential models, such as GRUs, to identify consistent patterns.

Despite the challenge, the consistent and decent performances of LR, RF, XGB-LR, and LR-RF+XGBoost models, particularly when paired with TF-IDF vectorizers, suggest their suitability for this particular task. The LR model using the TF-IDF vectorizer performs the best across all metrics among the models, concluding that traditional machine learning algorithms are highly relevant for identifying gender based on online writings in Manglish

language, especially in cases where data complexity and dataset volume are not fully aligned, which can cause models to fail to fully learn the patterns in the language.

D. eXplainable Artificial Intelligence (XAI)

The goal of this process is the XAI implementation (SHAP and LIME) to the best-fitted model, which in this case is LR with TF-IDF model, to produce and evaluate the output of the local explanations, especially for the misclassified instances, in order to identify the issues that cause the model to wrongly predict the gender of the writing. To enable future analysis, every detail of this procedure, including processing time, documentation, and scalability, was documented. In order to examine the misprediction of the gender, the performance can be analyzed through the confusion matrix, which also plays a crucial role for XAI by providing a clear and concise representation of a model's performance.

In Fig. 2, the confusion matrix for the model shows that there is a total of 552 instances which are predicted incorrectly by the models in which 291 instances are mistakenly predicted as male and 261 instances as female.

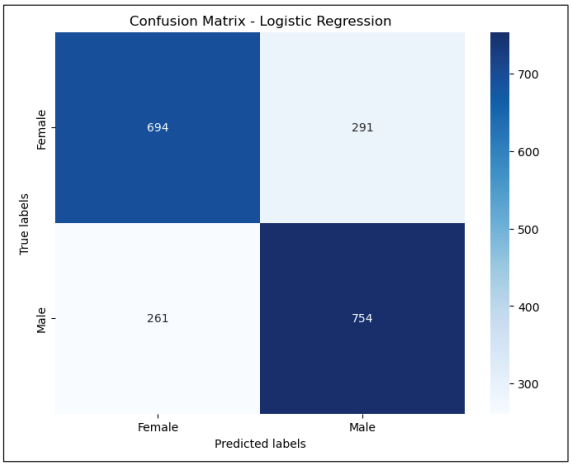

Fig. 2. Confusion matrix.

By scrutinizing these misclassifications, the goal is to pinpoint specific patterns, features, or words that might be influencing the inaccuracies. Once the issues can be identified, the model will then be modified accordingly in order to improve and enhance the performance. The original and modified model will then be compared to observe the significance of analysing the misclassified instances.

1) SHAP implementation: From Fig. 3, it can be seen that the words that are used daily such as *'je, 'iphone', 'makan', 'okay', 'tidur, 'suka'* and *'pakai'* have higher predictive values. This indicates that these common terms show up a lot in the dataset and are particularly important to the tweet context, which explains the higher weight given by the model. This implies that these words strongly influence the prediction of the gender for the written tweet. Also, it can be seen that slangs are also commonly used by internet users, such as *'je', 'pon'* which explain the higher predictive values compared to others. Malaysians, especially the millennial people in which born between 1981 to 1996 (ages 29 to 44 in 2025), love using shortform when expressing their thoughts online as to minimize the use of characters to avoid exceeding the limits set by X. Therefore, words such as *'mcm'* which actually means *'macam'* and *'yg'*, representing the word *'yang'*, can be seen in most of the texts since almost all of the chosen authors aged beyond which contradicts the behavior of young people these days who prefer writing each word as a whole instead of using shortform, explaining why those words also have higher predictive values for the model. This nuanced analysis not only sheds light on linguistic patterns but also prompts reflection on the evolving nature of language in the digital age.

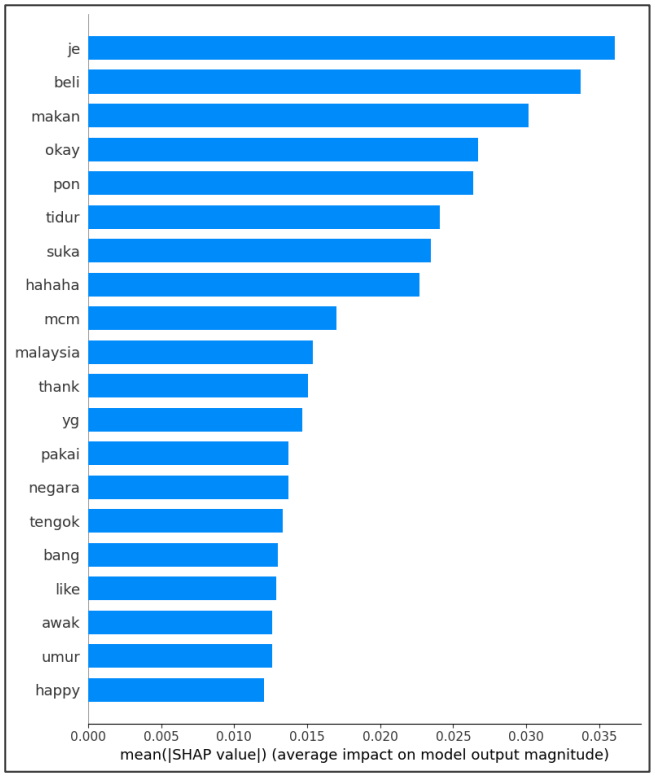

Fig. 3. SHAP value bar graph.

2) LIME implementation: After a few observations, it is obvious that some words are misclassified to a certain gender when the words are commonly used by the opposite gender. For example, the word 'member' (meaning friends in Manglish slang) as shown in Fig. 4. It is observed that the word is commonly used by males, as out of three instances that contain the word, only one instance belongs to a female, yet the word is associated with females, which causes misprediction for these instances.

Another misclassification would be the word 'power' (meaning strength or awesome in Manglish slang) which is also associated with females by the model when it is usually used by males. Out of 13 instances that contain the word 'power', only 5 instances are written by females, while the rest are by males, proving that the word is actually used by males.

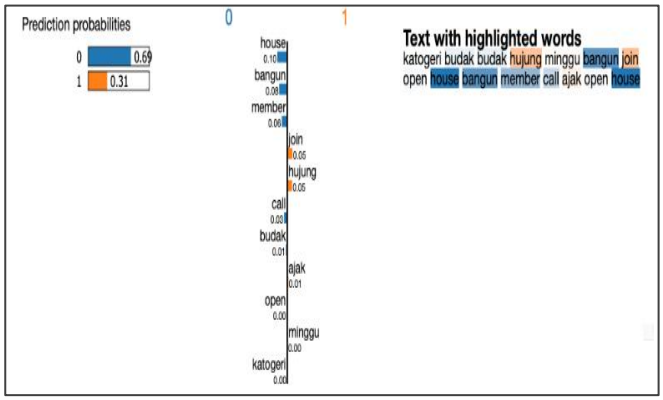

Fig. 4. LIME explanation.

This discrepancy in usage patterns contradicts the prediction of the model, highlighting a limitation in capturing the nuanced associations of certain words and emphasizing the importance of refining the understanding of the model for context and gender-related language nuances.

In the dataset, as mentioned, most of the authors are millennials or from older generations. Hence, given the cultural context and communication norms of these generations, the utilization of short forms was a common practice back then, where everyone used the same short form. Considering this socio-linguistic perspective, the use of short forms becomes a dynamic and evolving aspect of language rather than a static marker tied to a specific gender. Therefore, it is somehow irrelevant to associate those short forms with a certain gender since almost everyone uses it, especially for online communication. This can be seen by comparing a few instances – 301st and 366th observations are written by males, while 355th and 378th are by females. Both males and females use the exact short form of 'yang', which is 'yg' as that is what they are used to in informal or online interaction. This convergence in language usage reflects a commonality in the way individuals, irrespective of gender, adapt and adopt linguistic shortcuts for the sake of brevity and efficiency in digital communication.

E. Model Improvements

1) Shortforms and unnecessary words removal: Since many authors, irrespective of gender, commonly use the same short forms in their sentences, associating specific words with a particular gender would be meaningless. Thus, to ensure fairness and impartiality in the predictions of the model for both genders, it is advisable to remove such gender-agnostic words from consideration.

2) Associate words with the correct gender: To enhance the accuracy of the prediction, words that are inaccurately associated with a particular gender need to be reevaluated and reassigned accordingly. In this case, the words 'power', 'bro', and 'member', which are originally assigned to females by the model, are modified so that they can be associated with males instead.

3) Models comparison after modifications: Now that the dataset has been modified, the same process is done in the original model in order to make a comparison of the performance of the prediction. Here, four evaluation metrics are used, which are accuracy, precision, recall and F1-score. Table IV shows list of words to be removed.

TABLE IV. LIST OF WORDS TO BE REMOVED

Used Words	Root words	Meaning
Yg	Yang	Malay grammar
ahahah	-	Laughing
kt	Kat	At
hahahahahahahaha	-	Laughing
elehh	-	Sarcasm
hahahaha	-	Laughing
jgn	Jangan	Do not
hahshahahaha	-	Laughing
mmg	Memang	Really / Actually
tgh	Tengah	In the middle of
je	Sahaja	Only / Just
jee	Sahaja	Only / Just
hahahahaha	-	Laughing
tuh	Itu	That
pon	-	Also
hahaha	-	Laughing
mcm	Macam	Like
x	Tak	No
usbdjfjdjdndn	-	Gibberish
sbb	sebab	Because

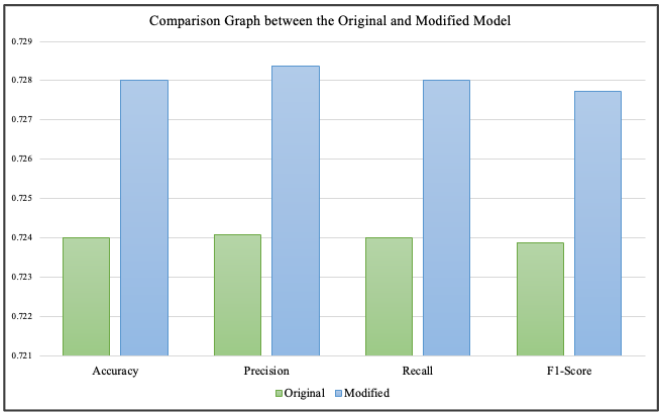

Fig. 5. Comparison graph for both models before and after modifications.

Fig. 5 shows that both models showcase almost similar performance for all the evaluation metrics. The accuracy levels of the two models are nearly identical, while the improved model shows a small improvement in accuracy, going from 72.4% to 72.8%, indicating that the predictions were generally true. With precision scores of 72.4% and 72.8%, respectively, both models demonstrate a reliable ability to recognize instances of a particular gender. For the recall, which gauges the capacity of the model to identify all relevant instances of a particular class, it shows that the new model's sensitivity has

slightly improved, going from 72.4% to 72.8%. Lastly, the F1-Score, which provides a fair assessment of a model's performance by taking into account both recall and precision, increases slightly from 72.4% to 72.8% in the updated model, making it in line with other indicators.

This implies that several misclassification-related issues that were found by XAI techniques have been addressed and solved properly, which eventually improves the ability of the model to correctly predict the gender of the written tweet. Even if the enhancements are subtle, they demonstrate how significant it is to analyze and continuously update models in light of knowledge obtained by XAI applications in order to maximize both the effectiveness and practical applicability of the model. These techniques make the variables impacting and influencing the model predictions more comprehensible, thus, making it possible to pinpoint particular traits or terms that lead to incorrect classifications. By enabling the users to make well-informed judgments on the individual instances through LIME, XAI improves model performance and fosters a deeper understanding on the data-driven decision-making process.

V. Conclusion

This investigation explored gender identification within Malay-English (Manglish) short texts, employing a comparative analysis of traditional machine learning (ML), shallow deep learning (DL), and deep sequential models, complemented by eXplainable Artificial Intelligence (XAI) for misclassification analysis. A significant contribution of this study is the creation of the first dedicated dataset for gender classification in Malay short texts, collected from 50 anonymized Malaysian public figures. The experimental results demonstrate that Word2Vec embeddings generally facilitate stronger performance across DL and deep sequential models, with Bi-CNN achieving the highest metrics: 0.722 accuracy, 0.727 precision, 0.722 recall, and 0.720 F1-score. This highlights the effectiveness of semantic embeddings in capturing the complexities of code-mixed language. In contrast, TF-IDF vectorization yielded substandard performance for most deep learning models, indicating an incompatibility between sparse feature representation and DL architectures. However, Logistic Regression, a traditional ML model, proved an exception, achieving robust performance with TF-IDF: 0.728 accuracy, precision, recall, and F1-score. This suggests that model complexity does not always equate to superior performance, and simpler models can be highly effective with suitable feature engineering. Crucially, the integration of XAI tools (SHAP and LIME) provided granular insights into model behavior, particularly concerning misclassifications. The analysis identified frequent short-form usage and word misassignment as primary contributors to prediction errors in Manglish texts. This proves that the use of explainable AI (XAI) tools offers useful insights into model interpretability and error sources, which are essential for improving algorithms and guaranteeing transparency real-world deployment. Iterative model refinements, directly informed by these XAI insights, led to a modest but significant improvement in accuracy from 72.4% to 72.8%. This underscores the value of XAI not just for interpretability but also as a practical guide for model optimization in challenging linguistic contexts.

The study encountered several unique challenges in handling the Manglish language, as the language itself is difficult to decipher. The major limitation that often sets the study back is the restriction imposed by hardware resources, specifically RAM capacity, which limits the amount of data that can be processed simultaneously. This constraint limits the models' ability to fully experiment on the dataset, preventing the study from realizing the full potential of the models. Additionally, the dynamic and informal nature of Manglish, which includes extensive use of slang, newly coined terms which was randomly created every other day, as well as the frequent code-switching between Malay and English within a single sentence, poses a very challenging and unique task for both traditional and deep learning, as well as for data preprocessing.

There are also some limitations to the XAI tools for both SHAP and LIME, in which these tools provide a post hoc analysis rather than a detailed understanding of the internal workings of the model. The interpretations generated by XAI may offer insights into associations and trends but might not reveal the underlying causal relationships. Especially for binary classification cases, it is necessary to analyze the instances individually first before the problems can be identified, which is somewhat very time-consuming for the users. Additionally, XAI tools may also not fully capture the external biases present in the training data. If the training data itself contains inherent biases, they may persist throughout the prediction process, thus leading to inaccurate performance. These bias issues are then expected to continue even during the XAI implementation as XAI tools might not be able to explicitly highlight or mitigate them completely as the tool itself are not competent enough to identify the biases directly.

Given these limitations, it will be more promising for future research to improve data preprocessing, such as applying custom tokenization, stemming, or even lemmatization techniques designed for Manglish, to better handle non-standard grammar, spelling variations, and mixed-language expressions. Further research can also focus on developing more specialized approaches tailored to the unique structure of Manglish, including the use of advanced sequence models such as LSTM networks, which will benefit stakeholders such as social media platforms, and AI developers seeking to build more inclusive and context-aware language technologies. Lastly, future research may also explore model optimization strategies, such as mini-batch training, to address the issue of hardware limitations and ensure that larger portions of the dataset can be utilized without overloading the system memory. Altogether, addressing these methodological and technical gaps will be beneficial for more advanced gender identification research in multilingual and informal languages, such as Manglish, as well as for applications involving social media analysis and user profiling where understanding nuanced language use and demographic attributes is essential.

References

[1] Alanazi, S. A. (2019). Toward identifying features for automatic gender detection: A corpus creation and analysis. IEEE Access, 7(Ii), 111931–111943. https://doi.org/10.1109/ACCESS.2019.2932026

[2] Vashisth, P., & Meehan, K. (2020). Gender Classification using Twitter Text Data. 2020 31st Irish Signals and Systems Conference, ISSC 2020. https://doi.org/10.1109/ISSC49989.2020.9180161

[3] Subon, F. (2013). Gender Differences in the Use of Linguistic Forms in the Speech of Men and Women in the Malaysian Context. IOSR Journal Of Humanities And Social Science, 13(3), 67–79. https://doi.org/10.9790/0837-1336779

[4] Tanuar, E., Abdurachman, E., & Gaol, F. L. (2020, November). Analysis of Gender Identification in Bahasa Indonesia using Supervised Machine Learning Algorithm. In 2020 3rd International Conference on Information and Communications Technology (ICOIACT) (pp. 421-424). IEEE.

[5] Dalyan, T., Ayral, H., & Özdemir, Ö. (2022). A Comprehensive Study of Learning Approaches for Author Gender Identification. Information Technology and Control, 51(3), 429–445. https://doi.org/10.5755/j01.itc.51.3.29907

[6] Asyafie, M. A., Harun, M., Shapiai, M. I., & Khalid, P. I. (2014, December). Identification of phoneme and its distribution of Malay language derived from Friday sermon transcripts. In 2014 IEEE Student Conference on Research and Development (pp. 1-6). IEEE.

[7] ElSayed, S., & Farouk, M. (2020). Gender identification for Egyptian Arabic dialect in twitter using deep learning models. Egyptian Informatics Journal, 21(3), 159–167. https://doi.org/10.1016/j.eij.2020.04.001

[8] Sharma, B., Sharma, L., Lal, C., & Roy, S. (2024). Explainable artificial intelligence for intrusion detection in IoT networks: A deep learning based approach. Expert Systems with Applications, 238, 121751.

[9] Geleta, R. R., Eckelt, K., Parada-Cabaleiro, E., & Schedl, M. (2023, September). Exploring intensities of hate speech on social media: A case study on explaining multilingual models with XAI. In Proceedings of the 4th Conference on Language, Data and Knowledge (pp. 532-537).

[10] Ali, T. S., Ali, S. S., Nadeem, S., Memon, Z., Soofi, S., Madhani, F., ... & Bhutta, Z. A. (2022). Perpetuation of gender discrimination in Pakistani society: results from a scoping review and qualitative study conducted in three provinces of Pakistan. BMC women's health, 22(1), 540.

[11] Maskat, R., Azman, N. A., Nulizairos, N. S. S., Zahidin, N. A., Mahadi, A. H., Norshamsul, S. R., et al. (2024). A bi-annotated Malay-English code-switching (Manglish) dataset of X posts for biological gender identification and authorship attribution. Data in Brief, 52, 110034. https://doi.org/10.1016/j.dib.2024.110034