Asia-Pacific Journal of Information Technology and Multimedia 15(1) 2026: 349 – 361
https://doi.org/10.17576/apjitm-2026-1501-22

Unifying Modalities: A Comparative Analysis of Bilinear Pooling Fusion Techniques for Multimodal Fake News Detection

Penyatuan Modaliti: Analisis Perbandingan Teknik Penggabungan Bilinear bagi Pengesanan Berita Palsu Bermultimod

*Idza Aisara Norabid[1], Masita Jalil[1], Rozniza Ali[1], Ahed Mleih Al-Sbou[2], Noor Hafhizah Abd Rahim[*1]*

[1]Artificial Intelligence Group, Faculty of Computer Science and Mathematics, Universiti Malaysia Terengganu, Kuala Nerus, 21030, Terengganu, Malaysia

[2]Department of Computer Science, Faculty of Information Technology, Al-Hussein Bin Talal University, Ma'an, Jordan

[*]Corresponding author: noorhafhizah@umt.edu.my

Received 26 October 2025
Accepted 28 April 2026, Available online 30 June 2026

ABSTRACT

The spread of fake news on social media has required the need for multimodal detection techniques that combine both textual and image data. In this paper, it presents a comparative study of fusion approaches for fine-grained multimodal fake news detection that apply BERT method for representing the textual features, while ResNet models (RestNet18 and ResNet50) for image features. The comparison study involves a few fusion techniques, namely Multimodal Factorized Bilinear Pooling (MFB), Multimodal Compact Bilinear pooling (MCB) and alongside their self-attention-enhanced variants. The experiments are conducted on nine subsets of the Fakeddit dataset which have various sizes, to evaluate the performance scalability. The findings show that bilinear pooling techniques perform better in accuracy specifically in larger datasets. Among the approaches tested, MFB consistently achieves strong and stable performance while MCB also performs well, although slightly lower than MFB across all experiments. In addition, ResNet50 tends to outperform ResNet18 when trained on larger datasets. To conclude, the main contribution of this study is the benefits and limitations of four fusion techniques which provide useful guidelines for developing a more reliable automated system of multimodal fake news detection.

Keywords**:** Deep learning, Fake news detection, Fine-grained, Fusion technique, Multimodal

ABSTRAK

Penyebaran berita palsu di media sosial memerlukan keperluan untuk teknik pengesanan multimodal yang menggabungkan data teks dan imej. Dalam kajian ini, ia membentangkan kajian perbandingan pendekatan gabungan untuk pengesanan berita palsu multimodal yang terperinci menggunakan kaedah BERT untuk mewakili ciri teks, manakala model ResNet

(RestNet18 dan ResNet50) untuk ciri imej. Kajian perbandingan ini melibatkan beberapa teknik gabungan, iaitu Pengumpulan Bilinear Berfaktor Multimodal (MFB), Pengumpulan Bilinear Padat Multimodal (MCB) dan varian yang dipertingkatkan perhatian kendiri. Eksperimen dijalankan pada sembilan subset set data Fakeddit yang mempunyai pelbagai saiz, untuk menilai kebolehskalaan prestasi teknik tersebut. Penemuan menunjukkan bahawa teknik pengumpulan bilinear menunjukkan prestasi yang lebih baik dari segi ketepatan khususnya dalam set data yang lebih besar. Antara pendekatan yang diuji, MFB secara konsisten mencapai prestasi yang kukuh dan stabil manakala MCB juga menunjukkan prestasi yang baik, walaupun sedikit rendah daripada MFB dalam banyak eksperimen. Selain itu, ResNet50 cenderung mengatasi ResNet18 apabila dilatih pada set data yang lebih besar. Secara keseluruhannya, kajian ini mengetengahkan kekuatan dan batasan teknik gabungan yang berbeza dan memberikan pandangan bermanfaat untuk membangunkan sistem pengesanan berita palsu multimodal yang lebih cekap dan tepat.

Kata kunci**:** Pembelajaran mendalam, Pengesanan berita palsu, Terperinci, Teknik gabungan, Bermultimod

INTRODUCTION

As our lives have moved increasingly online, we have seen a troubling rise in "information pollution". This pollution is a mess of fake news and digital noise that has become a defining challenge of our time (Yuan et al. 2023). In the context of fake news detection, most approaches primarily focus on the textual content, as it is the dominant form through which news is conveyed (Hua et al. 2023). In today's media-rich environment, news content often includes both text and images, which together shape the narrative. Image information and text information in news are complementary to each other and readers often interpret them in relation to one another. Hence, the fusion between text and image information is a crucial part of fake news detection (Jing et al. 2023).

This leads to the increase of researchers' interest in exploring the multimodal approaches to fake news detection that integrate both textual and visual features. For instance, Hua et al. (2023) introduced a contrastive learning-based multimodal framework that employed both back-translated text and entire-image representations. The authors' method demonstrated that multimodal representations improved the fake news detection performance. Study by H. Wang, Wang & Han (2022) also combines image-text features and emphasizes the semantic correlation by proposing Fake News Detection Framework (FNDF). In which, the authors' gained a notable performance that highlights the importance of deep cross-modal interactions.

Multimodal fusion refers to the process of combining information from multiple modalities into a unified representation. The modalities can be in the form of text data, visual data or audio data. The unique strength of each modality contributes to the enhancement of the overall model performance (Li & Tang 2024). Fusion usually involves integrating textual and visual content in the context of multimodal fake news detection. Thus, the way to fuse features effectively from different modalities becomes a key challenge to capture complementary patterns. Therefore, several fusion techniques have been proposed. It varies from simple techniques like concatenation to more advanced techniques such as bilinear pooling or attention mechanisms.

This paper presents a comparative analysis of fusion techniques for multimodal fake news detection. Also, it evaluates the performance of the techniques in capturing complementary patterns and handling fine-grained fake news labels. The goal of this paper is to identify the

strengths and limitations as well as the potential directions for future research by systematically analysing these fusion techniques. The analysis involves reviewing and benchmarking these fusion strategies. In which, it contributes to a deeper understanding of how fusion mechanisms affect multimodal fake news detection. It also can provide guidance in designing more interpretable models that are capable of handling fine-grained labels.

This paper is organised into several sections which are; 2) reviews related work on multimodal fake news detection and fusion techniques, 3) describes the dataset and experimental setup, 4) presents the comparative results, 5) concludes the paper and discusses directions for future research.

RELATED WORK

With the rising rate of misinformation on social media, the task of fake news detection has gained increasing attention. Early studies have adopted unimodal approaches that focused on textual analysis. Such examples are the DSS model, which analyzed propagation patterns, and a Tsetlin Machine-based framework for interpretable linguistic feature extraction (Bhattarai et al. 2022; Davoudi et al. 2022). Researchers have extended unimodal methods to image-based information as online news began to shift towards visual content. For instance, to extract textual topics from visual content, existing studies utilized RGB-based statistical analysis of headline images and Google Cloud Vision (Lisangan et al. 2022; Zannettou et al. 2020). Thus, recent research has moved towards multimodal approaches that combine both textual and visual information to enhance detection accuracy.

This multimodal evolution has led to the development of advanced models and fusion techniques that aim to integrate heterogeneous modalities more effectively. A simple yet widely adopted fusion technique is concatenation (Liu et al. 2023; Segura-Bedmar & Alonso-Bartolome 2022; Sengan et al. 2023; Wang et al. 2022). In which, this fusion technique has the ability to combine the extracted features from different modalities to form a single unified feature. However, simple concatenation often fails to capture the complex relationships between textual and visual data (Zou et al. 2024).

Furthermore, a more expressive fusion technique is bilinear pooling that can capture fine-grained interactions between modalities. Instead of simply combining the features, bilinear pooling computes the outer product of the text and image features. This technique allows every dimension of the textual features to interact multiplicatively with every dimension of the visual features. This results in a richer set of pairwise correlations. The simplest multimodal bilinear model is defined as follows:

$$z = x^T W y \tag{1}$$

where x and y represent the feature vectors from the text and image modalities respectively, W is a learnable weight matrix, and z is the resulting fused feature.

Yet, the full outer product can be computationally expensive (Zhang et al. 2020), so practical variants such as Multimodal Compact Bilinear pooling (MCB) (Fukui et al. 2016) and Multimodal Factorized Bilinear pooling (MFB) (Yu et al. 2017) are used to compress or factorize the bilinear interactions, making the technique more efficient and suitable for deep learning applications. The MCB projects the input feature vectors into a higher-dimensional

space using random projections and approximates the outer product through the Count Sketch and convolution operations:

$$MCB(x, y) = FFT^{-1}\left(FFT\left(\Psi(x)\right) \odot FFT\left(\Psi(y)\right)\right) \tag{2}$$

where $\Psi(\cdot)$ denotes the Count Sketch projection and $\odot$ represents element-wise multiplication. While MFB avoids the high cost of outer products by first projecting both modalities into a lower-dimensional space and then performing element-wise multiplication:

$$MFB(x, y) = \left(SumPool\left((U^T x) \circ (V^T y), k\right)\right) \tag{3}$$

where U and V are trainable projection matrices that map x and y to a common dimension, $\circ$ denotes element-wise multiplication and SumPool performs sum pooling over non-overlapping windows of size k.

These techniques have been implemented in recent studies to enhance multimodal fusion in fake news detection tasks (Guo et al. 2023; Kumari & Ekbal 2021; Nadeem et al. 2023). These existing work applied MFB or MCB to effectively integrate textual and visual features into a unified feature.

Moreover, attention mechanisms (self-attention and multihead attention) by Vaswani et al. (2017) have emerged as a useful tool for tasks involving multimodal data. It allows models to weigh the importance of features across different modalities. The self-attention consists of three components which are queries (Q), keys (K), and values (V). It computes the dot product of the query with each key to generate attention weights as follows:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{4}$$

Another variant for attention mechanism, multihead attention, applies multiple attention mechanisms in parallel. It enables the model to simultaneously capture different types of relationships from multiple representation subspaces. This approach gives the ability for models to focus on diverse and complementary aspects of textual and visual features during fusion. Existing studies from Kumari & Ekbal (2021) and Nadeem et al. (2023) introduced attention-driven frameworks that integrate textual and visual features to enhance cross-modal learning in fake news detection.

Most existing studies on multimodal fake news detection rely on fusion techniques such as concatenation or attention-based approaches commonly in binary classification settings. While some work has explored more expressive fusion techniques like MFB or MCB, those techniques are still often evaluated on binary classification. Hence, research investigating the use of bilinear pooling and attention mechanisms for fine-grained fake news classification is still limited. In order to address this limitation, this paper provides a comparative evaluation of MFB, MCB and their variants enhanced with self-attention. The textual and visual features are extracted from Bidirectional Encoder Representations from Transformers (BERT) and Residual Network (ResNet) architectures. The publicly available Fakeddit dataset (Nakamura et al. 2020) that provides fine-grained labels across varying degrees of misinformation is utilized in the series of experiments. The architecture is further discussed in the following section.

METHODOLOGY

This section presents the base framework for different types of fusion techniques in multimodal fake news detection. The framework involves four main phases; 1) data collection, 2) feature extraction, 3) feature fusion and 4) classification. As illustrated in Figure 1, the process begins by collecting text-image pairs from the Fakeddit dataset. The dataset contains around one million labelled Reddit posts that provides a richer labelling structure. This dataset offers six fine-grained categories, namely true, satire, misleading content, imposter content, false connection, and manipulated content.

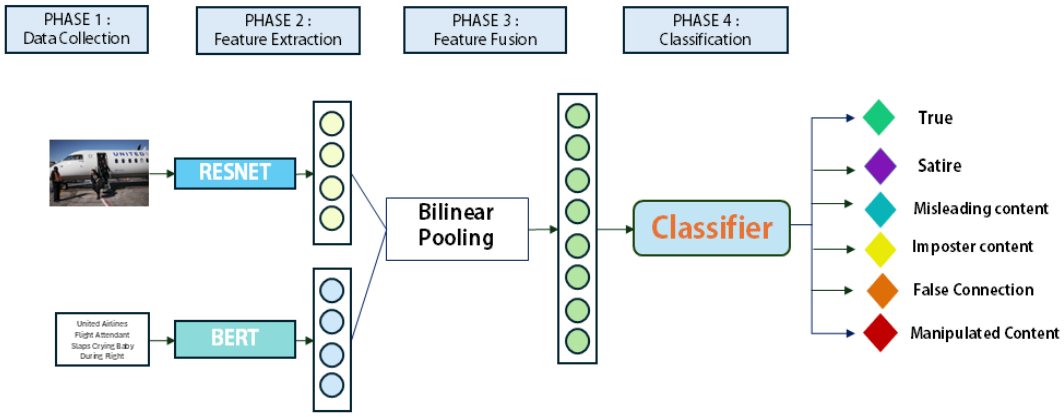

FIGURE 1. Framework for multimodal fake news detection

In the data collection phase, several subsets of the Fakeddit dataset were generated based on different conditions. Each subset consists of 10,000 samples of datasets. A total of nine subsets were created, increasing in increments of 10,000 samples. Each sample consists of textual data that correlates with its images that allows the analysis of multimodal detection. The various sizes of dataset supports a systematic performance of evaluation for different fusion techniques. Therefore, this represents both low-resource and high-resource training scenarios. Hoy & Koulouri (2022) emphasized that larger and more diverse datasets are important in improving the generalisation of models across various domains and temporal contexts.

The next phase is the feature extraction. In this phase, the textual modality is extracted using the BERT model (Devlin et al. 2019), which is capable of capturing deep context of relationship in the language. Meanwhile, the visual modality is extracted using the ResNet model that comes with two variants, ResNet18 and ResNet50 (He et al. 2016). ResNet18 offers lightweight architecture suitable for limited-resource environments, while ResNet50 is expected to capture more details about visual patterns. As mentioned by Zannettou et al. (2020), both BERT and ResNet are widely adopted due to the effective performance and general applicability across domains. Moreover, the potential of these models in cross-domain fusion tasks further supports the suitability in multimodal fake news detection (Ren 2024; Shaharudin et al. 2025).

The extracted text and image features are then passed to the feature fusion phase. The features are integrated using bilinear pooling technique. Two types of bilinear pooling are applied, Multimodal Factorized Bilinear pooling (MFB) and Multimodal Compact Bilinear pooling (MCB), alongside their self-attention-enhanced variants. These techniques enable for a richer interaction between modalities, going beyond simple concatenation. Specifically, MFB and MCB capture interactions between all dimensions of the text and image features. Meanwhile, the attention-enhanced versions aim to improve features by reweighting the important features.

FIGURE 2. Overview of baseline bilinear pooling fusion technique

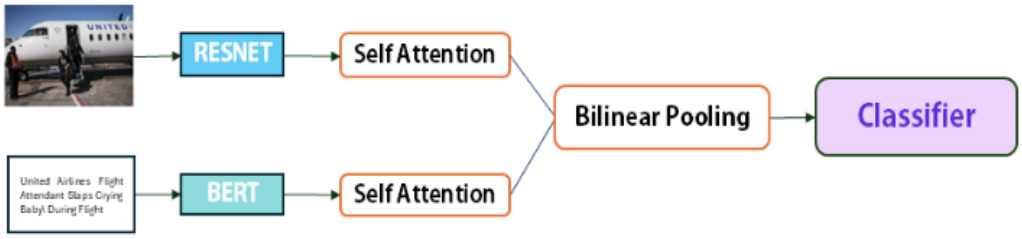

FIGURE 3. Overview of the enhanced version with self-attention

Figure 2 shows the baseline bilinear pooling fusion techniques (MFB and MCB) and Figure 3 presents the attention-enhanced versions, where self-attention is applied to the individual modality features before passing it to the fusion process. This attention mechanism is intended to highlight important features in both modalities, potentially improving classification performance. Finally, in the classification phase, the fused features are used to predict the misinformation category of each post according to the six-class labelling scheme in Fakeddit.

EXPERIMENT SETUP

All series of experiments were carried out in Python using the PyTorch deep learning framework. All training and evaluation were conducted in a CPU-based environment, an Intel Core i7 and 16GB of RAM to simulate a resource-constrained setup. The Fakeddit dataset was used as the primary benchmark for all experiments, supporting fine-grained fake news classification with six distinct labels. For feature extraction, two modalities were extracted: textual features using BERT and visual features using either ResNet18 or ResNet50.

Four different fusion techniques were employed in an early fusion setting in order to compare the performance of the techniques. The textual and visual features were combined before passing to a classifier. The techniques include: MFB, MCB, MFB with self-attention and MCB with self-attention. To ensure a fair evaluation, the dataset was split into 80:20 ratio for each experimental run. Each model was trained for up to 10 epochs using the Adam optimizer with a learning rate of 1e-4 and a batch size of 32. Adam optimizer was chosen because of its efficiency to optimize deep learning models across various domains (Mohammed & Nazri 2025). Additionally, cross-entropy loss was used as the objective function and early stopping was applied based on validation loss to prevent overfitting. Last but not least, the model performance was evaluated using classification accuracy.

RESULTS AND DISCUSSION

This section presents the experimental results and discusses the performance of various fusion techniques applied to multimodal fake news detection using the Fakeddit dataset. The evaluation focuses on how different fusion strategies, MFB, MFB with self-attention (Att-MFB) and MCB with self-attention (Att-MCB) perform across multiple dataset sizes and model configurations. Performance is assessed using accuracy as the primary metric. The experimental results are summarized in the following tables. Table 1. presents the performance of various fusion techniques using the BERT and ResNet18 combination, while Table 2. reports the corresponding results using BERT and ResNet50. Each table compares the accuracy achieved across multiple dataset sizes for the four fusion techniques.

TABLE 1. Results on BERT and ResNet18 across all fusion techniques

Fusion Technique	10k	20k	30k	40k	50k	60k	70k	80k	90k
MFB	0.60	0.65	0.63	0.66	0.69	0.69	0.66	0.70	0.72
Att-MFB	0.59	0.63	0.60	0.67	0.67	0.68	0.70	0.70	0.71
MCB	0.61	0.73	0.68	0.65	0.68	0.71	0.71	0.71	0.71
Att-MCB	0.68	0.65	0.66	0.68	0.66	0.68	0.63	0.70	0.73

TABLE 2. Results on BERT and ResNet50 across all fusion techniques

Fusion Technique	10k	20k	30k	40k	50k	60k	70k	80k	90k
MFB	0.60	0.68	0.68	0.67	0.71	0.72	0.73	0.81	0.71
Att-MFB	0.65	0.71	0.65	0.72	0.70	0.66	0.70	0.80	0.69
MCB	0.71	0.62	0.69	0.71	0.71	0.82	0.81	0.73	0.74
Att-MCB	0.52	0.66	0.56	0.69	0.76	0.71	0.65	0.68	0.69

Figure 4 below illustrates the accuracy performance of the BERT + ResNet18 model across different dataset sizes using two fusion techniques involving MFB and Att-MFB. The highest accuracy is observed with the MFB at 90k samples, reaching 0.72. Meanwhile, the lowest accuracy occurs with Att-MFB at 30k samples, scoring 0.59. Although Att-MFB initially performed slightly lower than MFB, it gradually catches up as the dataset size increases. It reached the highest score at 0.71 with 90k samples, almost equivalent to MFB score. This pattern may indicate that the self-attention mechanism in Att-MFB requires more data to generalize effectively. This is because self-attention has the ability to reweight feature interactions when the model has access to sufficient samples to learn complex cross-modal patterns. The narrowing performance gap at higher dataset sizes highlights the potential of attention-enhanced pooling, particularly for high-resource environments. However, it suggests that the added complexity of attention may not always translate to significant gains.

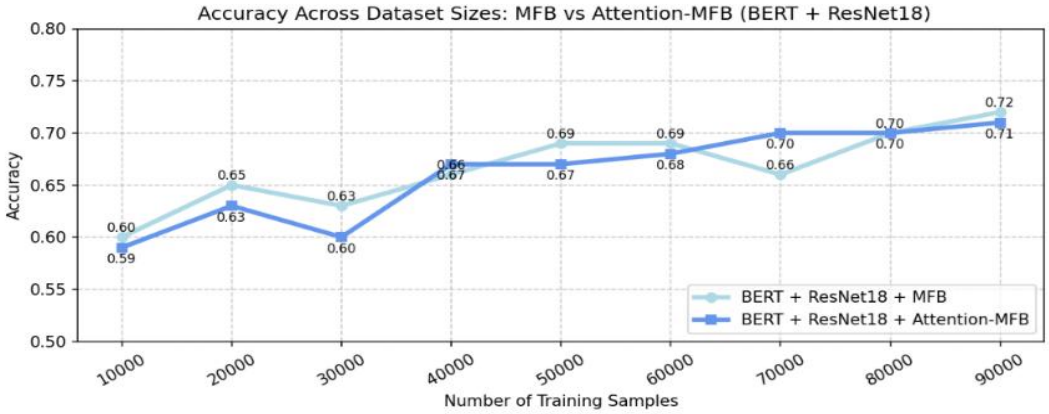

FIGURE 4. Comparison of model Bert + ResNet18 with MFB and Attention-MFB

Furthermore, Figure 5 shows the accuracy performance of the BERT + ResNet50 with MFB and Att-MFB. The results show that accuracy improves as the training dataset size increases with some fluctuations observed at certain points. Initially, at 10k and 20k samples, the Att-MFB fusion slightly outperforms the MFB. The accuracy score reaches at 0.65 and 0.71 compared to the MFB with 0.60 and 0.68. Conversely, MFB demonstrates steadier growth in performance as the dataset increases. It reached the highest accuracy at 0.81 with 80k samples, slightly ahead of Att-MFB at 0.80. This pattern may imply that Att-MFB can offer better performance when data is scarce, though it may also introduce instability in the learning process. Meanwhile, MFB proves to be more dependable for a larger dataset where consistency takes priority.

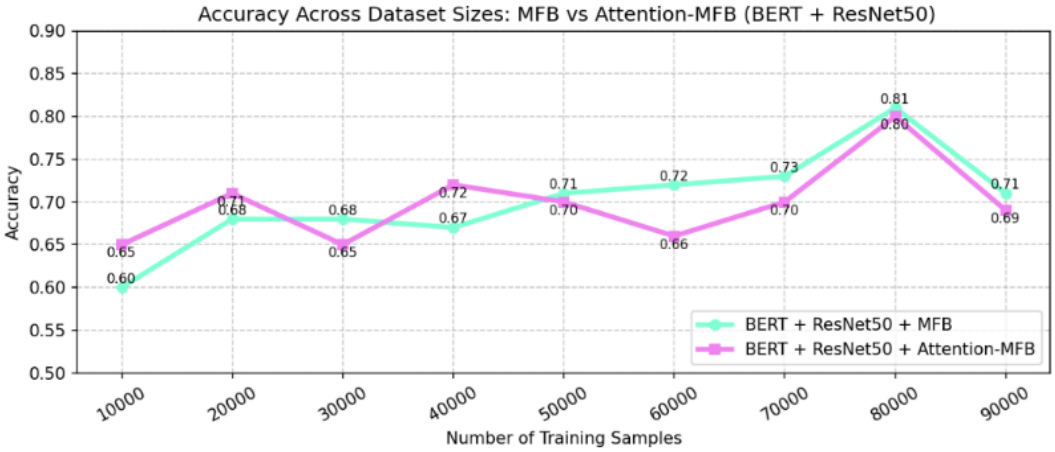

FIGURE 5. Comparison of model Bert + ResNet50 with MFB and attention-MFB

Figure 6 and Figure 7 present a side-by-side comparison of the two ResNet models, ResNet18 and ResNet50, using the same fusion techniques, MFB and Att-MFB. Both models show a positive trend in performance as the training dataset size increases. It suggests that the more training data is used, the more beneficial it is to the model performance. ResNet50 consistently achieves higher accuracy than ResNet18 with the most significant margin observed at 80k samples according to both figures. The highest accuracy is observed where BERT + ResNet50 + MFB achieves 0.81 accuracy. The deeper architecture of ResNet50 may influence the performance gap. The additional layer of the model gives it a greater ability to extract complex visual patterns from larger datasets. On the other hand, Figure 7 displays the performance trend using the Att-MFB fusion technique. Interestingly, while ResNet50 manage to achieve high

accuracy at 0.80 with 80k samples, it is less consistent throughout the training compared to MFB. Attention-based fusion may require more training data to stabilize and generalize effectively, which is reflected in the fluctuating pattern.

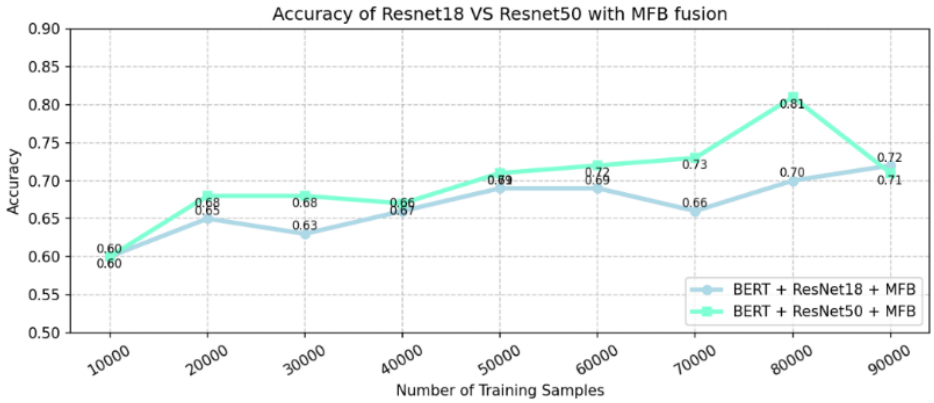

FIGURE 6. Comparison of model Bert + ResNet18 and ResNet50 with MFB fusion

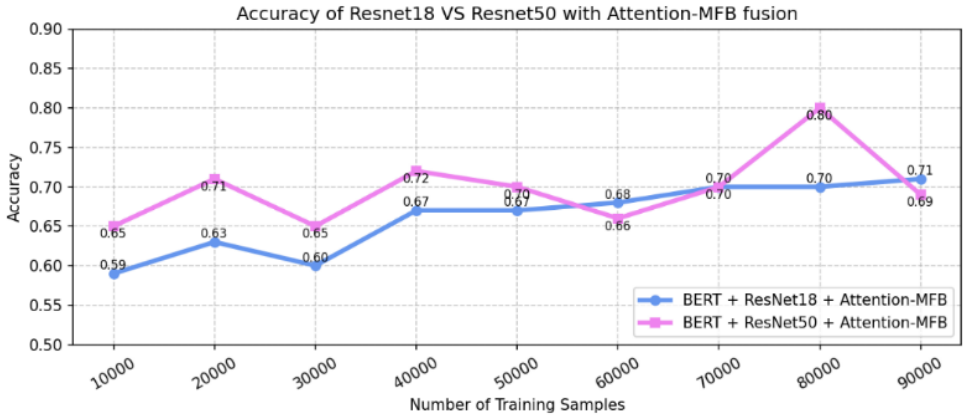

FIGURE 7. Comparison of model Bert + ResNet18 and ResNet50 with Attention-MFB fusion

The results in Figure 8 indicate that Bert + ResNet18 with MCB achieves the highest peak at 0.73 with 20k samples. A plateau performance is also observed at 0.71 in the larger datasets starting from 60k samples. Meanwhile, Att-MCB demonstrates more performance variation, reaching the highest accuracy at 0.73 with 90k samples. This indicates that the addition of attention can enhance the model's ability to focus on more informative features. Yet, the improvements are not consistently significant, suggesting that incorporating attention may have its benefits depending on the data distribution or size. The fluctuating performance of Att-MCB may also reflect its higher representational complexity. Without sufficient data volume or careful tuning, that complexity can work against the model. The attention mechanism enables the model to dynamically reweight the feature, however it may increase the risk of overfitting and introduce instability during training.

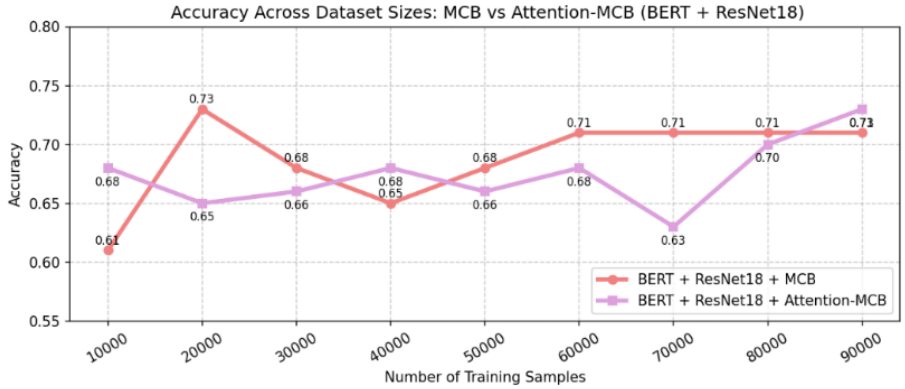

FIGURE 8. Comparison of model Bert + ResNet18 with MCB and Attention-MCB fusion

In comparison of MCB and Att-MCB fusion techniques using ResNet50, MCB exhibits more stable performance as shown in Figure 9. The highest recorded accuracy across all experiments at 0.82 with 60k samples. Although Att-MCB achieves a reasonable score at 0.76 with 50k samples, it still performs slightly lower in several other dataset sizes compared to MCB. The attention-based fusion theoretically intended to provide more refined cross-modal feature interactions. However, in practice, the benefits appear inconsistent across experiments. The deeper architecture of ResNet50 may already capture complex patterns, possibly reducing the added value of attention mechanisms. As a result, the training instability of Att-MCB may require more careful tuning to achieve steady improvements.

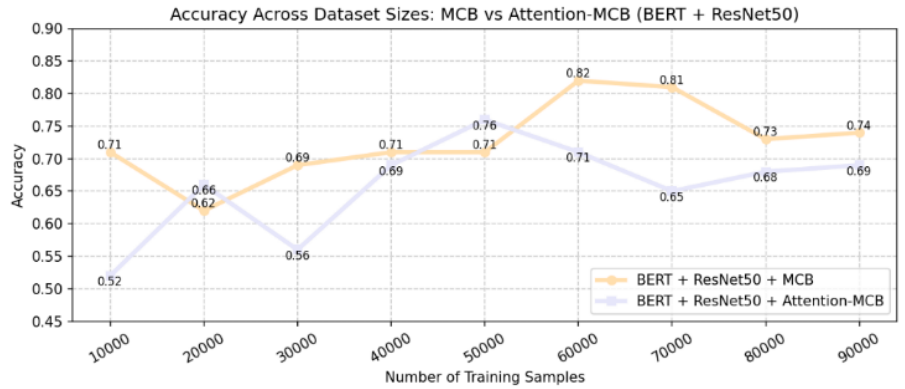

FIGURE 9. Comparison of model Bert + ResNet50 with MCB and Attention-MCB fusion

From the overall results, ResNet50 consistently achieves better performance than ResNet18, likely due to the model's deeper architecture to capture complex visual patterns. While attention-enhanced techniques like Att-MCB and Att-MFB show some potential, the performance remains inconsistent and may introduce training instability. In contrast, standard MFB offers more stable and reliable results across dataset sizes, making it a practical choice for scalable fake news detection.

CONCLUSION

This paper presented a comparative analysis of four fusion techniques for multimodal fake news detection using both text and image modalities. Namely, MFB, MCB, Attention-based for MFB and Attention-based for MCB. The experiments were conducted by combining the extracted features from BERT and ResNet models. The Fakeddit dataset is chosen to set up a fine-grained setting. Then, the performance of each fusion technique was evaluated by varying

the dataset sizes. The findings show that MFB consistently offers stable and high performance, especially on larger datasets. This technique also demonstrated stronger performance than MCB, although MCB still offered reliable results. Meanwhile, the self-attention-enhanced variants (Att-MFB, Att-MCB) show some potential but suffer from performance fluctuations and are more sensitive to the changes in dataset sizes. Moreover, models that used ResNet50 generally achieved better results than ResNet18, especially when paired with MFB. These results highlight the importance of selecting fusion techniques that can handle complex unified features while still maintaining stable performance. This paper contributes to a better understanding of the way fusion techniques influence multimodal fake news detection. It also offers practical guidance for selecting suitable fusion techniques based on dataset size and model architecture.

ACKNOWLEDGEMENT

The Malaysian Government through the Fundamental Research Grant Scheme (FRGS), Ministry of Higher Education Malaysia, has funded this research project under grant number FRGS/1/2023/ICT02/UMT/03/1.

REFERENCES

Bhattarai, B., Granmo, O.-C. & Jiao, L. 2022. Explainable Tsetlin Machine Framework for Fake News Detection with Credibility Score Assessment. Dlm. Calzolari (pnyt.), Bechet (pnyt.), Blache (pnyt.), Choukri (pnyt.), Cieri (pnyt.), Declerck (pnyt.), Goggi (pnyt.), et al. (pnyt.). *LREC 2022: Thirteen International Conference on Language Resources And Evaluation*, pp.4894–4903.

Davoudi, M., Moosavi, M. R. & Sadreddini, M. H. 2022. DSS: A hybrid deep model for fake news detection using propagation tree and stance network. *Expert Systems with Applications*, *198*. doi:10.1016/j.eswa.2022.116635.

Devlin, J., Chang, M. W., Lee, K. & Toutanova, K. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, *1*(Mlm), 4171–4186.

Fukui, A., Park, D. H., Yang, D., Rohrbach, A., Darrell, T. & Rohrbach, M. 2016. Multimodal compact bilinear pooling for visual question answering and visual grounding. *EMNLP 2016 - Conference on Empirical Methods in Natural Language Processing, Proceedings*, 457–468. doi:10.18653/v1/d16-1044.

Guo, Y., Ge, H. & Li, J. 2023. A two-branch multimodal fake news detection model based on multimodal bilinear pooling and attention mechanism. *Frontiers in Computer Science*, *5*(April). doi:10.3389/fcomp.2023.1159063.

He, K., Zhang, X., Ren, S. & Sun, J. 2016. Deep Residual Learning for Image Recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, hlm.Vol. 2016-Decem, 770–778. IEEE. doi:10.1109/CVPR.2016.90.

Hoy, N. & Koulouri, T. 2022. Exploring the Generalisability of Fake News Detection Models. *IEEE International Conference on Big Data (Big Data), Osaka, Japan*, 5731–5740. doi:10.1109/BigData55660.2022.10020583.

Hua, J., Cui, X., Li, X., Tang, K. & Zhu, P. 2023. Multimodal fake news detection through data augmentation-based contrastive learning. *Applied Soft Computing*, *136*. doi:10.1016/j.asoc.2023.110125.

Jing, J., Wu, H., Sun, J., Fang, X. & Zhang, H. 2023. Multimodal fake news detection via progressive fusion networks. *Information Processing and Management*, *60*(1), 103120. doi:10.1016/j.ipm.2022.103120.

Kumari, R. & Ekbal, A. 2021. AMFB: Attention based multimodal Factorized Bilinear Pooling for multimodal Fake News Detection. *EXPERT SYSTEMS WITH APPLICATIONS*, *184*. doi:10.1016/j.eswa.2021.115412.

Li, S. & Tang, H. 2024. Multimodal Alignme.nt and Fusion: A Survey 1–20.

Lisangan, E. A., Tungadi, A. L. & Wibowo, F. 2022. Fake News Detection: An Image-Based Semi-Automated Method Using Statistic Feature. *AIP Conference Proceedings*, *2578*(November). doi:10.1063/5.0106217.

Liu, P., Qian, W., Xu, D., Ren, B. & Cao, J. 2023. Multi-Modal Fake News Detection via Bridging the Gap between Modals. *ENTROPY*, *25*(4). doi:10.3390/e25040614.

Mohammed, A. G. & Nazri, M. Z. A. 2025. Heart Disease Prediction Using Artificial Neural Network with ADAM Optimization and Harmony Search Algorithm. *Asia-Pacific Journal of Information Technology and Multimedia*, *14*(1), 194–218. doi:10.17576/apjitm-2025-1401-12.

Nadeem, M. I., Ahmed, K., Li, D., Zheng, Z., Alkahtani, H. K., Mostafa, S. M., Mamyrbayev, O. et al. 2023. EFND: A Semantic, Visual, and Socially Augmented Deep Framework for Extreme Fake News Detection. *SUSTAINABILITY*, *15*(1). doi:10.3390/su15010133

Nakamura, K., Levy, S. & Wang, W. Y. 2020. r / Fakeddit : A New Multimodal Benchmark Dataset for Fine-grained Fake News Detection (May), 6149–6157.

Ren, J. 2024. Multimodal Sentiment Analysis Based on BERT and ResNet.

Segura-Bedmar, I. & Alonso-Bartolome, S. 2022. Multimodal Fake News Detection. *Information (Switzerland)*, *13*(6). doi:10.3390/info13060284.

Sengan, S., Vairavasundaram, S., Ravi, L., AlHamad, A. Q. M., Alkhazaleh, H. A. & Alharbi, M. 2023. Fake News Detection Using Stance Extracted Multimodal Fusion-Based Hybrid Neural Network. *Ieee Transactions on Computational Social Systems*. doi:10.1109/TCSS.2023.3269087.

Shaharudin, S. N., Yussof, W. N. J. H. W., Hitam, M. S., Awalludin, E. A., Ismail, N. B. & Ibrahim, M. E. S. C. 2025. E-Swish Activations Function in ResNet Architectures for Enhanced Sea Turtle Individual Recognition. *Asia-Pacific Journal of Information Technology and Multimedia*, *14*(1), 295–310. doi:10.17576/apjitm-2025-1401-16

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. et al. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*, *2017-Decem* (Nips), 5999–6009.

Wang, H., Wang, S. & Han, Y. 2022. Detecting fake news on Chinese social media based on hybrid feature fusion method. *EXPERT SYSTEMS WITH APPLICATIONS*, *208*. doi:10.1016/j.eswa.2022.118111.

Wang, J., Mao, H. & Li, H. 2022. FMFN: Fine-Grained Multimodal Fusion Networks for Fake News Detection. *Applied Sciences (Switzerland)*, *12*(3). doi:10.3390/app12031093.

Yu, Z., Yu, J., Fan, J. & Tao, D. 2017. Multi-modal Factorized Bilinear Pooling with Co-attention Learning for Visual Question Answering. *Proceedings of the IEEE International Conference on Computer Vision*, *2017-Octob*, 1839–1848. doi:10.1109/ICCV.2017.202.

Yuan, L., Jiang, H., Shen, H., Shi, L. & Cheng, N. 2023. Sustainable Development of Information Dissemination: A Review of Current Fake News Detection Research and Practice. *Systems*, *11*(9), 458. doi:10.3390/systems11090458.

Zannettou, S., Caulfield, T., Bradlyn, B., De Cristofaro, E., Stringhini, G. & Blackburn, J. 2020. Characterizing the use of images in state-sponsored informationwarfare operations by russian trolls on Twitter. *Proceedings of the 14th International AAAI*

Conference on Web and Social Media, ICWSM 2020, (Icwsm), 774–785. doi:10.1609/icwsm.v14i1.7342.

Zhang, C., Yang, Z., He, X. & Deng, L. 2020. Multimodal Intelligence: Representation Learning, Information Fusion, and Applications 1–15. doi:10.1109/JSTSP.2020.2987728.

Zou, T., Qian, Z., Li, P. & Zhu, Q. 2024. PVCG: Prompt-Based Vision-Aware Classification and Generation for Multi-Modal Rumor Detection. *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 11036–11040. doi:10.1109/icassp48485.2024.10447285.