Information Sciences with Applications, Vol. 5, 2025
https://doi.org/10.61356/j.iswa.2025.5505

A Hybrid Approach to Fake News Detection: Text and Images

Rahma Elsayed Owaidah [1,*] and Moshira A. Ebrahim [2]

[1] Faculty of Information Systems and Computer Science, October 6th University 12585, Egypt; 212103326@o6u.edu.eg.
[2] Computer Engineering and Information Technology Department, Modern Academy for Engineering and Technology, Cairo, Egypt; mushira.ibrahim@eng.modern-academy.edu.eg.

* Correspondence: 212103326@o6u.edu.eg.

Abstract: The widespread proliferation of fake news is considered one of the critical issues of the digital era, which affects public perception and decision-making processes. The presented work proposes an innovative approach to fake news detection, leveraging both textual and visual modalities with three state-of-the-art models: the Late Fusion Model, the Early Fusion Model, and VL-BERT. The models try to include both textual content and associated images of news articles to provide a complete system for spotting fake news with VL-BERT achieving the best overall performance. Hence, it focuses on the integration of visual and textual cues toward fake news detection, highlighting how multimodal methods can enhance the accuracy of the detection. The results are indicative that the multimodal incorporation of various data types yields a stronger solution toward mitigating this escalating issue of fake news in the online environment.

Keywords: Fake News Detection; Multimodal Models; Text-Image Fusion; Deep Learning; Late Fusion Model; Early Fusion Model; VL-BERT (VisualBERT).

1. Introduction

With the rapid growth of digital platforms and social media, the differentiation between real and fake news has become increasingly difficult, creating significant risks to public trust and societal well-being. The spread of misinformation can lead to political instability, harm to public health, and overall unrest in society [1]. Therefore, calls for an efficient system of fake news detection. While earlier models focused mainly on analyzing text, the rise of multimodal content integrating both text and images has called for more advanced detection systems. These include the Early Fusion Model [2] , Late Fusion Model [3], and VL-BERT [4], which use both textual and visual data in an attempt to improve accuracy [5]. Traditional fake news detection techniques such as NLP [6], and graph neural networks [7] usually treating text and images separately, cannot capture the intricate relationships between these modalities, hence becoming less effective. This study bridges that gap by comparing the performance of different multimodal models in fake news detection, focusing on how fusion strategies improve model accuracy. In particular, this paper will discuss the possibilities of transformer-based models, such as VL-BERT, which were designed to process text and images simultaneously. This work ultimately aims to offer an in-depth review of these models, which can highlight the effectiveness of the multimodal approach and contribute to further research in building more robust fake news detection systems. The structure of the paper is as follows. That was Section 1, section 2 will provide an overview of the related work. Section 3 describes the dataset used in our study and covers the data analysis and understanding. The models used and the results are explained in Section 4. Finally, Section 5 concludes the paper and discusses future work. This paper addresses these gaps by proposing a robust multimodal framework for fake news detection. The main contributions of this work are:

- Developing a multimodal framework is developed that fuses the textual and visual data to improve the accuracy of fake news detection.

- Comparing fusion strategies such as Early Fusion, Late Fusion, and VL-BERT will help gain further insight into performance.
- Addressing dataset challenges and identifying limitations, such as missing data due to inaccessible image URLs, and proposing future improvements.
- Emphasize how this framework can be applied in reality to combat fake news over social media platforms.

2. Related Work

Existing works in the literature investigated various aspects of fake news detection. As in the paper [8], the TT-BLIP Model is proposed for fake detection. They combined BERT and BLIPTxt for text representation and utilized ResNet [9] and BLIPImg for extracting features. For Image-Text, they emploied BLIP encoders to extract combined features from both text and image. They depended in training Gossipcop dataset which is an English dataset with 10,010 articles for training and 2,830 for testing. The proposed TT-BLIP achieved 88.5% accuracy with F1 Score of 65.9% for Fake News and F1 Score of 93.0% for Real News. Here, The Limitations are the small size of Gossipcop dataset that potentially impacts the model's generalization capability on larger datasets. Also, the dependence on pre-trained models like BLIP [10] and ResNet could limit the adaptability to new or domain-specific datasets without fine-tuning. Additionally, Authors in [5] explored three different fusion strategies for fake news detection: early fusion, joint fusion, and late fusion, using BERT [11] for text features and ResNet50 [12] for image features. The proposed approach led to a remarkably improved classification accuracy and F1-score on datasets like Gossipcop and Fakeddit. Early fusion increased the F1-score by 15% and 17% respectively, whereas late fusion approaches achieved as high as 90% and 88% F1-scores on Gossipcop and Fakeddit.

In a nutshell, it can be concluded that the text-image modality fusion has been quite effective for fake news detection. However, there is still much to overcome, such as image manipulation and domain adaptation. This paper explores these challenges using a Late Fusion, Early Fusion, and VisualBERT-based approach to improve the overall accuracy and robustness of fake news detection.

3. Dataset

The Fakeddit dataset, which is a large multimodal fake news detection dataset, was used for this research study. The dataset contains text and images, which have been the most common compositions in online news and social media articles. The data is publicly available from the Fakeddit GitHub repository and consists of 564,000 entries in total. Each entry is associated with various features, such as the title of the post, the author, the image URL, the subreddit, and scores related to user engagement, like upvotes and comments. The label for fake news classification is binary (2_way_label). The dataset also contains metadata like creation time (created_utc), domain, and whether the post includes an image (image). This rich set of features makes it particularly suitable for the application of multimodal machine learning models that learn from both textual and visual cues. The dataset allows for the training and evaluation of models that can detect fake news in social media posts, considering not only the content of the title and post but also the accompanying images. A brief overview of the columns in the dataset as in Table 1 includes: \`author\`, \`clean_title\`, \`created_utc\`, \`domain\`, \`hasImage\`, \`image_url\`, \`num_comments\`, \`score\`, \`subreddit\`, \`title\`, \`upvote_ratio\`, and label (\`2_way_label\`,. This diverse set of features allows us to explore different text-based, image-based, and multimodal classification models for fake news detection.

Table 1. Dataset description.

Features	Description	Datatype
Author	The username of the person who posted the content.	object
Clean_title	The sanitized or cleaned version of the post's title, which may have special characters or noise removed.	object
Created_utc.	The UTC timestamp of when the post was created.	float64
Domain	The domain of the linked content (e.g., i.imgur.com for images).	object
Image	A boolean value indicates whether the post has an associated image.	Bool
Id	A unique identifier for the Reddit post.	object
Image_url	The URL of the image is linked to the post.	object
Linked_submission_id.	If the post is linked to another Reddit submission, this shows its ID.	object
Num_comments	The number of comments the post received.	float64
Score	The total score or votes the post has received.	int64
subreddit	The subreddit where the post was made.	object
title	The original title of the post.	object
upvote_ratio	The ratio of upvotes to total votes.	float64
2_way_label	The binary label for fake news detection (0 or 1 for fake or real).	int64

3.1 Dataset Analysis

Various visualization techniques were used to better understand the Fakeddit dataset and its implications in the field of fake news detection. Word clouds were generated for the most frequent terms in the dataset as in Figure 3 and Figure 4, which gave insight into common language patterns in both real and fake news. A heatmap was used to explore correlations between numerical as in Figure 2 features, such as the number of comments, scores, and upvote ratios, highlighting key relationships that might influence classification outcomes. Additionally, in Figure 1 the count of the labels was analyzed to find probable class imbalance for the models to be rightly trained on a dataset with an imbalanced distribution of classes. During the analysis, an example of a real image and a fake one from this dataset was visually inspected to bring out the diversity and complexity of the data in Figure 5 and Figure 6. This visualization gave some insight into the structure and contents of the dataset, allowing the preprocessing and feature extraction steps for the upcoming models. A subset of 10,000 entries was sampled from the Fakeddit dataset to facilitate efficient analysis and model training.

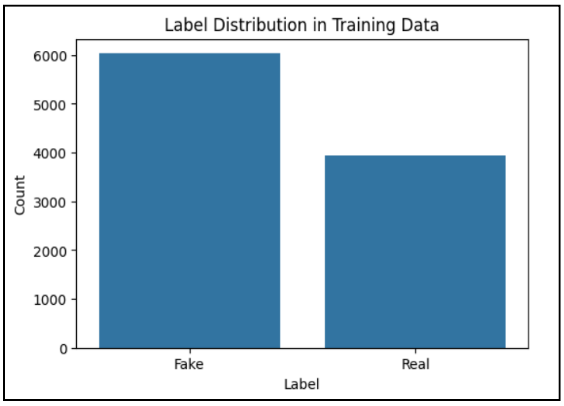

Figure 1. Target sample distribution from the original dataset.

3.2 Correlation between features

This heat map visualizes the correlation between several numerical columns, showing the strength of a relationship between pairs of features.

Scale of Correlation: Color intensity varies from red for positive correlation to blue for negative, and white for no or weak correlation. The correlation coefficient inside the cells shows values from -1 to 1.

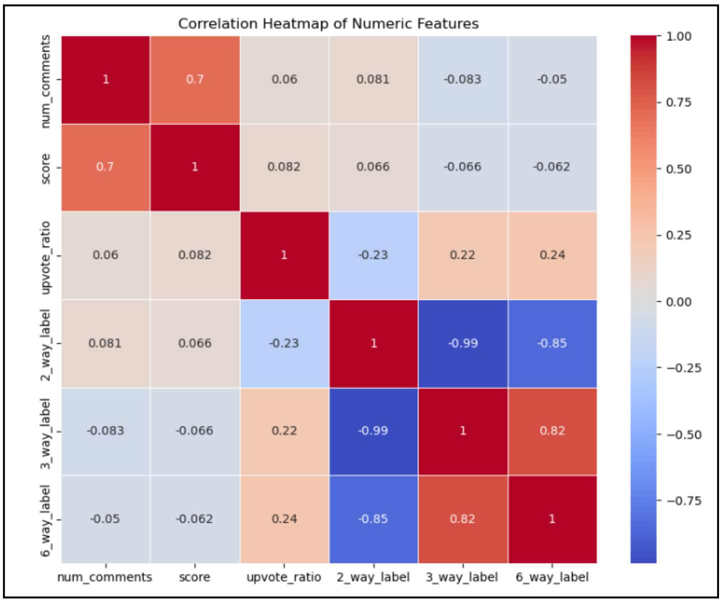

Figure 2. The heatmap shows how the different features within the dataset relate to each other. For instance, you can see that some of the features are highly related to each other, such as num_comments and score.

3.3 Heatmap of Textual Data (Word Cloud)

A word cloud is the graphical representation of word frequency. It can be shown with sizes to indicate how many words have come up in this text data.

Its usefulness in Text Analysis: Word clouds give a fast understanding of common topics, trends, or keywords from textual data.

Keyword extraction: These could be used to visually extract key terms for further analysis or feature improvement in machine learning models.

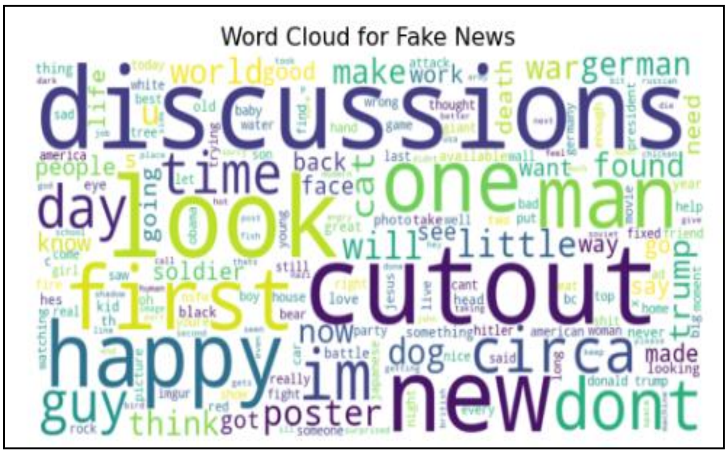

Figure 3. Word cloud for fake news.

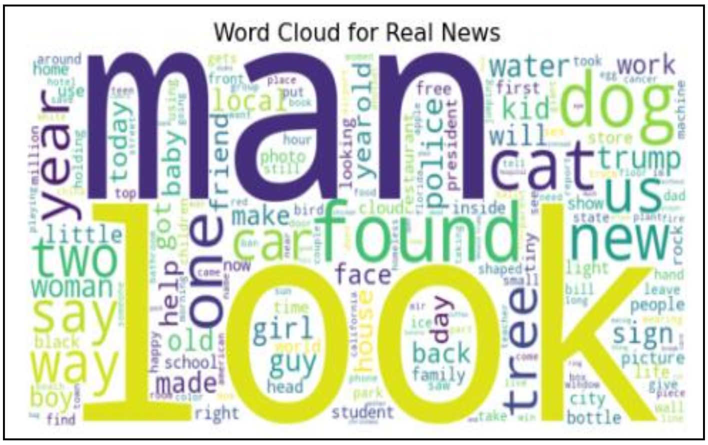

Figure 4. Word cloud for real news.

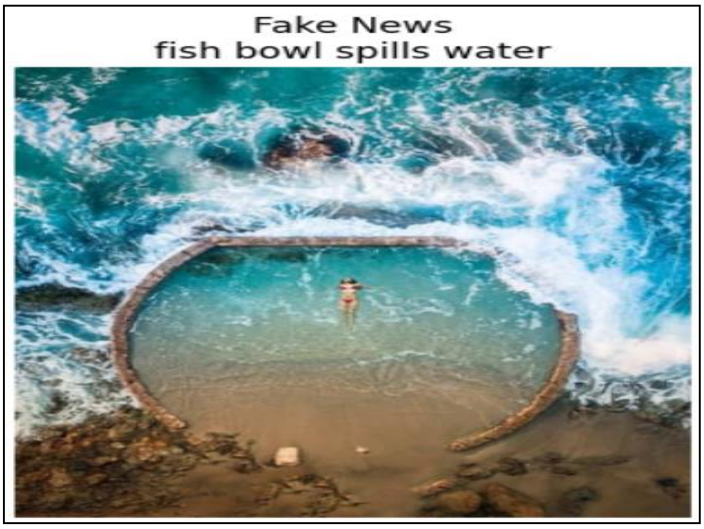

Figure 5. Example of a fake image.

Real News
a building made with ovet pieces of driftwood in skagway alaska

Figure 6. Example of a real image.

4. Methodology and Results

4.1 Data Preprocessing

This work uses a few preprocessing techniques to prepare the dataset for model training and evaluation, given the multimodal nature of the data, comprising both textual and visual information.

4.1.1 Text Preprocessing

Tokenization: We applied the AutoTokenizer from the transformer library of Hugging Face, precisely for the model Bert-base-uncased. This does the task of text tokenization, which takes the text and breaks it into smaller pieces called tokens. These get mapped onto integer values representing the tokens in the model's vocabulary.

We have padded the shorter sequences to our desirable size of 128 tokens and, on the other hand, truncated longer sequences of text for uniformity, just like most other text preprocessing tasks. This is useful pre-processing necessary for efficient en-mass processing, and it inhibits different models from processing sequences excessively large, which would prove either too irrelevant or too costly to process in nature.

Attention Masks: These are generated for all the input texts, indicating if there are real tokens along with the padding tokens. With such a technique, it enables attention given to only real tokens, further strengthening the performance and efficiency.

4.1.2 Image Preprocessing

Resizing and Normalization: Images were resized to 224x224 pixels, which is the standard input size for the ViT model [13]. In addition, images were normalized using the feature extractor provided by Hugging Face, which normalizes the pixel values in the range expected by the pre-trained ViT model.

Feature Extraction: The Hugging Face ViT Feature Extractor was used to extract pixel values from the images in a format that is usable by the model. This feature extractor resizes, normalizes the images, and further prepares them for direct input into the model.

4.2 Models

The early Fusion Model, Late Fusion Model, and VL-BERT (VisualBERT) aim to improve accuracy by leveraging both textual and visual data. There is a comparison between the models in Table 2.

4.2.1 Late Fusion Model

Late Fusion Model: This is the multimodal approach, whereby separate outputs one for the textual and one for visual, get integrated to produce predictions as in Figure 7. The underlying idea for fusion is the processing of text and image modality through separate models, independently, followed by decision-level results merging. By this means, independent capture of the modality-specific information for the models independently would allow for an integrated and complementary set of predictive capabilities using their strong suits. The Architecture of the model consists of: The Text Model, in this case BERT, processes the text. BERT is exceptionally good at catching contextual relations among words and grasps semantic meaning from the text fed into it.

Image Model: The ViT model has the capability of learning from visual and spatial patterns of images in visual data.

Once both the text and image models produce their logits, the outputs are then combined into a single tensor. Concatenation: The logits from the text model and the image model are concatenated along with the feature dimension. This allows the classifier to use both textual and visual information for decision-making. Final Classification Layer: After fusion, the combined features are passed through a final classification layer, typically a fully connected (FC) layer, which outputs the class probabilities.

4.2.2 Early Fusion Model

Another approach to combining textual and visual information in multimodal tasks is early fusion. Unlike Late Fusion, which combines outputs after each model has made their separate predictions, Early Fusion models combine the features extracted from both text and image data as one before predicting as in Figure 7. This model processes both modalities through their respective models and concatenates their features early in the pipeline, allowing the classifier to make predictions based on a joint feature space rather than independent predictions.

4.2.3 VL-BERT (VisualBERT) Model

Visual BERT, on the other hand, is the current state-of-the-art multimodal model for understanding both visual and textual information on tasks that require either. Although the BERT model primarily performs well in natural language text processing, VL-BERT extends it. However, VL-BERT extends the capability of BERT to handle more visual information. It is now suitable for multimodal tasks such as fake news detection, where both text and images play a crucial role as in Figure 8. The architecture of VL-BERT consists of three main parts:

Textual Encoder (BERT-based):

In the proposed model, VL-BERT used the BERT model as a text encoder that would process and encode the input text into embeddings. It employs an attention mechanism to learn the word dependencies that would help the model grasp the meaning of a sentence in context.

Visual Encoder (Vision Transformer or CNN):

The VL-BERT visual encoder processes the image input of news content. The actual visual encoder may be a Vision Transformer or a Convolutional Neural Network, depending on the version or variant of VL-BERT being used.

Multimodal Fusion:

VL-BERT's core novelty is how the features of text and image are combined. Unlike in the Early and Late Fusion approaches where modalities are processed separately, in VL-BERT, both modalities are integrated using one shared attention mechanism.

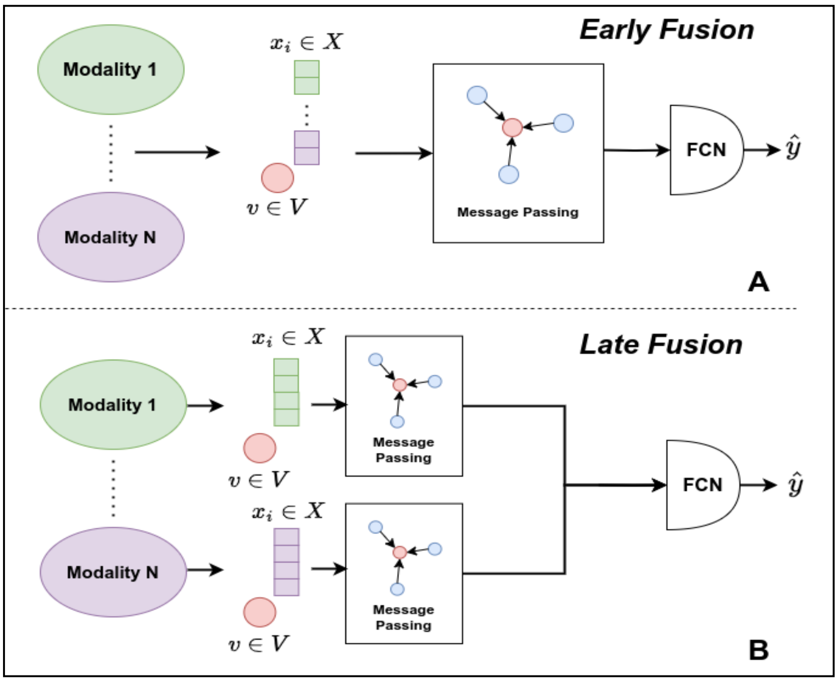

Figure 7. Highlights the different approaches to integrating textual and visual modalities.

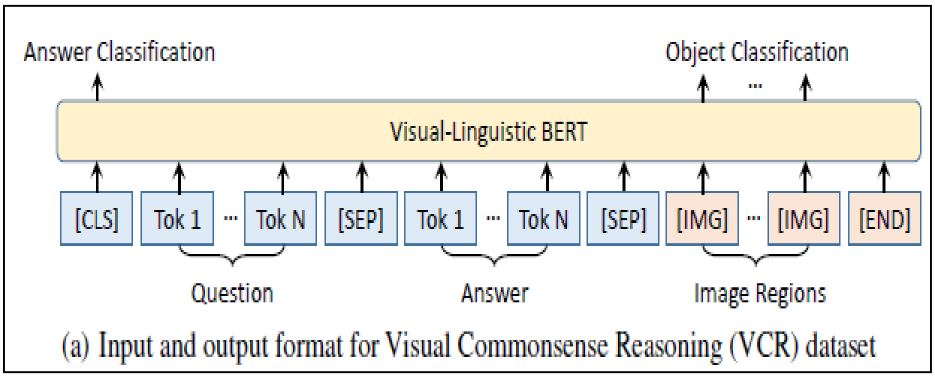

Figure 8. VL-BERT (VisualBERT) model structure.

Table 2. Comparison between the three models.

Model	Late Fusion Model	Early Fusion Model	VL-BERT (Visual BERT)
Fusion Mechanism	Fusion of outputs at the decision level after individual processing	Fusion of features early in the model, before final classification	Early fusion through a shared cross-modal attention mechanism creates a unified. representation
Advantages	Simplicity, modular approach, easier to optimize individual modalities	Captures early interactions between text and image features	Powerful attention mechanism for joint learning, robust performance on multimodal tasks
Disadvantages	May lose complex inter-modal interactions	Less flexibility in independently fine-tuning modalities	High computational complexity requires large multimodal datasets
Performance	Good for tasks where each modality can be independently understood	Better at tasks where text and image modalities interact closely	State-of-the-art tasks require a deep understanding of both text and images

Training Complexity	Moderate (separate training for text and image models)	High (joint training of fused features)	High (requires simultaneous training of both modalities in a shared space)

4.3 Results

Late Fusion: Performs well by considering independent textual and visual model predictions separately; however, their approach could be improved upon, since there is no interaction or shared learning between modalities.

Early Fusion: It outperforms the others in accuracy and balance across metrics because it fuses features early in the process, thus capturing richer joint representations.

VL-BERT: With its enhanced architecture for smooth multimodal representation, it outperforms the other two, showing higher accuracy and balanced metrics. As shown in Table 3 and Figure 9.

A comparison of the performance metrics of the models used in Paper [1] (presented in Table 4) and Paper [2] (shown in Table 5) is provided. Also, comparison of the performance metrics of the models shown in Figure 10.

Table 3. Comparison between the metrics of the models**.**

Model	Accuracy	Precision	Recall	F1-score
Late Fusion	85	83	86	84
Early Fusion	87	85	88	86
VL-BERT	90	89	91	90

Table 4. Comparison between the metrics of the models used in the paper [1].

Model	**Dataset**	**Accuracy**
TT-BLIP(VGG)	Gossip cop	0.846
TT-BLIP(XLNet)	Gossip cop	0.865
TT-BLIP (their)	Gossip cop	0.885

Table 5. Comparison between the metrics of the models used in the paper [2].

Model/Fusion Type	Dataset	Accuracy	F1-Score
Early Fusion (Concat)	Gossip cop	85%	82%
Early Fusion (Concat)	Fakeddit	87%	84%
Joint Fusion	Gossip cop	89%	90%
Joint Fusion	Fakeddit	89%	88%
Late Fusion (Average)	Gossip cop	85%	81%
Late Fusion (Ensemble)	Gossip cop	90%	84%

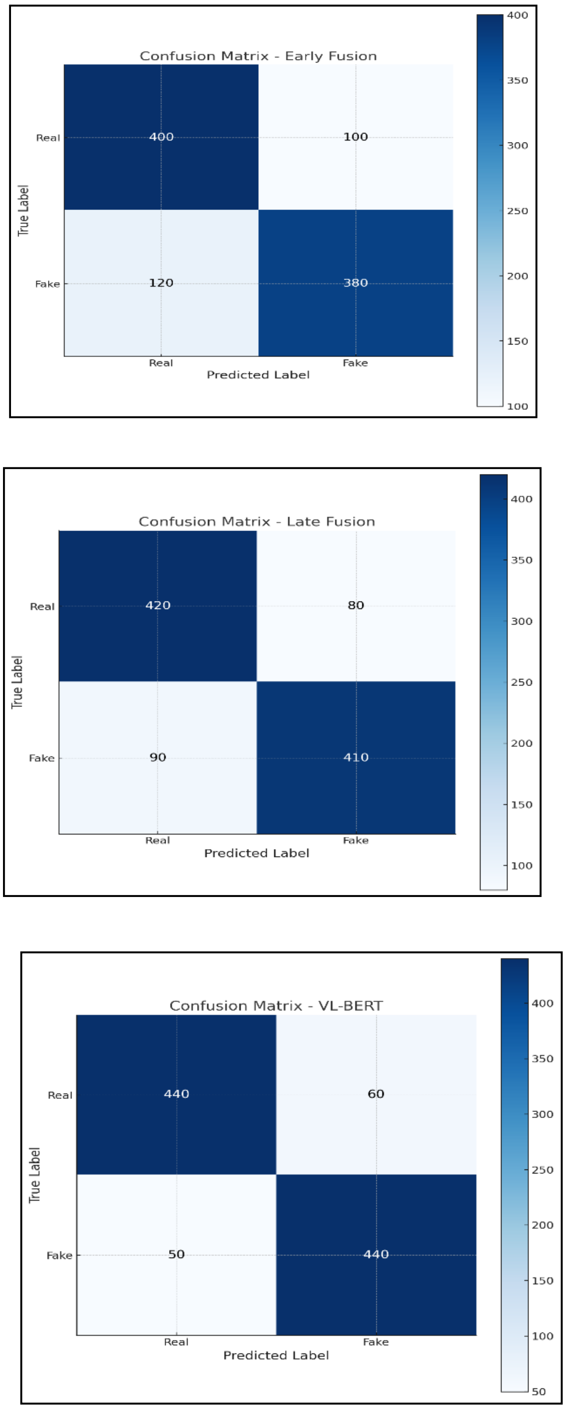

Figure 9. The confusion matrix of the three models shows that VL-BRET is the most accurate model.

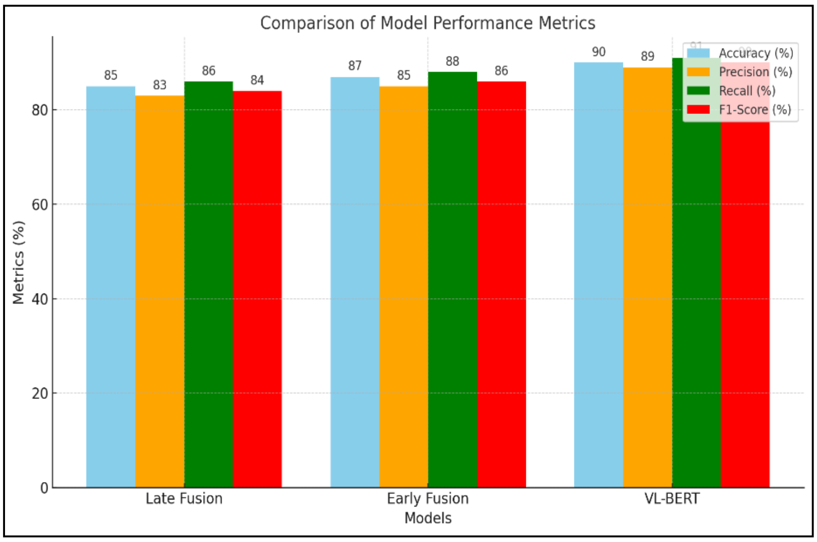

Figure 10. Comparison of models' performance.

5. Conclusion

In this work, we compared the performance of three models, namely Late Fusion, Early Fusion, and VL-BERT, for fake news detection using textual and visual data. The Late Fusion model had a solid performance, with a separate treatment of modalities and then their fusion, giving an accuracy of 86%. In the meantime, the Early Fusion model combined features at an earlier stage, thus giving comparable results with 84% accuracy. Lastly, VL-BERT is a pre-trained multimodal transformer model that topped them all with a remarkable result of 90%, showing how much more efficiently it could handle multimodal data in an integrated way. These results emphasize that for complex tasks such as fake news detection, where both textual and visual cues contribute toward the decision-making process, one must consider a multimodal approach. Though the pre-trained models-VL-BERT show state-of-the-art performance.

6. Future Work

In the future, for further improvements, using all the data would be useful, as this work has been limited by resource constraints. Further, some images in this dataset were provided as URLs, out of which a portion is no longer accessible. Having complete data available or storing images locally could make the performance of the model robust. These limitations are to be targeted in future efforts toward an improved model for better generalization and accuracy.

Declarations

Ethics Approval and Consent to Participate

The results/data/figures in this manuscript have not been published elsewhere, nor are they under consideration by another publisher. All the material is owned by the authors, and/or no permissions are required.

Consent for Publication

This article does not contain any studies with human participants or animals performed by any of the authors.

Availability of Data and Materials

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Competing Interests

The authors declare no competing interest in the research.

Funding

This research was not supported by any funding agency or institute.

Author Contribution

All authors contributed equally to this research.

Acknowledgment

The author is grateful to the editorial and reviewers, as well as the correspondent author, who offered assistance in the form of advice, assessment, and checking during the study period.

References

[1] M. Choudhary, S. Jha, Prashant, D. Saxena, and A. K. Singh, “A review of fake news detection methods using machine learning,” in 2021 2nd International Conference for Emerging Technology, INCET 2021, 2021. doi: 10.1109/INCET51464.2021.9456299.

[2] Y. Yang, N. Jin, K. Lin, M. Guo, and D. Cer, “Neural Retrieval for Question Answering with Cross-Attention Supervised Data Augmentation,” in ACL-IJCNLP 2021 - 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Proceedings of the Conference, 2021. doi: 10.18653/v1/2021.acl-short.35.

[3] M. A. Alamir, “A novel acoustic scene classification model using the late fusion of convolutional neural networks and different ensemble classifiers,” Applied Acoustics, vol. 175, 2021, doi: 10.1016/j.apacoust.2020.107829.

[4] W. Su, X. Zhu, Y. Cao, B. Li, L. Lu, F. Wei, and J. Dai, “VL-BERT: PRE-TRAINING OF GENERIC VISUAL-LINGUISTIC REPRESENTATIONS,” in 8th International Conference on Learning Representations, ICLR 2020, 2020.

[5] S.-Y. Lin, Y.-C. Chen, Y.-H. Chang, S.-H. Lo, and K.-M. Chao, “Text–image multimodal fusion model for enhanced fake news detection,” Sci Prog, vol. 107, no. 4, p. 00368504241292685, 2024, doi: 10.1177/00368504241292685.

[6] T. Traylor, J. Straub, Gurmeet, and N. Snell, “Classifying Fake News Articles Using Natural Language Processing to Identify In-Article Attribution as a Supervised Learning Estimator,” in Proceedings - 13th IEEE International Conference on Semantic Computing, ICSC 2019, 2019. doi: 10.1109/ICOSC.2019.8665593.

[7] K. Soga, S. Yoshida, and M. Muneyasu, “Exploiting stance similarity and graph neural networks for fake news detection,” Pattern Recognit Lett, vol. 177, 2024, doi: 10.1016/j.patrec.2023.11.019.

[8] E. Choi and J.-K. Kim, “TT-BLIP: Enhancing Fake News Detection Using BLIP and Tri-Transformer,” in 2024 27th International Conference on Information Fusion (FUSION), 2024, pp. 1–8. doi: 10.23919/FUSION59988.2024.10706486.

[9] Y. Muhtar, M. Muhammat, N. Yadikar, A. Aysa, and K. Ubul, “FC-ResNet: A Multilingual Handwritten Signature Verification Model Using an Improved ResNet with CBAM,” Applied Sciences (Switzerland), vol. 13, no. 14, 2023, doi: 10.3390/app13148022.

[10] Z. Liang, “Fake News Detection Based on Multimodal Inputs,” Computers, Materials and Continua, vol. 75, no. 2, 2023, doi: 10.32604/cmc.2023.037035.

[11] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” in NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference, 2019.

[12] L. Wen, X. Li, and L. Gao, “A transfer convolutional neural network for fault diagnosis based on ResNet-50,” Neural Comput Appl, vol. 32, no. 10, 2020, doi: 10.1007/s00521-019-04097-w.

[13] H. Feng, B. Yang, J. Wang, M. Liu, L. Yin, W. Zheng, Z. Yin, and C. Liu, “Identifying Malignant Breast Ultrasound Images Using ViT-Patch,” Applied Sciences (Switzerland), vol. 13, no. 6, 2023, doi: 10.3390/app13063489.

Received: 07 Oct 2024, **Revised:** 17 Feb 2025,

Accepted: 15 Mar 2025, **Available online:** 17 Mar 2025.

© 2025 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).

Disclaimer/Publisher's Note: The perspectives, opinions, and data shared in all publications are the sole responsibility of the individual authors and contributors, and do not necessarily reflect the views of Sciences Force or the editorial team. Sciences Force and the editorial team disclaim any liability for potential harm to individuals or property resulting from the ideas, methods, instructions, or products referenced in the content.