Hybrid Deep Learning Models Based on Transformers for Fake News Detection in Albanian and English News

Elton Tata
Faculty of Contemporary Sciences and Technologies, South East European University, Tetovo, Republic of North Macedonia

Jaumin Ajdari
Faculty of Contemporary Sciences and Technologies, South East European University, Tetovo, Republic of North Macedonia

Xhemal Zenuni
Faculty of Contemporary Sciences and Technologies, South East European University, Tetovo, Republic of North Macedonia

Mentor Hamiti
Faculty of Contemporary Sciences and Technologies, South East European University, Tetovo, Republic of North Macedonia

ABSTRACT
In the era of social media, news propagation has shifted to digital platforms enabling faster and more extensive information distribution. Fake News Detection (FND) has gained considerable attention, with various studies, but still, it remains a complex task. This study providesan approach that combines several deep learning architectures. First, the Albanian and IFND datasets are collected and loaded. Subsequently, the CoreNLP toolbox is utilized for data pre-processing and feature extraction, followed by the application of the RMS-BERT-CapsNet (Root Mean Square - Bidirectional Encoder Representations from Transformers - Capsule Network) framework. A Deep-Shallow multimodal fusion approach based on Variational Autoencoder (VAE) used to fit and encode the textual and visual data. Performance metrics such as accuracy, loss, F1 score, training loss, and validation loss are used to evaluate the efficiency of the proposed methodology

General Terms
Fake News Identification

Keywords
Fake News, IFND, BERT, BiLSTM.

1. INTRODUCTION
In our rapidly evolving technological world, advanced tools are being developed to better understand human behaviour and the ways we share information. One such challenge is the spread of fake news. In a globally connected society, where ideas travel across continents in seconds, distinguishing fact from fiction has never been more critical. This article explores how emerging technologies are helping to identify fake news, ensuring the integrity of the information we share [1]. Several social networking systems, including Facebook, Instagram, Twitter, and others, have gained a lot of popularity in recent years because they make it simple to obtain information and offer a prompt forum for information exchange. Thus, social media has become an important centre in the spread of fake news, and researchers are rightly paying attention to the spread and access of fake news to online social networking platforms [2]. The widespread use of fake news, whether it comes from human or computer sources, has an adverse effect on society and individuals in both political and social domains. As a result, automated systems for identifying fake news have become essential. However, fake news makes up a significant portion of the information circulating on the internet, which is deliberately produced to attract viewers and shape opinions and choices made by individuals in order to raise money from clicks and influence important events such as political elections [3]. Researchers have investigated the issue of fake news in detail, which is difficult and demands fact-checking. Towards efficient fake news identification, several datasets (such as LIAR, Fake News Corpus, and Fake or Real news) have been made accessible to the world [4].

The current studies on the detection of fake news using deep learning algorithms have had remarkable success with a variety of social network news elements, including textual content, user features, and user comments. However, fake information is not the intended use case for context learning. For instance, a study on fake news detection prioritized analysing the attributes of the initial news spreader while disregarding subsequent user comments [5]. Distinguishing genuine news from fake within the vast sea of publicly available information remains a formidable challenge. Usually, the purpose of the content written in a simple English is to mislead reliable online users. Fake news can have a negative impact on society, so it must be detected using procedures that are reliable and highly effective.To achieve these goals, a fake news detection technique is proposed that uses genetic research and deep learning for news classification [6]. Spreading of fake news can have several detrimental effects, including financial loss and social criticism. Recently, there have been a lot of very influential fake news stories that propagate fake information about various events and purposes, such as about cyberattacks, impacts on the presidential elections, terrorist plots, natural disasters, and extreme weather [7]. Fake news can have political purposes, as was demonstrated in the 2016 United States elections, where its impact increased significantly compared to other real election-related news. To test this impact, are tested the LIAR dataset and the LIAR PLUS dataset using Triple branch BERT, comparing its accuracy to support vector machines (SVM), logistic regression (LR), and a bidirectional short-term memory (Bi-LSTM) neural network [8]. According to the New York Times, fake news is "a fabricated article produced to deceive" and is distributed in various forms on social platforms. For their identification and classification, a model (FNDNet) has been proposed to automatically learn discriminative features for the classification of fake news by combining a deep convolutional neural network (CNN) to extract several features in each layer [9].This highlights the need for adequate research to increase the validity of the material shared online on these social media sites. Consequently, it is essential that this issue is addressed responsibly and automatically without requiring direct human involvement. Given the volume of data that is constantly being posted on social media, as well as the ambiguity and complexity of the news's language and substance, classifying and detecting fake news is one of the most difficult challenges [10].

A. Motivation & Objective

The main motive of this study is to provide suitable and efficient ways to challenge the problem of fake news spread in the digital era, having a substantial impact on the field of fake news identification. The following are the specific objectives of this research,

- To improve the performance of the system and to present an effective processing method that improves the capacity and evaluates news in both Albanian and English languages.
- To improve detection accuracy by extracting a balanced set of features by implementing optimal training methods leading to faster and more accurate model convergence.
- Development of a hybrid method for feature extraction from textual and visual data.
- By using the benefit of both transformer and recurrent-based architecture this method seeks to increase the classification accuracy.

B. Research Contribution

The highlights of the research work are illustrated below,

- To ensure the most accurate assessment of fake news, two datasets, one in Albanian and one in English, were taken into consideration to evaluate the model and compare the results.
- To extract the important characteristics from the data used the RMS-BERT-CapsNet algorithm.
- In order to enhance the identification of fake news, this article presents a unique VAE-based Deep-Shallow multimodal fusion technique.
- Continual Learning with “Gradient Episodic Memory (GEM) and Elastic Weight Consolidation (EWC)” are both used for the purpose of providing flexibility in training.
- Hybrid categorization is achieved by combining the “Vision Transformer (ViT)” and “Bidirectional Long Short-Term Memory (Bi-LSTM)” architectures.

C. Research Organization

The following sections comprise the rest of this document: Section II provides a literature survey of the earlier researches and related to this work, the main statements of the issues that are addressed in the present work. Section III presents the research methodology for the proposed study, and consists of a pseudocode, a mathematical representation, and a protocol. In section IV, the experimental results are presented together with a comparison of the recommended and ongoing tasks. Section V offers a conclusion to the recommended study.

2. RELATED WORK AND PROBLEM STATEMENT

Detecting fake news is a difficult and challenging task, as it is associated with a number of characteristics of the news, such as its authenticity, author's intention, content and form. These techniques, along with the various methods implemented for detecting fake news, are described in the previous work [11]. In the article [12], researchers suggested an “ensemble-based deep learning model” that uses the “LIAR” dataset to categorize news as authentic or fake. Because of the characteristics of the dataset, two models based on deep learning were used. “Bi-LSTM-GRU”-dense deep learning model was used for the text parameter "statement," and dense deep learning method was utilized for the other parameters.The paper [13], suggested the “ICNN-AEN-DM Web-Informed-Augmented Fake News Detection Model”, which uses stacked-layers of convolutional neural networks and deep autoencoders. The data are obtained from reliable sources' websites through online searches, which validate or refute the statements made in the news article. Next, a deep autoencoder-staked CNN layer was built to develop an algorithm using a stochastic deep learning basis.For multimodal fake news detection, the researchers in the study [14], suggested using a “Complementary Attention Fusion with an Optimized Deep Neural Network (CAF-ODNN)”. To capture complex correlations, the method used photo descriptions to encode the photos. The technique to represent good results introduces image captions to represent images semantically.To improve feature extraction, an optimized deep neural network (ODNN) that uses compositional learning is implemented. The efficiency of the model can be reduced if both image and text data of a news story are not taken into account. The research work [15] proposes an automatic technique for identifying fake news by combining text and images to produce a feature vector that is comprehensive and has exceptional data richness. The suggested method preserves the semantic linkages between words while extracting textual attributes using the “BERT (Bidirectional Encoder Representations from Transformers)” paradigm. Unlike the “convolutional neural network (CNN)”, the proposed “CapsNet (Capsule Neural network)” model extracts the maximum illuminating properties of visual data. The suggested model incorporates textual and visual elements from news stories; however, it does not examine user behavioural traits or information about user profiles.The study [16], examines the whole justification for fake news identification. The study also focuses a major emphasis on abilities, traits, taxonomy, classifications of fake information, and strategies for recognizing fake news. This study identified fake news using unpredictable hidden language analysis. In specific, the study provides a complete virtual analysis of the several research works that have inspired this issue by describing the underlying philosophy of the relevant work. In an effort to stop the propagation of misinformation, this research [17], offers an automated technique for recognizing fake news. The proposed effective multimodal fake news detection creates a multimodal feature vector with a high content density by combining textual and auto visual features. A "multimodal factorized bilinear union" is used to combine the collected information in order to have a better performance of the model. Furthermore, visuals and accompanying text have a strong connection, complicated or changed pictures combined with written explanations might occasionally fool the model and cause possible mistakes in analysis. In the study [18], the researchers have used transfer learning to annotate four datasets for fake news and rumour identification with their corresponding emotion class labels. To detecting fake news and rumours, they demonstrate the relationship between a text's validity and its underlying emotion. This leads them to propose a multi-task framework that predicts the content's originality and sentiment to identify misleading news and rumours.

2.1 Problem Statement

This session presents the findings and solutions identified in various studies.

Specific Research Works & Issues: The author of the paper [19] offers a machine-learning method to detect fake news to stop misinformation spreading. The suggested multimodal EFND (Extreme Fake News Detection) develops a multiple-modal characteristic vector featuring a high density of information by combining context-specific, social settings, and graphic information from articles in newspapers and social media. The characteristics that have been acquired are combined via the use of a multimodal factored bilinear pooling technique in order to intensify their correlation while offering a more accurate shared representation.

- Nevertheless, the proposed technique is not tested and calibrated for other languages than English. This may affect the efficiency of the system that detects fake information.

An effective framework for the identification of fake news that is multimodal and incorporates proper multimodal feature fusion is proposed in the paper [20]. The purpose of this framework is to achieve an effective multimodal common depiction by using details from text as well as imagery and working to maximize the connection between the two types of data. Their study shows that when text and image are used together, the model's performance may be enhanced. The model examines both the text and the image of the article to assess whether it is authentic or potentially fake, determining if it should be integrated into the network. The main limitation of this research is,

- The proposed method's difficulty in extracting robust, invariant features from complex images hampers its ability to accurately identify fake news.

The research paper, [21] provides a fusion approach where textual and visual information are integrated in order to create a more reliable model for detecting fake news. To achieve optimal conditions for the most efficient result, it has been tested on different combinations of information. For this purpose, three well-known data sets (Weibo, MediaEval and CASIA) have been used. For learning text features, pre-trained models of Electra and XLnet have been used. While for learning visual features, ELA has been used. The "LIME (Local Interpretable Model-agnostic Explanations)" is used to identify the super pixels that contribute to the interpretability of the given model, hence improving its interpretability.

- If the visual data is removed from its context, the relationship between the image and the text is lost, which leads to a decrease in the performance of the model.

The research, [22] propose a "deep learning" based algorithm for detecting fake news. They have experimented with many methodologies, including machine learning, deep learning, and transformers. The results obtained indicate that deep learning models exhibited superior performance compared to other methods. The findings of this study revealed that it is possible to properly identify fake news by combining a broader language feature collection with machine learning or deep learning models.

- However, it has encountered difficulty in maintaining high performance due to the high growth and exchange of news among social media users.

In [23], a hybrid architecture using Bi-GRU and Bi-LSTM deep learning techniques for fake news identification is proposed. TF-IDF and Count Vectorises approaches were used to perform the feature extraction process. These techniques used as well as Weakly supervised SVM techniques provided an accuracy of 90%. The main issues of this research are given below,

- From the results obtained by making a comparison CNN performed poorly compared to the other models presented in this paper. Bi-LSTM and BiGRU with weakly supervised SVM have high performance in the classification of fake news using large amounts of data.

Research Solutions

The suggested model integrates a number of cutting-edge methods for a wide range of languages. Initially, efficient data preparation is conducted using the CoreNLP toolbox to ensure cross-lingual compatibility.To enhance model performance, key features are extracted from the data using the RMS-BERT-CapsNet technique. A novel VAE-based Deep-Shallow multimodal fusion strategy is then introduced to improve fake news detection. Additionally, the application of Elastic Weight Consolidation (EWC) and Gradient Episodic Memory (GEM) facilitates flexible training and supports continual learning. Finally, the integration of Bidirectional Long Short-Term Memory (Bi-LSTM) and Vision Transformer (ViT) architectures results in robust and accurate hybrid categorization.

3. PROPOSED METHOD

The proposed method is designed to develop an effective and robust system for detecting fake news across different types of language by leveraging the deep learning method as shown in Fig. 1. The following crucial processes are carried out as a part of this strategy;

- Data collection
- Language-Agnostic Pre-processing
- Advance feature extraction
- Deeper Multimodal Fusion
- Adaptable training with Continual Gradient Episodic
- Hybrid classification architecture

A. Data collection

The proposed method utilizes the IFND and Albanian datasets for multilingual and multimodal fake news detection. The IFND dataset contains text and images, while the Albanian dataset contains only text. The data in this dataset relates to news published between 2013 and 2021. The content of the dataset is scraped using "Parse Hub". The fake news content of the dataset is enhanced by using an intelligence augmentation technique. It confirms that research on this time-consuming exploration problem can be driven by the availability of such a large dataset, which will improve prediction algorithms. The data is consistent with binary categorization. Each news item in the dataset is labelled as true or fake. Another unique feature of this dataset is that it is collected from several news data sources. Furthermore, this dataset is open and accessible to the entire research community. Albanian data is used in the model to extract and validate text content, which enhances the accuracy of fake news prediction.

B. Language-Agnostic Pre-processing

For data processing for normalization and cleaning with the proposed method, uses the CoreNLP toolkit. As a natural language processing tool, it enables the extraction of textual features for further analysis, including attributes of quotes and relationships, redaction, token and sentence boundaries, numerical and temporal values, named entities, sentiment, parts of speech, dependency and constituency analysis. Maintaining textual consistency by linking pronouns and referring terms, sentiment analysis can detect emotionally charged language, which is often a sign of fake news driven by emotions. These pre-processed features are then fed into deep learning algorithms such as RMS-BERT-CapsNet, VAE, EWC-GEM Continuous Learning, and Bi-LSTM-ViT for fake news classification, leading to a more accurate classification result. The data in the datasets used are structured in a uniform manner to enable us to perform accurate and efficient analyses.

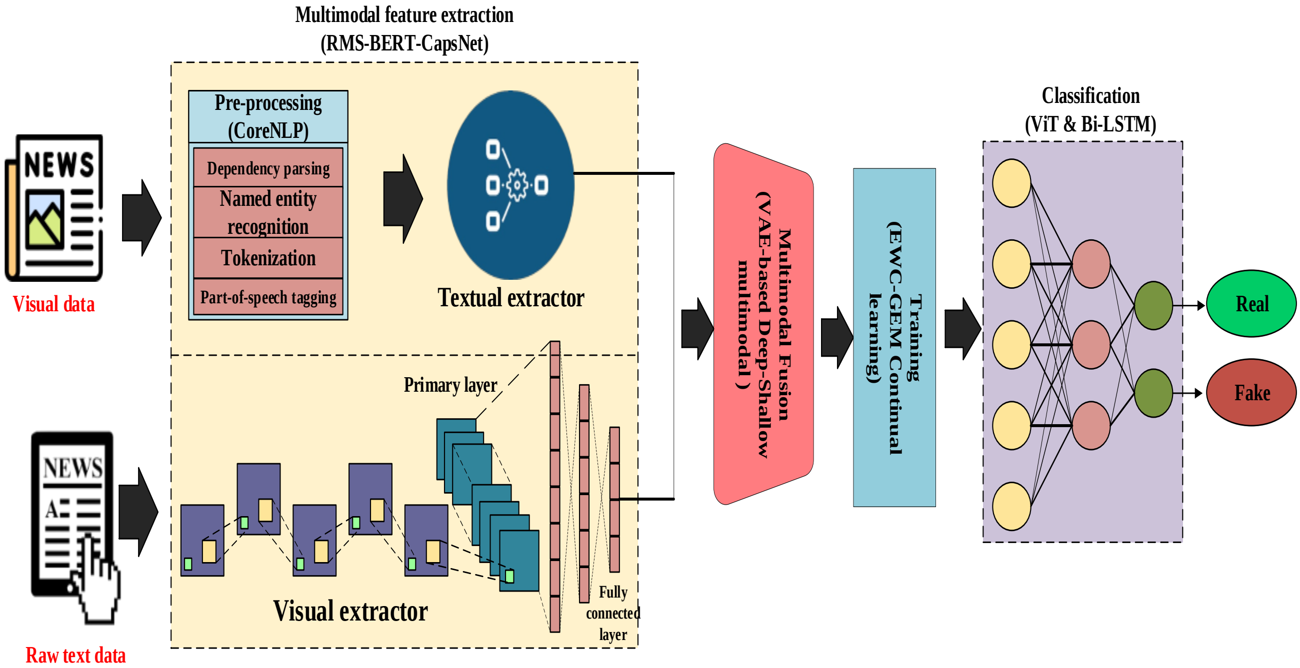

Fig 1: Overall architecture of Proposed method

C. Advance feature extraction

This research proposes a novel RMS-BERT-CapsNet (Root Mean Square Propagation–Bidirectional Encoder Representations from Transformers–Capsule Network) method for the feature extraction process. By effectively capturing the context and complexities of the language, the proposed method provides a comprehensive understanding.

a. Root mean Square propagation.

Root mean square (RMS) is used as an optimization technique to adjust model parameters during training, rather than directly determining "features" from an input such as a news article. During training, the optimizer (RMS in this case) adjusts the model parameters in such a way that the model learns to extract features from the input news text (in this case, BERT inputs, followed by Capsule Network feature extraction. Gradients are calculated during backpropagation to indicate how much each parameter of the model should change to minimize the loss (e.g., the error between predicted and actual labels). These gradients are smoothed over time, adjusting the model weights to improve its performance in extracting features from the input text.

The purpose of using RMSProp is to:

- Efficiently optimize the parameters (weights) of the neural network during training.

- Adjust the learning rate for each parameter based on the magnitude of the gradients.

- Improve the model's ability to extract meaningful features from text and images, leading to better classification.

By Using RMSProp, the code ensures that the model learns effectively from the data, adjusting the weights in order to improve performance and minimize classification error.

b. Bidirectional Encoder Representations from Transformers (BERT)

BERT (Bidirectional Encoder Representations from Transformers) is used to extract features from text data. Specifically, BERT helps convert raw input text into high-level feature representations that can be fed into a neural network for further processing and classification of fake news.

- To convert the input text into a format that BERT can process, used Bert Tokenizer. This involves tokenizing the text (breaking it down into individual words or sub words) and converting it into a sequence of input IDs.
- To structure the input data, BERT leverages superior arguments such as "CLS (classification token at the beginning) and SEP (separation token)". In BERT, the first token [CLS] is a special token that is used to represent the entire sentence or document. The feature representation for this token is used as a feature vector.
- BERT generates contextual embedding that capture the meaning of each word in relation to other words in the sentence, which is important for understanding the nuances of the text.
- After extracting features from BERT, the code uses these features as input to the Capsule Network (defined by the Capsule Network class). The capsule network is designed to capture hierarchical

relationships between features, further improving the model's ability to classify news as true or fake.

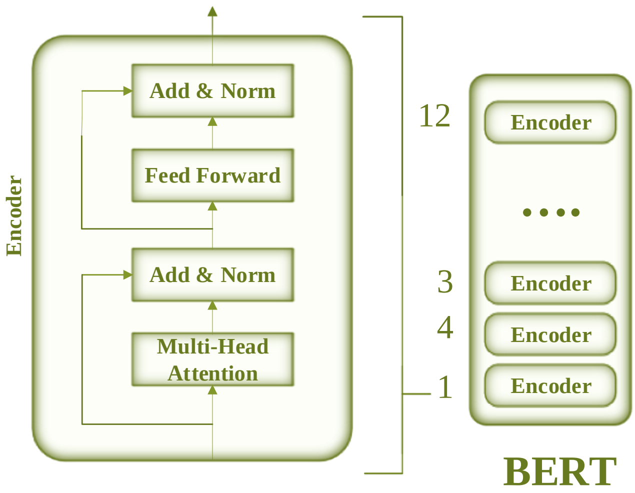

Fig. 2 Architecture of BERT

As shown in Fig.2 a sequential list of words from the text documents that have been integrated into vectors acts as the input for the textual feature extractor. Here we $\mathbb{K}$th dimensional word embedded vector with inserted text i is denoted as follows,

$$\Gamma = [\Gamma^0, \Gamma^1, \dots, \Gamma^n] \quad (1)$$

Here, n refers to the amount of words in the text Γ^0 denotes the inserting of [CLS], and a feature vector is denoted as Γ_v.

$$\Gamma_v = [\Gamma_v^0, \ \Gamma_v^1, \dots, \Gamma_v^n] \quad (2)$$

BERT is used to extract high-quality, context-aware features from text data, which are then used for news classification.

c. Capsule Network

Capsule Networks play a critical role in learning hierarchical relationships and spatial dependencies in input data. Capsule Networks (CapsNets) are a type of neural network architecture designed to address the limitations of traditional networks such as Convolutional Neural Networks (CNNs).Capsule Networks are used to process text features extracted from BERT and improve the task of fake news classification by:

- Capturing spatial relationships between features.
- Learning hierarchical structures in the data.
- Using a routing mechanism to focus on the most relevant feature combinations for classification.

A capsule network has several components, each playing a specific role:

1. Primary capsule layer

The primary capsule layer processes the output from the convolutional layer. Each capsule in the layer is a small Conv1D network followed by a nonlinear transformation.

2. Digit Capsule Layer

The "digit capsules" (or output capsules) represent the final representation of the features used for classification. Each capsule in this layer corresponds to a "Real" or "Fake" class.

3. Pumpkin Function

The pumpkin function ensures that the output of each capsule is a vector of size less than 1.

The input to the Capsule Network is the feature vector produced by BERT. It enhances the model's ability to learn complex hierarchical and spatial relationships within text data, improving the accuracy of the final classification for fake news identification. Caps-Nets capture the structural connections and associations between words, phrases, and sentences, providing a unique approach to extracting information for fake news identification. As a result, this improves the accuracy and reliability of the model in fake news identification.

D. Deeper Multimodal Fusion

In this research, is proposed a multimodal representation approach based on Variational Autoencoders (VAE). VAE consists of two main components: the encoder and the decoder. It is designed to compress image data into a latent representation and reconstruct the original image. The approach focuses on exploring clustering techniques to identify high-density areas in news items that are considered subjects.Then, using statistical metrics such as clustering integrity and separability to characterize the structural strengths, features are extracted from each cluster.

Examine a clustering with $\mathcal{K}$ clusters, that is $\mathcal{P} = \mathcal{Q}_1 \cup \mathcal{Q}_2, \cup \dots \cup \mathcal{Q}_{\mathcal{K}}$, where $\mathcal{Q}_j$ is a cluster of documents and $2 \le \mathcal{K} < d$. The coefficient is then used to determine if the news is centred around a single topic or a combination of concepts.

$$\mathbb{s}(\sigma_i, \mathcal{K}) = \frac{\beta(\sigma_i) - \alpha(\sigma_i)}{\max(\alpha(\sigma_i), \beta(\sigma_i))} \quad (3)$$

These values range from -1 to +1. A high number means that the news is mostly related to one issue and poorly aligned with other clusters. By combining the coefficient values during the examination of each article under different classification settings, it becomes possible to capture information about the concentration of fake news.Fake news often combines misleading or manipulated images with textual content. By analysing the images, the model can identify visual inconsistencies or features that may be associated with the fraud. These features are learned from the VAE Encoder

The purpose of this code is to leverage a multimodal approach to enhance the accuracy of detecting fake news by analysing both textual and visual content, combining their strengths, and addressing the limitations of single-modal methods.

E. Adaptable training with Continual Gradient Episodic

After completing the multimodal fusion process, the EWC-GEM Continuous Learning method (Continuous Learning with Gradient Episodic Memory and Elastic Weight Consolidation) is proposed for training the deep learning model. This study examines two widely used approaches, outlined as follows:

Gradient Episodic Memory (GEM) ensures that the model learns new fake news patterns without forgetting the old ones. It does this by modifying the gradients so that learning a new task does not degrade performance on previously learned tasks$t-1$.The main tasks of GEM in the proposed model are to store important samples from previous tasks (memory buffer), compute gradients for new and old tasks separately, compare new and old gradients to prevent conflicting updates, and ensure learning without catastrophic forgetting by allowing continuous learning.

Elastic Weight Consolidation (EWC) used to prevent catastrophic forgetting by ensuring that important weights

learned on previous tasks do not change drastically when learning a new task. This is achieved by calculating the Fisher Information Matrix (FIM) to measure the importance of each parameter and adding a tuning penalty to the loss function.
he two datasets are treated as separate tasks, denoted as$(\gamma_1 \; and \; \gamma_2)$. Following the model's normal training procedure for the first task, GEM and EWC are applied during the training of the second task.

$$\min_{\psi} \sum_{(\mathfrak{u}_i,\mathfrak{v}_i)\epsilon\gamma_2} loss(\mathcal{F}(\mathfrak{u}_i;\psi),\mathfrak{v}_i) = \sum_{(\mathfrak{u}_i,\mathfrak{v}_i)\epsilon\mathbb{M}} loss\big(\mathcal{F}(\mathfrak{u}_j;\psi),\mathfrak{v}_j\big) \le \sum_{(\mathfrak{u}_i,\mathfrak{v}_i)\epsilon\mathbb{M}} loss\big(\mathcal{F}(\mathfrak{u}_j;\psi_1),\mathfrak{v}_i\big) \quad (4)$$

The model is denoted as FFF, with ψ1\psi_1ψ1 and MMM representing the parameters associated with the task upon completion of the first task. Equation (4) defines the optimization problem under the GEM framework.

$$\sum_{(\mathfrak{u}_i,\mathfrak{v}_i)\epsilon\gamma_2} loss(\mathcal{F}(\mathfrak{u}_i;\psi),\mathfrak{v}_i) + \frac{\omega}{2}\mathrm{Z}\big(\psi - \psi^*_{\gamma_2}\big)^2 \quad (5)$$

Here, ω is the regularisation weight, Z denoted as fisher information matrix, and $\psi^*_{\gamma_2}$ is the variables of the "Gaussian distribution" which is employed through "EWC" to estimate the subsequent $\mathbb{P}(\psi|\gamma_2)$, then the equation (5) shows the EWC's loss function. In the proposed model, two continuous learning techniques "Gradient Episodic Memory (GEM) and Elastic Weight Consolidation (EWC)" are used to maintain optimal performance on both new and current datasets.

F. Hybrid classification architecture

LSTM (Long Short-Term Memory) is used as part of the BiLSTM (Bidirectional LSTM) layer in the proposed model to process sequential data, especially the output from Vision Transformer (ViT). The purpose of LSTM is to process the output sequences for images provided by Vision Transformers (ViT) to capture temporal or sequential dependencies in image features. LSTM is bidirectional, meaning it processes the input sequence in both forward and backward directions to capture dependencies in both directions (past-to-future and future-to-past). The final output of LSTM (after processing the sequence) is used for further processing. In this case, only the output from the last time step is used for news classification.

a. Encoder and Decoder

In the proposed model, encoders refer to the components of the model that process and transform the raw input data (image and text) into feature representations suitable for classification. Specifically, the two encoders in this model are: Vision Transformer (ViT) for image encoding and BERT for text encoding. Both of these encoders use the transformer architecture to learn the relationships in the image and text data, which is essential for the model's ability to understand both modalities and classify the input as true or fake.The resultant sequence is produced by the decoder, usually using the context information that the encoder has delivered. In the conventional sequence-to-sequence framework, the decoder is usually a unidirectional LSTM, meaning it only processes the outputsequence in the forward direction (from the first token to the last). The decoder step is executed with a softmax activation, which is equated to,

$$softmax(x_i) = \frac{\exp(x_i)}{\sum_j \exp(x_j)} \quad (6)$$

The decoder applies Softmax to convert the logit to class probabilities. CrossEntropyLoss is used as a loss function to measure how well the model's predictions match the actual labels. It is used to update the weights so the model improves over time.

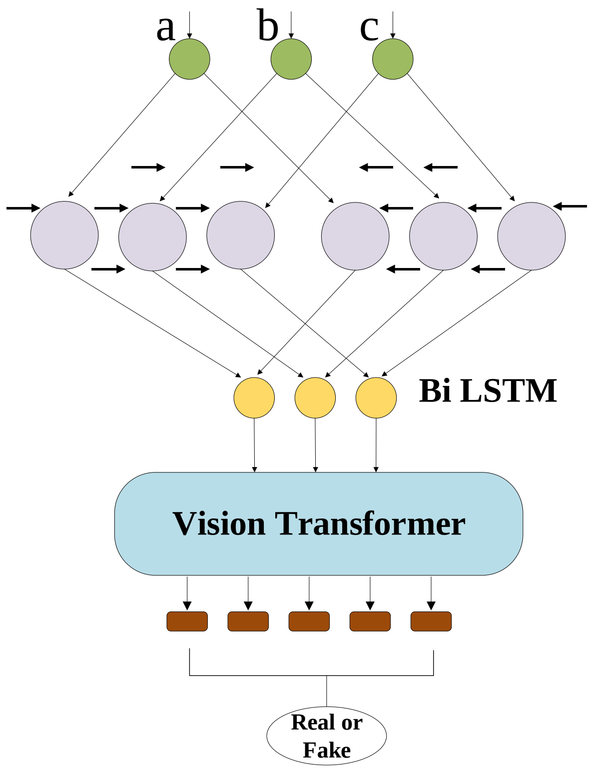

Fig. 3 Hybrid classification architecture of BiLSTM-ViT

Vision Transformer (ViT): It is designed to handle the image data using transformer-based architectures. Unlike the traditional "convolutional neural networks (CNNs)", ViTs divide the images into squares and treat them like sequences. By combining both vision transformers with Bi-directional LSTM (ViT- Bi LSTM) can achieve greater accuracy. Fig.3 represents the hybrid classification architecture. By evaluating input in both "forward and backward directions", the "Bidirectional Long Short-Term Memory (Bi-LSTM)" version of the regular "Long Short-Term Memory (LSTM)" network improves the model's capacity to extract contextual data from sequential data. This method combines the advantages of recurrent neural networks and transformers to provide an effective structure for recognizing fake news.

4. EXPERIMENTAL RESULTS

The experimental examination of the proposed Leveraging deep learning architecture for Identifying Fake news is shown in this section. This experimental result proves that the proposed method enhances the classification accuracy of fake news detection. This subsection contains the study report, and comparative analysis.

A. Comparative analysis

In this analysis, the proposed method is compared with existing methods EFND (Extreme Fake News Detection framework) [26], Stacked GRU (Gated Recurrent Unit) [29], and Local Interpretable Model-agnostic Explanations (LIME) [28] methods under various metrics to analyze the accuracy, loss, F1-score. In this section, the efficiency of the proposed method is analyzed by comparing it with existing methods under two scenarios, as outlined below.Scenario 1 is evaluated with English dataset (IFND) and Scenario 2 is evaluated with Albanian datasets.

a. Number of epochs vs. Accuracy (%)

The graph for the number of epochs and accuracy in fake news detection shows how model performance is evaluated during training. This illustrates the need for early prevention or regularization strategies to minimize overfitting and enhance the classification process of detecting fake news.

$$Accuracy(n) = A_{\infty} - \frac{\mathbb{W}}{1+e^{-r(n-n_0)}} \quad (7)$$

Here, A_{∞} is the maximum accuracy achievable, $\mathbb{W}$ refers to the difference between initial accuracy and maximum accuracy, r denotes the constant that controls the steepness of the curve, and n_0 is the epoch where accuracy growth starts to slow down.

Scenario: 1

In this scenario, the performance of the proposed model was evaluated against EFND and LIME across different epochs. The accuracy (%) outcomes are displayed in Fig. 4.

Table 1.Accuracy Comparison for Scenario 1 (English Dataset - IFND)

X-axis (Number of Epochs)	Y-axis Accuracy (%)		
	Proposed	**EFND**	**LIME**
2	65	60	52
4	72	67	59
6	83	72	68
8	90	78	72
10	95	81	78

Table 1 shows that the proposed model consistently outperformed both EFND and LIME at every epoch. The proposed model's accuracy at two epochs was 65%, which was 13% better than LIME and 5% better than EFND.The proposed model consistently outperformed both EFND and LIME across all epochs, achieving a peak accuracy of 95% at 10 epochs, significantly higher than the competitors. The increasing accuracy values demonstrate the model's strong learning capability and robustness.

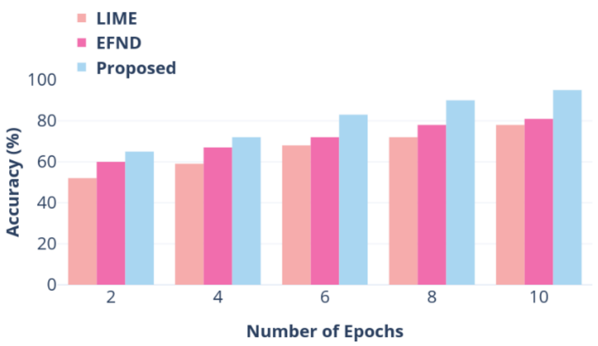

Fig. 4 Accuracy percentage for scenario 1

Scenario: 2

In the second scenario, the proposed model was evaluated against EFND and Stacked GRU.The accuracy results across a number of epochs are illustrated in Fig. 5.

Table 2 Accuracy Comparison for Scenario 2 (Albanian Dataset)

X-axis (Number of Epochs)	Y-axis Accuracy (%)		
	Proposed	**EFND**	**Stacked GRU**
2	70	67	52
4	73	70	60
6	80	71	65
8	85	73	69
10	90	75	70

The proposed approach outperforms EFND and Stacked GRU from the outset of training, as Table 2 demonstrates. The suggested model's 70% accuracy at two epochs was noticeably better than Stacked GRU's 52% and EFND's 67%. The proposed approach outperformed both EFND (75%) and Stacked GRU (70%) at 10 epochs with an accuracy of 90%. In this scenario, the proposed model again demonstrated superior performance, maintaining higher accuracy across all epochs. It consistently outperformed both EFND and Stacked GRU, showing the model's robustness even with a different dataset.

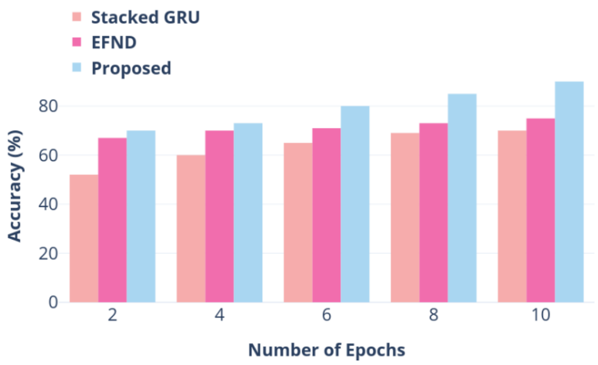

Fig. 5 Accuracy percentage for scenario 2

b. Number of epochs vs. Loss (%)

This Loss function calculates the variance between the actual distribution and the anticipated probability distribution, with lower values indicating better model performance. The equation for the loss function is given below,

$$Loss = -\frac{1}{n}\sum_{i=1}^{n}\sum_{c=1}^{c} g_{ic}\log(\hat{g}_{ic}) \quad (8)$$

Here, n is denoted as the number of samples, c refers to the number of classes, g_{ic} is the actual label and $\hat{g}_{ic}$ refers to the predicted probability of sample i being in class c.

Scenario: 1

In this scenario, the performance of the proposed model was evaluated against EFND and LIME across different epochs. The output outcomes are illustrated in Fig. 6.

Table 3 Loss Comparison for Scenario 1 (English Dataset - IFND)

X-axis (Number of Epochs)	Y-axis Loss (%)		
	Proposed	**EFND**	**LIME**
2	25	34	37
4	27	29	32

6	24	36	40
8	26	41	47
10	28	45	52

The proposed model continuously showed reduced loss values in comparison to both EFND and LIME. Starting with a loss of 25% at two epochs, it maintained lower loss values across all epochs. At 10 epochs, the loss value of the proposed model was 28%, which is significantly lower than EFND (45%) and LIME (52%). This consistent reduction in loss demonstrates the model's superior learning ability.

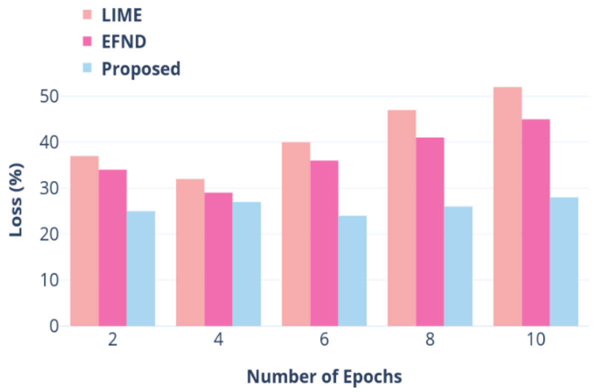

Fig. 6 Loss percentage for Scenario 1

Scenario: 2
In scenario 2, the loss values of the proposed model were compared against EFND and Stacked GRU. The loss percentage values are depicted in Fig. 7.

Table 4 Loss Comparison for Scenario 2 (Albanian Dataset)

X-axis (Number of Epochs)	**Y-axis Loss (%)**		
	Proposed	**EFND**	**Stacked GRU**
2	27	47	54
4	32	36	40
6	25	40	48
8	30	43	47
10	34	49	53

The proposed model maintained consistently lower loss values compared to EFND and Stacked GRU. At two epochs, it achieved a 27% loss, far below EFND (47%) and Stacked GRU (54%). The suggested model, developed a loss of 27% at 2 epochs, which is much less than EFND's 47% and Stacked GRU's 54%. In contrast to EFND's 49% and Stacked GRU's 53%, By 10 epochs, the loss for the proposed model was 34%, maintaining superior generalization performance across all epochs.

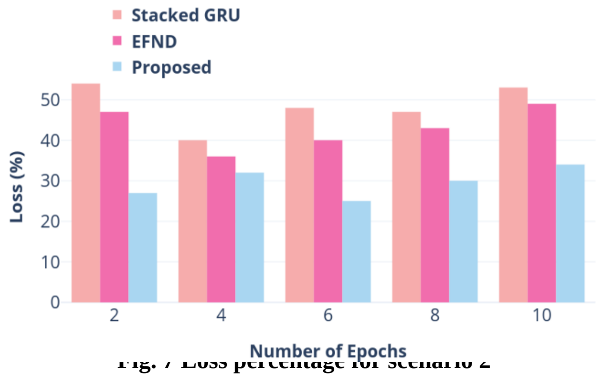

Fig. 7 Loss percentage for Scenario 2

This reduction in loss illustrates the superior generalization ability of the proposed model. The graph shows the capability of the proposed method to maintain a low error rate with increasing epochs highlighting its effectiveness in fake news detection.

c. Number of epochs vs. F1-score (%)

The F1 score is an important evaluation metric for classification problems like fake detection, especially in imbalanced datasets. Since it shows a better balance between precision and recall.The F1-score is calculated as,

$$F1 = 2 \times \frac{Precision \ \times Recall}{Precision \ + Recall} \quad (9)$$

Scenario: 1
In this scenario, the F1-score for the proposed model was compared with EFND, and LIME approaches which were measured across different epochs to ensure how effectively these models perform while predicting balanced memory and accuracy in fake news detection. Fig. 8 show the data and graph.

Table 5 F1-Score Comparison for Scenario 1 (English Dataset - IFND)

X-axis (Number of Epochs)	**Y-axis F1-score (%)**		
	Proposed	**EFND**	**LIME**
2	30	27	23
4	47	39	30
6	54	39	34
8	60	56	47
10	70	65	52

The proposed model consistently achieved higher F1-scores than EFND and LIME at all epochs. Starting at 30% at two epochs, it reached a peak of 70% at 10 epochs. This superior performance reflects the model's balanced ability to achieve high precision and recall.

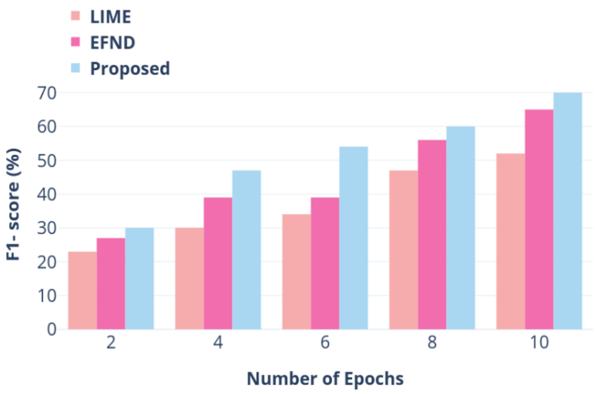

Fig. 8 F1-score percentage for scenario 1

Scenario: 2
In this scenario 2, the F1-score for the suggested model was equated with EFND and Stacked GRU models in different epochs. The outcomes are illustrated in Fig.9.

Table 6 F1-Score Comparison for Scenario 2 (Albanian Dataset)

X-axis (Number of Epochs)	Y-axis F1-score (%)		
	Proposed	EFND	Stacked GRU
2	33	30	26
4	40	36	30
6	53	47	42
8	66	53	47
10	72	68	55

The proposed model demonstrated higher F1-scores than both EFND and Stacked GRU across all epochs. Starting with an F1-score of 33% at two epochs, it peaked at 72% at 10 epochs. This result confirms the model's robust classification capability for fake news detection.

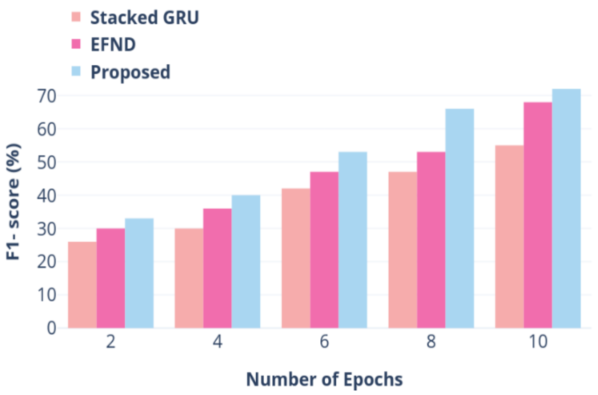

Fig. 9 F1-score percentage for scenario 2

The constant improvement in F1-scores showed that the suggested model could learn quickly, as evidenced by its capacity to achieve improved precision and recall even after fewer training epochs. This suggests that, in comparison to EFND and LIME, the suggested design performs better at identifying real positives and lowering fake positives.

B. Research summary

Initially, the IFND and Albanian Datasets are collected and loaded according to the respective scenarios. Data pre-processing is then implemented using the CoreNLP toolkit, which specifies the annotators for tokenization, sentence splitting, and part-of-speech tagging. Following this, a feature extraction process is carried out using the RMS-BERT-CapsNet method by implementing the CapsNet layer, consisting of the Conv1d, PrimaryCaps, and DigitCaps layers. This process extracts higher-level features from the BERT embeddings and the last hidden state through the CapsNet model, normalizing the extracted features. The textual and visual data are subsequently encoded and stored using the VAE-based Deep-Shallow multimodal fusion method. The training process is supported by the implementation of EWC-GEM Continuous Learning. Finally, a 'Vision Transformer with a Bidirectional Long Short-Term Memory' algorithm is applied to improve classification accuracy. This method integrates the strengths of Vision Transformers with the sequential modeling capabilities of BiLSTMs, enhancing classification performance in vision tasks.

5. CONCLUSION

In this paper, a wide-ranging novel approach for identifying fake news is proposed using a Multimodal-based RMS-BERT-CapsNet model, enabling comprehensive feature extraction and reliable fake news identification. An innovative VAE-based Deep-Shallow multimodal fusion technique is developed as part of this research to enhance the accuracy of fake news recognition. To offer flexibility in training, EWC-GEM Continual Learning is employed. The Vision Transformer and Bidirectional Long Short-Term Memory (Bi-LSTM-ViT) architectures are combined to achieve hybrid categorization. When compared to existing methods, this study demonstrates a significant improvement in evaluation parameters. In conclusion, this approach is presented as a reliable method for identifying fake news across multiple languages and various digital platforms. This paper proposed an advanced multimodal RMS-BERT-CapsNet architecture for fake news detection, integrating VAE-based deep-shallow multimodal fusion, EWC-GEM continual learning, and hybrid classification using BiLSTM-ViT. Experimental results indicate that the proposed model consistently outperformed traditional methods, including EFND, LIME, and Stacked GRU, across multiple evaluation metrics.The proposed approach's ability to effectively integrate textual and visual information, coupled with continual learning, contributed to its superior performance. Future work may explore the application of this model to other languages, the incorporation of user behavior analysis, and further optimization for real-time fake news detection on social media platforms.
This research contributes to the field of fake news detection and offers a scalable solution to combat disinformation in the era of rapidly spreading digital news.

6. REFERENCES

[1] Chauhan, T., & Palivela, H. (2021). Optimization and improvement of fake news detection using deep learning approaches for societal benefit. *International Journal of Information Management Data Insights*, *1*(2), 100051.

[2] Kaliyar, R. K., Goswami, A., & Narang, P. (2021). FakeBERT: Fake news detection in social media with a

BERT-based deep learning approach. *Multimedia tools and applications*, *80*(8), 11765-11788.

[3] Umer, M., Imtiaz, Z., Ullah, S., Mehmood, A., Choi, G. S., & On, B. W. (2020). Fake news stance detection using deep learning architecture (CNN-LSTM). *IEEE Access*, *8*, 156695-156706.

[4] Jain, V., Kaliyar, R. K., Goswami, A., Narang, P., & Sharma, Y. (2022). AENeT: an attention-enabled neural architecture for fake news detection using contextual features. *Neural Computing and Applications*, *34*(1), 771-782.

[5] Albahar, M. (2021). A hybrid model for fake news detection: Leveraging news content and user comments in fake news. *IET Information Security*, *15*(2), 169-177.

[6] Okunoye, O. B., & Ibor, A. E. (2022). Hybrid fake news detection technique with genetic search and deep learning. *Computers and Electrical Engineering*, *103*, 108344.

[7] Hanshal, O. A., Ucan, O. N., & Sanjalawe, Y. K. (2023). RETRACTED ARTICLE: Hybrid deep learning model for automatic fake news detection. *Applied Nanoscience*, *13*(4), 2957-2967.

[8] Mehta, D., Dwivedi, A., Patra, A., & Anand Kumar, M. (2021). A transformer-based architecture for fake news classification. *Social network analysis and mining*, *11*, 1-12.

[9] Kaliyar, R. K., Goswami, A., Narang, P., & Sinha, S. (2020). FNDNet–a deep convolutional neural network for fake news detection. *Cognitive Systems Research*, *61*, 32-44.

[10] Mallik, A., & Kumar, S. (2024). Word2Vec and LSTM-based deep learning technique for context-free fake news detection. *Multimedia Tools and Applications*, *83*(1), 919-940.

[11] E. Tata, J. Ajdari and N. Besimi, "Fake News Detection: A Comprehensive Survey," *2023 46th MIPRO ICT and Electronics Convention (MIPRO)*, Opatija, Croatia, 2023, pp. 309-314, doi: 10.23919/MIPRO57284.2023.10159859.

[12] Aslam, N., Ullah Khan, I., Alotaibi, F. S., Aldaej, L. A., & Aldubaikil, A. K. (2021). Fake detect: A deep learning ensemble model for fake news detection. *complexity*, *2021*(1), 5557784.

[13] Ali, A. M., Ghaleb, F. A., Mohammed, M. S., Alsolami, F. J., & Khan, A. I. (2023). Web-informed-augmented fake news detection model using stacked layers of convolutional neural network and deep autoencoder. *Mathematics*, *11*(9), 1992.

[14] Luvembe, A. M., Li, W., Li, S., Liu, F., & Wu, X. (2024). CAF-ODNN: Complementary attention fusion with optimized deep neural network for multimodal fake news detection. *Information Processing & Management*, *61*(3), 103653.

[15] Palani, B., Elango, S., & Viswanathan K, V. (2022). CB-Fake: A multimodal deep learning framework for automatic fake news detection using capsule neural network and BERT. *Multimedia Tools and Applications*, *81*(4), 5587-5620.

[16] Mishra, S., Shukla, P., & Agarwal, R. (2022). Analyzing machine learning enabled fake news detection techniques for diversified datasets. *Wireless Communications and Mobile Computing*, *2022*(1), 1575365.

[17] 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, 110125.

[18] Choudhry, A., Khatri, I., Jain, M., & Vishwakarma, D. K. (2022). An emotion-aware multitask approach to fake news and rumor detection using transfer learning. *IEEE Transactions on Computational Social Systems*, *11*(1), 588-599.

[19] Nadeem, M. I., Ahmed, K., Li, D., Zheng, Z., Alkahtani, H. K., Mostafa, S. M., ... & Abdel Hameed, H. (2022). EFND: A semantic, visual, and socially augmented deep framework for extreme fake news detection. *Sustainability*, *15*(1), 133.

[20] Kumari, R., & Ekbal, A. (2021). Amfb: Attention-based multimodal factorized bilinear pooling for multimodal fake news detection. *Expert Systems with Applications*, *184*, 115412.

[21] Zamil, Y. K., & Charkari, N. M. (2023). Combating Fake News on Social Media: A Fusion Approach for Improved Detection and Interpretability. *IEEE Access*.

[22] Amer, E., Kwak, K. S., & El-Sappagh, S. (2022). Context-based fake news detection model relying on deep learning models. *Electronics*, *11*(8), 1255.

[23] Syed, L., Alsaeedi, A., Alhuri, L. A., & Aljohani, H. R. (2023). Hybrid weakly supervised learning with deep learning technique for detection of fake news from cyber propaganda. *Array*, *19*, 100309.