

Received 27 April 2024, accepted 24 May 2024, date of publication 27 May 2024, date of current version 4 June 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3406258

 SURVEY

# A Review of Deep Learning Techniques for Multimodal Fake News and Harmful Languages Detection

ENIAFE FESTUS AYETIRAN<sup>1,2</sup> AND ÖZLEM ÖZGÖBEK<sup>1</sup>

<sup>1</sup>Department of Computer Science, Norwegian University of Science and Technology, 7491 Trondheim, Norway

<sup>2</sup>Department of Computer Science, Achievers University, Owo 341104, Nigeria

Corresponding author: Eniafe Festus Ayetiran (eniafe.ayetiran@ntnu.no)


The work of Eniafe Festus Ayetiran was supported by ERCIM “Alain Bensoussan” Fellow.

**ABSTRACT** The detection of fake news and harmful languages has become increasingly important in today’s digital age. As the prevalence of fake news and harmful languages continue to increase, so also is the correspondent negative impact on individuals and the society. Researchers are exploring new techniques to identify and combat these issues. Deep neural network (DNN) has found a wide range of applications in diverse problem domains including but not limited to fake news and harmful languages detection. Fake news and harmful languages are currently increasing online and the mode of dissemination of these contents is fast changing from the traditional unimodal to multiple data forms including texts, audios, images and videos. Multimedia contents containing fake news and harmful languages pose more complex challenges than unimodal contents. The choice and efficacy of the fusion methods of the multimedia contents is one of the most challenging. Our area of focus is multimodal techniques based on deep learning that combines diverse data forms to improve detection accuracy. In this review, we delve into the current state of research, the evolution of deep learning techniques that have been proposed for multimodal fake news and harmful languages detection and the state-of-the-art (SOTA) multimedia data fusion methods. In all cases, we discuss the prospects, relationships, breakthroughs and challenges.

**INDEX TERMS** Fake news, abusive language, deep learning, hate speech, harmful languages, deepfake, offensive language, toxic language, online trolling, cyberbullying, cyberaggression, extremism, multimedia data fusion.

## I. INTRODUCTION

In recent years, the proliferation of fake news and harmful languages across various online platforms especially social media has become a serious issue of concern. As technology advances, so do the techniques used to deceive and harm through these platforms. Detecting and mitigating the spread of such contents has become a priority for researchers and governments. Disinformation has been described as any deliberate fabrication of false information and presented as real in order to mislead readers [9]. Fake news is a particular form of disinformation that has been deliberately fabricated

The associate editor coordinating the review of this manuscript and approving it for publication was Angel F. Garcia-Fernandez .

with emotionally-charged contents, imitating mainstream news with the main goal of deceiving the reading audience. Often misconstrued as fake news is misinformation, which also contains incorrect information and also misleads readers but which is unintentional. Harmful languages on the other hand involves the use of languages that attacks the dignity of an individual, a group of persons or a community. Harmful languages are so-called because they involve target entities which are directly or indirectly harmed. Diverse works in literature have tried to categorize harmful languages based on how harsh and extreme the contents are. Some of these categorizations include but not limited to the following; hate speech, offensive language, cyberbullying, abusive language, radicalization, extremism among others.

For instance, Law Insider,<sup>1</sup> defines abusive language as “the use of remarks intended to be demeaning, humiliating, mocking, insulting, or belittling that may or may not be based on the actual or perceived race, color, religion, sex, national origin, sexual orientation, or gender identity of an individual”. It also defines offensive language as “any utterance which is blasphemous, obscene, indecent, insulting, hurtful, disgusting, morally repugnant, or which breaches commonly accepted standards of decent and proper speech”. These diverse categorizations are often used interchangeably and are sometimes difficult to differentiate. An example of this difficulty can be noticed in the attempt of [110] to differentiate between hate speech and offensive language wherein they explained that hate speech is directed at a particular individual or a group based certain attributes while offensive language is not directed at neither an individual nor a group. However, in reality it is known fact that both can be directed at an individual and/or a group. Both fake news and harmful languages are serious societal problems and propagators are now using multiple data forms to spread these contents. The data medium of spread of fake news and harmful languages has shifted from the traditional text and now includes data forms such as video, audio, images and memes.

Technology has advanced rapidly in recent years, giving rise to numerous possibilities and opportunities. One of the notable advancements is in the field of deep learning, a subfield of Machine Learning that deals with complex problems and large datasets using multiple neural network layers. Deep learning techniques have revolutionized various fields, including natural language processing (NLP) and computer vision. One area that has benefited greatly from this advancement is the detection of fake news and harmful languages in multimedia data. By combining textual and visual information, researchers have been able to develop powerful models capable of identifying deceptive and harmful languages in online platforms.

Deep learning techniques have gained significant attention in recent times due to their ability to handle complex data and extract patterns. They have become increasingly important in the field of natural language processing. These techniques have shown great promise in various applications, such as computer vision, language understanding, and sentiment analysis. In recent years, there has been a growing interest in leveraging deep learning for detecting multimodal fake news and harmful languages. This is due to the rising concerns about the spread of false information and harmful languages on social media. This field combines natural language processing, computer vision and audio analysis to analyze different types of data sources and uncover deceptive or harmful languages. It is important to conduct reviews that explore different approaches and models used in this domain in order to track progress and identify challenges. By examining various techniques on deep learning techniques

for multimodal fake news and harmful languages detection, researchers can identify the strengths and limitations of existing methods, gain further insights and propose better innovative solutions. In this review, we explore the different multimodal fusion approaches and deep learning models used for multimodal fake news and harmful languages detection.

The rest of this article is structured in the following order: Section II briefly discusses the interconnection between fake news and harmful languages, highlighting the common effect of both. In section III, we survey related reviews on multimodal fake news and harmful languages detection. Section IV presents an overview of data modalities. Section V presents an overview of multimedia data fusion techniques. A survey of deep learning techniques and experimental datasets for fake news and harmful languages detection are discussed in section VI. Section VII discusses the evaluation metrics used in fake news and harmful languages detection. Section VIII presents the challenges in multimodal content understanding in the context of fake news and harmful languages detection while Section IX identifies future research direction. Lastly, Section X concludes the paper.

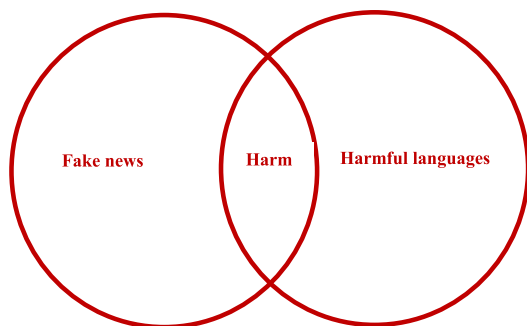
## II. FAKE NEWS AND HARMFUL LANGUAGES; THE COMMON EFFECT

[9] posit that both fake news and any form of harmful languages are societal menaces with a common effect of causing harm to the entity or group of entities they are being directed. Though fake news may not be directed at specific targets, they are however harmful because a disinformed entity (individual, group or community) constitutes danger to the entity and the larger society. Both fake news and hate speech for instance, can easily cause emotional torture to their victims. Reference [9] further opine that fake news and harmful languages are often intertwined in the sense that hatred mostly breeds fake news. In some cases, harmful languages are inserted in fake news. Furthermore, fake news about a person or race can also trigger reactions that can lead to the use of harmful language. In consequence, harm to a person or group of persons can lead to other effects such as protests, litigations among other actions. The resultant effect is illustrated with FIGURE 1. As a result of the identified common effect, techniques and evaluation used in both tasks are similar. It is therefore logical to review research works on both together.

## III. RELATED REVIEWS AND CONTRIBUTIONS

There have been quite a few surveys/reviews on deep learning for the detection of fake news [3], [41], [67] and harmful languages [44]. To the best of our knowledge, standard reviews on multimodal deep learning studies for fake news detection is currently limited to one [25] and none for harmful languages except a survey on general multimodal hate speech detection [19]. This is partly because multimodal studies for both tasks are still emerging rather than being extensive. Reference [25] surveyed deep learning techniques

<sup>1</sup>Lawinsider.com



**FIGURE 1.** Common effect of fake news and harmful languages.

used in multimodal fake news detection specifically focusing on state-of-the-art works, datasets and fusion techniques. On harmful languages, [19] presented a general review of multimodal hate speech detection. In contrast, we focus on deep learning methods for harmful languages encompassing hate speech, cyberbullying, offensive language etc. In a slight departure from the prior reviews, we expand the reviewed deep learning techniques based on current research, categorize them using the proper perspectives, identifies their strengths and weaknesses and the challenges. Furthermore, on multimedia data fusion strategies, we define two major categorization criteria. On the categorization based on model development stage, we expand the scope from the two identified by [25] to four and explores the details of each technique. On the categorization based on how features or weights are combined, we also discuss new techniques not discussed in prior works. Furthermore, we went further to explore a joint training technique in which each modality contributes to the overall decision of a deep neural model by jointly optimizing a shared objective function. In order to stimulate researchers to gain further insights on the tasks, we severally discuss the techniques used in the two tasks.

#### IV. OVERVIEW OF DATA MODALITIES

In this section, we explore the possible data forms used in multimedia content understanding including for fake news and harmful languages detection. These modalities include text, image, audio, memes and video.

##### A. TEXT

Text has been the traditional medium for the propagation of fake news and harmful languages. This situation, we believe is due to the fact that text is the easiest and most readily available tool for propagating these contents. It involves combination of alphabets, numbers and special characters to convey statements and/or questions which can be understood by both humans and machines. They are represented by text document formats such as txt, doc, pdf etc. Some of the deep learning methods which rely only on text for fake news and harmful languages include [58], [71], [89], [92] and [50], [74], [85], [111] respectively.

##### B. IMAGE

Image is also one of the modalities used for fake news and harmful languages propagation. Their use is not however as prevalent as texts. An image is a still photographic representation of an entity in time. In most cases, images often complement texts. In news articles for instance, the purpose is to capture a scene illustrating the subject/content of the news. In multimodal fake news detection, image can be an important component for detecting fakeness by determining if the image has been distorted or if the image illustration does not agree with the news subject or content. The latter is especially applicable in deepfake detection. One of the very few works on fake news detection using deep learning is the work of [64] and [102] who developed a framework for fake news detection through image analysis using a transformer.

##### C. MEME

A meme is a visual representation of either an image, a short animated image or video meant to convey humour and often used on social media. Some of the images in datasets used for fake news and harmful languages are actually memes with some having texts inserted within them [52], [76], [93], [99].

##### D. AUDIO

An audio is a sound recording of events mainly voice used for auditory purposes. Audio has been utilized in tasks covering both fake news and harmful languages to change a real narrative with falsehood with the intention of propagating hate, propaganda, extremism among others. Most of the applications of audio are on spoof and deepfake [4]. Audio-based learning techniques for fake news detection include [11], [98], [101], [104]

##### E. VIDEO

A video is a continuous or moving visual representation of a scene or sequence of events. Usage of videos for the detection of fake news and harmful languages are also not as common as text. This perhaps due to the fact that videos considerably translates to reality and the ease with which tampered videos can be detected. Videos are sometimes accompanied by corresponding audio components at a point in time. Some of the rare works based on deep learning include ones for detecting offensive video [2] and inappropriate contents (extremism, hate speech etc.) [5] in videos.

#### V. OVERVIEW OF DEEP LEARNING-BASED MULTIMEDIA DATA FUSION TECHNIQUES

Multimedia data fusion is the process of combining individual data forms as a single contiguous structured data. The techniques for combining multimedia data cut across the tasks of fake news and harmful languages. We therefore discuss them in these contexts and with reference to relevant works on deep learning for both tasks. A typical multimodal fusion merges data either at the level of raw data, features

or layer weights as illustrated in FIGURE 2, where  $m$  is the number of data media involved.

We categorize fusion techniques of multimedia data according to the following criteria:

- Stage in the model development where the fusion takes place
- How features or weights are combined in the model

#### A. DATA FUSION BASED ON STAGE OF OCCURRENCE IN THE MODEL DEVELOPMENT

We extended the two types identified by [25] to three with additional details. The following are the types:

- **Early fusion:** Early fusion otherwise known as data-level or feature-level fusion occurs at the very early stage of model development before any network layer. It occurs at the raw or preprocessed data stage or just when features have been extracted from the raw data. An illustration of early fusion is presented in FIGURE 3. Dotted arrows denote possibility of fusion of the source data forms. A major challenge in fusing raw or preprocessed data is how to combine the heterogeneous data forms in a way that can be processed by the deep neural network. Fusion is easier for features. One strategy to overcome the challenge is to employ tools which combines computer vision and machine learning to unify the data [9].
- **Intermediate Fusion:** Intermediate fusion merges layer weights at any point within the network layers and produces a single fused weight that is passed to an activation function for decision making. It should be noted that intermediate fusion can be progressive, in which multiple fusions occurs within a multilayer network. FIGURE 4 shows an illustration of intermediate fusion. Dotted arrows indicate fusion can take place from any of the layers.
- **Late Fusion:** Late fusion is also known as decision-level fusion. In late fusion, a full model for each modality is first developed. The outputs (decisions or probabilities) of the individual models are then merged. FIGURE 5 illustrates late fusion. For a multimedia content with  $m$  number of modalities, a deep neural model with  $n$  number of layers combines the outputs of  $m$  number of models corresponding to each modality to a single output.
- **Hybrid Fusion:** Hybrid fusion combines the characteristics of two or more of early, intermediate and late fusion.

#### B. DATA FUSION BASED ON HOW FEATURES OR WEIGHTS ARE COMBINED IN THE MODEL

These categories of fusion perform some arithmetic or merging operations on features or weights. Each of these techniques are applicable at either some or all stages of a deep neural model. These techniques require that the dimensions of the individual features or weights be uniform otherwise dimension normalization needs to be done to make them

uniform. The following are the fusion techniques based how data features or layer weights are merged:

- **Concatenation:** Concatenation is by far still the most popular choice of multimodal fusion. Let  $x_i$  denote a data modality and  $n$  the total number of modalities to be fused, the concatenation operation on the set of modalities  $x_1, \dots, x_n$  is given by  $y$ , the fused data as presented in equation (1):

$$y = x_1 \oplus x_2 \oplus \dots \oplus x_n \quad (1)$$

Concatenation operation can be applied at early and intermediate stages of model development. For each modality with dimension  $m$ , the concatenation operation results in  $m \times n$  features or weights.

- **Summation:** Summation operation [8] adds features or weights of each modalities under consideration. Let  $x_i$  denote a data modality and  $n$  the total number of modalities to be fused, the addition operation on the set of modalities  $x_1, \dots, x_n$  is given by  $y$ , the fused data as depicted in equation (2):

$$y = x_1 + x_2 + \dots + x_n \quad (2)$$

Addition operation on a set of features or weights return the same dimension as those of the uniform individual modalities. Summation operation can be applied at all the stages of the model development.

- **Multiplication:** The multiplication operation when performed on a set of features or weights gives the product of the elements. Let  $x_i$  denote a data modality and  $n$  the total number of modalities to be fused, the multiplication operation on the set of modalities  $x_1, \dots, x_n$  is given by  $y$ , the fused data as presented in equation (3):

$$y = x_1 \odot x_2 \odot \dots \odot x_n \quad (3)$$

The dimension of the resulting features or weights remain same after multiplication operation. The multiplication operation can also be applied at all stages of a deep neural model development. However, the summation operation can be used as a form of voting technique in late fusion, in which case the decisions (probabilities) produced by individual models developed from each of the modalities are summed to obtain a final decision.

- **Average:** The average operation [8] computes the mean of corresponding features or weights belonging to the modalities under consideration. Let  $x_i$  denote a data modality and  $n$  the total number of modalities to be fused, the average operation on the set of modalities  $x_1, \dots, x_n$  is given by  $y$ , the fused data as shown in equation (4):

$$y = \frac{1}{n} \sum_{i=1}^n x_i \quad (4)$$

Average operation is also applicable at all stages of a deep neural model development. The dimension of

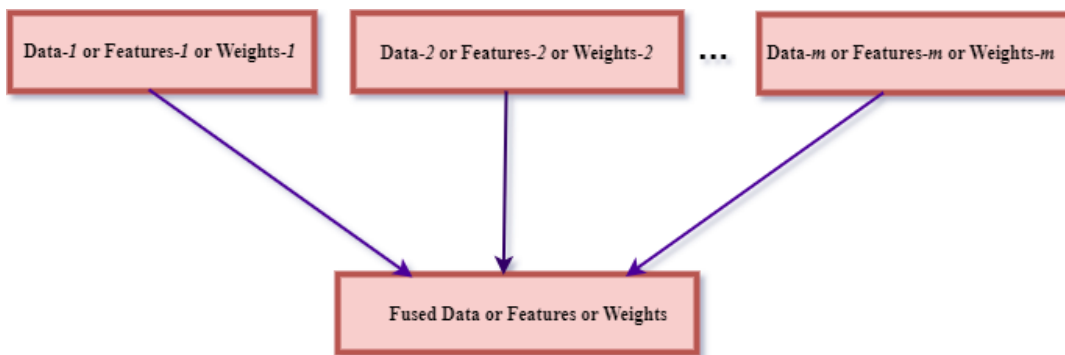


FIGURE 2. Multimedia data fusion.

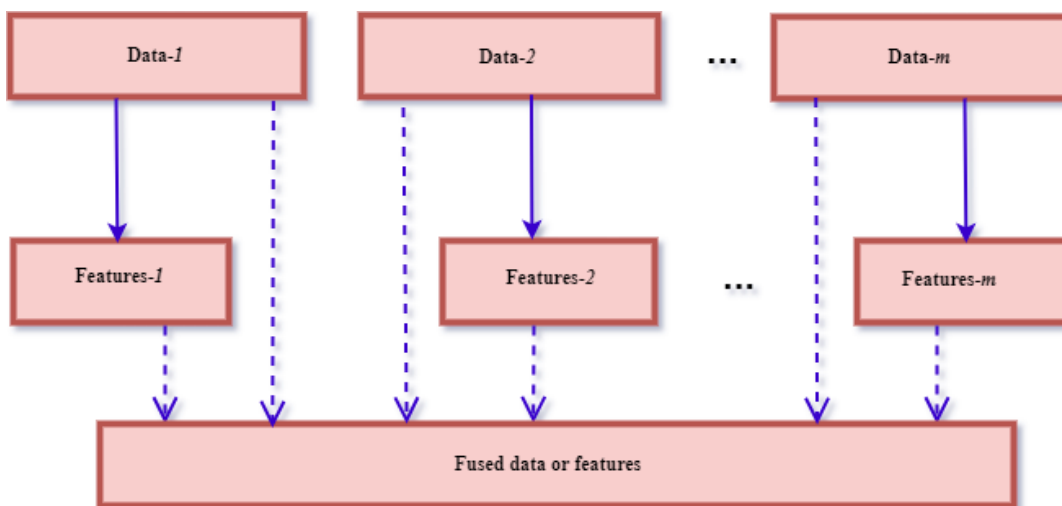


FIGURE 3. Early fusion.

the resulting features or weights remain same as the original dimensions of individual source data. However, the average operation can be used as a form of voting technique in late fusion, in which case the average of decisions (probabilities) produced by individual models from each of the modalities is computed, on which basis the final decision is decided.

- **Maximum:** The maximum operation [8] compares and obtains the maximum of the feature vectors or weights of two or more modalities. Let  $x_i$  denote a data modality and  $n$  the total number of modalities to be fused, the maximum operation on the set of modalities  $x_1, \dots, x_n$  is given by  $y$ , the fused data as shown in equation (5):

$$y = \max(x_1, x_2, \dots, x_n) \quad (5)$$

Similar to the average and summation operations, the maximum operation can also be used as a form of majority voting technique in late fusion in which case a model from among the individual models with the highest probability is selected as the final decision.

- **Mode:** The mode is the most frequent class predictions determined based on the output probabilities from models of individual modalities. It is used for majority voting where the final class prediction is determined by a simple majority based on the mode of the multimodal class predictions.

### C. JOINT MULTIMODAL TRAINING OF MODELS

Some works have also explored the possibility of joint training using multiple models, one for each modality under consideration. This approach is a form of late fusion. However, the output from individual layers are not directly merged but features are trained jointly with individual or shared objective function which the models seek to optimize. In other words, each model contributes directly to the final decision. In some cases, it is possible to assign weights to each model. This weight determines the proportion of contribution to the overall decision. Reference [7] applied joint training of aspect-level and document-level models for aspect-based sentiment classification. In the study, the author used a document level weight  $\lambda$  to assign the



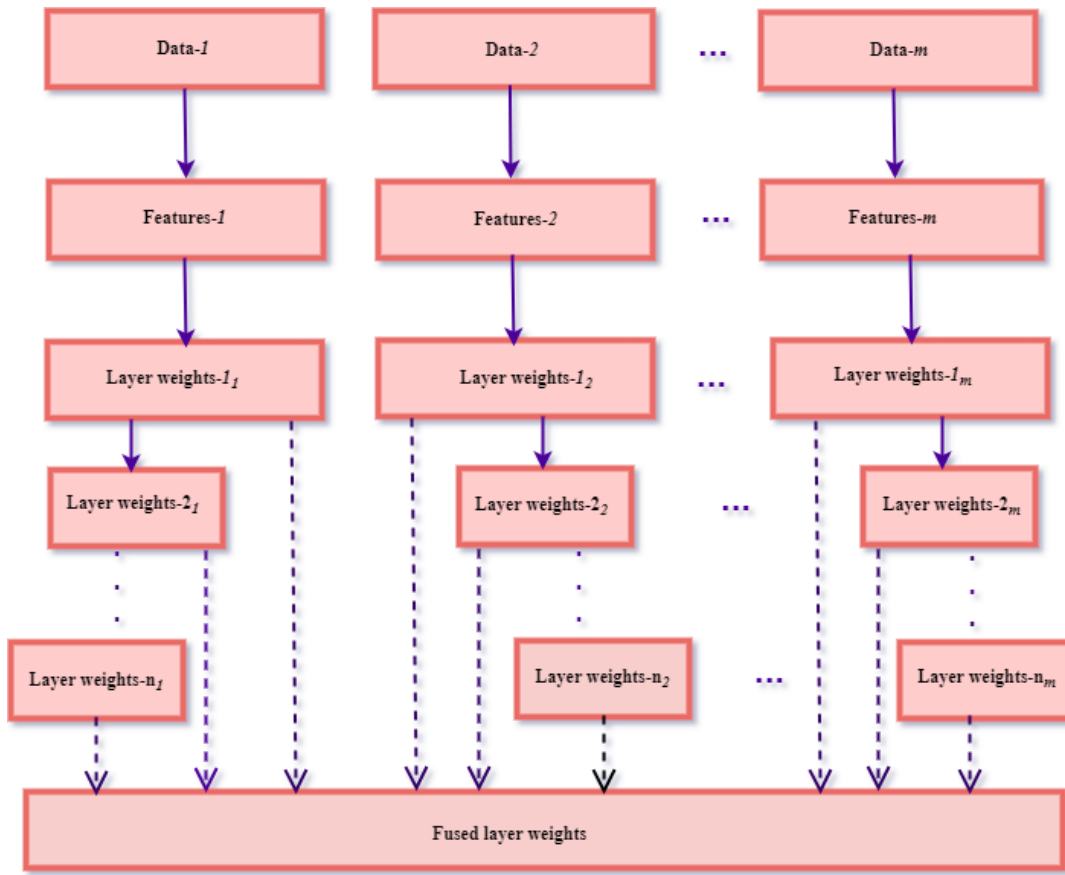


FIGURE 4. Intermediate fusion.

contributing weight of the document-level model to the overall model. Results show that this technique is effective as it achieved state-of-the-art performance. In another work, [107] optimized a joint objective for parameter learning. They particularly designate hate speech and sarcasm as primary and auxiliary tasks and treated as separate models for the purpose of cross-task transfer learning.

## VI. SURVEY OF DEEP LEARNING TECHNIQUES AND DATASETS FOR MULTIMODAL FAKE NEWS AND HARMFUL LANGUAGES DETECTION

In this section, we discuss state-of-the-art multimodal deep learning techniques for fake news and harmful languages detection. We do not intend to delve into the core details of the algorithms themselves to prevent proliferation of literature which is already sufficient on each of the algorithms. We discuss these works severally for fake news and harmful languages detection according to the combination of modalities used. These algorithms comprises several categories of neural networks ranging from Convolutional Neural Network (CNN), Recurrent Neural Network (RNN), Graph Neural Network (GNN), attention-based network, Generative Adversarial Network (GAN), transformers and hybrid methods. Some of the prominent algorithms used for

these tasks include but not limited to ones based on Convolutional Neural Network [31], traditional Recurrent Neural Network [84], Long Short-Term Memory (LSTM) [39], Bidirectional LSTM (BiLSTM) [36], Gated Recurrent Unit (GRU) [21], Bidirectional Gated Recurrent Unit (BiGRU) and attention-based approaches [10], [62]. We further present overviews of datasets used in both tasks.

### A. DEEP LEARNING TECHNIQUES FOR MULTIMODAL FAKE NEWS DETECTION

In this section, we discuss relevant works on deep learning for multimodal fake news detection. TABLE 1 shows the surveyed models for fake news detection, the underlying deep learning architecture, the modalities involved, the fusion technique(s) and the experimental datasets used.

#### 1) TEXT AND IMAGE

The use of image and text for fake news detection using deep learning techniques is by far the most commonly used combination of modalities. The work of [103] is identified to be one of the earliest attempts on multimodal fake news detection. In their work, they developed an end-to-end architecture based on a neural model called EANN. The architecture comprises three primary components namely

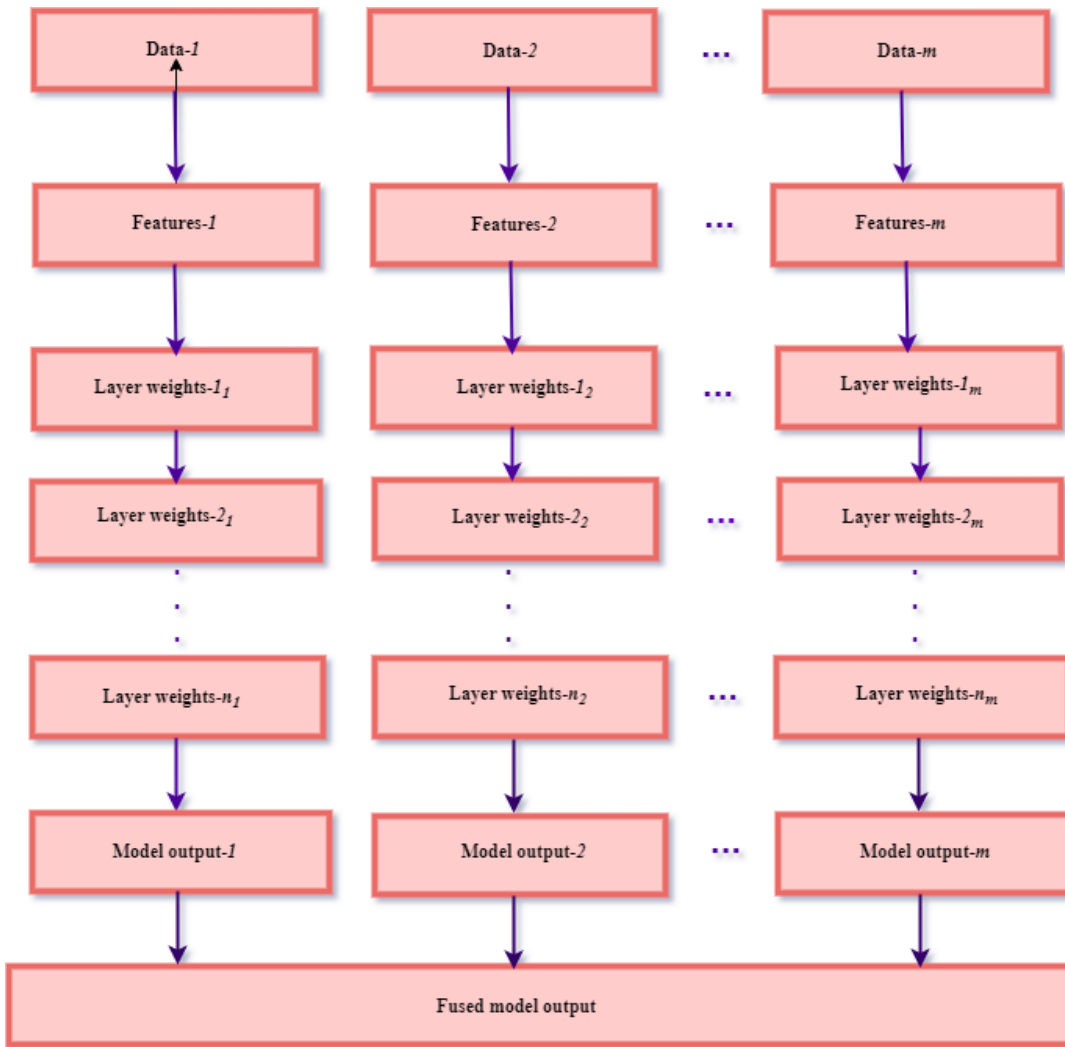


FIGURE 5. Late fusion.

a multimodal feature extractor, a fake news detector and an event discriminator. The multimodal feature extractor component comprises two sub-components; a text features extractor and a visual features extractor. Each of the components interrelate for multimodal fake news detection. They evaluated this framework on two standard datasets which depict improvements in performance over the baselines. Similar to EANN is MVAE [51]. MVAE used variational autoencoder for classifying multimodal news contents as real or fake. MVAE consists of an encoder, a decoder and a fake news detector. Both the encoder and the decoder each comprises a text and visual extractor. While the encoder basically encodes the multimodal inputs and outputs a shared representation of learnt features as latent vectors, the decoder reconstructs the latent vectors. The decoder receives the encoded representations as input, which consequently passes its own output to the fake news detector as input. The fake news detector classifies news contents based on the encoded

representations together with the sum of reconstructed and Kullback-Leibler divergence losses. In terms of performance, MVAE outperforms EANN and other baselines with significant improvements. SpotFake, proposed by [96] employs pretrained transformers to incorporate contextualized textual information and image recognition into multimodal fake news detection. Precisely, they employ Bidirectional Encoder Representations from Transformers (BERT) [27] and VGG-19 [94] to extract textual and image features which were combined for the classification. The goal is to achieve a pure classifier without any other underlying task. Experimental results produced by SpotFake on the same evaluation datasets as EANN and MVAE show marginal gain on only one of the datasets. In order to validate a created dataset named News-Bag [47] experimented with some reproduced deep neural techniques and models. MVAE is one of the top performing according to the presented results. Reference [114] opines that identification of irrelevant images in news contents can

**TABLE 1. Models based on deep learning for multimodal fake news detection. keys: T&I - text and image, T&V - text and video, A&V - audio and video, T&A&V - text, audio and video.**

Work/model	Algorithm/DNN Architecture	Modalities	Fusion technique(s)	Dataset(s)
EANN [103]	Adversarial Neural Network	T&I	Late	Twitter [13], [14] and Weibo [46]
MVAE [51]	Variational autoencoder	T&I	Late/Summation	Twitter [13], [14] and Weibo [46]
SpotFake [51]	Fully-connected neural network	T&I	Intermediate/Concatenation	Twitter [13], [14] and Weibo [46]
[47]	Variational autoencoder	T&I	Late/Summation	NewsBag [47]
SAFE [114]	TextCNN/CNN	T&I	Late/Summation	PolitiFact [93] and GossipCop [93]
[33]	Word embeddings and visual transformer	T&I	Early	PolitiFact [93] and GossipCop [93]
[34]	Textual and visual transformer	T&I	Intermediate/Concatenation	GossipCop [93]
[65]	Hierarchical attention network	T&I	Late/Summation	Fake news datasets <sup>2</sup> <sup>3</sup> <sup>4</sup>
[95]	Sentence transformer and EfficientNets	T&I	Intermediate/Summation	Twitter [13], [14] and Weibo [46]
MCNN [106]	Attention-based BiGRU	T&I	Late/Summation	MC [106], Twt [13], [14], PFact [93], TI [109]
MCAN [105]	Co-attention networks	T&I	Intermediate (progressive)/Concatenation	Twitter [13], [14] and Weibo [46]
EM-FEND [78]	Co-attention transformer	T&I	Intermediate/Concatenation	TI-CNN [109] and Weibo [46]
CAFE [18]	Attention ResNet and multichannel CNN	T&I	Intermediate/Multiplication	Twitter [13], [14] and Weibo [46]
TTEC [42]	Contrastive learning	T&I	Early/Concatenation	ReCOVery [113]
MMFN [116]	Textual and visual Transformers	T&I	Early/Concatenation	GCop [93], Twitter [13] and Weibo [46]
Inter-modal [9]	Attention-based BiLSTM-CNN	T&I	Intermediate/Concatenation	PolitiFact [93]
UCNet [72]	LSTM	T&V	Intermediate/Concatenation	VAVD [72] and FVC [73]
FANVM [22]	Adversarial networks	T&V	Early/Concatenation	MYVC [22], VAVD [72] and FVC [73]
FVDM [23]	TextCNN/Attention-BiLSTM	T&V	Intermediate/Concatenation	MYVC [22], VAVD [72] and FVC [73]
[66]	CNN/Memory Fusion Network	A&V	Late	DFDC [28] and TIMIT [87]
[24]	Dissonance score	A&V	Late	DFDC [28] and TIMIT [87]
[116]	CNN/RNN	A&V	Late	Subset of DFDC [28] and FF++ [83]
[48]	Pretrained CNN-based models	A&V	Intermediate/Concatenation	FakeAVCeleb [49]
AVFakeNet [43]	Dense swin transformer network	A&V	Late/Maximum	FakeAVCeleb [49]
AVForensics [117]	CNN encoder/Contrastive learning	A&V	Intermediate (Progressive)	FF++ [83], DFDC [28], DF [45] and FS [57]
[86]	EfficientNet/Time-Delay Neural Network	A&V	Hybrid	FkAVCD [49], DFDC [28] and TIMIT [87]
Multimodaltrace [82]	MLP-Mixer layers	A&V	Intermediate/Summation	FkAVCD [49], PDD [88] and WLDD [1]
TikTec [90]	Co-attention fusion	T&A&V	Intermediate (Progressive)	COVID-19 Video Dataset [90]
SV-FEND [77]	Transformers	T&A&V	Intermediate (Progressive)/Concatenation	FakeSV dataset [77]
NEED [79]	Transformers	T&A&V	Intermediate/Concatenation	FakeSV dataset [77]

help can detect fake news in multimedia contents. Based on this notion, they developed a model named SAFE. SAFE is based on similarity between text and corresponding images in multimedia news contents. SAFE extended a CNN-based method for textual and visual features extraction. Images are first converted to text using image captioning. The main crux of SAFE is the computation of similarity between text and image features which serves as main optimization values for learning other model parameters. In another work, [33] leverages textual, visual and semantic information for fake news classification using a neural classifier. Textual and semantic features include embeddings of posts and

sentiments respectively while visual features comprise image tags and local binary patterns (LBP). Experiments were carried using three datasets. In a later work, [34] extended their earlier work using multiple image information as visual features. In contrast to the earlier use of word embeddings and image tags/local binary patterns, they instead use BERT [27] and VGG-16 [94] for text and image representations respectively. The crux in both works is the computation of semantic similarity between textual and visual features. Reference [65] proposed an approach based on Hierarchical Attention Network (HAN), image captioning and forensic analysis to tackle the task of fake news detection with



specific focus on fake images in multimedia news contents. In addition they used headlines matching news contents with other algorithms such as Noise Variance Inconsistency (NVI) and Error Level Analysis (ELA) specifically to detect fake images. The end product is an ensemble method which combines these algorithms having being tested severally. Overall result shows that the ensemble method outperforms individual algorithms, also against the considered baselines. Still on fake images in multimedia news contents, [95] proposes an approach to detect fake news using a variant of CNN (EfficientNetB0 [100]) and sentence transformer (RoBERTa [59]) for visual and text representations respectively. While their approach achieves a superior performance over baselines on English dataset, it produces inferior performance on the Chinese dataset. Multimodal Consistency Neural Network (MCNN) [106] is a related network-based work. MCNN consists of five subnetworks namely; a text feature extraction component, a visual semantic feature extraction component, a visual tampering feature extraction component, a similarity measurement component and a multimodal fusion component. Experiments with MCNN on four benchmark datasets depict improvements over compared baselines. In another work, Multimodal Fusion with Co-attention Networks (MCAN) [105] was proposed. MCAN includes a co-attention block, a co-attention layer and multiple co-attention stacking on spatial-domain, frequency-domain and textual features. MCAN outperforms compared baselines on two domain datasets. Reference [78] posits that modeling images' semantics as a supplement to text constrains their performance. As a result of this notion, they propose a architecture called "EM-FEND" based on three text-image correlation clues that can improve multimodal fake news detection if exploited. These clues include entity inconsistency, mutual enhancement and text complement. "EM-FEND" fuses multimodal features based on these clues using multimodal co-attention mechanism. They experimented on two datasets covering two languages and reported state-of-the-art results. Cross-modal Ambiguity Learning (CAFE) model was proposed by [18]. CAFE comprises three modules namely; a cross-modal alignment module, a cross-modal ambiguity learning module and a cross-modal fusion module. The main goal of CAFE is adaptive aggregation of unimodal features and cross-modal correlations. CAFE was evaluated on two datasets and shows improvements over compared baselines. A model (TTEC) based on contrastive learning, back-translation and multi-head attention was introduced [42]. TTEC utilized BERT for back-translation of the text modality while contrastive learning was utilized for image modeling. The entire methodology involves text data augmentation and back-translation, multimodal information encoding (text and image), data fusion, followed by joint learning of fused feature representation with multi-head attention and contrastive learning. The multi-head attention serves as the main learning model while the contrastive learning serves as an auxiliary model to enhance the effectiveness of training. A dataset based on COVID-19 was

used for experiment. Experimental results improve compared baselines. The compared baselines included a reproduction based on LSTM, CNN and SAFE, a prior work evaluated on a different dataset. In a distinct approach from prior works, [116] distinguishes between fine-grained and coarse-grained multimodal information. They proposed a model in which they combined these two distinct multimodal information and referred to it as multi-grained multi-modal fusion network (MMFN). For each of the modalities, both the fine-grained and coarse-grained features are first encoded and then fused. The textual features are encoded by BERT while the visual features are encoded by Swin Transformer [60]. The final fusion is derived by merging the encoded multi-grained features for both modalities achieved through the use ViLBERT [61]. The final classifier is trained as a concatenation of unimodal text representation, unimodal image representation and fused multimodal representation. MMFN predominantly outperforms considered baselines across three datasets.

## 2) TEXT AND VIDEO

Metadata, transcripts and other forms of text together with videos have been explored for multimodal fake news detection. Reference [72] used LSTM on a number of features extracted from videos which they called "comment embedding". A weight between 0 and 1 is obtained for each comment through sigmoid activation. The weight is then multiplied by a 300-dimension "comment embedding" to obtain a "unified comments embedding". The third phase of their approach is a concatenation of "unified comments embedding" with extracted simple features. The output of the concatenation is passed through layers of network which they named Unified Comments Net (UCNet). They experimented with a dataset (VAVD) they created for this purpose with an existing dataset (FVC) [73]. Reference [22] proposed a model based on topic modeling and adversarial neural networks. In the proposed model, comments and titles/descriptions of videos were encoded, distribution of topics between the comments and the titles/descriptions were then computed using Latent Dirichlet Allocation (LDA) [12] model with Gibbs sampling. The goal of the topic distribution is to compute the differences in stance. An adversarial network which comprises two modules was then applied. These modules include a fake news video encoder and a topic discriminator. The model reports improvements over existing systems spanning four datasets. In another study, [23] combines domain knowledge with fusion of text and video features. It basically uses text to validate the genuineness of video contents. Domain knowledge is built from the training set using Pearson correlation coefficient to select features that differentiate between fake and real videos. Furthermore, these features were ranked according to how probable they are likely to be fake. Comments based on likes and domain knowledge were encoded using BERT while titles/descriptions were encoded using CNN. The

videos on the other hand were encoded with VGG-19. The final model is a linear combination of the various encoded representations. The results were compared with three baselines with improvements over them. Furthermore, they presented variants of the model using different embeddings. Also, series of ablation studies were conducted to show the effect of the various embeddings.

### 3) AUDIO AND VIDEO

Videos are often accompanied by corresponding audio. Modeling these two modalities for fake news detection have also been explored. Most of the application areas are in deepfake detection. Reference [66] proposed a deep learning model powered by siamese network and triplet loss function. The model also incorporated affective computing into the training and classification pipeline. Quite a number of features were extracted and used for the training. These include the use of various layers of CNN for audio/video features, Memory Fusion Network (MFN) among others. Reference [24] hypothesizes that distortion of at least one of two modalities will cause disagreement between them. They posit that this hypothesis can be exploited to detect fakeness in an audio-visual content. They therefore computed the Modality Dissonance Score (MDS) between the modalities. MDS simply computes dissimilarity score between audio and visual segments in a audio-enabled video. The audio features were extracted by Mel-Frequency Cepstral Coefficients (MFCC) while the visual features were extracted by 3-dimensional ResNet [37]. It achieved the best performance on one of the three benchmarks. In this method, latent features were first extracted from both audio and video through a variant of CNN and then fed into a recurrent layer. The classifier compares the effect of cross-entropy and Kullback-Leibler (KL) divergence loss functions. Series of known neural classifiers were used for experiments including some ablation experiments. Reference [115] proposed an audio-visual technique which leverages the intrinsic synchronization between audio and video for deepfake detection. This technique employs joint training technique. The authors evaluated this approach on two datasets curated from existing datasets.

Another work conducted several experiments and reported that purely multimodal approach to detection of deepfake audio and video in multimedia data performed worst when compared to unimodal and ensemble approaches [48]. These experiments were carried out using variants of Convolutional Neural Network on FakeAVCeleb dataset [49]. FakeAVCeleb comprises both deepfake audios and videos. The report shows that ensemble method performs best followed by the unimodal method. Two variants of the ensemble method were presented namely soft-voting and hard-voting. The difference between the two variants is that the final decision was determined by average and majority votes from the two modalities used for soft-voting and hard-voting respectively. AVFakeNet [43] uses Dense Swin Transformer Network

(DST-Net) for audio-visual classification of deepfake videos. AVFakeNet is a unified framework which consists of a number of blocks. Evaluation of the framework using FakeAVCeleb dataset [49] reveals better performance when compared to other reproduced CNN-based models. AVForensics [117] is a transformer-based audio-visual framework for deepfake detection. AVForensics is a dual-phase framework primarily for deepfake videos detection but driven by the audio component of the multimedia content. It uses joint audio-visual contrastive learning in training and classified videos into real or fake categories. Experimental results prove the efficacy of the approach in comparison with a number of compared approaches. With the conviction that producing realistic video sequence with inconsistent modalities, [86] proposed a time-aware neural model to detect deepfake videos. In their technique, they trained separate models for each modality. The model involves layers of CNN, siamese network and attention. Based on experimental findings, they concluded that the multimodal approach is better than unimodal one. Multimodaltrace [82] is one of the latest works based on deep learning and which consider audio-visual modalities for deepfake detection. Multimodaltrace considers spectral and spatio-temporal features in audio and visual modalities respectively using Multi-Layer Perceptron (MLP) mixer layers. It consists of six blocks and made use of joint training in the development pipeline. A contrasting characteristic of Multimodaltrace is that the problem is formulated as a multiclass multilabel classification problem. Experimental results shows that several algorithms and different techniques were used.

### 4) TEXT, AUDIO AND VIDEO

Apart from audio and video, audio or video transcription can also be done to have an equivalent text modality. With a specific focus on detecting misleading videos related to COVID-19 using multimedia contents, [90] introduced a model (named TikTec) with the aim of answering two important research questions as follow: (i) How to aggregate heterogeneous information covering several modalities in videos and (ii) How to extract information from misleading and manipulated multimedia contents including videos. TikTec primarily consists of a Caption-guided Visual Representation Learning (CVRL) component, an Acoustic-aware Speech Representation Learning (ASRL) component, a Visual-speech Co-attentive Information Fusion (VCIF) and a Supervised Misleading Video Detection (SMVD) module. The CVRL leverages the captions on video frames and/or audio-transcriptions for visual representation learning. A bidirectional GRU is used to encode the semantic information in these captions and transcriptions. The ASRL learns the features of the audio in the videos. The audio segments are transformed into vectors using Mel-Frequency Cepstral Coefficients. A corresponding text in each audio segment is combined with that particular segment to form a hybrid text-audio representation. In the VCIF module on the other hand, a co-attention map is used to fuse the frames

and speech features extracted from the video and audio contents respectively. Finally, the SMVD module uses a neural classifier to predict whether a multimedia-based video is misleading or not.

A recent work (FakeSV) [77] addresses two major issues. First is the issue of inadequate datasets for multimodal studies of fake news detection. The second is adequate usage of modalities for fake news detection. They addressed these issues by creating a large-scale multimodal dataset called FakeSV and usage of the available modalities and attributes in the developed dataset for fake news detection experiments. They consider multimedia data for fake news detection from three perspectives of news content, social context and propagation. They employ BERT, VGGish [38] and VGG-19 to extract features from text, audio and video respectively. To obtain spatio-temporal and multi-granularity information, the videos were considered at both frame and clip levels. The resulting multimodal classifier consists of two cross-modal transformers. They conducted extensive experiments and compare the performance of their model with state-of-the-art models covering the perspectives earlier stated. Their model outperforms the compared SOTA models. They prove the efficacy of their approach through ablation experiments some of which include consideration of individual modalities. Reference [79] followed up on prior works to detect correlations among videos which emanated from the same event. This, according to them can either be complementary or contradictory and therefore, can serve as a mechanism to evaluate them. Based on this assumption, they introduced “Neighbor-Enhanced fakeE news video Detection” (NEED) framework. NEED extracted features from related multimedia contents including title, comments, user profile, keyframes, video clips and audio. These features were then aggregated using Graph Attention (GAT) network. The classifier employs an attention module on the constructed event graph produced from graph network. Evaluation includes the main NEED model and ablation studies. The ablation studies are based on graph aggregation and debunking rectification used severally. The results show the effectiveness of NEED.

## B. MULTIMODAL DATASETS FOR FAKE NEWS DETECTION

We present an overview of the experimental datasets that have been used for multimodal fake news detection. These datasets have been organized according to the modalities involved. The fake news datasets used by [65] as presented in TABLE 1 are available at<sup>2,3,4</sup>

### 1) TEXT AND IMAGE

The following datasets for fake news detection consist of text and image modalities:

- **Twitter dataset:** “Image-verification corpus” datasets for MediaEval2015 and MediaEval2016 [13], [14] often referred to as Twitter datasets have been used as benchmark datasets for the “Verifying Multimedia Use” task in MediaEval 2015 and 2016 workshops. The MediaEval2015 dataset consist of 11 events as training set, comprising a total 5,008 real and 6,840 fake tweets. The test set consist of 1,217 real and 2,564 fake tweets. In MediaEval2015, the number of fake tweets is higher by 192 due to a number of rumor tweets that were included but discarded in the final dataset. With respect to MediaEval2016, the training and test set of MediaEval2015 were combined in one set which served as training set and a new set of 1,107 (real) and 1,121 (fake) posts for testing.
- **Weibo dataset:** The Weibo dataset [46] in which real news component were collected between May, 2012 to January, 2016 from trustworthy news sources in China including Xinhua News Agency while the fake news part were crawled from other sources and verified by the rumor debunking platform of a microblogging website called Weibo. The original Weibo dataset comprises 4,749 and 4,779 fake and real news respectively making a total of 9,528 news with images. In the experiment conducted by the authors, the training set consists of 3,749 rumor (fake) and 3,783 non-rumor (real) news while the test set consists of 1,000 rumor (fake) and 996 non-rumor (real) news. Weibo now has several versions based on the period of collection. For instance, [106] used a version collected between May 2012 to November 2018.
- **PolitiFact:** PolitiFact dataset is part of FakeNewsNet [93], a repository of news data which fact-checks political reports and issues. It has been collected from the website<sup>5</sup> of the organization. It consists of three contexts; news content, social context and spatio-temporal information. Categories labeling have been carried out by human annotators as part of the dataset development. The news content component comprises mainly the news headline and body. PolitiFact consists of news articles that were published from May, 2002 to July, 2018. It comprises 1056 news articles with 624 real news and 432 fake news. 948 instances of the entire news have textual information out of which 420 are fake while 528 are real. The number of news with visual (image) content is 783 comprising 336 fake and 447 real news respectively.
- **GossipCop (GCop):** The Gossipcop dataset is also part of FakeNewsNet [93]. Gossipcop is a website<sup>6</sup> that also fact-checks news reports but focuses on entertainment and celebrity news. The news articles in this dataset were published between July, 2000 to December, 2018. It shares the same characteristics with PolitiFact in

<sup>2</sup><https://www.kaggle.com/datasets/jrvvika/fake-news-detection>

<sup>3</sup><https://www.kaggle.com/datasets/pontes/fake-news-sample>

<sup>4</sup><https://drive.google.com/file/d/0B3e3qZpTccsMFo5bk9Ib3VCc2c/view>

<sup>5</sup><https://www.politifact.com/>

<sup>6</sup><https://www.gossipcop.com/>

terms of annotation, contexts and contents. Gossipcop has 22,140 news articles with 16,817 real news and 5,323 fakes news. The number of news with textual information is 21,641 out of which 4,947 are fake while 16,694 are real. 18,417 instances have visual contents comprising 1,650 fake and 16,767 real instances. This statistics show that GossipCop is a class-imbalanced dataset.

- **r/Fakeddit**: r/Fakeddit [68] consists 1,063,106 multiple categories of news instances. It is one of the largest multimedia collections on fake news. In addition to the text and image modalities, metadata and comments are also part of the contents. r/Fakeddit has been thoroughly annotated and organized into fine-grained binary, 3-way and 6-way categories.
- **NewsBag**: NewsBag [47] has 215,000 news instances. This comprises 15,000 fake news and 200,000 real news. The real news have been curated from the Wall Street Journal while the fake news were crawled from The Onions; an American digital media and news organization. To cater for the class-imbalance, a new version was created (NewsBag++) in which the fake category was increased to 389,000. Another set for testing was created separately which consists of 11,000 real articles and 18,000 fake news.
- **TI-CNN dataset**: TI-CNN dataset [109] was developed for the validation of a CNN-based model. It comprises 8,074 real and 11,941 fake news. The real news are collected from trustworthy sources such as Washington Post etc. while the fake news are crawled from websites contained in Risdal's collection of fake news published on Kaggle<sup>7</sup>
- **ReCOVeRy**: ReCOVeRy [113] focuses on the reliability of news pertaining to COVID-19. The news were published between January to May, 2020. ReCOVeRy comprises 2,029 news articles out of which 1,364 are labeled as reliable with 665 labeled as unreliable. In contrast to most dataset, all the instances of the dataset have accompanying images.

## 2) TEXT AND VIDEO

Text and video modalities are core components of the following experimental datasets for fake news detection:

- **Volunteer Annotated Video Dataset (VAVD)**: As the name suggests, VAVD [72] was created through volunteering efforts of 20 participants. Over a 100,000 videos and comments uploaded on Youtube between September 2013 and October 2016 were collected and annotated into categories. The annotations were carried out through a two-round annotation process into three categories namely "Legitimate Spam" and "Not Sure".
- **Fake Video Corpus (FVC)**: FVC [73] was developed as part of InVID project. It consists of videos and their metadata. It has several versions as a result of continuous

expansion with the latest version having 2458 real and 3957 fake videos.

- **Misleading Youtube Video Corpus (MYVC)**: MYVC [22] is a product of collection of real and fake news for popular fact-checking websites and Youtube. It consists of 902 fake and 903 real news contents. In the experiment conducted by the authors ([22]), this dataset was merged with FVC and VAVD.

## 3) AUDIO AND VIDEO

We briefly describe the following datasets which are based on audio and video modalities:

- **Deepfake Detection Challenge (DFDC) Preview dataset**: The DFDC dataset [28] was developed to evaluate submissions for the Deepfake Detection Challenge. The raw data were collected by direct filming of people who agreed to participate. The final dataset was split into training, validation and test sets. The training set consists of 119,154 ten seconds video clips with 486 unique subjects out of which 100,000 contains deepfakes. The validation set comprises 4,000 ten seconds video clips, out of which 2000 clips contains deepfakes covering 214 unique subjects. The test set has 10,000 ten seconds video clips out of which 5,000 contains deepfakes.
- **VidTIMIT Audio-Video dataset (TIMIT)**: The VidTIMIT (or TIMIT) dataset [87] comprises video and corresponding audio recordings of 43 people, reciting short sentences. The dataset was collected in three sessions, with an average delay of seven days between Session 1 and 2, and six days between Session 2 and 3. The sentences were chosen from the test part of the TIMIT corpus [32]. There are ten sentences per person. The first six sentences are assigned to Session 1. The next two sentences are assigned to Session 2 and the remaining two to Session 3. The first two sentences for all persons are the same, with the remaining eight generally different for each person.
- **FakeAVCeleb dataset (FkAVCD)**: FakeAVCeleb dataset [49] comprises 500 real and 19,500 fakes videos making a total of 20,000 videos. Each video has an accompanying audio. One striking characteristic of this dataset is that it can be used for diverse classification problem because of its mixture of real and fake modalities. The possible combination are Real-Audio/Real-Video, Fake-Audio/Real-Video, Real-Audio/Fake-Video and Fake-Audio/Fake-Video.
- **World Leaders Deepfake Dataset (WLDD)**: [1] created this dataset with a focus on world leaders each of which is referred to as a "person of interest" (POI). The raw videos were downloaded from Youtube. The entire dataset consists of parts; the real part comprising 30,683 ten seconds clips involving 1,004 unique people. The fake part comprising comedic impersonators for each POI, face-swap deep fakes, lip synchronization deep and puppet master deep fakes.

<sup>7</sup><https://www.kaggle.com/datasets/mrisdal/fake-news>



- **Presidential Deepfakes Dataset (PDD):** The Presidential Deepfakes Dataset [88] contains 32 videos of the two most recent United States presidents; Joe Biden and Donald Trump. Each video in the dataset features one of the two presidents expressing a political opinion in a formal environment. The contents of 16 out of the 32 were modified to create the deepfakes. The modified features include the audio and visual (video) modalities.

#### 4) TEXT, AUDIO AND VIDEO

The following datasets containing text, audio and video modalities are used for fake news detection experiments:

- **COVID-19 Video Dataset:** COVID-19 Video Dataset [90] is a collection of videos related to COVID-19 collected from TikTok. The videos have accompanying metadata, video descriptions and audio contents. The dataset contains 226 misleading and 665 reliable videos. The ground-truth labels were done based on majority vote by human annotators.
- **FakeSV Dataset:** FakeSV Dataset [77] contains 1,827 fake and 1,827 real news instances which were collected from Douyin<sup>8</sup> and Kuaishou.<sup>9</sup> Both Douyin and Kuaishou are Chinese apps for sharing users' short videos. Each news instance comprises user, title, metadata and video with accompanying audio.

### C. DEEP LEARNING TECHNIQUES FOR MULTIMODAL HARMFUL LANGUAGES DETECTION

In this section, we discuss relevant works on deep learning for multimodal harmful languages detection. TABLE 2 shows the surveyed models for harmful languages detection, the underlying deep learning architecture, the modalities involved, the fusion technique(s) and the experimental datasets used.

#### 1) TEXT AND IMAGE

Reference [40] investigated cyberbullying detection using both textual and image modalities. This study is one of the earliest multimodal network-based approaches on this task. Logistic regression classifier trained with a forward feature selection is the technique employed for multilabel classification of contents. The classes include "Cyberbullying", "Non-cyberbullying", "Cyberaggression" and "Non-cyberaggression". A dataset collected from Instagram was used for experiments and evaluation of the model. For sarcasm detection in multimedia tweets, [16] developed a model which combines images' attributes, images and texts using bidirectional LSTM network. In order to take into account the importance of each modality, a representation fusion was introduced as part of the model development pipeline. According to the authors, this representation fusion was inspired by attention mechanism. Automated hate speech detection was studied by [108] with multimodal techniques

involving text and images. TextCNN [53] was used for text representation while a pretrained CNN-based model was used for image representation, the outputs of which were fused. They experimented with a number of multimodal fusion approaches including concatenation, addition and attention mechanism. Evaluation reports on the experiments did not show any tangible gain in fusing the two modalities. What can be referred to as a truly standard benchmark dataset for multimodal hate speech classification was developed by [35] which they named MMHS150K. MMHS150K is a large scale collection of tweets from Twitter and annotated for hate speech task. Evaluation of models on the dataset considers image-inserted texts in addition to the main text and image modalities. They experimented widely with diverse families of models. Reference [35] reported that multimodality did not achieve tangible improvement when compared with unimodal models. Reference [15] investigated the role of semantics and multimodality for both implicit and explicit hate speech detection. A subset of MMHS150K [35] was sampled to verify the validity of their hypothesis. They concluded that the multimodal model achieves best result when compared to other unimodal models. CapsNet-ConvNet [55] combines capsule network used with dynamic routing algorithm and deep Convolution Neural Network for cyberbullying detection. Modalities used by CapsNet-ConvNet include text, image and image-inserted text. The prediction component of the model is a late fusion of predictions from text and image-based models. CapsNet-ConvNet outperforms three other reproduced machine learning algorithms. A recent work of [107] leverages domain knowledge transfer for multimodal hate speech detection. The authors posit that there is a high interconnection between hate speech and sarcasm and therefore designate them as primary and auxiliary tasks for the purpose of cross-domain transfer learning. The model consists mainly of adaptation modules namely; semantic, definition and domain adaptation modules. Parameters learning is achieved through a joint optimization of the objective function by the domain models. Experiments show efficacy of the approach across the utilized datasets. A similar work [30] also uses transfer learning but however combines it with LSTM-based model for hate speech identification in multimodal fashion. They directly benchmark the performance of their model against the models of [35] and other reproduced works. Experiments were carried out using a minute subset of the MMHS150K dataset [35]. The results show marginal improvements on two common metrics.

#### 2) TEXT AND MEMES

To the best of our knowledge, the work of [99] is the first work to experiment on a truly multimedia contents for offensive language detection. A dataset named MultiOFF was developed for this purpose using existing meme data collection and experimented with some known neural classifiers. In their study, multimodal experiments show very little improvements over unimodal experiments when the same algorithms are

<sup>8</sup><https://www.douyin.com/>

<sup>9</sup><https://www.kuaishou.com/new-reco>



**TABLE 2. Models based on deep learning for multimodal harmful languages detection. keys: T&I - text and image, T&M - text and Meme, T&A&V - text, audio and video.**

Work/model	Algorithm/DNN Architecture	Modalities	Fusion Techniques(s)	Dataset(s)
[40]	Logistic regression	T&I	-	CID [40]
[16]	ResNet/BiLSTM	T&I	Hybrid/Average	SD [16]
[108]	TextCNN/ResNet	T&I	Intermediate/Concatenation/Summation	FHD [108]
[35]	CNN-based pretrained models, LSTM	T&I	Hybrid/Concatenation	MMHS150K [35]
[15]	Transformers, LSTM/CNN/Average Pooling	T&I	Early	MMHS150K [35]
[55]	Capsule network/CNN	T&I	Late	MMD [55]
CDKT [107]	Contrastive attention/Cross-domain knowledge transfer	T&I	Late/Summation	SD [16] and HMCD [52]
[30]	LSTM/transfer learning	T&I	Intermediate/Concatenation	Subset of MMHS150K [35]
Inter-modal [9]	Attention-based BiLSTM-CNN	T&I	Intermediate/Concatenation	MMHS150K [35] and MultiOFF [99]
[99]	BiLSTM, Stacked LSTM, Visual Transformer	T&M	Early/Concatenation	MultiOFF [99]
[52]	Text and visual transformers	T&M	Late	HMCD [52]
DisMultiHate [56]	Transformers/Attention-based encoders	T&M	Late/Summation	MultiOFF [99] and HMCD [52]
[75]	Transformers/CNN-based pretrained models	T&M	Late	HarMeme [75]
MOMENTA [76]	CLIP [80], attention mechanism	T&M	Intermediate/Concatenation	HarMeme [75] and Harm-P [76]
[63]	Transformers/GRU, CLIP [80]	T&M	Early/Concatenation	MultiBully [63]
MeBERT [112]	Transformers, Stacked LSTM, Attention etc.	T&M	Intermediate/Multiplication	MAD [91] and MultiOFF [99]
[20]	CNN/Global average pooling, visual attention	T&M	Early/Concatenation	MultiOFF [99], MMHS150K [35], HMCD [52]
MemeFier [54]	CLIP [80] encoding	T&M	Late/Majority voting	MultiOFF [99], HMCD [52], MAD [91]
[97]	Several machine learning algorithms	T&A&V	-	VMSD [81]
[26]	Text and vision transformers	T&A&V	Intermediate/Concatenation	HATEMM Dataset [26]

used. Curiously, some unimodal experiments with different algorithms outperform multimodal experiments. As part of the hateful memes challenge competition, [52] developed a dataset of multimedia memes to identify hateful memes. They presented a number of models based on known neural architectures. These models were evaluated based on defined benchmarks. These models comprises both unimodal and multimodal approaches and include Text BERT, Visual BERT, Image-Region, ViLBERT, Late Fusion, Concat BERT among others. Reference [56] proposed a technique called DisMultiHate to disentangle target entities in multimedia memes for hate speech detection. The proposed technique consists of three modules namely data pre-processing, text representation learning and visual representation learning modules. The text and visual representation learning use encoder and attention-based encoder respectively. DisMultiHate uses a regression layer to generate the probability of a multimedia content being hate or not. Experimental evaluation of the method on MultiOFF [99] improved performance over compared baselines. As a way of evaluating a dataset (HarMeme) they created, [75] experimented with several unimodal and multimodal models. Some of the unimodal models are Text BERT, VGG19, DenseNet among others. The multimodal models include but not limited to Concat BERT, Late Fusion and ViLBERT CC. Reference [76] developed MOMENTA, a framework for identification of

harmful memes and the target entities. It uses Google's Vision API to extract image-inserted texts. The extracted text and images were then encoded with a pre-trained visual-linguistic model and VGG-19 respectively. A key component of MOMENTA is the fusion of intra-modal and cross-modal attention. It outperforms majority of the compared baselines. Reference [63] proposes a model which considers sentiment, emotion and sarcasm for detecting cyberbullying in multimedia memes. ResNet-50 [37] and BERT [27] were used for representation of image and text features respectively. A core part of the model is an inter-modal attention layer. In order to evaluate the model, a dataset was created with which compared models were also benchmarked. The compared models also comprises those of ablation experiments. MeBERT is another work [112] which uses external knowledge-base to enhance semantic representation for the detection of offensive memes. It concatenates global texts and images features based on attention mechanism for the task. Experiments on two public datasets shows the effectiveness of the technique. MSKAV [20] is a multimodal deep learning model developed to capture hateful information in memes with specific application to hate speech and offensive language. The authors introduced several attention blocks in the model development. The performance of MSKAV were compared with those of ablation experiments. A recent work on detection of multimodal hate speech and offensive

language is MemeFier [54], a deep learning framework for classifying memes. It incorporates external knowledge into features' encoding. A key component of MemeFier is dual-stage, alignment-aware fusion of modalities. Experiments on three datasets show MEMEFIER outperforms baselines on two of the three datasets.

### 3) TEXT, AUDIO AND VIDEO

In a seminal application of machine learning to multimedia contents involving text audio and video, [97] supplemented traditional text content with audio and visual (video) contents for the detection of cyberbullying. They utilized some possible features which have been identified in prior literature namely channel capacity, arousal, affect and cognition. They applied about five machine learning algorithms using the identified features. The authors concluded that audio and video are an important component of cyberbullying detection and that their use as supplement to text greatly enhanced cyberbullying detection. As a way of validating HateMM dataset [26], experiments on fusion of diverse deep learning models were conducted. For each modality, a number of neural architectures were used for feature representation. Element-wise multiplication of features from BERT, Vision Transformer (ViT) [29] and MFCC produced the best performance.

## D. MULTIMODAL DATASETS FOR HARMFUL LANGUAGES DETECTION

### 1) TEXT AND IMAGE

- **Cyberbullying Incidents Dataset (CID):** Cyberbullying Incidents Dataset [40] was created from raw 25,000 Instagram public user profiles involving media objects. Due to the cost of annotation, only a subset of the media sessions were labeled. This subset includes 3,165 unique media sessions out of which 697 contains profane words. The categories in the dataset are “Cyberbullying”, “Non-cyberbullying”, “Cyberaggression” and “Non-cyberaggression”.
- **Facebook Hate Dataset (FHD):** The Facebook Hate Dataset [108] is a collection of hate posts reported by users over a seven month period. Every instance of the data contains some piece of text and an image. The dataset was split into train/development and test sets. The train/development set contains 320,000 positive (not hate) and 58,000 negative (hate) samples. The test set on the other hand, contains 42,000 positive (not hate) and 11,000 negative (hate) samples.
- **MMHS150K:** MMHS150K [35] is a dataset for hate speech detection which has been developed from a large-scale collection of tweets from Twitter. The dataset consists of 150,000 samples, on which standard annotations have been carried out by human annotators. The annotated data consists of 112,845 “Not-hate” samples and 36,978 “Hate” samples. The categories are “No attacks to any community”, “homophobic”, “sexist”, “racist”, “religion based attacks” and “attacks

to other communities”. The other five categories apart from “No attacks to any community” are “Hate” labels. The dataset was further split into test, validation and training sets consisting of 10,000, 5,000 and 135,000 samples respectively. Each data sample has a text and associated image and tangible number of the images have texts inserted within them.

- **Deciphering Implicit Hate Dataset (DIHD):** Deciphering Implicit Hate [15] Dataset contains 5,000 instances which have been taken from MMHS150K [35] and re-annotated into four categories namely “Hateful”, “Counterspeech”, “Reclaimed” and “None” (Not-hate). These categories each contains 1,850, 113, 366, 2,671 instances respectively.
- **Mix-modal Dataset (MMD):** The Mix-modal Dataset [55] is a dataset for cyberbullying detection which contains 10,000 instances drawn from Youtube, Instagram and Twitter. It consists of text and images although the authors recognizes some of the images as infographics. The categories are “Bullying” and “Non-bullying” and contain 5,700 and 4,000 instances respectively. In the “Bullying” category, there are 1,260 image-only, 3,000 text-only and 1,440 infographic instances. On the other hand, the “Non-bullying” category contains 740, 3,000 and 560 instances for image-only, text-only and infographic instances respectively.
- **Sarcasm Dataset (SD):** The Sarcasm Dataset [16] was built through preprocessing of a collection of tweets with images. The dataset consists of positive (sarcastic) and negative (non-sarcastic) categories. It has been split into training, validation and test sets. The training set consists of 8,642 positive and 11,174 negative instances. The validation comprises 959 positive and 1,451 negative instances. The number of positive and negative instances in the test set are 959 and 1450 respectively.
- **Hateful Meme Challenge Dataset (HMCD):** The Hateful Meme Challenge Dataset [52] was build by a third-party firm for the challenge organizers. The standard phases involved the development are filtering, meme construction, hatefulness rating and benign confounders. Inter-annotator agreement was also reached on instances where the annotators had disagreements.

### 2) TEXT AND MEME

Combination of text and meme has wider use in harmful languages detection than fake news detection. Some of the datasets containing these two modalities are described as follow:

- **MultiOFF:** MultiOFF dataset [99] was developed from a collection of memes from social media which were annotated for offensiveness or otherwise. It is an extension of an existing dataset about 2016 United States Presidential Election. In all, MultiOFF contains 743 instances split into training, validation and test sets. The composition of modalities involve only memes but the image-text were extracted to serve as the text

modality. The training set contains 187 “Offensive” and 258 “Non-offensive” samples respectively. Both the validation and test sets are each made up of 59 “Offensive” and 90 “Non-offensive” samples respectively.

- **Hateful Meme Challenge Dataset:** Hateful Meme Challenge Dataset [52] was designed to make it difficult for unimodal models to succeed. The development process consists of standard procedures for dataset creation including inter-annotator agreement. In all, it has 10,000 memes covering five types which include multimodal hate, unimodal hate, benign image, benign text and random not-hateful samples.
- **HarMeme:** HarMeme [75] is a collection of COVID-19 related memes from social media. It contains a total of 3,544 samples annotated into three categories namely “Very Harmful”, “Partially Harmful” and “Harmless”. The general train, development and test split of the dataset are 3,013, 177 and 354. The primary targets of the contents are individuals, organizations, communities and societies.
- **Harm-P Dataset:** Harm-P Dataset [76] is an extension of HarMeme with addition of more data related to United States’ politics. It has 3,552 instances. The general train, validation and test splits are 3,020, 177 and 355. The categories and targets are same as that of HarMeme.
- **MultiBully:** MultiBully [63] is a mixture of different types of harmful languages including Cyberbully, Harmfulness and Sarcasm. It also comprises sentiment and emotions types. MultiBully covers English and Hindi languages and consists of 5,854 instances split into train, validation and test sets. Each of the harmful languages and sentiment/emotion types has different categories.
- **Memotion Analysis Dataset (MAD):** Memotion Analysis Dataset [91] encompasses three tasks namely sentiment, humour and emotion intensity classification. In the humour classification task which is the task related to harmful languages, the labeled categories are “Sarcastic”, “Humorous”, “Motivation” and “Offensive” meme. The entire dataset consists of approximately 10,000 instances.

### 3) TEXT, AUDIO AND VIDEO

- **Vine Media Session Dataset (VMSD):** Vine Media Session Dataset [81] is a dataset for Cyberbullying detection which has been collected from Vine<sup>10</sup> video sessions. The final dataset contains 969 videos being the result of data collected from 59,560 users and filtered in such a way that each instance belong to a unique user and contain all the media sessions. The videos have accompanying audio and other attributes such as user information, comments etc.

<sup>10</sup>Vine was an American short-form video hosting service which is no longer in existence.

- **HATEMM Dataset:** HATEMM Dataset [26] is a product of raw data collected from BitChute,<sup>11</sup> a social video hosting platform. After undergoing the standard dataset creation procedure, the final annotated HATEMM dataset contains 1,083 videos of approximately 43 hours content. The length of each video is between 2.40 and 2.56 minutes. Furthermore, each video has an associated transcribed text and audio. The categorization is a 2-way approach and the categories are “Hate” and “Not hate”.

## E. DEEP LEARNING TECHNIQUE FOR MULTIMODAL FAKE NEWS AND HARMFUL LANGUAGES DETECTION

### 1) TEXT AND IMAGE

To the best of our knowledge, the work of [9] is the first unified deep learning model for the detection of both fake news and harmful languages. In their work, they exploit the advances in deep learning and computer vision to unify the modalities in order to mitigate the effect of heterogeneity and semantic gap inherent in multimodal content understanding. In addition to the use of text and image modalities, the work also exploits text inserted in images. The major hypothesis in the model is that the modalities can complement each other for detection accuracy through inter-modal attention. The experiments on harmful languages cover hate speech and offensive language. They reported performance improvements on the state-of-the-art across the three tasks.

## VII. EVALUATION METRICS

Both fakes news and harmful languages detection have mostly been formulated as classification tasks [6], [69], [70], therefore the same metrics can be employed in evaluating them. We briefly give a simple definition of the following basic terms with abbreviations useful for describing the main evaluation metrics.

- **False Positives (FP):** False positives is the number of samples predicted as positive while the actual values are negative.
- **True Positives (TP):** True positives is the number of samples whose actual values are positive and correctly predicted as positive
- **True Negatives (TN):** True negatives is the number of samples whose actual values are negative and correctly predicted as negative
- **False Negatives (FN):** False negatives is the number of samples predicted as negative while the actual values are positive.

The following metrics have used to evaluate prior works on either one or both tasks.

<sup>11</sup><https://www.bitchute.com/>

### 1) ACCURACY

Accuracy is the percentage of correct predictions out of the total samples as defined by equation (5):

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (6)$$

### 2) PRECISION

Precision is the fraction of correctly predicted samples out of the total predicted samples. It is defined by equation (6):

$$Precision = \frac{TP}{TP + FP} \quad (7)$$

### 3) RECALL (TRUE POSITIVE RATE (TPR) OR SENSITIVITY)

Recall is the fraction of correctly predicted samples out of the total available samples. Recall is defined by equation (7):

$$Recall = \frac{TP}{TP + FN} \quad (8)$$

### 4) F1

F1 Score is the harmonic mean of precision and recall. F1 is computed using equation (8):

$$F1 = 2 * \frac{precision * recall}{precision + recall} \quad (9)$$

### 5) AREA UNDER CURVE (AUC)

To have an understanding of AUC, it is necessary to first understand the following additional metrics:

- **False Positive Rate (FPR):** The computation of False Positive Rate is given by equation (9):

$$FPR = \frac{FP}{TN + FP} \quad (10)$$

- **Receiver Operating Characteristic (ROC) Curve:** The ROC curve is derived from the plot of TPR against FPR. It is a metric which allows the measure of a classification problem at a certain threshold.

Therefore, the AUC is used as a summary of the ROC curve and it is used to evaluate a model's ability to distinguish between positive and negative classes. The higher a model's AUC, the better the model's performance.

## VIII. CHALLENGES OF MULTIMODAL CONTENT UNDERSTANDING: CONTEXT OF DEEP LEARNING, FAKE NEWS AND HARMFUL LANGUAGES

Some of the current challenges in multimodal fake news and harmful languages detection include the following although some are also the case in unimodal detection:

- **Dataset:** Standard datasets for both fake news and harmful languages are currently inadequate. Some of the datasets are noisy and are not annotated through standard procedures. Another challenge on dataset is the issue of imbalance modalities. Some instances of the datasets have one modality without other modalities.
- **Multilingualism:** Most of the reviewed deep neural classifiers are trained monolingual classifiers. However,

it is a fact that languages have their own peculiarities. Work on multilingual deep neural classifiers for both tasks is an area that needs to be explored. This is particularly more important in the sense that some world events attracts contributions from across the globe in which users for instance tweets in their native languages. A typical example is found in the Twitter datasets [13], [14]. Dealing with this kind of dataset requires either translation to a uniform language or the development of a multilingual classifier. The former needs to further grapple with challenges associated with machine translation.

- **Multiclass classification:** Majority of the works on multimodal fake news and harmful languages detection are formulated and adapted as binary classification problem even when using datasets with multiple classes. However, in reality some of these problems are best addressed with multiclass classification.
- **Heterogeneity gap:** The heterogeneity gap [17] applies to deep multimodal content understanding in general. Heterogeneity gap refers to the peculiarities and unique distribution of features of individual data modality when represented by deep neural networks. Despite the efforts to mitigate this gap, much still needed to be done for effective prediction outcomes in multimodal .fake news and harmful languages detection.
- **Semantic gap:** The semantic gap [17] is also a general problem in deep multimodal content understanding. Correlations among textual and visual features in multimedia contents is one that is difficult to capture. In the case of fake news and harmful languages detection, despite the introduction of measures such as attention mechanism, more effective approaches still need to be looked into.

## IX. FUTURE RESEARCH DIRECTION

As discussed earlier in section VIII that missing modalities is one of the challenges in multimodal fake news and harmful languages detection. The creation of large-scale datasets in these areas which is representative of all modalities is required. This will enable studies on the best techniques to represent them and the effect of each in the performance of deep learning models.

Research on multilingual classifiers for these tasks is one that cannot be overemphasized. Language barrier is an important problem which needs to be addressed in this domain. Research efforts in this direction is one which will benefit the field immensely.

Furthermore, advances on how to best bring out the contributory ability of the modalities, perhaps in fusion strategies is still an open research area. This is because the issue of heterogeneity and semantics of features of different data forms is still not fully resolved and require further exploration.



## X. CONCLUSION

In this paper, we reviewed the state-of-the-art deep learning-based works on multimodal fake news and harmful languages detection. The twin menaces of fake news and harmful languages being serious societal problems, we identify the interrelationships and the common effect of both. The possible data modalities for the two tasks were comprehensively discussed. In contrast to prior works, we have categorized reviews of techniques according to the data modalities involved. Furthermore, we delved into in depth details of data fusion strategies, introducing additional possible strategy for modeling multimodality in deep neural classifiers. The current challenges and possible future directions were also discussed.

## REFERENCES

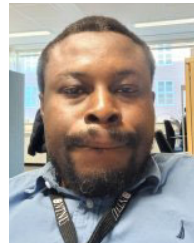
- [1] S. Agarwal, H. Farid, Y. Gu, M. He, K. Nagano, and H. Li, "Protecting world leaders against deep fakes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Long Beach, CA, USA, Jun. 2019, pp. 38–45.
- [2] C. Alcantara, V. Moreira, and D. Feijo, "Offensive video detection: Dataset and baseline results," in *Proc. 12th Lang. Resour. Eval. Conf.*, F. N. Bechet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis, Eds. Marseille, France, 2020, pp. 4309–4319. [Online]. Available: <https://aclanthology.org/2020.lrec-1.531/>
- [3] J. Alghamdi, Y. Lin, and S. Luo, "A comparative study of machine learning and deep learning techniques for fake news detection," *Information*, vol. 13, no. 12, p. 576, Dec. 2022, doi: [10.3390/info13120576](https://doi.org/10.3390/info13120576).
- [4] Z. Almutairi and H. Elgibreen, "A review of modern audio deepfake detection methods: Challenges and future directions," *Algorithms*, vol. 15, no. 5, p. 155, May 2022, doi: [10.3390/a15050155](https://doi.org/10.3390/a15050155).
- [5] V. Anand, R. Shukla, A. Gupta, and A. Kumar, "Customized video filtering on YouTube," 2019, *arXiv:1911.04013*.
- [6] E. F. Ayetiran, "An index-based joint multilingual/cross-lingual text categorization using topic expansion via BabelNet," *TURKISH J. Electr. Eng. Comput. Sci.*, vol. 28, no. 1, pp. 224–237, Jan. 2020, doi: [10.3906/elk-1901-140](https://doi.org/10.3906/elk-1901-140).
- [7] E. F. Ayetiran, "Attention-based aspect sentiment classification using enhanced learning through CNN-BiLSTM networks," *Knowl.-Based Syst.*, vol. 252, Sep. 2022, Art. no. 109409, doi: [10.1016/j.knsys.2022.109409](https://doi.org/10.1016/j.knsys.2022.109409).
- [8] E. F. Ayetiran, P. Sojka, and V. Novotný, "EDS-MEMBED: Multi-sense embeddings based on enhanced distributional semantic structures via a graph walk over word senses," *Knowl.-Based Syst.*, vol. 219, May 2021, Art. no. 106902, doi: [10.1016/j.knsys.2021.106902](https://doi.org/10.1016/j.knsys.2021.106902).
- [9] E. F. Ayetiran and Ö. Özgöbek, "An inter-modal attention-based deep learning framework using unified modality for multimodal fake news, hate speech and offensive language detection," *Inf. Syst.*, vol. 123, Jul. 2024, Art. no. 102378, doi: [10.1016/j.is.2024.102378](https://doi.org/10.1016/j.is.2024.102378).
- [10] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *Proc. 3rd Int. Conf. Learn. Represent.*, Y. Bengio and Y. LeCun, Eds., San Diego, CA, USA, 2015, pp. 1–15.
- [11] D. M. Ballesteros, Y. Rodriguez-Ortega, D. Renza, and G. Arce, "Deep4SNet: Deep learning for fake speech classification," *Exp. Syst. Appl.*, vol. 184, Dec. 2021, Art. no. 115465, doi: [10.1016/j.eswa.2021.115465](https://doi.org/10.1016/j.eswa.2021.115465).
- [12] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, Jan. 2003. [Online]. Available: <http://jmlr.org/papers/v3/blei03a.html>
- [13] C. Boididou, K. Andreadou, S. Papadopoulos, D. Dang-Nguyen, G. Boato, M. Riegler, and Y. Kompatsiaris, "Verifying multimedia use at MediaEval 2015," in *Proc. MediaEval*, M. A. B. Larson, M. Ionescu, X. Sjoberg, J. Anguera, M. Poignant, M. Riegler, C. Eskevich, C. Hauff, R. F. E. Sutcliffe, G. J. F. Jones, Y. Yang, M. Soleymani, and S. Papadopoulos, Eds., Wurzen, Germany, 2015, pp. 1–3. [Online]. Available: <https://ceur-ws.org/Vol-1436/Paper4.pdf>
- [14] C. Boididou, S. Papadopoulos, D. Dang-Nguyen, G. Boato, M. Riegler, S. E. Middleton, A. Petlund, and Y. Kompatsiaris, "Verifying multimedia use at mediaeval 2016," in *Proc. MediaEval*, G. Gravier, C. Demarty, H. Bredin, B. Ionescu, C. Boididou, E. Dellandrea, J. Choi, M. Riegler, R. F. E. Sutcliffe, I. Szoke, G. J. F. Jones, and M. A. Larson, Eds., Hilversum, The Netherlands, 2016, pp. 1–10. [Online]. Available: [https://ceur-ws.org/Vol-1739/MediaEval\\_2016\\_paper\\_3.pdf](https://ceur-ws.org/Vol-1739/MediaEval_2016_paper_3.pdf)
- [15] A. Botelho, S. Hale, and B. Vidgen, "Deciphering implicit hate: Evaluating automated detection algorithms for multimodal hate," in *Proc. Findings Assoc. Comput. Linguistics: ACL-IJCNLP*, 2021, pp. 1896–1907, doi: [10.18653/v1/2021.findings-acl.166](https://doi.org/10.18653/v1/2021.findings-acl.166).
- [16] Y. Cai, H. Cai, and X. Wan, "Multi-modal sarcasm detection in Twitter with hierarchical fusion model," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, Florence, Italy, 2019, pp. 2506–2515, doi: [10.18653/v1/p19-1239](https://doi.org/10.18653/v1/p19-1239).
- [17] W. Chen, W. Wang, L. Liu, and M. S. Lew, "New ideas and trends in deep multimodal content understanding: A review," *Neurocomputing*, vol. 426, pp. 195–215, Feb. 2021, doi: [10.1016/j.neucom.2020.10.042](https://doi.org/10.1016/j.neucom.2020.10.042).
- [18] Y. Chen, D. Li, P. Zhang, J. Sui, Q. Lv, L. Tun, and L. Shang, "Cross-modal ambiguity learning for multimodal fake news detection," in *Proc. ACM Web Conf.*, F. Laforest, R. Troncy, E. Simperl, D. Agarwal, A. Gionis, I. Herman, and L. Medini, Eds., Lyon, France, 2022, pp. 2897–2905, doi: [10.1145/3485447.3511968](https://doi.org/10.1145/3485447.3511968)
- [19] A. Chhabra and D. K. Vishwakarma, "A literature survey on multimodal and multilingual automatic hate speech identification," *Multimedia Syst.*, vol. 29, no. 3, pp. 1203–1230, Jun. 2023, doi: [10.1007/s00530-023-01051-8](https://doi.org/10.1007/s00530-023-01051-8).
- [20] A. Chhabra and D. K. Vishwakarma, "Multimodal hate speech detection via multi-scale visual kernels and knowledge distillation architecture," *Eng. Appl. Artif. Intell.*, vol. 126, Nov. 2023, Art. no. 106991, doi: [10.1016/j.engappai.2023.106991](https://doi.org/10.1016/j.engappai.2023.106991).
- [21] K. Cho, B. van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder–decoder for statistical machine translation," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, Doha, Qatar, 2014, pp. 1724–1734, doi: [10.3115/v1/d14-1179](https://doi.org/10.3115/v1/d14-1179).
- [22] H. Choi and Y. Ko, "Using topic modeling and adversarial neural networks for fake news video detection," in *Proc. 30th ACM Int. Conf. Inf. Knowl. Manag.*, G. Demartini, G. Zuccon, J. S. Culpepper, Z. Huang, and H. Tong, Eds. 2021, pp. 2950–2954, doi: [10.1145/3459637.3482212](https://doi.org/10.1145/3459637.3482212).
- [23] H. Choi and Y. Ko, "Effective fake news video detection using domain knowledge and multimodal data fusion on YouTube," *Pattern Recognit. Lett.*, vol. 154, pp. 44–52, Feb. 2022, doi: [10.1016/j.patrec.2022.01.007](https://doi.org/10.1016/j.patrec.2022.01.007).
- [24] K. Chugh, P. Gupta, A. Dhall, and R. Subramanian, "Not made for each other- audio-visual dissonance-based deepfake detection and localization," in *Proc. 28th ACM Int. Conf. Multimedia*, C. W. Chen, R. Cucchiara, X. Hua, G. Qi, E. Ricci, Z. Zhang, and R. Zimmermann, Eds., Seattle, WA, USA, Oct. 2020, pp. 439–447, doi: [10.1145/3394171.3413700](https://doi.org/10.1145/3394171.3413700).
- [25] C. Comito, L. Caroprese, and E. Zumpano, "Multimodal fake news detection on social media: A survey of deep learning techniques," *Social Neww. Anal. Mining*, vol. 13, no. 1, p. 101, Aug. 2023, doi: [10.1007/s13278-023-01104-w](https://doi.org/10.1007/s13278-023-01104-w).
- [26] M. Das, R. Raj, P. Saha, B. Mathew, M. Gupta, and A. Mukherjee, "HateMM: A multi-modal dataset for hate video classification," in *Proc. Int. AAAI Conf. Web Social Media*, vol. 17, 2023, pp. 1014–1023.
- [27] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Human Lang. Technol.*, J. Burstein, C. Doran, and T. Solorio, Eds., Minneapolis, MN, USA, 2019, pp. 4171–4186, doi: [10.18653/v1/n19-1423](https://doi.org/10.18653/v1/n19-1423).
- [28] B. Dolhansky, R. Howes, B. Pflaum, N. Baram, and C. Canton Ferrer, "The deepfake detection challenge (DFDC) preview dataset," 2019, *arXiv:1910.08854*.
- [29] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16 × 16 words: Transformers for image recognition at scale," in *Proc. 9th Int. Conf. Learn. Represent.*, 2021, pp. 1–22.
- [30] V. Dwivedy and P. K. Roy, "Deep feature fusion for hate speech detection: A transfer learning approach," *Multimedia Tools Appl.*, vol. 82, no. 23, pp. 36279–36301, Sep. 2023, doi: [10.1007/s11042-023-14850-y](https://doi.org/10.1007/s11042-023-14850-y).



- [31] K. Fukushima and S. Miyake, "Neocognitron: A new algorithm for pattern recognition tolerant of deformations and shifts in position," *Pattern Recognit.*, vol. 15, no. 6, pp. 455–469, Jan. 1982, doi: [10.1016/0031-3203\(82\)90024-3](https://doi.org/10.1016/0031-3203(82)90024-3).
- [32] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. L. Dahlgren, V. Z. Dahlgren, and V. Zue, "TIMIT acoustic-phonetic continuous speech corpus," Tech. Rep., 1993, doi: [10.35111/17gk-bn40](https://doi.org/10.35111/17gk-bn40).
- [33] A. Giachanou, G. Zhang, and P. Rosso, "Multimodal fake news detection with textual, visual and semantic information," in *Proc. 23rd Int. Conf., P. Sojka, I. Kopecek, K. Pala, and A. Horak, Eds., Brno, Czech Republic. Berlin, Germany: Springer, 2020*, pp. 30–38, doi: [10.1007/978-3-030-58323-1\\_3](https://doi.org/10.1007/978-3-030-58323-1_3).
- [34] A. Giachanou, G. Zhang, and P. Rosso, "Multimodal multi-image fake news detection," in *Proc. IEEE 7th Int. Conf. Data Sci. Adv. Analytics (DSAA)*, G. I. Webb, Z. Zhang, V. S. Tseng, G. Williams, M. Vlachos, and L. Cao, Eds., Oct. 2020, pp. 647–654, doi: [10.1109/DSAA49011.2020.00091](https://doi.org/10.1109/DSAA49011.2020.00091).
- [35] R. Gomez, J. Gibert, L. Gomez, and D. Karatzas, "Exploring hate speech detection in multimodal publications," in *Proc. IEEE Winter Conf. Appl. Comput. Vis., Snowmass Village, CO, USA, Mar. 2020*, pp. 1459–1467, doi: [10.1109/WACV45572.2020.9093414](https://doi.org/10.1109/WACV45572.2020.9093414).
- [36] A. Graves, N. Jaitly, and A.-R. Mohamed, "Hybrid speech recognition with deep bidirectional LSTM," in *Proc. IEEE Workshop Autom. Speech Recognit. Understand., Olomouc, Czech Republic, Dec. 2013*, pp. 273–278, doi: [10.1109/ASRU.2013.6707742](https://doi.org/10.1109/ASRU.2013.6707742).
- [37] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Las Vegas, NV, USA, Jun. 2016*, pp. 770–778, doi: [10.1109/CVPR.2016.90](https://doi.org/10.1109/CVPR.2016.90).
- [38] S. Hershey, S. Chaudhuri, D. P. W. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, M. Slaney, R. J. Weiss, and K. W. Wilson, "CNN architectures for large-scale audio classification," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process., New Orleans, LA, USA, Mar. 2017*, pp. 131–135, doi: [10.1109/ICASSP.2017.7952132](https://doi.org/10.1109/ICASSP.2017.7952132).
- [39] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997, doi: [10.1162/neco.1997.9.8.1735](https://doi.org/10.1162/neco.1997.9.8.1735).
- [40] H. Hosseinmardi, R. I. Rafiq, R. Han, Q. Lv, and S. Mishra, "Prediction of cyberbullying incidents in a media-based social network," in *Proc. IEEE/ACM Int. Conf. Adv. Social Netw. Anal. Mining, R. Kumar, J. Caverlee, and H. Tong, Eds., San Francisco, CA, USA, Mar. 2016*, pp. 186–192, doi: [10.1109/ASONAM.2016.7752233](https://doi.org/10.1109/ASONAM.2016.7752233).
- [41] L. Hu, S. Wei, Z. Zhao, and B. Wu, "Deep learning for fake news detection: A comprehensive survey," *AI Open*, vol. 3, pp. 133–155, Jan. 2022, doi: [10.1016/j.aiopen.2022.09.001](https://doi.org/10.1016/j.aiopen.2022.09.001).
- [42] J. Hua, X. Cui, X. Li, K. Tang, and P. Zhu, "Multimodal fake news detection through data augmentation-based contrastive learning," *Appl. Soft Comput.*, vol. 136, Mar. 2023, Art. no. 110125, doi: [10.1016/j.asoc.2023.110125](https://doi.org/10.1016/j.asoc.2023.110125).
- [43] H. Ilyas, A. Javed, and K. M. Malik, "AVFakeNet: A unified end-to-end dense Swin transformer deep learning model for audio-visual deepfakes detection," *Appl. Soft Comput.*, vol. 136, Mar. 2023, Art. no. 110124, doi: [10.1016/j.asoc.2023.110124](https://doi.org/10.1016/j.asoc.2023.110124).
- [44] M. S. Jahan and M. Oussalah, "A systematic review of hate speech automatic detection using natural language processing," *Neurocomputing*, vol. 546, Aug. 2023, Art. no. 126232, doi: [10.1016/j.neucom.2023.126232](https://doi.org/10.1016/j.neucom.2023.126232).
- [45] L. Jiang, R. Li, W. Wu, C. Qian, and C. C. Loy, "DeeperForensics-1.0: A large-scale dataset for real-world face forgery detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Seattle, WA, USA, Jun. 2020, pp. 2886–2895.
- [46] Z. Jin, J. Cao, H. Guo, Y. Zhang, and J. Luo, "Multimodal fusion with recurrent neural networks for rumor detection on microblogs," in *Proc. ACM Multimedia Conf.*, Q. Liu, R. Lienhart, H. Wang, S. K. Chen, S. Boll, Y. P. Chen, G. Friedland, J. Li, and S. Yan, Eds., Mountain View, CA, USA, Oct. 2017, pp. 795–816, doi: [10.1145/3123266.3123454](https://doi.org/10.1145/3123266.3123454).
- [47] S. Jindal, R. Sood, R. Singh, M. Vatsa, and T. Chakraborty, "Newsbag: A multimodal benchmark dataset for fake news detection," in *Proc. Workshop Artif. Intell. Saf., 34th AAAI Conf. Artif. Intell.*, H. Espinoza, J. Hernandez-Orallo, X. C. Chen, S. S. Oheigeartaigh, X. Huang, M. Castillo-Effen, R. Mallah, and J. A. McDerimid, Eds., New York City, NY, USA, Feb. 2020, pp. 138–145. [Online]. Available: <https://ceur-ws.org/Vol-2560/paper27.pdf>
- [48] H. Khalid, M. Kim, S. Tariq, and S. S. Woo, "Evaluation of an audio-video multimodal deepfake dataset using unimodal and multimodal detectors," in *Proc. 1st Workshop Synth. Multimedia - Audiovisual Deepfake Gener. Detection*, S. Winkler, W. Chen, A. Dhall, and P. Korshunov, Eds., Oct. 2021, pp. 7–15, doi: [10.1145/3476099.3484315](https://doi.org/10.1145/3476099.3484315).
- [49] H. Khalid, S. Tariq, M. Kim, and S. S. Woo, "FakeAVCeleb: A novel audio-video multimodal deepfake dataset," in *Proc. NIPS*, J. Vanschoren and S. Yeung, Eds., Dec. 2021, pp. 1–22.
- [50] S. Khan, M. Fazil, V. K. Sejwal, M. A. Alshara, R. M. Alotaibi, A. Kamal, and A. R. Baig, "BiCHAT: BiLSTM with deep CNN and hierarchical attention for hate speech detection," *J. King Saud Univ.-Comput. Inf. Sci.*, vol. 34, no. 7, pp. 4335–4344, Jul. 2022, doi: [10.1016/j.jksuci.2022.05.006](https://doi.org/10.1016/j.jksuci.2022.05.006).
- [51] D. Khattar, J. S. Goud, M. Gupta, and V. Varma, "MVAE: Multimodal variational autoencoder for fake news detection," in *Proc. World Wide Web Conf.*, L. Liu, R. W. White, A. Mantrach, F. Silvestri, J. J. McAuley, R. Baeza-Yates, and L. Zia, Eds., San Francisco, CA, USA, 2019, pp. 2915–2921, doi: [10.1145/3308558.3313552](https://doi.org/10.1145/3308558.3313552).
- [52] D. Kiela, H. Firooz, A. Mohan, V. Goswami, A. Singh, P. Ringshia, and D. Testuggine, "The hateful memes challenge: Detecting hate speech in multimodal memes," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., 2020, pp. 1–14.
- [53] Y. Kim, "Convolutional neural networks for sentence classification," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, Doha, Qatar, 2014, pp. 1746–1751, doi: [10.3115/v1/d14-1181](https://doi.org/10.3115/v1/d14-1181).
- [54] C. Koutlis, M. Schinas, and S. Papadopoulos, "MemeFier: Dual-stage modality fusion for image meme classification," in *Proc. ACM Int. Conf. Multimedia Retr.*, I. Kompatsiaris, J. Luo, N. Sebe, A. Yao, V. Mazaris, S. Papadopoulos, A. Popescu, and Z. H. Huang, Eds., Thessaloniki, Greece, Jun. 2023, pp. 586–591, doi: [10.1145/3591106.3592254](https://doi.org/10.1145/3591106.3592254).
- [55] A. Kumar and N. Sachdeva, "Multimodal cyberbullying detection using capsule network with dynamic routing and deep convolutional neural network," *Multim. Syst.*, vol. 28, pp. 2043–2052, Dec. 2022, doi: [10.1007/s00530-020-00747-5](https://doi.org/10.1007/s00530-020-00747-5).
- [56] R. K. W. Lee, R. Cao, Z. Fan, J. Jiang, and W. Chong, "Disentangling hate in online memes," in *Proc. ACM Multimedia Conf.*, 2021, pp. 5138–5147, doi: [10.1145/3474085.3475625](https://doi.org/10.1145/3474085.3475625).
- [57] L. Li, J. Bao, H. Yang, D. Chen, and F. Wen, "FaceShifter: Towards high fidelity and occlusion aware face swapping," 2019, *arXiv:1912.13457*.
- [58] X. Li, P. Lu, L. Hu, X. Wang, and L. Lu, "A novel self-learning semi-supervised deep learning network to detect fake news on social media," *Multimedia Tools Appl.*, vol. 81, no. 14, pp. 19341–19349, Jun. 2022, doi: [10.1007/s11042-021-11065-x](https://doi.org/10.1007/s11042-021-11065-x).
- [59] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A robustly optimized BERT pretraining approach," 2019, *arXiv:1907.11692*.
- [60] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," 2021, *arXiv:2103.14030*.
- [61] J. Lu, D. Batra, D. Parikh, and S. Lee, "ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks," in *Proc. Adv. Neural Inf. Process. Syst.*, H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alche-Buc, E. B. Fox, and R. Garnett, Eds., Vancouver, BC, Canada, 2019, pp. 13–23.
- [62] T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, L. Marquez, C. Callison-Burch, J. Su, D. Pighin, and Y. Marton, Eds., Lisbon, Portugal, Sep. 2015, pp. 1412–1421, doi: [10.18653/v1/d15-1166](https://doi.org/10.18653/v1/d15-1166).
- [63] K. Maity, P. Jha, S. Saha, and P. Bhattacharyya, "A multitask framework for sentiment, emotion and sarcasm aware cyberbullying detection from multi-modal code-mixed memes," in *Proc. 45th Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, E. Amigo, P. Castells, J. Gonzalo, B. Carterette, J. S. Culpepper, and G. Kazai, Eds., Madrid, Spain, JJUL. 2022, pp. 1739–1749, doi: [10.1145/3477495.3531925](https://doi.org/10.1145/3477495.3531925).
- [64] E. Masciari, V. Moscatto, A. Picariello, and G. Sperli, "Detecting fake news by image analysis," in *Proc. 24th Symp. Int. Database Eng. Appl.*, B. C. Desai and W. Cho, Eds., Aug. 2020, p. 27, doi: [10.1145/3410566.3410599](https://doi.org/10.1145/3410566.3410599).
- [65] P. Meel and D. K. Vishwakarma, "HAN, image captioning, and forensics ensemble multimodal fake news detection," *Inf. Sci.*, vol. 567, pp. 23–41, Aug. 2021, doi: [10.1016/j.ins.2021.03.037](https://doi.org/10.1016/j.ins.2021.03.037).

- [66] T. Mittal, U. Bhattacharya, R. Chandra, A. Bera, and D. Manocha, "Emotions don't lie: An audio-visual deepfake detection method using affective cues," in *Proc. 28th ACM Int. Conf. Multimedia*, C. W. Chen, R. Cucchiara, X. Hua, G. Qi, E. Ricci, Z. Zhang, and R. Zimmermann, Eds., Seattle, WA, USA, Oct. 2020, pp. 2823–2832, doi: [10.1145/3394171.3413570](https://doi.org/10.1145/3394171.3413570).
- [67] M. F. Mridha, A. J. Keya, M. A. Hamid, M. M. Monowar, and M. S. Rahman, "A comprehensive review on fake news detection with deep learning," *IEEE Access*, vol. 9, pp. 156151–156170, 2021, doi: [10.1109/ACCESS.2021.3129329](https://doi.org/10.1109/ACCESS.2021.3129329).
- [68] K. Nakamura, S. Levy, and W. Y. Wang, "r/Fakeddit: A new multimodal benchmark dataset for fine-grained fake news detection," in *Proc. 12th Lang. Resour. Eval. Conf.*, N. Calzolari, F. Bechet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, and S. Piperidis, Eds., Marseille, France, May 2020, pp. 6149–6157. [Online]. Available: <https://aclanthology.org/2020.lrec-1.755/>
- [69] V. Novotný, E. Festus Ayetiran, M. Stefánik, and P. Sojka, "Text classification with word embedding regularization and soft similarity measure," 2020, *arXiv:2003.05019*.
- [70] T. M. Oladele and E. F. Ayetiran, "Social unrest prediction through sentiment analysis on Twitter using support vector machine: Experimental study on Nigeria's #EndSARS," *Open Inf. Sci.*, vol. 7, no. 1, Mar. 2023, Art. no. 20220141, doi: [10.1515/opis-2022-0141](https://doi.org/10.1515/opis-2022-0141).
- [71] B. Palani and S. Elango, "BBC-FND: An ensemble of deep learning framework for textual fake news detection," *Comput. Electr. Eng.*, vol. 110, Sep. 2023, Art. no. 108866, doi: [10.1016/j.compeleceng.2023.108866](https://doi.org/10.1016/j.compeleceng.2023.108866).
- [72] P. Palod, A. Patwari, S. Bahety, S. Bagchi, and P. Goyal, "Misleading metadata detection on YouTube," in *Proc. 41st Eur. Conf. IR Res.*, L. Azzopardi, B. Stein, N. Fuhr, P. Mayr, C. Hauff, and D. Hiemstra, Eds., Cologne, Germany, Berlin, Germany: Springer, 2019, pp. 140–147, doi: [10.1007/978-3-030-15719-7\\_18](https://doi.org/10.1007/978-3-030-15719-7_18).
- [73] O. Papadopoulou, M. Zampoglou, S. Papadopoulos, and I. Kompatsiaris, "Invid fake video corpus 2018 (version 1)," Tech. Rep., 2018, doi: [10.5281/zenodo.2535479](https://doi.org/10.5281/zenodo.2535479).
- [74] G. K. Pitsilis, H. Ramampiaro, and H. Langseth, "Effective hate-speech detection in Twitter data using recurrent neural networks," *Int. J. Speech Technol.*, vol. 48, no. 12, pp. 4730–4742, Dec. 2018, doi: [10.1007/s10489-018-1242-y](https://doi.org/10.1007/s10489-018-1242-y).
- [75] S. Pramanick, D. Dimitrov, R. Mukherjee, S. Sharma, M. S. Akhtar, P. Nakov, and T. Chakraborty, "Detecting harmful memes and their targets," in *Proc. Findings Assoc. Comput. Linguistics: ACL-IJCNLP*, C. Zong, F. Xia, W. Li, and R. Navigli, Eds. 2021, pp. 2783–2796, doi: [10.18653/v1/2021.findings-acl.246](https://doi.org/10.18653/v1/2021.findings-acl.246).
- [76] S. Pramanick, S. Sharma, D. Dimitrov, M. S. Akhtar, P. Nakov, and T. Chakraborty, "MOMENTA: A multimodal framework for detecting harmful memes and their targets," in *Proc. Findings Assoc. for Comput. Linguistics*, M. Moens, X. Huang, L. Specia, S. W. Yih, Eds. Dimitrov, Dominican Republic, 2021, pp. 16–20, doi: [10.18653/v1/2021.findings-emnlp.379](https://doi.org/10.18653/v1/2021.findings-emnlp.379).
- [77] P. Qi, Y. Bu, J. Cao, W. Ji, R. Shui, J. Xiao, D. Wang, and T. Chua, "FakeSV: A multimodal benchmark with rich social context for fake news detection on short video platforms," in *Proc. 37th AAAI Conf. Artif. Intell. (AAAI)*, *34th Conf. Innov. Appl. Artif. Intell. (IAAI)*, *13th Symp. Educ. Adv. Artif. Intell. (EAAI)*, B. Williams, Y. Chen, and J. Neville, Eds., Washington, DC, USA, 2023, pp. 14444–14452. [Online]. Available: <https://ojs.aaai.org/index.php/AAAI/article/view/26689>
- [78] P. Qi, J. Cao, X. Li, H. Liu, Q. Sheng, X. Mi, Q. He, Y. Lv, C. Guo, and Y. Yu, "Improving fake news detection by using an entity-enhanced framework to fuse diverse multimodal clues," in *Proc. CM Multimedia Conf.*, H. T. Shen, Y. Zhuang, J. R. Smith, Y. Yang, P. Cesar, F. Metzger, B. Prabhakaran, Eds. Oct. 2021, pp. 1212–1220, doi: [10.1145/3474085.3481548](https://doi.org/10.1145/3474085.3481548).
- [79] P. Qi, Y. Zhao, Y. Shen, W. Ji, J. Cao, and T. Chua, "Two heads are better than one: Improving fake news video detection by correlating with neighbors," in *Proc. Findings Assoc. Comput. Linguistics*, A. Rogers, J. L. Boyd-Graber, and N. Okazaki, Eds., Toronto, ON, Canada, 2023, pp. 11947–11959, doi: [10.18653/v1/2023.findings-acl.756](https://doi.org/10.18653/v1/2023.findings-acl.756).
- [80] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *Proc. 38th Int. Conf. Mach. Learn.*, M. Meila and T. Zhang, Eds. Jul. 2021, pp. 8748–8763. [Online]. Available: <http://proceedings.mlr.press/v139/radford21a.html>
- [81] R. I. Rafiq, H. Hosseinmardi, R. Han, Q. Lv, S. Mishra, and S. A. Mattson, "Careful what you share in six seconds: Detecting cyberbullying instances in vine," in *Proc. IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, J. Pei, F. Silvestri, J. Tang, Eds., Paris, France, Aug. 2015, pp. 617–622, doi: [10.1145/2808797.2809381](https://doi.org/10.1145/2808797.2809381).
- [82] M. A. Raza and K. M. Malik, "Multimodaltrace: Deepfake detection using audiovisual representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Vancouver, BC, Canada, Jun. 2023, pp. 993–1000, doi: [10.1109/cvprw59228.2023.00106](https://doi.org/10.1109/cvprw59228.2023.00106).
- [83] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "FaceForensics++: Learning to detect manipulated facial images," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Seoul, Korea (South), Oct. 2019, pp. 1–11, doi: [10.1109/ICCV.2019.00009](https://doi.org/10.1109/ICCV.2019.00009).
- [84] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, vol. 1, D. E. Rumelhart and J. L. McClelland, Eds. Cambridge, MA, USA: MIT Press, 1987, pp. 318–362.
- [85] S. Sadiq, A. Mehmood, S. Ullah, M. Ahmad, G. S. Choi, and B.-W. On, "Aggression detection through deep neural model on Twitter," *Future Gener. Comput. Syst.*, vol. 114, pp. 120–129, Jan. 2021, doi: [10.1016/j.future.2020.07.050](https://doi.org/10.1016/j.future.2020.07.050).
- [86] D. Salvi, H. Liu, S. Mandelli, P. Bestagini, W. Zhou, W. Zhang, and S. Tubaro, "A robust approach to multimodal deepfake detection," *J. Imag.*, vol. 9, no. 6, p. 122, Jun. 2023, doi: [10.3390/jimaging9060122](https://doi.org/10.3390/jimaging9060122).
- [87] C. Sanderson and B. C. Lovell, "Multi-region probabilistic histograms for robust and scalable identity inference," in *Proc. 3rd Int. Conf.*, M. Tistarelli and M. S. Nixon, Eds., Alghero, Italy, Berlin, Germany: Springer, 2009, pp. 199–208, doi: [10.1007/978-3-642-01793-3\\_21](https://doi.org/10.1007/978-3-642-01793-3_21).
- [88] A. Sankaranarayanan, M. Groh, R. Picard, and A. Lippman, "The presidential deepfakes dataset," in *Proc. 1st Workshop Adverse Impacts Collateral Effects Artif. Intell. Technol.*, Montreal, CA, USA, 2021, pp. 1–16. [Online]. Available: <https://ceur-ws.org/Vol-2942/paper3.pdf>
- [89] I. K. Sastrawan, I. P. A. Bayupati, and D. M. S. Arsa, "Detection of fake news using deep learning CNN–RNN based methods," *ICT Exp.*, vol. 8, no. 3, pp. 396–408, Sep. 2022, doi: [10.1016/j.icte.2021.10.003](https://doi.org/10.1016/j.icte.2021.10.003).
- [90] L. Shang, Z. Kou, Y. Zhang, and D. Wang, "A multimodal misinformation detector for COVID-19 short videos on tiktok," in *Proc. IEEE Int. Conf. Big Data*, Y. Chen, H. Ludwig, Y. Tu, U. M. Fayyad, X. Zhu, X. Hu, S. Byna, X. Liu, J. Zhang, S. Pan, V. Papalexakis, J. Wang, A. Cuzzocrea, and C. Ordonez, Eds., Orlando, FL, USA, Dec. 2021, pp. 899–908, doi: [10.1109/BigData52589.2021.9671928](https://doi.org/10.1109/BigData52589.2021.9671928).
- [91] C. Sharma, D. Bhageria, W. Scott, S. Pykl, A. Das, T. Chakraborty, V. Pulabaigari, and B. Gamback, "SemEval-2020 task 8: Memonition analysis—The visuo-lingual metaphor!" in *Proc. Int. Committee Comput. Linguistics*, A. Herbelot, X. Zhu, A. Palmer, N. Schneider, J. May, E. Shutova, Eds. Dec. 2020, pp. 759–773, doi: [10.18653/v1/2020.semeval-1.99](https://doi.org/10.18653/v1/2020.semeval-1.99).
- [92] K. Shu, L. Cui, S. Wang, D. Lee, and H. Liu, "DEFEND: Explainable fake news detection," in *Proc. 25th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, A. Teredesai, V. Kumar, Y. Li, R. Rosales, E. Terzi, G. Karypis, Eds., Anchorage, AK, USA, Jul. 2019, pp. 395–405, doi: [10.1145/3292500.3330935](https://doi.org/10.1145/3292500.3330935).
- [93] K. Shu, D. Mahudeswaran, S. Wang, D. Lee, and H. Liu, "FakeNewsNet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media," *Big Data*, vol. 8, no. 3, pp. 171–188, Jun. 2020, doi: [10.1089/big.2020.0062](https://doi.org/10.1089/big.2020.0062).
- [94] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. 3rd Int. Conf. Learn. Represent.*, Y. Bengio and Y. LeCun, Eds., San Diego, CA, USA, 2015, pp. 1–15.
- [95] B. Singh and D. K. Sharma, "Predicting image credibility in fake news over social media using multi-modal approach," *Neural Comput. Appl.*, vol. 34, no. 24, pp. 21503–21517, Dec. 2022, doi: [10.1007/s00521-021-06086-4](https://doi.org/10.1007/s00521-021-06086-4).
- [96] S. Singhal, R. R. Shah, T. Chakraborty, P. Kumaraguru, and S. Satoh, "SpotFake: A multi-modal framework for fake news detection," in *Proc. IEEE 5th Int. Conf. Multimedia Big Data (BigMM)*, Singapore, Sep. 2019, pp. 39–47, doi: [10.1109/BIGMM.2019.00-44](https://doi.org/10.1109/BIGMM.2019.00-44).
- [97] D. Soni and V. K. Singh, "See no evil, hear no evil: Audio-visual-textual cyberbullying detection," in *Proc. ACM Human-Comput. Interact.*, vol. 2, pp. 1–26, doi: [10.1145/3274433](https://doi.org/10.1145/3274433).

- [98] N. Subramani and D. Rao, "Learning efficient representations for fake speech detection," in *Proc. 34th AAAI Conf. Artif. Intell. (AAAI), 32nd Innov. Appl. Artif. Intell. Conf. (IAAI), 10th AAAI Symp. Educ. Adv. Artif. Intell. (EAAI)*, New York, NY, USA, Feb. 2020, pp. 5859–5866. [Online]. Available: <https://ojs.aaai.org/index.php/AAAI/article/view/6044>.
- [99] S. Suryawanshi, B. R. Chakravarthi, M. Arcan, and P. Buitelaar, "Multimodal meme dataset (MultiOFF) for identifying offensive content in image and text," in *Proc. 2nd Workshop Trolling, Aggression Cyberbullying, Publisher Eur. Lang. Resour. Assoc.*, Marseille, France, 2020, pp. 32–41. [Online]. Available: <https://aclanthology.org/2020.trac-1.6>
- [100] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. 36th Int. Conf. Mach. Learn.*, vol. 97, K. Chaudhuri and R. Salakhutdinov, Eds., Long Beach, CA, USA, Jun. 2019, pp. 9–15. [Online]. Available: <http://proceedings.mlr.press/v97/tan19a.html>
- [101] R. Wang, F. Juefei-Xu, Y. Huang, Q. Guo, X. Xie, L. Ma, and Y. Liu, "DeepSonar: Towards effective and robust detection of AI-synthesized fake voices," in *Proc. 28th ACM Int. Conf. Multimedia*, C. W. Chen, R. Cucchiara, X. Hua, G. Qi, E. Ricci, Z. Zhang, and R. Zimmermann, Eds., Seattle, WA, USA, Oct. 2020, pp. 1207–1216, doi: [10.1145/3394171.3413716](https://doi.org/10.1145/3394171.3413716).
- [102] Y. Wang, X. Jin, and X. Tan, "Pornographic image recognition by strongly-supervised deep multiple instance learning," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Phoenix, AZ, USA, Sep. 2016, pp. 4418–4422, doi: [10.1109/ICIP.2016.7533195](https://doi.org/10.1109/ICIP.2016.7533195).
- [103] Y. Wang, F. Ma, Z. Jin, Y. Yuan, G. Xun, K. Jha, L. Su, and J. Gao, "EANN: Event adversarial neural networks for multi-modal fake news detection," in *Proc. 24th ACM SIGKDD Int. Conf. Knowledge Discovery Data Mining*, Y. Guo and F. Farooq, Eds., London, U.K., Aug. 2018, pp. 849–857, doi: [10.1145/3219819.3219903](https://doi.org/10.1145/3219819.3219903).
- [104] R. L. P. C. Wijethunga, D. M. K. Matheesha, A. A. Noman, K. H. T. De Silva, M. Tissera, and L. Rupasinghe, "Deepfake audio detection: A deep learning based solution for group conversations," in *Proc. 2nd Int. Conf. Advancements Comput. (ICAC)*, vol. 1, Dec. 2020, pp. 192–197.
- [105] Y. Wu, P. Zhan, Y. Zhang, L. Wang, and Z. Xu, "Multimodal fusion with co-attention networks for fake news detection," in *Proc. Findings Assoc. Comput. Linguistics*, C. Zong, F. Xia, W. Li, and R. Navigli, Eds. 2021, pp. 2560–2569, doi: [10.18653/v1/2021.findings-acl.226](https://doi.org/10.18653/v1/2021.findings-acl.226).
- [106] J. Xue, Y. Wang, Y. Tian, Y. Li, L. Shi, and L. Wei, "Detecting fake news by exploring the consistency of multimodal data," *Inf. Process. Manag.*, vol. 58, no. 5, Sep. 2021, Art. no. 102610, doi: [10.1016/j.ipm.2021.102610](https://doi.org/10.1016/j.ipm.2021.102610).
- [107] C. Yang, F. Zhu, G. Liu, J. Han, and S. Hu, "Multimodal hate speech detection via cross-domain knowledge transfer," in *Proc. 30th ACM Int. Conf. Multimedia*, J. Magalhaes, A. D. Bimbo, S. Satoh, N. Sebe, X. Alameda-Pineda, Q. Jin, V. Oria, and L. Toni, Eds., Lisboa, Portugal, 2022, pp. 4505–4514, doi: [10.1145/3503161.3548255](https://doi.org/10.1145/3503161.3548255).
- [108] F. Yang, X. Peng, G. Ghosh, R. Shilon, H. Ma, E. Moore, and G. Predovic, "Exploring deep multimodal fusion of text and photo for hate speech classification," in *Proc. 3rd Workshop Abusive Lang. Online*, Florence, Italy, 2019, pp. 11–18. [Online]. Available: <https://aclanthology.org/W19-3502>
- [109] Y. Yang, L. Zheng, J. Zhang, Q. Cui, Z. Li, and P. S. Yu, "TI-CNN: Convolutional neural networks for fake news detection," 2018, *arXiv:1806.00749*.
- [110] S. Yuan, A. Maronikolakis, and H. Schütze, "Separating hate speech and offensive language classes via adversarial debiasing," in *Proc. 6th Workshop Online Abuse Harms (WOAH)*, 2022, pp. 1–10. [Online]. Available: <https://aclanthology.org/2022.woah-1.1>
- [111] Z. Zhang and L. Luo, "Hate speech detection: A solved problem? The challenging case of long tail on Twitter," *Semantic Web*, vol. 10, no. 5, pp. 925–945, Sep. 2019, doi: [10.3233/sw-180338](https://doi.org/10.3233/sw-180338).
- [112] Q. Zhong, Q. Wang, and J. Liu, "Combining knowledge and multi-modal fusion for meme classification," in *Proc. Conf. MMM*, Phu Quoc, Vietnam, Berlin, Germany: Springer, 2022, pp. 599–611, doi: [10.1007/978-3-030-98358-1\\_47](https://doi.org/10.1007/978-3-030-98358-1_47).
- [113] X. Zhou, A. Mulay, E. Ferrara, and R. Zafarani, "ReCOVeRY: A multimodal repository for COVID-19 news credibility research," in *Proc. 29th ACM Int. Conf. Inf. Knowl. Manag.*, M. d' Aquin, S. Dietze, C. Hauff, E. Curry, and P. Cudre-Mauroux, Eds. Oct. 2020, pp. 3205–3212, doi: [10.1145/3340531.3412880](https://doi.org/10.1145/3340531.3412880).
- [114] X. Zhou, J. Wu, and R. Zafarani, "SAFE: Similarity-aware multi-modal fake news detection," in *Proc. 24th Pacific-Asia Conf.*, H. W. Lauw, R. C. Wong, A. Ntoulas, E. Lim, S. Ng, and S. J. Pan, Eds., Singapore. Berlin, Germany: Springer, 2020, pp. 354–367, doi: [10.1007/978-3-030-47436-2\\_27](https://doi.org/10.1007/978-3-030-47436-2_27).
- [115] Y. Zhou and S.-N. Lim, "Joint audio-visual deepfake detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Montreal, QC, Canada, Oct. 2021, pp. 14780–14789, doi: [10.1109/ICCV48922.2021.01453](https://doi.org/10.1109/ICCV48922.2021.01453).
- [116] Y. Zhou, Y. Yang, Q. Ying, Z. Qian, and X. Zhang, "Multi-modal fake news detection on social media via multi-grained information fusion," in *Proc. ACM Int. Conf. Multimedia Retr.*, I. Kompatsiaris, J. Luo, N. Sebe, A. Yao, V. Mazaris, S. Papadopoulos, A. Popescu, and Z. H. Huang, Eds., Thessaloniki, Greece, 2023, pp. 343–352, doi: [10.1145/3591106.3592271](https://doi.org/10.1145/3591106.3592271).
- [117] Y. Zhu, J. Gao, and X. Zhou, "AVForensics: Audio-driven deepfake video detection with masking strategy in self-supervision," in *Proc. ACM Int. Conf. Multimedia Retr.*, I. Kompatsiaris, J. Luo, N. Sebe, A. Yao, V. Mazaris, S. Papadopoulos, A. Popescu, and Z. H. Huang, Eds., Thessaloniki, Greece, 2023, pp. 162–171, doi: [10.1145/3591106.3592218](https://doi.org/10.1145/3591106.3592218).



**ENIAFE FESTUS AYETIRAN** is a Senior Lecturer in Computer Science and has over a decade of university teaching and research experience. He obtained a Bachelor of Science degree in Computer Science (2006) awarded by Adekunle Ajasin University, a Master of science degree in computer science (2011) awarded by University of Ibadan. In 2017, under the auspices of the European Commission's Erasmus Mundus (Erasmus+) doctoral fellowship, he obtained a Ph.D degree in computer science awarded by University of Luxembourg and another Ph.D in an interdisciplinary field of legal informatics jointly awarded by University of Bologna, University of Turin, Autonomous University of Barcelona, Tilburg University and Mykolas Romeris University. As part of his doctoral studies, at various times, he was at the Universities of Bologna and Turin, Autonomous University of Barcelona and University of Luxembourg. He was a postdoctoral fellow at the Computer Science Department of Norwegian University of Science and Technology. He was also a postdoctoral fellow at the Faculty of Informatics of Masaryk University, Brno in Czech Republic. He is passionate about quality teaching and cutting-edge research. His research interests include natural language processing, machine learning, and computational social science including fakes news and hate speech detection. He is a Co-Organizer of Workshop on Advances in Disinformation Detection (WADD 2023). He is a Reviewer for several journal including but not limited to Knowledge-Based Systems, Neurocomputing Journal, Neural Networks, PeerJ Computer Science, Applied Sciences, Engineering Application of Artificial Intelligence, Neural Networks, and Scientific African.



**ÖZLEM ÖZGÖBEK** is currently an Associate Professor with the Department of Computer Science, Norwegian University of Science and Technology. She is involved in many academic organizing committees including: News Recommendation and Analytics (INRA) Workshop Series, Women In RecSys, NewsImages Challenge, Norwegian Big Data Symposium (NOBIDS). She is the Project Coordinator of International Work-Integrated-Learning in Artificial Intelligence (IWIL AI) Project. Her research interests include recommender systems, privacy issues in recommender systems, and disinformation detection for online news.