Journal of Information & Optimization Sciences
ISSN 0252-2667 (Print), ISSN 2169-0103 (Online)
Vol. 44 (2023), No. 3, pp. 301–314
DOI : 10.47974/JIOS-1342

FakeExpose: Uncovering the falsity of news by targeting the multimodality via transfer learning

Sakshi Kalra *
Chitneedi Hemanth Sai Kumar
Yashvardhan Sharma
Department of Computer Science & Information Systems
BITS Pilani
Pilani Campus
Pilani 333031
Rajasthan
India

Gajendra Singh Chauhan
Department of Humanities and Social Sciences
BITS Pilani
Pilani Campus
Pilani 333031
Rajasthan
India

Abstract

Social media for news utilization has its own pros and cons. There are several reasons why people look for and read news through internet media. On the one hand, it is easier to access, and on the other, social media's dynamic content and misinformation pose serious problems for both government and public institutions. Several studies have been conducted in the past to classify online reviews and their textual content. The current paper suggests a multimodal strategy for the (FND) task that covers both text and image. The suggested model (FakeExpose) is created to automatically learn a variety of discriminative features, instead of relying on manually created features. Several pre-trained words and image embedding models, such as DistilRoBERTa and Vision Transformers (ViTs) are used and fine-tined for the best feature extraction and the various word dependencies. Data augmentation is used to address the issue of pre-trained textual feature extractors not processing a maximum of 512 tokens at a time. The accuracy of the presented model on PolitiFact and GossipCop is 91.35 percent and 98.59 percent, respectively, based on current standards. According to our knowledge, this is the first attempt to use the FakeNewsNet

* *E-mail:* p20180437@pilani.bits-pilani.ac.in

repository to reach the maximum multimodal accuracy. The results show that combining text and image data improves accuracy when compared to utilizing only text or images (Unimodal). Moreover, the outcomes imply that adding more data has improved the model's accuracy rather than degraded it.

Subject Classification: *68T50.*

Keywords: *FND, Transfer learning, Deep learning (DL), Social media analytics, Multimodal, Fact-checking.*

1. Fake News Problem

Social media is a powerful tool for spreading news because of its low cost, quick dispersal, and ease of connecting to the whole world from one location [1]. The role of traditional media consisting of newspapers and televisions on how we acquire and take news has turned out to be insignificant than within the past. In fact, the rise of social networking structures has played a significant part in this change. For example, as of January 2020, 3.80 billion individuals around the world use social media. The proportion of consumption has increased significantly by 9% in the past year. Smart devices are currently being used by more than 5.19 billion people worldwide. Fake news has an ability to put immensely negative effects on people and the community because of its widespread.

1.1 *Fake News Classification*

Classifying fake news involves figuring out how true a piece of news is. Fact-checking is one strategy for preventing the spread of false information; expert-based fact-checking has high accuracy but is not scalable; the other strategy is crowd sourced fact-checking, which has a high possibility of losing accuracy but scales well. Thus, manual FND is no longer necessary, making way for automatic FND [3]. Various fact-checking websites are available for verifying the online reviews. Figure 1. illustrate some examples of claims fact-checked by Politifact.com.

Figure 1

Claim fact-checked by PolitiFact.com

1.2 *Fake News Detection Challenge*

Previous research had focused on using manual features or unimodal DL features. The issue is that it completely disregards the interdependency of multi-modal material in tweets. Tweets with visual content, such as videos and graphics, may attract user's consideration higher than tweets with textual content. As a result, some research [6], [7], [8], [9] focuses specifically on the multimodal content of tweets for FND. In earlier studies, several ML and DL-based techniques have been used for the FND task. However, they were quite effective but unable to succeed due to a lack of contextual information in the text and a failure to capture several visual features. To deal with the problems listed above, we proposed a multi-modal fusion architecture (FakeExpose) which fuses text and image content in tweets and provide a combined model of FND.

The rest of this work is set up like this: Section 2 talks about the literature review, and Section 3 talks about the multi-modal dataset used in this study. In section 4, we talk about the proposed model architecture and its details. In Section 5, the details of the experiment are explained, and in Section 6, the work is summed up.

2. Related Work

FND aims to figure out whether or not a piece of information is true [10], [11]. This problem is usually referred to as binary text classification since the majority of the literature defined it.

2.1 *Fake News Detection based on Text Analysis*

Traditional ML-based methods for detecting fake news perform effectively, but DL-based methods have great advantages over ML-based methods. In [14], RNN-based systems can automatically determine the semantic content and temporal relationship of words in sentences. Despite this, they are unable to learn phrases that include a large number of words. To resolve the issue of long-range dependency, Chen et. al. employed a self-attention-based architecture [15]. There have been few architectures based on Generative Adversarial Networks (GANs) [16] for the FND task. The concept of feature extraction via transfer learning using transformer-based models has recently been explored in this problem [17].

2.2 *Fake News Detection based on Multimodal Analysis*

Multimodal FND includes images and it is differing from the textual based FND. Few researchers have combined text and image features to figure out the credibility of news [20]. Nakamura et. al. [22] suggested a novel multimodal architecture using both text and image features. Pre-trained models such as InferSent and BERT were employed for extracting the text features. VGG16, ResNet 50, and EfficentNet were employed for extracting the visual features, with ResNet50 being the most favorable, followed by VGG16, and then EfficientNet. The authors of [3] introduced a multi-modal Event Adversarial Neural Network (EANN) to evaluate the textual and graphical content of tweets and construct event-invariant feature representations using the adversarial network. In [4] Singhal et. al. suggested a novel multimodal architecture that has been created using the cutting-edge pre-trained models on text and images, named XLNet and VGG19 and achieved the noticeable accuracy on the FakeNewsNet repository. Table 1 gives the comparative analysis of various multimodal architectures trained on the FakeNewsNet repository.

Table 1

Comparative Analysis of various multimodal architectures trained on FakeNewsNet Repository

Models	Modality	PolitiFact	GossipCop
EANN [3]	Multimodal	0.74	0.86
MVAE [6]	Multimodal	0.673	0.775
SpotFake [5]	Multimodal	0.721	0.807
SpotFake+ [4]	Multimodal	0.846	0.856

3. Dataset Used

The FakeNewsNet dataset is used in our work as it contains large texts, a wide range of categories, and binary-class label data from the PolitiFact.com and GossipCop.com. This dataset performs better than several pre-existing datasets in terms of the variety of information it includes. As far as we are aware of, the FakeNewsNet dataset contains news content features that includes linguistic and image features, social context features such as user and network related features, and spatial-Temporal features that can help any researcher in quickly complete the FND task. The statistics related to the data repository are employed in Table 2.

Table 2

Dataset Statistics

Dataset	Size	No. of Classes	Modality	Source	Data Category
FakeNewsNet (PolitiFact) (GossipCop)	**485** **12840**	**2**	**TEXT+IMAGE**	**Multisource**	**Variety**

4. Our Architecture

This study proposed data augmentation techniques and various architectures for Unimodal classification and Multimodal classification by fusing both text and image features in the following sub-sections.

4.1 *Data Augmentation*

An exploratory data analysis (EDA) has performed to get useful information about the dataset. It was found that a few logo-based small images are included in it and could not be used for the image feature extraction. Also, the number of words per record is relatively high, pre-trained word embedding models such as various transformer-based models can only process 512 tokens at a time. The effective preprocessing of the data is performed to address these two issues as shown in Figure 2 (Left) and for the data augmentation in order to get the better accuracy. The records of the data repository are distributed into two categories depending on the kind of image (images with logos and regular images). One idea is to skip the logo images; however, this will result in a loss of information because these images contain some text in the form of sentences related to it. So np.zeros ((224,224,3)) function is used to replace logo images with blank images programmatically in order to preserve the text. For regular images, sentences longer than 500 words are again divided into two sentences, one with the words ranging from 0-500 and other with from 250-750, as the pre-trained models can only process 512 tokens at a time. Sentence 1 contains the original image of the record and sentence 2 contains the 45° left rotated image. As seen in Figure 2 (Right), the class label for both sentences will be the same (as the original).

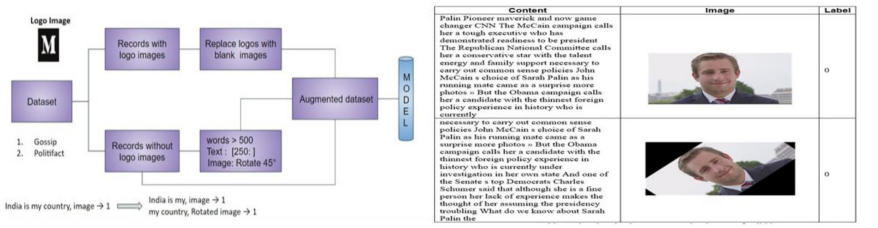

Figure 2

Data Preprocessing (Left) and Data Augmentation in case of a Regular Image (Right)

4.2 *Motivations for Transfer Learning*

A variety of neural network based techniques is used to process natural languages. These techniques should be able to learn word representations and long-range word dependencies during the feature extraction phase. RNNs are feed-forward neural networks that have been ruled out over time and are used to handle sequential data with a defined ordering, but it has some disadvantages: RNNs are so slow that they are trained using a simplified version of backpropagation, which is extremely time-consuming in terms of hardware resources and these are not able to handle long sequences. The transformer-based models, which use encoder-decoder architecture, are released in 2017. These models are created in a manner that the input sequence can be transmitted parallelly which makes these models different from the RNNs. Transfer learning mimics the human brain's behavior. It's a deep learning (and machine learning) based process of passing information from one model to the next. With transfer learning, we can use all or a portion of a previously learned model to tackle new problems.

4.3 *Motivations Related to DistilRoBERTa for Text Classification*

The DistilRoBERTa architecture is derived from the Roberta base model after distillation. The benefit of distillation is that it reduces the time required for training and running the models; for example, on a specific task, the time spent on training is 15 minutes and 59 seconds using distilled models but the time is nearly doubled i.e. 29 minutes and 59 seconds when a non-distilled model is used. Figure 3 depicts a contrast between regular models and distil models. The reduced time is due to the exclusion of extraneous parameters that are not used in the model's inference phase.

4.4 *Motivations Related to Vision Transformers for Image Classification*

Neural networks have improved their ability to interpret natural languages with time. These networks can create fixed or variable-length vector-space representations and also combine input from close words to determine meaning in a given context. Up until now, (CNNs) have been the de facto model for processing image data. It has a few drawbacks that motivate other researchers to come up with better solutions. CNNs have limitations in their ability to capture pixel-level dependencies across the spatial and temporal domains due to their operation on a fixed-sized window. Also, after the training process, the filter weights in CNNs remain fixed, stopping the process from dynamically adjusting to input changes. Vision Transformers (ViT's) [30] can be the solution to overcome the shortcomings of CNNs. The internal model architectures of ViT's and CNN's are compared to determine the major differences between them: If we examine how local and global spatial information is used, we find that ViT's use more global information at lower levels than CNNs, resulting in giving better features. ViT's have more uniform representations and result in a higher similarity between representations obtained in shallow and deep layers than CNNs.

ViT's preserve more spatial information when compared to CNNs and may take advantage of a big amount of data to learn accurate intermediate representations. ViT's can be used to overcome this problem in which the entire image is divided into numerous equal-sized patches, and encoded

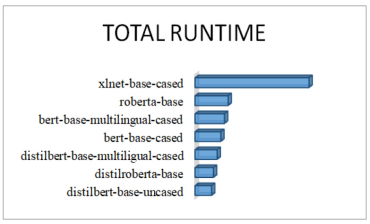

Figure 3

Comparison of Runtime for Distilled models vs. Regular models

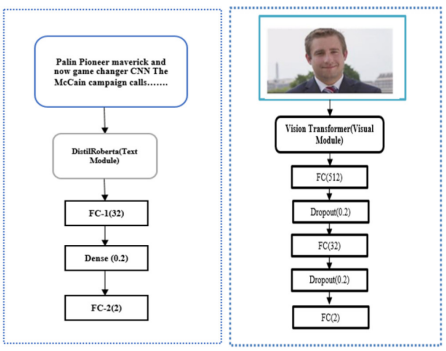

Figure 4

DistilRoBERTa (Left) and Vision Transformers (ViT's) (Right)

similarly to word embedding's. However, unlike the BERT architecture, the values are the learnable parameters in this case. The classified output is obtained after passing these values through self-attention layers and neural networks.

4.5 *Hyperparameters Tuning*

In statistical terms, Hyperparameters are those parameters that are derived from a prior distribution. In a significant way, these parameters contribute in the efficient training of any ML and DL-based algorithms and have immediate control over the behavior of the training algorithm. Thus these parameters have a big influence on the model's performance. Table 3 demonstrates the number of Hyperparameters used for the proposed multimodal architecture.

Table 3

Hyperparameters Statistics

Hyperparameter	Description or value
Dense Layers	3
Dropout rate	0.2
Loss function	sparse_categorical_crossentropy
Optimizer	Adam, Adagrad, SGD
Activation function	Relu
Learning Rate	6e-6, tried with many
Beta-1	0.9
Beta-2	0.99
Epochs	10
Batch size	8

4.6 *Model Construction*

The first step in the model generation is the collection of data. The model input is a multi-modal tweet (Tw = T, I), with T and I refers the tweet's text and visual content. The model output is the FND label, which is either a Real (R) or a Fake (F). The proposed model has three components: a text module, a visual module, and a multimodal fusion module.

4.6.1 Text Module

The text module can be expressed using equation (1), in which F_T (Intermediate Features-2) as shown in Figure 5 represents the text

module's output (text feature representation) and θ_T represents the text module's parameters. The complete architecture for text classification using DistilRoBERTa is shown in Figure 4 (Left).

$$F_T = f_T(T; \theta_T), \tag{1}$$

4.6.2 Visual Module

The visual module can be expressed using equation (2), in which F_I (Intermediate Features-1) as shown in Figure 6 represents the visual module's output (image feature representation) and θ_I represents the visual module's parameters. The complete architecture for image classification using Vision Transformers (ViT's) is shown in Figure 4 (Right).

$$F_I = f_i(I; \theta_I), \tag{2}$$

4.6.3 Multimodal Fusion Module

The multimodal fusion module takes two inputs: a previously extracted image feature representation F_I (Intermediate Features-1) and a text feature representation F_T (Intermediate Features-2) as shown in Figure 5. The multimodal fusion module can be expressed using equation (3), with the Result representing the output of the multimodal module.

Result: FC3(FC2(F(FC1(Intermediate Feature-1, Intermediate Feature-2))) (3)

Intermediate Feature-1 represents the image feature vector from Vision Transformers (ViT's), Intermediate Feature-2 represents the text feature vector from DistilRoBERTa, FC-1 denotes the Fully Connected Layer-1 refers to the dense layer that helps in changing the dimension to 768, FC-2 denotes the Fully Connected Layer-2 refers to the dense layer that helps change the dimension to 32, FC-3 denotes the Fully Connected Layer-3 refers to the softmax layer that helps in giving the final categorical classification (either Fake or Real) and F is a function that takes the output of FC-1 and fuses it with FT (Intermediate Feature-2)).

5. Experimental Analysis

The main challenge of generating multimodal architectures is to maintain the features related to different modalities while combining relevant features from different modalities.

5.1 *Unimodal (Text based Results)*

The textual-based model construction starts by taking the tokens as input and feeding this input to the pre-trained DistilRoBERTa. Textual useful features are retrieved and fed into a dense layer. Then the logits are computed using the output from the last dense layer, which has the same number of neurons and activation functions as the number of classes. After feeding these logits into the softmax, we get the probabilities. We train the model using the Adam optimizer and a sparse categorical cross-entropy loss, with a learning rate of 0.000006 per iteration. In this proposed work, four different textual based data architectures are tested: DistilRoBERTa, Distill BERT, XLNet, and Distil GPT2, and found that the DistilRoBERTa outperforms all these architectures. Table 4 shows the comparative analysis of four-different text based architectures using the accuracy as an evaluation parameter.

Table 4

Comparative Analysis of four-different Text based Architectures

Modality	Models	PolitiFact	GossipCop
Unimodal (TEXT)	XLNet+Dense	0.74	0.836
	DistilRoBERTa	**0.8942**	**0.8862**
	Distil BERT	08075	0.8576
	Distil GPT2	0.7596	0.8198

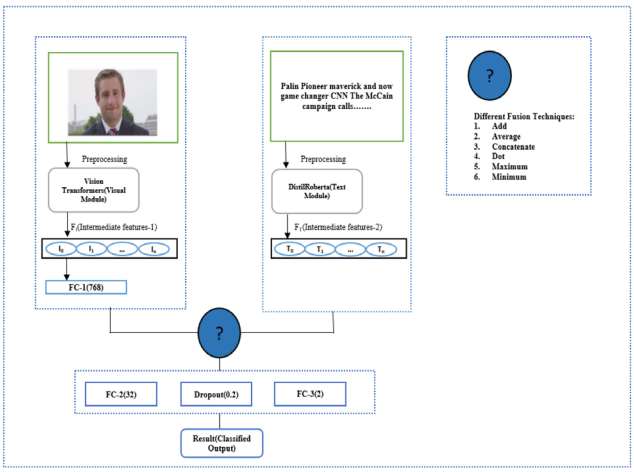

Figure 5

Multimodal Fusion Module

5.2 *Unimodal (Image based Results)*

Three different image based architectures are used for the image feature extraction: VGG19, VGG 16, and Vision Transformers (ViT's), and found that ViT's outperforms all these architectures. Three distinct image-based designs are compared and contrasted in Table 5.

Table 5

Comparison of three distinct image-based architectures

Modality	Models	PolitiFact	GossipCop
Unimodal (IMAGE)	VGG19	0.654	0.80
	VGG16	0.6932	0.8091
	ViTs	**0.7614**	**0.8188**

5.3 *Multimodal (Text + Image based Results)*

Multiple text and image-based architectures are employed for multimodal FND task, and the concatenation of DistilRoBERTa + ViT's provided the best accuracy out of all these multimodal architectures. Figure provides the comparative analysis of the results of proposed multimodal architecture (FakeExpose) with the current cutting-edge multimodal architectures trained on the FakeNewsNet dataset.

Figure 6 gives the comparative analysis of (FakeExpose) with different pre-existing multimodal architectures in the form of a Bar chart.

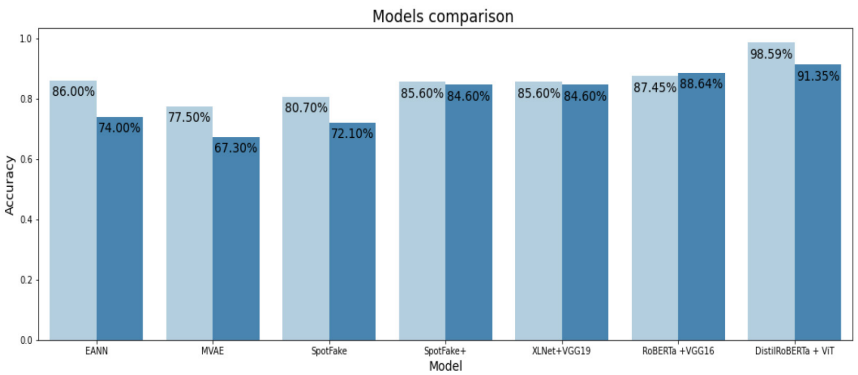

Figure 6

Comparative Analysis of Various Pre-Existing Multimodal Architectures with (FakeExpose)

5.4 *Various Multimodal Fusion Methods*

Various fusion methods such as concatenating and taking maximum of both the features are tried for the fusion of the extracted features from both the modalities. Concatenate works better on the proposed multimodal architecture. Table 7 listed the comparison of the outcomes of different fusion methods.

Table 6

Comparison of the Outcomes for Different Fusion Methods

Fusion/ Dataset	Add	Average	Concatenate	Dot	Maximum	Minimum
GossipCop	0.9718	0.9789	**0.9859**	**0.9859**	0.9648	0.9789
PolitiFact	**0.9135**	0.9038	**0.9135**	0.9038	0.9038	0.9038

6. Conclusions and Future Work

In this paper, a multimodal architecture (FakeExpose) is suggested for the FND task. Numerous text and image-based architectures are employed in which DistilRoBERTa and Vision Transformers (ViT's) is giving the best results for text and image classification. The idea of augmenting the data helped in improving the accuracy rather than degrading it. Several Transformer-based architectures are applied to the same dataset by other researchers, each with its own set of pros and cons. The findings indicate that the suggested approach offers the best multimodal accuracy of 91.35% on PolitiFact and 98.59% on GossipCop. The outcomes reveal that multimodal produces better results than unimodal. Future work in this field would be to make models more robust to new data with multiple classes. It is also possible to employ echo chambers to uncover fake news. The use of social context features can also be an important area of research.

References

[1] K. Shu, L. Cui, S. Wang, D. Lee, and H. Liu, "dEFEND: Explainable Fake News Detection," Proc. 25th ACM SIGKDD Int. Conf. Knowl. Discov. data Min., pp. 395-405 (2019), doi: 10.1145/3292500.3330935.

[2] X. Zhou, R. Zafarani, K. Shu, and H. Liu, "Fake News: Fun-damental Theories, Detection Strategies and Challenges," vol. 19 (2019), doi: 10.1145/3289600.3291382.

[3] Y. Wang et. al., "EANN: Event adversarial neural networks for multi-modal fake news detection," Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min., pp. 849-857 (Jul. 2018), doi: 10.1145/3219819.3219903.

[4] S. Singhal, A. Kabra, M. Sharma, R. R. Shah, T. Chakraborty, and P. Kumaraguru, "SpotFake+: A multimodal framework for fake news detection via transfer learning (student abstract)," AAAI 2020 - 34th AAAI Conf. Artif. Intell., pp. 13915-13916 (2020), doi: 10.1609/aaai.v34i10.7230.

[5] S. Singhal, R. R. Shah, T. Chakraborty, P. Kumaraguru, and S. Satoh, "SpotFake: A multi-modal framework for fake news detection," Proc. - 2019 IEEE 5th Int. Conf. Multimed. Big Data, BigMM 2019, pp. 39-47 (Sep. 2019), doi: 10.1109/BIGMM.2019.00-44.

[6] D. Khattar, M. Gupta, J. S. Goud, and V. Varma, "MvaE: Multimodal variational autoencoder for fake news detection," in The Web Conference 2019 - Proceedings of the World Wide Web Conference, WWW 2019, pp. 2915-2921 (May 2019). doi: 10.1145/3308558.3313552.

[7] V. R. Suri, B., Taneja, S., Aggarwal, S., & Sharma, "Fake news detection tool (FNDT): Shield against sentimental deception," *J. Inf. Optim. Sci.*, vol. 41, no. 6, pp. 1513-1524 (2020), [Online]. Available: https://doi.org/10.1080/02522667.2020.1802125

[8] S. K. Malhotra, P., & Malik, "Fake news detection using supervised machine learning techniques.," *J. Inf. Optim. Sci.,* vol. 43, no. 1, pp. 7-15 (2022).

[9] X. Ma and E. Hovy, "End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF," 54th Annu. Meet. Assoc. Comput. Linguist. ACL 2016 - Long Pap., vol. 2, pp. 1064-1074 (Mar. 2016), doi: 10.48550/arxiv.1603.01354.

[10] T. Chen, L. Wu, X. Li, J. Zhang, H. Yin, and Y. Wang, "Call AAention to Rumors: Deep AAention Based Recurrent Neural Networks for Early Rumor Detection", Accessed: Sep. 28, 2022. [Online]. Available: www.dailymail.co.uk/news/article-2313652/AP-Twiier-hackers-break-news-

[11] J. MA et. al., "Detecting rumors from microblogs with recurrent neural networks," Proc. 25th Int. Jt. Conf. Artif. Intell. (IJCAI 2016), pp. 3818–3824, Jul. 2016, Accessed: Jul. 29, 2022. [Online]. Available: https://ink.library.smu.edu.sg/sis_research/4630.

[12] H. Jwa, D. Oh, K. Park, J. M. Kang, and H. Lim, "exBAKE: Automatic Fake News Detection Model Based on Bidirectional Encoder Representations from Transformers (BERT)," Appl. Sci. 2019, Vol. 9, Page 4062, vol. 9, no. 19, p. 4062 (Sep. 2019), doi: 10.3390/APP9194062.

[13] R. Rajesh Kumar, E., Rama Rao, K. V. S. N., Nayak, S. R., & Chandra, "Suicidal ideation prediction in twitter data using machine learning techniques.," *J. Interdiscip. Math.*, vol. 23, no. 1, pp. 117-125 (2020), [Online]. Available: https://doi.org/10.1080/09720502.2020.1721674.

[14] K. Nakamura, S. Levy, and W. Y. Wang, "r/Fakeddit: A new multimodal benchmark dataset for fine-grained fake news detection," Lr. 2020 - 12th Int. Conf. Lang. Resour. Eval. Conf. Proc., pp. 6149-6157 (2020).

[15] 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.

[16] M. Raghu, T. Unterthiner, S. Kornblith, C. Zhang, and A. Dosovitskiy, "Do Vision Transformers See Like Convolutional Neural Networks?," Adv. Neural Inf. Process. Syst., vol. 34, pp. 12116-12128 (Dec. 2021).