Detecting Misinformation On Social Media Using Neural Networks

905F3: Wider Topics in Data Science

Disseration (3000 Words) Week 11

Candidate Number: XXXXXX

Word Count: 3298

Contents

Abstract

Misinformation is a major challenge for democratic societies, with social media platforms struggling to manage its volume and impact. Artificial neural networks (ANNs) offer promising solutions for misinformation detection. This essay explores different domains of misinformation — text, images, multimodal, propagation through social networks — highlighting ANNs' versatility, accuracy, and scalability for misinformation detection. Beginning with an introduction to ANNs, the essay then examines each domain, discussing how ANN methods can be adapted to the task and evaluating their effectiveness. Finally, the essay explores the current use of ANNs by social media companies for misinformation detection, the recent trajectory being adopted by major social media platforms, and suggests future directions.

Introduction

One's right to participate in democracy should not depend on making "correct" decisions. Democracy can be defended on various grounds: "consent of the governed," protection from power misuse, equality, liberty, etc. However, *healthy* democracies depend on informed decision-making. Access to clear, verifiable information is essential for effective participation, extending beyond politics into areas like healthcare (e.g., vaccination intent – – Loomba et al., 2021).

Misinformation (misleading information) and disinformation (misinformation spread to intentionally deceive) have harmful effects. A literature review by Adams et al. (2023) highlights increased distrust in media and democracy, polarisation, political disengagement, vaccine hesitancy, use of harmful treatments, climate change denial and political extremism. Disinformation is frequently used by illiberal political actors to spread propaganda (e.g., Russian disinformation in the Baltics - Morkūnas, 2023; worldwide increases in domestic terrorism - Piazza, 2022).

A survey of 1,993 UK adults found 6% reported never encountering misinformation on social media, with 38% encountering it "many times" (Enock et al., 2024). Similarly, 73% of 9,680 US respondents saw inaccurate news about the 2024 election "at least somewhat often" (Shearer et al., 2024). Globally, 87% of 8,000 respondents across 16 countries believed misinformation had a major impact on politics (Quétier-Parent et al., 2023). Vosoughi et al. (2018) found political fake news spread faster than true information on Twitter from 2006-2017. While some respondents may have misjudged or believe misinformation (Lyons et al., 2021), there is political will to address it. A YouGov poll showed 66% of UK adults felt social media firms should be held accountable for the 2024 riots incited by far-right users falsely claiming the Southport stabber was a Muslim asylum seeker (Smith, 2024). A 2023 US survey of 5,115 found 55% supported government restrictions on false information online, up from 39% in 2018 (Liedke, 2023). During COVID-19, 33% of 1,006 UK respondents wanted social media sites to prevent disinformation, while 55% supported government intervention (Open Knowledge Foundation, 2020).

Misinformation is harmful for democratic societies, making detection and interventions crucial in the social-media era. Artificial neural networks (ANNs) provide effective techniques for detecting misinformation. This essay outlines and evaluates these techniques, focusing on how ANN architectures can adapt to various misinformation-detection tasks. It then discusses the current use of ANNs by social media companies for misinformation detection, the recent trajectory being adopted by major social media platforms, and suggests future directions.

Neural Networks: An Overview

Artificial neural networks (ANNs) simplistically model biological neural networks. ANNs consist of layers of nodes (neurons). Each node is connected to every node in the previous and next layers via edges (synapses). **Figure 1** shows a simple ANN.

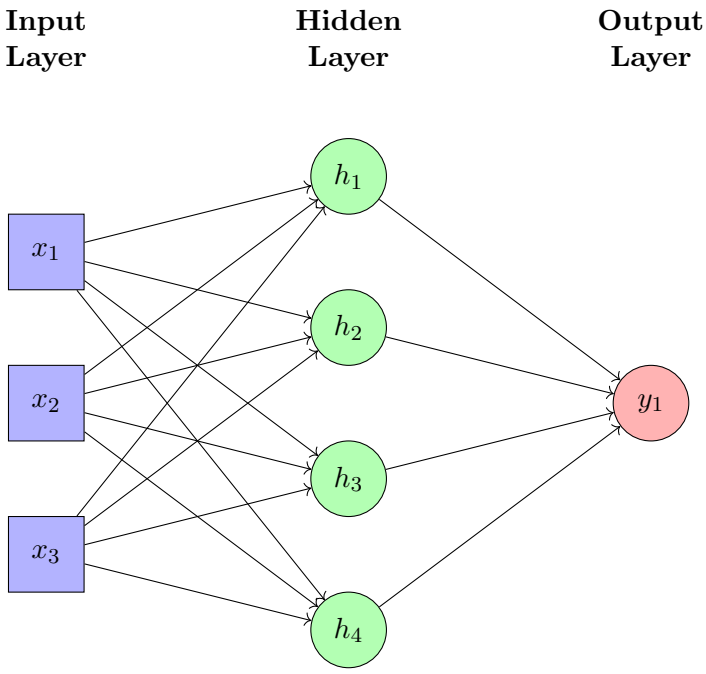

Figure 1: A simple neural network.

The **input layer** represents each numerically-encoded feature in the training data. In a classic example, each x_i might denote the grayscale intensity of a pixel in an image of hand-drawn digits (Deng, 2012). Each x_i is fully connected to every $h_j^{(l)}$ node in the l-th **hidden layer**.

Each connection has a weight $w_{ji}^{(l)}$ and a bias term b_j (initialised randomly), representing the "strength" between nodes $n_i^{(l)}$ and $n_j^{(l+1)}$ in layer l. For node $h_j^{(l)}$, the weighted inputs are summed and passed through an activation function $\phi(x)$. This function introduces non-linearity, allowing the network to approximate more complex functions, improving model accuracy. The **rectified linear unit (ReLU)** $max(0, x)$ is typically used for hidden layers, while softmax is used for the output layer to generate a probability distribution (e.g., predicting digits 0-9). The choice of activation function is task-dependent (Dubey et al., 2022).

The output of each node can be expressed as:

$$a_j = \phi\left(\sum_i w_{ji}x_i + b_j\right)$$

Each input value x_i passes through each layer, activating subsequent nodes until reaching the **output layer**. For example, with hand-drawn digits, the output layer could have 10 nodes, each representing a digit from 0-9, showing the probability of the image being that digit.

Simply passing data through an ANN (forward propagation) doesn't allow it to "learn" – **backpropagation** is essential. This algorithm requires a loss function (e.g., mean squared error, the average squared difference between predictions and true values) and utilises **gradient descent** to minimize it. By evaluating how the loss changes with the activation of the last hidden layer, the weight W_{old} is updated to W_{new} by subtracting the derivative $\frac{\delta C_0}{\delta W^{(L)}}$ (scaled by the **learning rate** η):

$$W_{new} = W_{old} - \eta\left(\frac{\delta C_0}{\delta W^{(L)}}\right)$$

This process is propagated back layer-by-layer to the input layer. Then, a new training sample is propagated forward, and its error is propagated back again, and so on. We stop this process after either: passing through the dataset some number of times (**epochs**), when the error stops updating, or when a desired model accuracy is achieved — again, stopping rules are task-dependent. The resulting ANN will have weights and biases that enable it to make accurate predictions on unseen data.

Next, we'll see how modifications of this basic ANN approach can help detect misinformation content on social media.

Detecting Misinformation

Text

To identify text-based misinformation, we should capture language's properties. Language is sequential, thus word order alters meaning. Word order then must be preserved. Word meaning also depends on context; for instance, "bat" can refer to a baseball bat, a mammal, or as in "I didn't bat an eyelid". Each conveys a different meaning. A basic ANN needs additional mechanisms to model word sequences and context.

Recurrent neural networks (RNNs) adapt ANNs to textual data by using a hidden state h_t at each node, which stores previous outputs and combines them with the next sample to generate a new output. This enables information to carry across iterations. RNNs can be unidirectional (see **Figure 2**) or bidirectional, processing both past and future data points (Schuster and Paliwal, 1997). Although more difficult to train during backpropagation, RNNs apply well to text by sequentially considering both prior and subsequent words. The recurrence relation for the hidden state (unidirectional) is:

$$h_t = \phi(W_h h_{t-1} + W_x x_t + b)$$

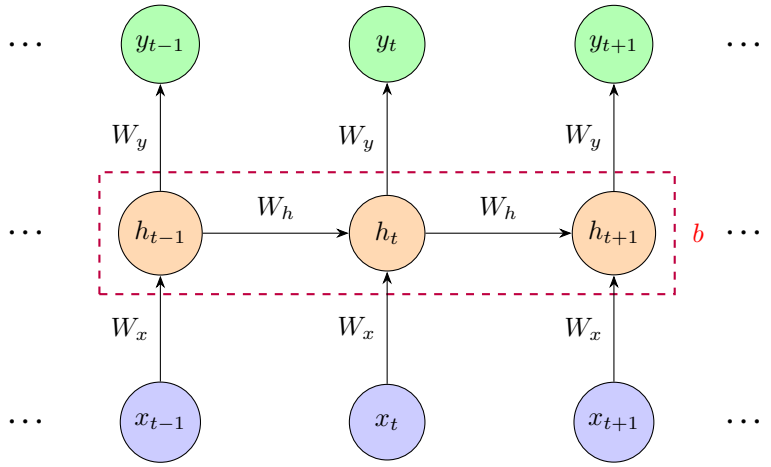

Figure 2: A Simple RNN Node Evolving Over Time

One problem with this approach is that as sequential data grows, the RNN struggles to "remember" earlier data (the vanishing gradient problem). While ReLU can help, enabling RNNs to retain long-term information would improve text processing.

Long short-term memory (LSTM) (Hochreiter and Schmidhuber, 1997) adds **cell states** to each node, which include 3 **gates**: a **"forget" gate** controls how much prior context is "for-

gotten", an **"input" gate** regulates how much new information is stored, and an **"output" gate** determines how much stored information is outputted. Like RNNs, LSTMs can be applied bi-directionally (Schuster and Paliwal, 1997).

LSTM is widely used for misinformation detection. Bahad et al. Bahad et al. (2019) compared four deep learning models on two fake news datasets ($N_1 = 6{,}335$, $N_2 = 4{,}009$). As shown in tables 3(a) and (b), the bi-directional LSTM (BiLSTM) outperformed CNN, "vanilla" RNN, and unidirectional LSTM in training, validation, and testing, except in dataset 1, where the unidirectional LSTM slightly outperformed BiLSTM in testing accuracy. The authors suggest that the RNN on dataset 1 likely suffered from the vanishing gradient problem (p.81), unlike BiLSTM. Additionally, Figure 5 of the article shows BiLSTM had the fastest convergence to long-term training accuracy.

While language is sequential, restricting models to treat it as such may not be ideal. The **Bidirectional Encoder Representations from Transformers (BERT)** model (Devlin et al., 2019) addresses this by processing the entire text sequence simultaneously through **attention** layers. The input text is split into **tokens**, each encoded with an **embedding** $\vec{E}_i$, initialised randomly then learned via backpropagation. These embeddings are multiplied by three weight vectors in the attention layer: $\vec{W}_Q$, $\vec{W}_K$, and $\vec{W}_V$, producing the **query** $\vec{Q}_i$, **key** $\vec{K}_i$, and **value** $\vec{V}_i$ vectors. The query vector represents how a token "listens" to others, the key vector indicates how it "wants to be listened to" and the value vector holds the information shared during attention.

A token's attention is found by taking its $\vec{Q}_i$ and computing the dot product with its and other tokens' $\vec{K}_i$. These are converted into probabilities using softmax, then multiplied by their respective $\vec{V}_i$. The results are normalised and summed for the attention output. The weight vectors are optimised via backpropagation. BERT uses multiple attention layers connected by feed-forward networks and can be fine-tuned for specific tasks, like classification.

Rakib Mollah et al. (2023) trained a **RoBERTa** (Liu et al., 2019) model – an optimised pre-trained version of BERT – on the WELFake dataset containing 72,134 labelled news articles. RobERTa achieved 0.9976 accuracy – just 34 mistakes from 14,308 test samples. While impressive, RoBERTa is computationally expensive, consisting of 124,645,632 parameters. Moreover, BERT models can struggle with complex language. Kim et al. (2022) trained a BERT model on two X (Twitter) datasets: one for general COVID-19 misinformation and one focused on garlic as a COVID-19 treatment. While BERT outperformed previous models (Word2Vec & FastText) on the garlic-specific dataset, it struggled with the general one, particularly with sarcasm. While BERT can perform well with misinformation classification, computational costs and challenges in understanding complex aspects of language remain as issues.

Images

To detect misinformation in images or videos, ANNs can be adapted. **Convolutional Neural Networks (CNNs)** enhance standard ANN architecture by adding **convolutional** and **pooling layers**. Convolutional layers use small filters (e.g., 3x3 pixels) to scan for patterns, such as parts of a tail to detect cats. The kernel compares its pattern with the image's pixels, creating a **feature map**. The next convolutional layer searches for larger structures, like the whole tail. Pooling layers

aggregate the feature map, which is passed to a fully-connected layer for classification. CNNs learn through backpropagation to adjust kernel weights for image classification.

CNNs can detect misinformation in images, not just text as with RNNs and BERT. Ghai et al. (2021) evaluated the VGG16 CNN model (Simonyan and Zisserman, 2015), pre-trained on the ImageNet dataset (>14,000,000 images). VGG16 was then trained and validated on four datasets (DVMM, CASIAv1.0, CASIAv2.0, IMD2020), containing authentic and tampered versions of a total of 20,128 images. The validation accuracies were 0.83, 0.76, 0.90, and 0.94 for each dataset. However, confusion matrices for CASIAv1.0 and IMD2020 revealed difficulty in classifying tampered images, though authentic ones were correctly identified.

Transformers have found recent applications in image classification, with **Vision Transformers (ViTs)** (Dosovitskiy et al., 2021) adapted to process images. Unlike regular transformers that use 1D positional encodings for text tokens, ViTs apply the attention mechanism to image sections (e.g., 16x16 pixels) with 2D positional encodings. Lamichhane (2025) used a ViT on a dataset of 30,000 images, half real and half generated by a generative adversarial network (GAN)[1] The ViT achieved 0.98 accuracy, outperforming CNN models such as ResNet (He et al., 2015), Xception (Chollet, 2017), DenseNet (Huang et al., 2018), and VGG. The authors also found a ViT with 12 layers to be an optimal balance between accuracy and computational cost.

Multimodal Misinformation

Misinformation on social media is often multimodal, with users posting videos that include visual and audio data, captions, comments, and metadata (likes and shares). Models used in real-world contexts could benefit from capturing this multimodality.

Shang et al. (2021) developed the **TikTok misinformation detection framework (TikTec)** to detect COVID-19 misinformation in TikTok videos by analysing frames, audio, captions, and metadata. TikTec has four modules: Caption-guided Visual Representation Learning (CVRL) identifies misinformation features from frames using its caption and audio; Acoustic-aware Speech Representation Learning (ASRL) detects misinformation in the audio; Visual-speech Co-attentive Information Fusion (VCIF) models the relationship between frames and audio to identify correlations; and Supervised Misleading Video Detection (SMVD) – an ANN – classifies videos as "true" or "false" based on VCIF.

TikTec was tested on 891 pre-labelled English TikTok videos and compared to five common methods for video misinformation detection, such as a CNN-LSTM model and ResNet. It outperformed these methods with an overall accuracy of 0.7231. The authors also assessed TikTec's performance with various modules removed (e.g., without the CVRL module) and found it still outperformed alternative methods. While TikTec may not yet be ideal for practical use due to accuracy and generalisability to other languages, its architecture offers a promising approach for developing video misinformation detectors. Notably, Shang et al. (2025) incorporated a self-attention mechanism from transformers into the **MultiTec** model, replacing the VCIF module

[1]GANs involve two neural networks: a generator that creates fake images and a discriminator that detects them, each improving through competition. GANs are often used to create *deepfakes*.

from Shang et al. (2021), and applied it to two COVID-19-related datasets of TikTok and YouTube Shorts videos. MultiTec outperformed other models, including TikTec, with an accuracy of 0.7442. The authors also noted that "videos containing satire or educational content debunking myths sometimes triggered false positives" (Shang et al., 2025, p.12), highlighting challenges in detecting complex language features.

Misinformation Spread

Rather than focusing on just a social media post's content, we can examine its relationship to the wider social network. For instance, we could model the network as a graph, with the post as a node connected to other nodes representing users and other posts. Similar posts could be linked through edges (e.g., shared users, topics) – a static approach. Furthermore, over time this graph could evolve, adding connections with users who interact with the post – a dynamic approach. **Graph neural networks (GNNs)** can be used to represent social media posts in these ways.

GNNs use graph structures with nodes (containing feature data, e.g., user info, post metadata) and edges (relationships between nodes, e.g., a user ($node_1$) likes ($edge_{1,2}$) a post $node_2$) as inputs. Layers correspond to increasing **neighbour** levels: the first layer includes **immediate neighbours**, the next adds **2-hop neighbours**, and so on. For example, a GNN could classify posts as "fake" or "true" by learning from post content and its relationships. During forward propagation, a node's information is updated using its features and aggregated neighbour features (e.g., weighted average). Backpropagation minimises a loss function (e.g., cross-entropy loss), adjusting weights for aggregation and node features. This enables GNNs to analyse both content and relationships in the context of misinformation. Variations like Graph Attention Networks (GATs) use a self-attention mechanism to fine-tune node connection weights.

Song et al. (2022) develop a Dynamic Graph Neural Network for Fake News detection (DGNF), consisting of two modules: a "structure-aware" module capturing graph structure over time using static snapshots, and a "temporal-aware" module tracking variations between snapshots. DGNF has two variants: "temporal self-attention" (DGNF-tsn), using transformers, and "temporal convolutional" (DGNF-tcn), using CNN kernels. Tested on three datasets from Twitter and Weibo, DGNF models nodes (tweets, retweets, replies) and edges (retweet/reply behaviours) with timestamps. Comparing DGNF to 11 models, including GAT, a graph convolutional network (GCN), and a recursive neural network (RvNN), both variants performed best in detecting real and fake news, with DGNF-tsn superior overall – the authors argue this is due to its flexibility in modelling node interactions over time, relative to DGNF-tcn. This demonstrates how ANNs can process complex data structures, modelling misinformation as a distinct object within social networks.

Discussion

This essay explored how ANNs detect misinformation, from RNNs, LSTM, and BERT for text interpretation, CNNs and ViTs for image pattern recognition, the integration of multimodal data and the use of GNNs to model misinformation in social media networks. Despite the segmentation, these methods often overlap, with CNNs and GNNs also applicable to text, such as CNNs using kernels to identify sentence substructures and GNNs representing sentences as graphs. The key takeaway is that ANN methods are versatile and effective for misinformation detection. We'll now explore how social media companies leverage ANNs for misinformation detection.

In a 2020 blog post, Facebook AI (now Meta AI) revealed that **ObjectDNA** (Meta, 2020), based on the **Object Embeddings for Spliced Image Retrieval (OE-SIR)** (Chen et al., 2021) model, was used to detect misinformation on Facebook. OE-SIR, built on the **Faster R-CNN** (Ren et al., 2016) architecture. Chen et al. (2021) passed spliced images as queries and asked each evaluated model to retrieve the correct authentic image using two databases: COCO-Fake and PIR. The models aimed to return all relevant images, with recall at K (R@K) measuring the proportion of the relevant images that appear in the top K results. OE-SIR outperformed other models at R@1 and R@10 and outperformed in F1 score for an image splicing localisation task (locating spliced objects in an image). This approach enabled Facebook AI to identify "misinformation objects" directly in the image, avoiding the need to aggregate across each image individually, as would be required without object embeddings.

Despite ANNs' effectiveness, some social media companies are shifting from ANN-driven detection to crowdsourced, community-driven approaches. In early 2025, Meta's CEO Mark Zuckerberg announced the replacement of independent fact-checkers with "community notes", similar to X's approach (McMahon et al., 2025). Community notes, corrections to misleading posts added by users based on bipartisan agreement, appear reasonably effective. A study of 205 random notes on X found 97% accurately corrected COVID-19 misinformation (Allen et al., 2024). Additionally, Renault et al. (2024) found that adding a community note to a misinformation post reduced retweets by half and increased tweet deletion probability by 80%. The rationale from Meta is that traditional fact-checkers are politically biased: "Experts, like everyone else, have their own biases and perspectives. This showed up in the choices some made about what to fact check and how." (Kaplan, 2025). Interestingly, research by Maldita, analysing 1,175,837 notes, found that fact checkers were the third most cited source, and notes based on fact checkers were more trustworthy and quicker to appear. Furthermore, Chuai et al. (2024) found that community note creation lags behind the rapid spread of misinformation on X. While community notes are accurate and help address misinformation, they're too slow and still heavily rely on fact-checkers. Instead, ANN methods could be more widely adopted, complementing, rather than replacing, community-driven approaches. ANNs can speed up misinformation detection. Zhou et al. (2024) propose **MUSE**, a large-language model that generates explanations of post accuracy within 2 minutes, outperforming GPT-4 and high-quality layperson responses (i.e. current community notes). This approach could then allow the community to rate and amend the generated notes if needed. Other proposals, like **supernotes**, generated from community notes, show higher user-rated helpfulness, source quality, and comprehensiveness (De et al., 2024). However, these promising approaches appear to conflict with the perspectives of social media companies.

Conclusion

In conclusion, misinformation on social media is a major challenge for society. ANN methods have demonstrated effectiveness in detecting misinformation via text, images, video, and social network propagation. While ANNs show promise in combating misinformation, current efforts by social media companies indicate that more must be done to promote ANN adoption in addressing this issue.

References

Adams, Z., Osman, M., Bechlivanidis, C. and Meder, B. (2023), '(Why) Is Misinformation a Problem?', *Perspectives on Psychological Science* **18**(6), 1436–1463.

Allen, M. R., Desai, N., Namazi, A., Leas, E., Dredze, M., Smith, D. M. and Ayers, J. W. (2024), 'Characteristics of X (Formerly Twitter) Community Notes Addressing COVID-19 Vaccine Misinformation', *JAMA* **331**(19), 1670–1672.

Bahad, P., Saxena, P. and Kamal, R. (2019), 'Fake News Detection using Bi-directional LSTM-Recurrent Neural Network', *Procedia Computer Science* **165**, 74–82.

Chen, B.-C., Wu, Z., Davis, L. S. and Lim, S.-N. (2021), 'Efficient Object Embedding for Spliced Image Retrieval'.

Chollet, F. (2017), 'Xception: Deep Learning with Depthwise Separable Convolutions'.

Chuai, Y., Pilarski, M., Renault, T., Restrepo-Amariles, D., Troussel-Clément, A., Lenzini, G. and Pröllochs, N. (2024), 'Community-based fact-checking reduces the spread of misleading posts on social media'.

De, S., Bakker, M. A., Baxter, J. and Saveski, M. (2024), 'Supernotes: Driving Consensus in Crowd-Sourced Fact-Checking'.

Deng, L. (2012), 'The MNIST Database of Handwritten Digit Images for Machine Learning Research [Best of the Web]', *IEEE Signal Processing Magazine* **29**(6), 141–142.

Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K. (2019), BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, *in* J. Burstein, C. Doran and T. Solorio, eds, 'Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)', Association for Computational Linguistics, Minneapolis, Minnesota, pp. 4171–4186.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J. and Houlsby, N. (2021), 'An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale'.

Dubey, S. R., Singh, S. K. and Chaudhuri, B. B. (2022), 'Activation functions in deep learning: A comprehensive survey and benchmark', *Neurocomputing* **503**, 92–108.

Enock, F., Bright, J., Stevens, F., Johansson, P. and Margetts, H. Z. (2024), 'How do people protect themselves against online misinformation? Attitudes, experiences and uptake of interventions amongst the UK adult population', *SSRN Electronic Journal* .

Ghai, A., Kumar, P. and Gupta, S. (2021), 'A deep-learning-based image forgery detection framework for controlling the spread of misinformation', *Information Technology & People* **37**(2), 966–997.

He, K., Zhang, X., Ren, S. and Sun, J. (2015), 'Deep Residual Learning for Image Recognition'.

Hochreiter, S. and Schmidhuber, J. (1997), 'Long Short-Term Memory', *Neural Computation* **9**(8), 1735–1780.

Huang, G., Liu, Z., van der Maaten, L. and Weinberger, K. Q. (2018), 'Densely Connected Convolutional Networks'.

Kaplan, J. (2025), 'More Speech and Fewer Mistakes'.

Kim, M. G., Kim, M., Kim, J. H. and Kim, K. (2022), 'Fine-Tuning BERT Models to Classify Misinformation on Garlic and COVID-19 on Twitter', *International Journal of Environmental Research and Public Health* **19**(9), 5126.

Lamichhane, D. (2025), 'Advanced Detection of AI-Generated Images Through Vision Transformers', *IEEE Access* **13**, 3644–3652.

Liedke, C. S. A. a. J. (2023), 'Most Americans favor restrictions on false information, violent content online'.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L. and Stoyanov, V. (2019), 'RoBERTa: A Robustly Optimized BERT Pretraining Approach'.

Loomba, S., de Figueiredo, A., Piatek, S. J., de Graaf, K. and Larson, H. J. (2021), 'Measuring the impact of COVID-19 vaccine misinformation on vaccination intent in the UK and USA', *Nature Human Behaviour* **5**(3), 337–348.

Lyons, B. A., Montgomery, J. M., Guess, A. M., Nyhan, B. and Reifler, J. (2021), 'Overconfidence in news judgments is associated with false news susceptibility', *Proceedings of the National Academy of Sciences* **118**(23), e2019527118.

McMahon, L., Kleinman, Z. and Subramanian, C. (2025), 'Meta to replace 'biased' fact-checkers with moderation by users', https://www.bbc.com/news/articles/cly74mpy8klo.

Meta (2020), 'Here's how we're using AI to help detect misinformation', https://ai.meta.com/blog/heres-how-were-using-ai-to-help-detect-misinformation/.

Morkūnas, M. (2023), 'Russian Disinformation in the Baltics: Does it Really Work?', *Public Integrity* **25**(6), 599–613.

Open Knowledge Foundation (2020), 'Opinion poll: Majority of Brits want government action against online disinformation', https://blog.okfn.org/2020/05/07/opinion-poll-majority-of-brits-want-government-action-against-online-disinformation/.

Piazza, J. A. (2022), 'Fake news: The effects of social media disinformation on domestic terrorism', *Dynamics of Asymmetric Conflict* **15**(1), 55–77.

Quétier-Parent, S., Lamotte, D. and Gallard, M. (2023), 'Elections & social media: The battle against disinformation and trust issues | Ipsos', https://www.ipsos.com/en/elections-social-media-battle-against-disinformation-and-trust-issues.

Rakib Mollah, M. A., Kabir, M. M. J., Kabir, M. and Reza, M. S. (2023), Detection of Fake News with RoBERTa Based Embedding and Modified Deep Neural Network Architecture, *in* '2023 26th International Conference on Computer and Information Technology (ICCIT)', pp. 1–6.

Ren, S., He, K., Girshick, R. and Sun, J. (2016), 'Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks'.

Renault, T., Amariles, D. R. and Troussel, A. (2024), 'Collaboratively adding context to social media posts reduces the sharing of false news'.

Schuster, M. and Paliwal, K. (1997), 'Bidirectional recurrent neural networks', *IEEE Transactions on Signal Processing* **45**(11), 2673–2681.

Shang, L., Kou, Z., Zhang, Y. and Wang, D. (2021), A Multimodal Misinformation Detector for COVID-19 Short Videos on TikTok, *in* '2021 IEEE International Conference on Big Data (Big Data)', pp. 899–908.

Shang, L., Zhang, Y., Deng, Y. and Wang, D. (2025), 'MultiTec: A Data-Driven Multimodal Short Video Detection Framework for Healthcare Misinformation on TikTok', *IEEE Transactions on Big Data* pp. 1–18.

Shearer, E., Lipka, M., Naseer, S., Tomasik, E. and Jurkowitz, M. (2024), 'Americans' Views of 2024 Election News'.

Simonyan, K. and Zisserman, A. (2015), 'Very Deep Convolutional Networks for Large-Scale Image Recognition'.

Smith, M. (2024), 'Two thirds of Britons say social media companies should be held responsible for posts inciting riots | YouGov', https://yougov.co.uk/politics/articles/50288-two-thirds-of-britons-say-social-media-companies-should-be-held-responsible-for-posts-inciting-riots.

Song, C., Teng, Y., Zhu, Y., Wei, S. and Wu, B. (2022), 'Dynamic graph neural network for fake news detection', *Neurocomputing* **505**, 362–374.

Vosoughi, S., Roy, D. and Aral, S. (2018), 'The spread of true and false news online', *Science* **359**(6380), 1146–1151.

Zhou, X., Sharma, A., Zhang, A. X. and Althoff, T. (2024), 'Correcting misinformation on social media with a large language model'.