MULTIMODAL FAKE NEWS DETECTION: A BENCHMARK OF STATE-OF-THE-ART APPROACHES

Lappeenranta–Lahti University of Technology LUT

Bachelor's Programme in Software Engineering, Bachelor's thesis

2025

Antti Keronen

Examiners: M.Sc. (Tech) Maksim Shchukin (Junior Researcher)

Doc. Dr. Damian Kedziora (Associate Professor)

ABSTRACT

Lappeenranta–Lahti University of Technology LUT

LUT School of Engineering Sciences

Software Engineering

Antti Keronen

Multimodal fake news detection: a benchmark of state-of-the-art approaches

Bachelor's thesis

2025

36 pages, 4 figures, 2 tables

Examiners: M.Sc. (Tech) Maksim Shchukin and Damian Kedziora (Ph.D)

Keywords: fake news, misinformation, multimodal analysis, text-image integration

This study examined multimodal fake news detection, focusing on posts that combine text and images. The research aimed to determine how effectively multimodal approaches detect fake news, how they integrate textual and visual information, and how they compare with unimodal methods that analyze only text or visual content. The analysis benchmarked multiple state-of-the-art models across two realistic datasets. The models were tested to see how well they can detect misleading content.

The results showed that multimodal approaches consistently outperform unimodal models, demonstrating that combining textual and visual information improves detection reliability. Methods that explicitly model interactions between text and images perform particularly well, while approaches that integrate information implicitly also benefit from capturing contextual dependencies. Visual content was found to provide complementary information, enhancing the overall performance of fake news detection.

TIIVISTELMÄ

Lappeenrannan–Lahden teknillinen yliopisto LUT

LUTin insinööritieteiden tiedekunta

Tietotekniikka

Antti Keronen

Multimodaalinen valeuutisten tunnistus: nykyaikaisten menetelmien vertailututkimus

Tietotekniikan kandidaatintyö

2025

36 sivua, 4 kuvaa, 2 taulukkoa

Tarkastajat: DI Maksim Shchukin ja Tohtori Damian Kedziora

Avainsanat: valeuutiset, harhaanjohtava tieto, multimodaalisuus, tekstikuvaintegraatio

Tässä tutkimuksessa tarkasteltiin multimodaalista valeuutisten tunnistusta, keskittyen julkaisuihin, joissa yhdistellään tekstiä ja kuvia. Tutkimuksen tavoitteena oli selvittää, kuinka tehokkaasti multimodaaliset menetelmät tunnistavat valeuutisia, miten ne yhdistävät tekstin ja visuaalisen sisällön ominaisuudet, ja kuinka ne vertautuvat unimodaalisiin menetelmiin, jotka analysoivat vain tekstiä tai kuvia. Tutkimuksessa vertailtiin useita nykyaikaisia malleja kahden media-aineiston pohjalta, Malleja testattiin arvioimaan niiden kykyä tunnistaa harhaanjohtavaa sisältöä.

Tulokset osoittivat, että multimodaaliset menetelmät päihittävät johdonmukaisesti unimodaaliset menetelmät, mikä osoittaa, että tekstin ja visuaalisen sisällön yhdistäminen parantaa tunnistuksen luotettavuutta. Menetelmät, jotka mallintavat tekstin ja kuvien väliset vuorovaikutukset eksplisiittisesti, toimivat erityisen hyvin, kun taas menetelmät, jotka integroivat tietoa implisiittisesti, hyötyvät myös kontekstuaalisista riippuvuuksista. Visuaalinen sisältö tarjosi täydentävää tietoa, mikä paransi yleistä valeuutisten tunnistuksen suorituskykyä.

SYMBOLS AND ABBREVIATIONS

Abbreviations

AI	Artificial Intelligence
BERT	Bidirectional Encoder Representations from Transformers
BLIP	Bootstrapped Language-Image Pretraining
CLIP	Contrastive Language-Image Pre-training
CNN	Convolutional Neural Network
GAN	Generative Adversarial Network
GRU	Gated Recurrent Unit
LSTM	Long Short-Term Memory
LLM	Large Language Model
LVLM	Large Vision-Language Model
NLP	Natural Language Processing
RNN	Recurrent Neural Network
SLM	Small Language Model

Table of contents

DECLARATIONS

Turnitin

The originality of this thesis has been reviewed with the Turnitin similarity checking service.

AI usage

The author of the thesis, Antti Keronen, used the following AI-tools during the preparation of the thesis:

1. ChatGPT (GPT-5.1)

    a. Purpose of use: Used as help to train and implement the benchmarked fake news detection models correctly.

    b. Purpose of use: To format the reference list

Responsibility

The author, Antti Keronen, takes full responsibility for the content of this thesis and has reviewed and edited the content generated by the possible use of AI tools.

1 Introduction

In September 2015, conspiracy theorists and bloggers falsely claimed that the Obama administration in US planned a military takeover of Texas through the Jade Helm 15 training exercise. Though the operation was routine, misinformation spread rapidly online, revealing how easily digital narratives can distort reality and inflame public fear. (Howerton, 2025). This thesis investigates multimodal fake news detection, focusing on posts that combine text and images. The topic is timely, because misinformation is spreading faster than facts, especially on social media platforms, shaping political decisions and public perception. Understanding and detecting multimodal misinformation is important to mitigate its real-world impacts and maintain information integrity in the digital world.

Previous studies have made significant progress in fake news detection, primarily focusing on unimodal approaches that analyse either text or images independently. Text-based methods using natural language processing (NLP) and transformer architectures have shown strong performance in identifying misleading information, while image-based approaches employing computer vision techniques have been effective in detecting manipulated visuals. More recent studies have investigated multimodal methods that integrate both textual and visual data, demonstrating that integrating multiple modalities can improve detection accuracy compared to single-modality methods. These studies provide a solid foundation, showing that leveraging the combination of text and images can improve the reliability of fake news detection models.

Despite a great progress in the research, several important challenges remain. Many studies continue to rely on small or specifically selected datasets that do not reflect the full diversity and complexity of media content. This limits how well models can generalize to practical situations. In addition, the connection between text and visual content is often simplified, with limited attention to cases where captions give misleading meaning to images or where the two do not fully match. Because of these limitations, we still lack a clear understanding of the performance of state-of-the-art methods on large and diverse collections of multimodal content. This creates a need for systematic benchmarks that test these models thoroughly.

This thesis addresses this need by benchmarking state-of-the-art models for multimodal fake news detection, from unimodal baselines to advanced multimodal fusion and large language models. The research problem focuses on how well these models detect fake news in realistic datasets and how they combine text and image information to improve accuracy. The study addresses three questions: How effective are the current state-of-the-art multimodal approaches? How do they integrate textual and visual cues? How do they compare with unimodal methods? While limited by code availability for the models and computational resources, the thesis aims to provide a clear evaluation of current techniques.

2 Background and Related work

The growth of digital media and social platforms has fundamentally transformed how information is generated and disseminated, leading to an unprecedented spread of false or misleading content online. Fake news has emerged as a significant challenge in today's information landscape, affecting public trust and the credibility of media. This chapter explores the concept of fake news, its various forms, and its broader societal impacts and challenges. Building on this, the chapter examines the evolution of fake news detection methods, highlighting the progression from unimodal approaches, focusing on either text or images alone, to more sophisticated multimodal systems that analyse combined text and visual content. It introduces key architectures, techniques, and representative models that underpin current state-of-the-art research in multimodal fake news detection. By reviewing both classical and modern methodologies, this chapter lays the theoretical and technical groundwork necessary to understand and compare the approaches that will be evaluated in upcoming chapters, thus connecting the emerging challenges of fake news with the technological advancements designed to combat it.

2.1 Fake News and Impact

The growth of digital media and social platforms has fundamentally changed how information is generated and shared, leading to an unprecedented spread of false or misleading content online. Fake news has become a major challenge in today's information landscape, affecting public trust and media credibility. This chapter explores what fake news is, the different forms it takes, and its broader societal impact and challenges.

In recent years, the term "fake news" has received growing attention, yet its exact definition remains debated. It is commonly understood as news-like content that intentionally spreads false or misleading information. What separates fake news from legitimate journalism is the purposeful intent to deceive. While mainstream media may occasionally publish inaccurate reports, their goal remains in telling the truth, while fake news intentionally ignores it and is defined as the deliberate presentation of false or misleading news designed to deceive (Gelfert, 2018). Zhou and Zafarani (2022) propose a similar definition as intentionally and

verifiably false news published by a news outlet. These definitions summarize the core features: falsity, intent to deceive, and presentation as news. So, fake news can be understood as intentionally deceptive content that mimics journalistic style, created to mislead audiences for political, ideological, or financial purposes (Zhou and Zafarani, 2020).

Fake news can appear in many different forms depending on its purpose, content, and the techniques used to mislead audiences. Each type has distinct characteristics, which affect how it spreads and how it can be detected. Fabricated news consists of completely invented stories presented as real journalism. These stories often replicate the appearance and voice of credible news sources, using bylines, quotes, and professional layouts to appear credible (Tandoc et al., 2018). Fabricated news spread on social media because its titles are short, simple, and easy to understand, while the article text is often repetitive, so people share it without reading carefully (Horne and Adali, 2017). Manipulated content involves genuine information that is altered or presented out of context to mislead audiences. Modern techniques include deep learning-based deepfakes, which create realistic but false audio or visual content. Manipulated visuals are particularly persuasive because people often trust visual information more than text. (Vaccari and Chadwick, 2020) Satirical or parody news imitate journalistic reporting for entertainment or social commentary. While not intended to deceive, satire can be misunderstood when it leaves outsides its original context, especially on social media where cues about tone or intent are often missing (Tandoc et al., 2018).

These types of fake news can also intersect. Fabricated stories may include manipulated visuals, and satire may be misinterpreted as genuine news. Understanding these forms is important for identifying, analysing, and countering fake news online.

Fake news has a great effect on individuals, society and political processes. Its spread can mislead audiences, affect opinions, and reduce trust in news resources. Because fake news often evokes strong emotional responses, it tends to spread quickly and widely (Del Vicario et al., 2016). Impact on individuals: Exposure to fake news can modify beliefs and alter the decision-making process of a human. People tend to share content that confirms their preexisting beliefs, thereby strengthening their biases and creating echo chambers (Del Vicario et al., 2016). Impact on society**:** On a societal level, fake news can affect public discourse and civic engagement. It promotes misinformation during elections and undermines confidence in public institutions (Allcott & Gentzkow, 2017).

Challenges in detection: Detecting fake news poses major difficulties because of its evolving nature and the diversity of forms it takes. According to UN ITU (2024) the rise of AI-generated deepfakes has outpaced current detection capabilities, often deceiving even trained professionals and complicating verification efforts. Also, the quick spread of deceptive information, particularly via social media platforms, frequently outpaces the verification capacity of fact-checkers and automated systems. This accelerated spread amplifies the impact of misinformation, letting it to influence large audiences before corrective interventions can be implemented. Challenges with regulations: Regulating fake news is difficult because it must balance controlling misinformation with protecting free speech, and the global nature of online content complicates enforcement (Jacobs, 2022).

The impact of fake news is amplified by digital platforms. Fake news is now seen as a serious threat to democracy, journalism, and free speech. It has made people trust governments less, and we still don't know how much harm it may cause in the future. (Zhou and Zafarani 2022)

2.2 Unimodal Fake News Detection

Before the rise of multimodal systems, fake news detection research focused mainly on single-source analysis, either textual or visual. Early unimodal approaches, such as textual classifiers, laid the foundation for modern misinformation detection by identifying deceptive patterns in language, paving the way for more advanced techniques (Shahi, 2025). Although such methods advanced understanding of how misinformation spreads in isolated formats, they proved insufficient for detecting the complex, cross-modal deceptions common in today's online environment (Shahi, 2025). This chapter examines the evolution of unimodal fake news detection, highlighting key developments in text-based and visual-based systems and the challenges that motivated the transition toward multimodal solutions.

Early research into fake news detection was mostly text-based, relying on linguistic cues and handcrafted textual features to distinguish between factual and deceptive information. Early models primarily applied standard machine learning algorithms, for example, support vector machines and logistic regression, and random forest to classify textual data before the adoption of deep-learning techniques. (Al-alshaqi et al., 2024) Studies during this period revealed that fake news typically relied on simple and repetitive content with shorter bodies and more attention-grabbing titles, relying on user's quick judgements rather than strong arguments to

persuade readers (Horne and Adalı, 2017). Sentiment analysis was also utilised to identify excessively positive or negative tones, which were common in misleading reports (Shahi, 2025). These early text-based systems still faced challenges when applied to the fast-changing online information ecosystem. They faced challenges in generalisation and were prone to overfitting, as datasets were very small or imbalanced and linguistic patterns varied across contexts (Hamed et al., 2023).

As the limitations of handcrafted features became increasingly evident, research in fake news detection shifted toward deep learning approaches. Standard Recurrent Neural Networks (RNNs) can model sequential data, yet they have difficulty capturing long-range dependencies because the gradients used during training often become too small or too large, making it difficult for the network to learn effectively over longer sequences. To address this issue, gated architectures like Long Short-Term Memory (LSTM) networks and Gated Recurrent Units (GRUs) were introduced. Their gating mechanisms and stable memory cells allow these models to capture long-term contextual information that standard RNNs cannot. Bidirectional variants further enhance performance by processing sequences in both directions, capturing richer context for tasks like text classification and fake news detection. (Bahad et al., 2019).

The evolution of natural language processing (NLP) further accelerated with the introduction of transformer-based architectures, particularly BERT (Bidirectional Encoder Representations from Transformers). BERT-based models have shown notable performance in fake news detection by learning bidirectional context and capturing deep semantic and syntactic relationships within text (Helal et al., 2024). Such models move beyond superficial linguistic cues and instead rely on learned embeddings that represent variative meaning, sarcasm, or contextual deception (Helal et al., 2024).

Unimodal text-based approaches still show notable weaknesses. Misinformation in the digital era often combines text with manipulated visual or multimedia content to increase credibility. Text-only models lack the ability to verify claims visually or detect inconsistencies between narrative and image content. For instance, a fake article might present a real photo from a different event to support false claims, an inconsistency that text-based analysis alone cannot detect. While textual analysis forms a foundational aspect of misinformation detection, its effectiveness is limited when deception involves multimodal factors such as images, videos, or embedded visual text.

The rapid spreading of image- and video-based misinformation on social media has made visual analysis a crucial component of fake news detection. Unlike textual misinformation, which relies on language manipulation, visual misinformation leverages the psychological credibility of images and videos to deceive audiences (Cao et al., 2020). Research into uni-modal visual detection has focused primarily on identifying manipulated, fabricated, or mis-leading media through pixel-level analysis, statistical inference, and deep learning-based visual representations (Shahi, 2025).

In the early stages, visual misinformation detection relied on digital image forensics. These methods focused on identifying tampering traces such as inconsistencies in compression ar-tefacts, lighting direction, or image metadata. Approaches like Error Level Analysis (ELA) and frequency-domain analysis were commonly used to highlight discrepancies introduced during image editing (Cao et al., 2020). These traditional techniques still proved inadequate for detecting subtle manipulations or artificial images produced by modern generative mod-els (Abbas & Taeihagh, 2024)**.**

The adoption of convolutional neural networks (CNNs) marked a major change. CNNs learn hierarchical image features directly from pixels, which makes them effective at finding ma-nipulation artifacts and visual patterns that handcrafted methods miss (O'Shea & Nash, 2015). One of the most influential architectures is VGG-19, a 19-layer CNN developed by the Visual Geometry Group at the University of Oxford (Simonyan & Zisserman, 2014). VGG-19 extends the VGG family's design principle of stacking small 3×3 convolutional filters to build very deep feature hierarchies. Also, there is multi-domain CNN frameworks that extract both pixel-level and frequency-domain features and then fuse them with attention have shown improved detection of tampered or misleading images (Qi et al., 2019). At the same time, the rapid growth of Generative Adversarial Networks (GANs) has produced highly realistic synthetic images and videos called deep fakes, which are hard to detect with simple forensic tests. Deepfake detection approaches employ AI-based frameworks that an-alyse images, audio, and video to automatically detect manipulated or artificially generated content. (Abbas & Taeihagh, 2024).

Visual-only systems still face obstacles. Images or videos can be authentic but shown with misleading captions or reused from other events or manipulated visuals may be paired with harmless text. Purely visual models cannot detect cross-modal mismatches or verify claims that require external knowledge, meaning they often miss cases where the accompanying

text contradicts the image (Shahi, 2025). These shortcomings motivate combining vision models with textual and contextual signals in multimodal systems.

2.3 Multimodal Fake News Detection

As misinformation has become more advanced, using text, images, and videos together to appear more believable. Research has moved from unimodal detection to multimodal. Multimodal methods combine information from several sources, such as text, visuals, and context, to better spot inconsistencies and signs of deception (Khattar et al., 2019). These systems can find relationships between different types of data that single-modality models often miss, giving a fuller picture of how misinformation spreads (Wang et al., 2022). This chapter looks at the main types of multimodal fake news detection approaches: early fusion, cross-modal fusion, and transformer- and LLM-based models. All the models represented are considered state-of-the-art models.

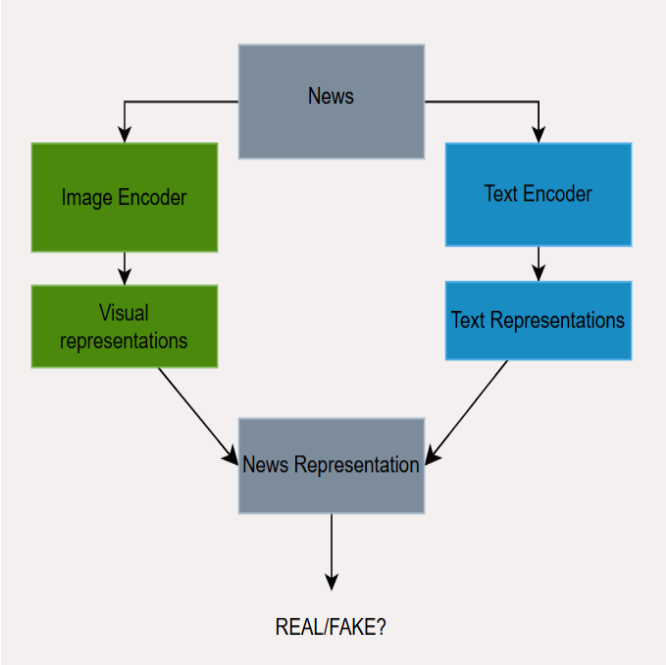

Figure 1: Simplified illustration of early fusion model (Singhal, 2019)

Early fusion, also called feature-level fusion, is a multimodal integration strategy where features extracted from different modalities, are combined into a single representation before being input into a classifier (Lin et al., 2024). As shown in Figure 1, this approach merges

the outputs of text and image encoders at an early stage, allowing the model to learn from both sources of information, image and text. This method leverages the complementary and redundant information from both modalities, enabling the model to capture cross-modal relationships that may indicate fake news (Khattar et al., 2019).

Early fusion can be implemented by concatenating feature vectors from text and images or mapping them into a shared embedding space (Lin et al., 2024). For instance, SAFE (Zhou et al., 2020) concatenates textual and visual features and measures their similarity to detect inconsistencies, improving the detection of misleading or manipulated content. MVAE (Khattar et al., 2019) uses a multimodal variational autoencoder to learn a shared representation of text and visual features, which is then fed into a classifier for fake news detection. SpotFake (Singhal et al., 2019) leverages pretrained language transformers and ImageNet models to extract textual and visual features, which are fused before classification. Similarly, a model EANN (Event adversarial neural networks for multi-modal fake news detection) (Wang et al., 2018) combines text and image features through early fusion before classification, forming a joint multimodal representation. What distinguishes EANN is its use of adversarial training to remove event-specific patterns, enabling the fused representation to generalize to unseen events.

The main advantage of early fusion is that it allows the classifier to directly model interactions between modalities, capturing joint patterns that may not be evident in unimodal analysis (Lin et al., 2024). However, it has notable challenges, including high-dimensional feature vectors, risk of overfitting, and the requirement for well-aligned multimodal features (Khattar et al., 2019).

Despite these challenges, early fusion remains widely used in multimodal fake news detection because it effectively exploits complementary information across modalities and serves as a baseline for more complex fusion methods (Hamed et al., 2023).

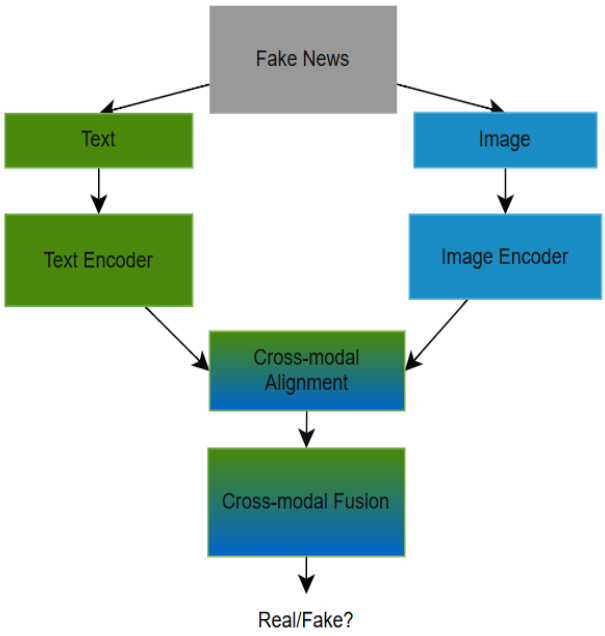

Figure 2: Simplified illustration of Cross-modal fusion (Chen et al., 2022)

Cross-modal fusion, also called intermediate or joint fusion, aims to model deep semantic interactions between both modalities by aligning and integrating them in a shared space (Chen et al., 2025a). As illustrated in Figure 2, this approach processes textual and visual features separately before combining them through cross-attention or other interaction mechanisms, enabling the model to learn complex relationships between modalities. Unlike early fusion, which simply concatenates features, cross-modal fusion focuses on capturing dynamic correlations and inconsistencies between modalities to improve fake news detection (Qiao et al., 2025).

Traditional approaches rely on pre-trained unimodal encoders, followed by fully connected layers to align features into a shared representation space (Li et al., 2025). However, these methods only align text and image features at a basic level and fail to capture deeper semantic connections between them (Li et al., 2025). To address these limitations, Li et al. (2025) proposed the Hierarchical Cross-Modal Interaction Network (HCMIN), which uses contrastive learning and dual attention to connect information within and between text and images. This helps the model better understand how well the text and image match or conflict, which is important for spotting fake news. Chen et al. (2025a) introduced Information-Level Multimodal Fusion for Fake News Detection (ILMF-FND), which refines unimodal features

before performing cross-modal alignment and fusion through low-rank tensor compression. Its modal adaptation module dynamically reweights unimodal and fused representations, enhancing detection accuracy while reducing computational complexity. Similarly, Chen et al. (2022) proposed CAFE (Cross Modal Ambiguity Learning for Multimodal Fake News Detection), which mitigates misclassification caused by inconsistent modalities by explicitly modelling text–image correlations. After this, Wang et al. (2025) developed a cross-modal contrastive learning framework that uses attention-guided alignment to achieve fine-grained correspondence and improved interpretability.

Qiao et al. (2025) proposed the Cross-modal Content Correlation Network (C3N), which captures text–image relationships through a convolutional operation on similarity matrices derived from CLIP embeddings. By analysing correlations rather than scalar similarities, C3N effectively detects inconsistencies that indicate manipulated or misleading content. These cross-modal fusion approaches demonstrate that modelling fine-grained cross modal dependencies and inconsistencies is crucial for reliable fake news detection.

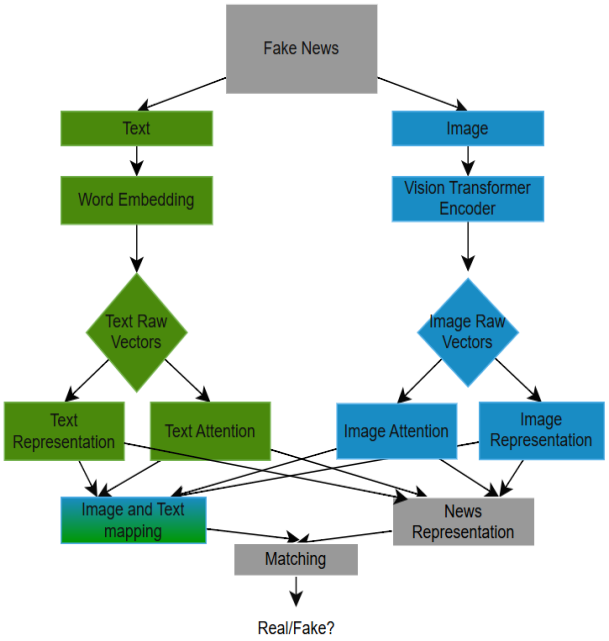

Figure 3: Simplified illustration of Transformer-based model (Yang et al., 2023)

Transformer-based architectures have become important to news detection because they effectively manage long-range dependencies and handle sequential data without CNN or RNN.

Unlike CNN or RNN, as shown in Figure 3, transformers use self-attention to learn contextual relationships within and across data modalities, which allows them to model complex dependencies between text and visual features. This capability makes them particularly suitable for multimodal misinformation analysis, where both textual and visual coherence are important (Nasser et al., 2025).

Several transformer-based models have been developed to enhance multimodal fusion for fake news detection. For instance, Choi and Kim (2024) proposed TT-BLIP, an end-to-end model that integrates BERT and BLIP encoders for textual and visual data, respectively. It employs a Multimodal Tri-Transformer architecture to fuse tri-modal features through multi-head attention mechanisms, effectively improving representation learning.

KGAlign, introduced by Zhang et al. (2025), extends the transformer framework by incorporating external knowledge graphs into multimodal fake news detection. KGAlign aligns textual, visual, and knowledge representations using transformer-based cross-modal attention, enriching semantic understanding beyond surface-level features. Another recent state-of-the-art approach is the TGA model (Yang et al., 2023), a multimodal transformer framework which extracts textual and visual features using separate transformer encoders and fuses them through an attention mechanism. It also leverages the degree of feature similarity between text and image modalities to refine the detection outcome, improving the identification of inconsistencies indicative of misinformation. (Yang et al., 2023)

Transformer-based models such as TT-BLIP (Choi and Kim, 2024) and KGAlign (Zhang et al., 2025) demonstrate the growing effectiveness of attention driven architectures in modelling complex multimodal relationships and identifying inconsistencies that signal misinformation

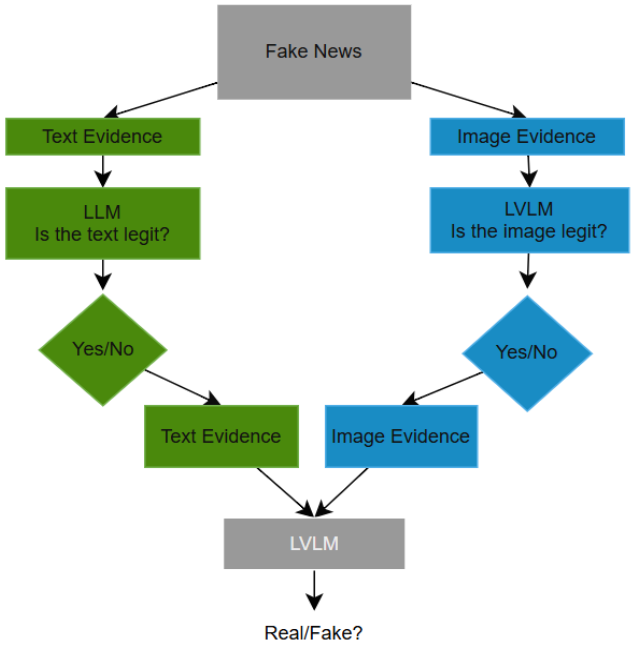

Figure 4: Simplified illustration of LLM/LVLM models (Tahmasebi et al., 2024)

Large Language Models (LLMs) have become an important area of development in multi-modal fake news detection because they can understand complex language patterns and use background knowledge more effectively than smaller pre-trained models. Traditional small language models (SLMs) such as BERT often struggle with domain-specific reasoning and cross-modal relationships, which limits their ability to detect misinformation accurately (Wang et al., 2024). To address these challenges, Wang et al. (2024) proposed the FND-LLM framework, which combines SLMs and LLMs to make use of both detailed feature extraction and broader reasoning. As shown in Figure 4, like other LLM models, the model similarly includes several coordinated branches for extracting linguistic and image-based information, a co-attention network for refining cross-modal relationships, a CLIP-based branch for aligning image and text features and an LLM branch that performs logical reasoning and external knowledge verification using models like GPT-4. This structure allows the system to validate news content and make final predictions with improved interpretability and accuracy.

Tahmasebi et al. (2024) introduced a Large Vision-Language Model for Fact Verification (LVLM4FV) that retrieves textual and visual evidence from multiple sources in a zero-shot setting, achieving better performance and generalization than supervised baselines. Hu et al.

(2024) proposed the Adaptive Rationale Guidance (ARG) model, in which LLMs support SLMs by providing multi-perspective reasoning explanations. Their study found that while LLMs alone may not outperform fine-tuned smaller models, they greatly improve interpretability and reasoning when used together in hybrid systems.

Another large language fake news detection model GLPN-LLM (Global Label Propagation Network with LLM-based Pseudo Labelling) is framework by Hu et al. (2025), which represents one of the most recent advances in multimodal fake news detection using large language models. A multimodal framework that integrates LLM capabilities through label propagation techniques. In this approach, LLMs generate pseudo labels for multimodal samples, which are then refined and propagated across the dataset using a mask-based global label propagation mechanism. This mechanism prevents label leakage during training by ensuring that samples do not propagate their own labels, while still allowing effective information sharing between instances. The synergy between LLM-based pseudo labelling and global label propagation significantly enhances detection accuracy and stability, even when individual pseudo labels are imperfect.

These studies demonstrate that LLMs and LVLMs improve fake news detection by combining content analysis with external reasoning and knowledge retrieval. This integration leads to more reliable, generalizable, and interpretable detection systems capable of handling both textual and visual misinformation.

3 Benchmark

This chapter presents a benchmark evaluation of state-of-the-art multimodal fake news detection models. The goal is to assess how well different categories of models, early fusion, cross-modal fusion, transformer-based, and large language model-based approaches perform in detecting misinformation across social media platforms. To provide a comprehensive perspective, these multimodal models are compared against text-only and visual-only baseline methods, highlighting the added value of integrating multiple modalities.

The evaluation is conducted using two widely adopted datasets in fake news research, Twitter and Weibo. These datasets contain diverse examples of multimodal fake news, including text-image combinations with varying levels of manipulation and contextual cues. Each model is tested under consistent experimental settings to ensure a fair comparison, focusing on key performance measures including precision, recall, F1 and accuracy. By conducting this benchmarking, the chapter seeks to determine how effectively multimodal approaches outperform unimodal methods, identify which fusion strategies and architectural designs get the most reliable detection results, and outline emerging trends that can guide future developments in multimodal fake news detection.

3.1 Benchmark Environment

All tests were conducted on a local workstation and, when necessary, verified using Google Colab. The local system specifications are as follows in Table 1:

Table 1: Experiment Environment

CPU	AMD Ryzen 7 7800x3d
GPU	Nvidia rtx 4070S
Memory	32gb ddr5 RAM
Operating System	Windows 10 64-bit

All models were executed in python environments using their respective framework versions as provided in the official implementations. Different update versions of Python, PyTorch, the Transformers library, and TensorFlow were used in this study. The environment was modified if it had gone out of date and was unimplementable.

3.2 Benchmark Datasets

Two publicly available multimodal datasets were used for training and evaluating the models. The Twitter dataset was collected in the context of the MediaEval 2015 workshop and comprises tweets with associated images or videos spread around major events. Each tweet is labelled as either real or fake. (Boididou et al., 2015) This dataset is one of the most widely used datasets in multimodal fake news detection setups. The Weibo dataset covers Chinese microblogs from May 2012 to January 2016. Fake news items are drawn from the official Weibo rumour-debunking system while real news posts are from authoritative sources (Jin et al., 2017). This dataset supports evaluation of cross-lingual and cross-platform rumour detection in a multilingual context.

3.3 Benchmark Models

For the benchmark study, several state-of-the-art fake news detection models were selected. The selection criteria were primarily based on the availability of code and the feasibility of testing the models. Adaptability to the chosen datasets was considered, ensuring that the models could handle the data format and languages.

The models selected to the benchmark:

Baseline models (To which the performance of the models is compared):

- BERT (Text-only)
- VGG-19 (Visual-only)

State-of-the-art models:

- Spotfake (A Multi-modal Framework for Fake News Detection) (2019)

- EANN: (Event-Adversarial Neural Networks for Multi-Modal Fake News Detection) (2018)
- CAFÉ (Cross Modal Ambiguity Learning for Multimodal Fake News Detection) (2022)
- ILMF-FND (Information-Level Multimodal Fusion) (2025)
- TGA (Multi-modal transformer) (2023)
- GLPN-LLM (Global Label Propagation Network with LLM-based Pseudo Labelling) (2025)

3.4 Benchmark Metrics

TP = True Positives

TN = True Negatives

FP = False Positives

FN = False Negatives

Eq 1. $$\boldsymbol{Precision} = \frac{\boldsymbol{TP}}{\boldsymbol{TP+FP}}$$

Eq 2. $$\boldsymbol{Recall} = \frac{\boldsymbol{TP}}{\boldsymbol{TP+FN}}$$

Eq 3. $$\boldsymbol{F1} = \frac{\boldsymbol{2 \times Pre \times Rec}}{\boldsymbol{Pre+Rec}}$$

Eq 4. $$\boldsymbol{Accuracy} = \frac{\boldsymbol{TP+TN}}{\boldsymbol{TP+TN+FP+FN}}$$

The evaluation metrics offer a detailed view of the models' performance. Precision (Eq 1.) indicates how many of the items labelled as positive are indeed positive, highlighting the model's effectiveness in avoiding false positive outcomes. Recall (Eq 2.) reflects the percentage of actual positive samples that the model correctly identifies, showing its effectiveness in reducing false negatives. Combining both precision and recall, the F1-score (Eq 3.) offers a single, balanced evaluation of performance. Accuracy (Eq 4.) measures the model's overall capability to make correct predictions across all classes, giving a general assessment

of its effectiveness. Together, these metrics deliver a complete picture of the strengths and limitations of the classification models.

4 Results

Table 2. Benchmark Results

Dataset	Approach	Model	Accuracy	Precision	Recall	F1
Weibo	Text-only	**BERT (2018)**	0.76	0.74	0.78	0.75
	Visual-only	**VGG-19 (2014)**	0.72	0.70	0.74	0.71
	Early Fusion	**SpotFake (2019)**	0.80	0.85	0.79	0.81
		EANN (2018)	0.80	0.77	0.82	0.79
	Cross Modal	**CAFE (2022)**	0.82	0.84	0.80	0.82
		ILMF-FND (2025	0.91	**0.94**	**0.87**	**0.88**
	Transformer	**TGA (2023)**	**0.92**	0.89	0.86	0.86
	LLM	**GLPN-LLM (2025)**	0.87	0.90	0.86	**0.88**
Twitter	Text-only	**BERT (2018)**	0.71	0.68	0.63	0.65
	Visual-only	**VGG-19 (2014)**	0.56	0.59	0.55	0.57
	Early Fusion	**SpotFake (2019)**	0.78	0.79	0.75	0.77
		EANN (2018)	0.75	0.72	0.77	0.74
	Cross Modal	**CAFE (2022)**	0.81	0.80	0.81	0.81
		ILMF-FND (2025)	0.88	**0.90**	0.91	**0.91**
	Transformer	**TGA (2023)**	**0.90**	0.85	**0.93**	0.89
	LLM	**GLPN-LLM (2025)**	0.88	0.86	0.92	0.89

Above in Table 2 are the benchmarking results. Baseline models were first evaluated to establish reference points. The text-only BERT model consistently outperformed the image-only VGG-19 across both datasets. On the Weibo dataset, BERT achieved an accuracy of 0.76 and an F1-score of 0.75, whereas VGG-19 only reached 0.72 on accuracy and 0.71 on F1. On Twitter, the performance of both baselines declined, reflecting the more diverse and less structured nature of Twitter content. BERT achieved an accuracy of 0.71 and an F1-score of 0.65, while VGG-19's performance dropped to 0.56 on accuracy and 0,57 on F1. These results highlight that textual information is generally more informative for fake news

detection, although relying only on text limits the overall accuracy and robustness of the models.

Early fusion models, which combine textual and visual features at an initial stage, improved performance over the unimodal baselines. SpotFake, an early fusion approach, achieved an F1-score of 0.81 on Weibo and 0.77 on Twitter, demonstrating that integrating multiple modalities provides complementary information useful for fake news detection. EANN obtained an F1-score of 0.79 on Weibo and 0.74 on Twitter, suggesting that early fusion strategies can consistently enhance performance. These improvements were more noticeable on Chinese dataset Weibo, indicating that the effectiveness of early fusion may depend on dataset characteristics like text length, language style, and image usage.

On cross-modal models, CAFE achieved an F1-score of 0.82 on Weibo and 0.81 on Twitter, outperforming the early fusion methods on both datasets. Out of the two, the more effective cross-modal model, ILMF-FND, achieved the higher scores across both datasets, with an F1-score of 0.88 on Weibo and 0.91 on Twitter. This indicates that capturing fine-grained relationships between textual and visual cues allows the model to leverage complementary information more effectively than simple feature fusion.

Transformer model TGA achieved an F1-score of 0.86 on Weibo, while on Twitter it reached 0.89. Although transformer models are effective at capturing contextual information, TGA's performance was slightly lower than ILMF-FND's, particularly on Weibo. Large language model GLPN-LLM also demonstrated strong performance, achieving an F1-score of 0.88 on Weibo and 0.91 on Twitter.

5 Discussion

The findings provide insight into the current state of fake news detection and highlight the strengths and limitations of different modelling strategies. One clear pattern is that multimodal approaches consistently outperform unimodal baseline models. Text-only Bert model generally surpasses image-only VGG-19, demonstrating that textual content contains more reliable cues for identifying misinformation. The inclusion of visual information further improves predictive performance, particularly when cross-modal interactions are explicitly modelled. ILMF-FND proves this by achieving the highest scores on both datasets on average, suggesting that the fine-grained integration of text and image enables the model to capture subtle signals that may not be evident from either modality alone.

The integration strategy of textual and visual cues is a key difference among models. Early fusion models, which combine features at the input stage, improve baseline performance but may fail to capture complex interdependencies between modalities. Cross-modal models, on the other hand, specifically look at how text and images work together, helping the system see when they support each other or give conflicting information. This results in more robust detection, as proved by ILMF-FND's performance. Transformer-based models and LLMs leverage attention mechanisms. They combine information on their own, understanding the context between text and images even without explicitly modelling the interactions. Their strong performance highlights the value of contextual learning, though they may still fall short of specialized cross-modal architectures when fine-grained interactions are critical.

Dataset characteristics also significantly influence model performance. Most of the models achieved higher scores on Weibo than on Twitter, likely due to differences in language complexity, post length, and visual content. Twitter posts are generally shorter, more heterogeneous, and may include memes or other unconventional visual content that is harder for models to interpret. This suggests that models trained and validated on a single platform may not generalize seamlessly across social media networks, emphasizing the need for broader evaluation.

This study confirms that advanced multimodal approaches provide benefits for fake news detection. Simple text or image analysis is insufficient for reliably identifying misinformation, while models that can capture cross-modal relationships or leverage pre-trained

knowledge offer the highest accuracy and robustness. The results also show that the type of dataset matters and that models need to adjust to the differences in content on each platform.

6 Conclusions & Future Work

This thesis has systematically evaluated the performance of state-of-the-art models for fake news detection by benchmarking the models on two different datasets, addressing the research questions set in the beginning of the study:

The findings demonstrate that multimodal approaches are highly effective in detecting fake news, consistently outperforming unimodal baselines. Among all the models, ILMF-FND achieved the highest performance on average with F1-scores of 0.88 on Weibo and 0.91 on Twitter. This shows that carefully modelling the interactions between text and images, rather than simply combining features, significantly improves detection accuracy. Transformer-based models and LLM, TGA and GLPN-LLM, also performed strongly, leveraging attention mechanisms and pre-trained contextual knowledge to capture relationships across modalities. While unimodal models, particularly Text-only BERT, provided a strong foundation, they were outperformed by multimodal architectures, demonstrating that visual cues add complementary information that enhances detection performance.

Regarding how textual and visual information is integrated, the study finds clear differences between approaches. Early fusion models like SpotFake and EANN combine text and image features at the input stage, which improves performance over unimodal models but cannot fully capture complex interactions between modalities. Cross-modal models, such as CAFE and especially ILMF-FND, explicitly model the relationships between text and images, allowing the system to detect when textual and visual content support each other or provide conflicting information. Transformer and LLM-based models integrate multimodal information implicitly through attention and contextual modelling, capturing dependencies without explicitly fusing features. These results indicate that models that carefully model cross-modal interactions provide the most robust and accurate fake news detection.

In comparing multimodal approaches to unimodal methods, the study confirms that text alone is more informative than images alone, but integrating visual content consistently improves performance. On both datasets, multimodal models outperformed Text-only BERT and Image-only VGG-19, highlighting the importance of combining modalities in a complementary way. This addresses the third research question by showing that while unimodal

approaches provide a baseline level of detection, multimodal models are clearly superior in accuracy, recall, and F1-score.

From a theoretical point of view, this thesis provides clear evidence that models combining text and images work better than those using only one type of data or combining them in a simple way. In practice, the results show that social media companies, fact-checkers, and other organizations could improve their fake news detection systems by using these multimodal models. On a wider level, better fake news detection can help stop the spread of misinformation, support more informed discussions, and reduce the negative effects of fake news on society.

Despite the outcomes, several limitations must be considered. The study only evaluated on two datasets, which may not capture the full diversity of social media content, including other languages, platforms, or non-standard visual formats such as memes. In addition, only a limited number of models were tested, even though there are many more approaches proposed in the literature. Many of these models could not be included because their implementations or pretrained weights were not publicly available, which restricted the scope of experimental comparison. The study also focused only on performance metrics and did not examine explainability, interpretability, or user trust. Factors that are increasingly important for practical deployment. Since the implementations used in this study were reproduced by the author with assistance from ChatGPT, there is a possibility of unintentional errors or deviations from the original methods, which may have influenced the results.

Future research should aim to overcome these limitations and continue advancing the field of multimodal fake news detection**.** One important direction is to evaluate models on a wider range of datasets, languages, and social media platforms to better understand how well they generalize across different contexts. This would help ensure that fake news detection systems are not biased toward specific cultures, languages, or content types. Including new types of visual data like GIF's, would also make the evaluations more representative of realistic misinformation. Improving computational efficiency is also a key area for future work. Current state-of-the-art multimodal models often require significant processing power, which limits their use in large-scale or real-time applications. Research into lightweight or optimized architectures could make it possible to deploy these systems on social media platforms where quick response times are critical. Future research should also focus on explainability and user trust. As fake news detection models are increasingly used in public systems, it is

important that their decisions are transparent and understandable to both users and moderators. Integrating explainable AI techniques would help reveal how models make decisions, allowing humans to verify results and increasing trust in automated systems. By addressing these areas, future studies can not only improve the technical performance of fake news detection models but also make them more ethical, accessible, and suitable for real-world use.

In conclusion, this thesis has successfully answered its research questions. Multimodal approaches are highly effective in detecting fake news, the integration of textual and visual cues is most effective when cross-modal interactions are modeled explicitly, and these approaches outperform unimodal methods consistently. By benchmarking multiple state-of-the-art models, the thesis provides insights into model strengths and limitations, emphasizes the importance of multimodal modeling, and lays a foundation for future research and practical applications to combat misinformation in modern social media environments.

References

Abbas, F. & Taeihagh, A. (2024), 'Unmasking deepfakes: A systematic review of deepfake detection and generation techniques using artificial intelligence', Expert Systems with Applications, vol. 252, part B, 124260. Available at: https://doi.org/10.1016/j.eswa.2024.124260

Al-Alshaqi, M., Rawat, D.B. & Liu, C., 2024. Ensemble Techniques for Robust Fake News Detection: Integrating Transformers, Natural Language Processing, and Machine Learning. Sensors, 24, 6062. Available at: https://doi.org/10.3390/s24186062

Allcot, H. & Gentxkov, M., 2017. Social Media and Fake News in the 2016 Election Journal of Economic Persperctives. Available at: https://doi.org/10.1257/jep.31.2.211

Bahad, P., Saxena, P. & Kama, R., 2019. Fake News Detection using Bi-directional LSTM-Recurrent Neural Network. In: International Conference on Recent Trends in Advanced Computing (ICRTAC). Procedia Computer Science, 165, pp.74–82.

Boididou, C., Andreadou, K., Papadopoulos, S., Dang-Nguyen, D. T., Boato, G., Riegler, M. & Kompatsiaris, Y. (2015). Verifying Multimedia Use at MediaEval 2015.

Cao, J., Qi, P., Sheng, Q., Yang, T., Guo, J. & Li, J., 2020. Exploring the role of visual content in fake news detection. *arXiv preprint arXiv:2003.05096*. Available at: https://doi.org/10.48550/arXiv.2003.05096

Chen, Y., Li, D., Zhang, P., Sui, J., Lv, Q., Tun, L. and Shang, L., 2022. Cross-modal ambiguity learning for multimodal fake news detection. In Proceedings of the ACM Web Conference 2022 (WWW '22), Virtual Event, April 25-29, 2022, pp. 2897-2905. Association for Computing Machinery. Available at: https://doi.org/10.1145/3485447.3511968

Chen, W., Cai, F., Guo, Y., Pan, Z., Chen, W. & Zhang, Y., 2025a. Contrastive Learning of Cross-Modal Information Enhancement for Multimodal Fake News Detection. Complex & Intelligent Systems,

Choi, E. & Kim, J.K., 2024. TT-BLIP: Enhancing Fake News Detection Using BLIP and Tri-Transformer. arXiv preprint arXiv:2403.12481.

Del Vicario, M., Bessi, A., Zollo, F., Petroni, F., Scala, A., Caldarelli, G., Stanley, H.E. & Quattrociocchi, W., 2016. The Spreading of Misinformation Online. Proceedings of the National Academy of Sciences of the United States of America, 113(3), pp.554–559. Available at: https://doi.org/10.1073/pnas.1517441113

Gelfert. A. 2018. Fake News: A Definition. Informal Logic. 38(1). 84-117. Available at: https://doi.org/10.22329/il.v38i1.5068

Hamed, S.K.H., Aziz, M.J.A. & Yaakub, M.R., 2023. A Review of Fake News Detection Approaches: A Critical Analysis of Relevant Studies and Highlighting Key Challenges Associated with the Dataset, Feature Representation, and Data Fusion. Heliyon, 9(10), e20382. Available at: https://doi.org/10.1016/j.heliyon.2023.e20382

Helal, N.A., Hassan, A., Badr, N.L. & et al., 2024. A contextual-based approach for sarcasm detection. Scientific Reports, 14, p.15415. Available at: https://doi.org/10.1038/s41598-024-65217-8

Horne, B.D. & Adali, S., 2017. This Just In: Fake News Packs a Lot in Title, Uses Simpler, Repetitive Content in Text Body, More Similar to Satire than Real News. Available at: https://doi.org/10.48550/arXiv.1703.09398

Howerton, G. (2025). "Ten years after Jade Helm, Texas is hooked on conspiracy theories." MSN News. Available at: https://www.msn.com/en-us/news/us/ten-years-after-jade-helm-texas-is-hooked-on-conspiracy-theories/ar-AA1JvJ61

Hu, B., Sheng, Q., Cao, J., Shi, Y., Li, Y., Wang, D. & Qi, P. 2024. Bad Actor, Good Advisor: Exploring the Role of Large Language Models in Fake News Detection. Available at: arXiv:2309.12247v2

Jacobs, L.G., 2022. Freedom of Speech and Regulation of Fake News. The American Journal of Comparative Law, 70(Suppl_1), pp.i278–i311. Available at: https://doi.org/10.1093/ajcl/avac010

Jin, Z., Cao, J., Guo, H. & Zhang, Y., 2017. Multimodal fusion with recurrent neural networks for rumor detection on microblogs. In: Proceedings of the 2017 ACM on Conference on Information and Knowledge Management (CIKM 2017). Available at: DOI: 10.1145/3123266.3123454

Khattar, D., Goud, J.S., Gupta, M. & Varma, V., 2019. MVAE: Multimodal Variational Autoencoder for Fake News Detection. In: The World Wide Web Conference (WWW '19). Association for Computing Machinery, New York, NY, USA, pp.2915–2921. Available at: https://doi.org/10.1145/3308558.3313552

Le Poidevin, O., 2025. UN report urges stronger measures to detect AI-driven deepfakes. *Reuters*, 12 July. Available at: https://www.reuters.com/business/un-report-urges-stronger-measures-detect-ai-driven-deepfakes-2025-07-11/

Li, Z., Li, X., Yang, J., Wang, X., Li, S. & Zhang, J., 2025. Hierarchical Cross-Modal Interaction Network for Multimodal Fake News Detection. Neurocomputing, 647, 130308. Available at: https://doi.org/10.1016/j.neucom.2025.130308

Lin, S.-Y., Chen, Y.-C., Chang, Y.-H., Lo, S.-H. & Chao, K.-M., 2024. Text–Image Multimodal Fusion Model for Enhanced Fake News Detection. Science Progress, 107(4). Available at: https://doi.org/10.1177/00368504241292685

Nasser, M., Arshad, N.I., Ali, A., Alhussian, H., Saeed, F., Da'u, A. & Nafea, I., 2025. A Systematic Review of Multimodal Fake News Detection on Social Media Using Deep Learning Models. Results in Engineering, 26, 104752 Available at: https://doi.org/10.1016/j.rineng.2025.104752

O'Shea, K. & Nash, R., 2015. An Introduction to Convolutional Neural Networks. Aberystwyth University / Lancaster University. Available at: https://arxiv.org/abs/1511.08458

Qi, P., Cao, J., Yang, T., Guo, J. & Li, J., 2019. Exploiting Multi-Domain Visual Information for Fake News Detection. Available at: https://arxiv.org/abs/1908.04472

Qiao, J., Li, X., Gao, C., Wu, L., Feng, J. & Wang, Z., 2025. Improving Multimodal Fake News Detection by Leveraging Cross-Modal Content Correlation. Information Processing & Management, 62(5), 104120. Available at: https://doi.org/10.1016/j.ipm.2025.104120

Shahi, G.K., 2025. Multimodal Misinformation Detection Using Early Fusion of Linguistic, Visual, and Social Features. In: Companion Publication of the 17th ACM Web Science Conference 2025 (WebSci Companion '25). Association for Computing Machinery, New York, NY, USA, pp.11–18. Available at: https://doi.org/10.1145/3720554.3733844

Simonyan, K. & Zisserman, A. (2014) Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

Singhal, S., Kabra, A., Sharma, M., Shah, R., Chakraborty, T. & Kumaraguru, P., 2020. SpotFake+: A Multimodal Framework for Fake News Detection via Transfer Learning (Student Abstract). Proceedings of the AAAI Conference on Artificial Intelligence, 34(10), pp.13915–13916. Available at: (PDF) SpotFake: A Multi-modal Framework for Fake News Detection

Tahmasebi, S., Müller-Budack, E. & Ewerth, R. (2024) Multimodal Misinformation Detection using Large Vision-Language Models. Available at: arXiv:2407.14321v1 [cs.CL], 19 Jul 2024.

Tandoc, E.C., Lim, Z.W. & Ling, R., 2017. Defining "Fake News": A Typology of Scholarly Definitions. Digital Journalism, 6(2). pp.1–17. Available at: https://doi.org/10.1080/21670811.2017.1360143

Vaccari, C. and Chadwick, A. 2020. Deepfakes and Disinformation: Exploring the Impact of Synthetic Political Video on Deception, Uncertainty, and Trust in News. Social Media + Society, 6, 1-13. Available at: https://doi.org/10.1177/2056305120903408

Wang, J., Mao, H. & Li, H., 2022. FMFN: Fine-Grained Multimodal Fusion Networks for Fake News Detection. Applied Sciences, 12(3), 1093. Available at: https://doi.org/10.3390/app12031093

Wang, J., Zhu, Z., Liu, C., Li, R. & Wu, X. (2024) LLM-Enhanced Multimodal Detection of Fake News. Published 24 Oct 2024.

Wang, Y., Ma, F., Jin, Z., Yuan, Y., Xun, G., Jha, K., Su, L. & Gao, J. (2018) EANN: Event Adversarial Neural Networks for Multi-Modal Fake News Detection. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD '18), London, United Kingdom, 19–23 August 2018. ACM, pp. –. Available at: DOI: 10.1145/3219819.3219903.

Yang, P., Ma, J., Liu, Y. & Liu, M., 2023. Multi-modal transformer for fake news detection. Available at: DOI: 10.3934/mbe.2023657

Zhang, L., Wu, Z., Li, Y. & Zhao, J., 2025. KGAlign: Knowledge-Grounded Multimodal Fake News Detection Using Transformer-Based Alignment. Available at: arXiv:2505.14714.

Zhou, X., Wu, J. & Zafarani, R., 2020. SAFE: Similarity-Aware Multi-Modal Fake News Detection. Available at: https://doi.org/10.48550/arXiv.2003.04981

Zhou, X. & Zafarani, R., 2022. Fake News: A Survey of Research, Detection Methods, and Opportunities. ACM Computing Surveys, 1(1), pp.1–40. Available at: https://arxiv.org/pdf/1812.00315v1