3rd Symposium on Security and Privacy in Speech Communication
19 August 2023, Dublin, Irland

How Risky is Multimodal Fake News Detection?

A Review of Cross-Modal Learning Approaches under EU AI Act Constrains

Razieh Khamsehashari[1]†, *Vera Schmitt*[1,2]†, *Tim Polzehl*[2], *Salar Mohtaj*[2], *Sebastian Möller*[1,2]

[1]Technische Universität Berlin
[2]German Research Center for Artificial Intelligence

razieh.khamsehashari@tu-berlin.de, vera.schmitt@tu-berlin.de, tim.polzehl@dfki.de

Abstract

Manual review methods have become insufficient when combating today's scale of online fake news, leading researchers to develop AI-based detection models, many of which struggle with, e.g., multimodal conflicts and ambiguity. Most promising models combine images and textual information in a cross-modal learning strategy. This work summarizes current multimodal fake news detection models, based on cross-modal learning. In order to evaluate if and how they can be applied in real-world use cases, we analyze best-performing models with respect to obligations like risk management, data governance, documentation, transparency, human oversight, and required accuracy, following the European Commission's AI Act. The analysis shows that the AI Act can be applied to a certain extent only, as the categories and their obligations are vaguely defined, leaving room for interpretation when translating the obligations into technical requirements.

Index Terms: Multimodal fake news detection, cross-modal learning, AI Act, risk-based approach, general-purpose AI

1. Introduction

The Ukraine war [1] and the *infodemic* during the Corona pandemic [2] provided concrete examples of the significant impact that fake news can have on shaping public opinions [3]. The widespread sharing of multimedia content on social media has led to the rapid spread of fake news, posing significant threats to societal stability and security. Consequently, there has been considerable research interest in the field of social forensics, specifically in the area of fake news detection [4]. Furthermore, the dissemination of false information has raised deeper concerns about the escalating political polarization, the erosion of democracy, and the declining trust in public institutions caused by the widespread prevalence of fake news [5, 6].

Given the enormous volume of social media posts generated by hundreds of millions of users, traditional manual review methods are impractical for effectively addressing the threat of fake news. Consequently, researchers in the field of computer science have dedicated their efforts to developing methods for detecting fake news [7, 4]. In recent years, various stakeholders, including industry, government institutions, and civil society, have collaborated to develop policies, automated detection tools, and enforcement frameworks to address deceptive actors and false content on the internet. Researchers have also proposed technological solutions for detecting fake news [8]. Existing methods primarily focus on combining textual and visual features but struggle to leverage multi-modal information at both detailed and broad levels effectively. Additionally, they encounter challenges due to a lack of correlation between different modalities or conflicting decisions made by each modality, leading to ambiguity [4]. Despite significant advancements in AI, it remains susceptible to attacks [9] and biases [10]. Moreover, the current rise of generative AI, such as ChatGPT, sparks the discussion about the rightful and ethical handling of fake news even more. Moreover, generative AI models show some of the major shortcomings of the proposed regulatory rules on AI from the European Commission (EC) to ensure the legal, effective, ethical and safe use of trustworthy AI systems, as pointed out by Reuters [11]. Two years ago, the EC first attempted to regulate AI [12] by proposing the Artificial Intelligence (AI) Act. Hereby, AI can be categorized according to different risk levels ranging from *minimal or no risk* to *unacceptable risk*. This risk-based approach was developed to address safety-critical applications to follow a set of obligations depending on the risk level they are categorized in. When considering current AI-based solutions to detect fake news in multimodal inputs, mainly focusing on image and text, it remains unclear which risk category they would apply to. Given that the handling of fake news intersects with fundamental human rights, including *freedom of expression*, it is vital to establish objective and transparent criteria for the detection and management of fake news.

Therefore, the contribution of this paper is (1) a review of the state-of-the-art best-performing multimodal fake news detection models for image and text, (2) an analysis of the risk level they would apply under the proposed AI Act of the EU, and (3) and analysis of the applicability and consequences of the risk-based categorization of AI systems. Thus, this paper tries to answer the following research questions:

1. RQ: What are the essential architectures and evaluation scores of current best-performing multimodal fake news detection AI models?
2. RQ: What risk categories do the models fall into and what are the consequences for societies and AI providers, e.g. with respect to risk management, data governance, documentation, transparency, human oversight, and required accuracy?

In the following, Section 2 the state-of-the-art multimodal fake news detection models are described by categorizing them into different modeling approaches. In Section 3 the best performing models are categorized in one of the risk categories proposed in the AI Act and obligations are described providers of such systems need to comply with. Finally, the findings are discussed and concluded in section 4.

2. Multimodal Fake News Detection

There are various modeling techniques for multimodal fake news detection when analyzing text and image inputs together. In the following section, we will (1) introduce a definition of

† Both authors contributed equally.

10.21437/SPSC.2023-1

fake news, (2) describe different learning approaches for the multimodal fake news detection task, and (3) compare different models based on their performances when trained on different datasets.

2.1. Definition of Fake News

Defining fake news is a complex task influenced by a country's constitutional and legal framework, culture, political context, and public awareness regarding the issue. The challenge lies in the fact that fake news often exists in a grey area of political expression that encompasses both misinformation and disinformation [13]. Additionally, the term fake news is frequently misused to discredit opinions and information that do not align with one's own perspective. This presents a challenge in implementing effective strategies to address fake news while safeguarding fundamental rights, such as freedom of expression (Article 11 in the EU Charter of Fundamental Rights), data protection (Articles 7 and 8), and media pluralism. In general, fake news can be classified into disinformation and misinformation. Misinformation refers to unintentionally false and inaccurate information shared by individuals [14]. Meanwhile, the EC High-Level Expert Group (HLEG) defines disinformation as "verifiable false or misleading information created, presented, and spread for economic gain or to intentionally deceive the public, potentially causing harm" [15]. In the following, the term fake news refers to this definition.

2.2. Multimodal Fake News Detection Approaches

In recent years, the rapid spread of fake news online has led to a growing interest in the automatic detection of multimodal fake news. In social media and also online media outlets, news are shared, combining image and also textual information. Thus, in most cases, fake news detection (FND) is a multi-modal problem. Recently, many existing approaches primarily focus on integrating unimodal features to produce multimodal news representations [20, 22, 31]. However, the effective aggregation of features from different modalities to enhance the overall performance is still an open question. Cross-modal learning is essentially better for capturing cross-modal consistency and feature correlations between different modalities to achieve accurate fake news detection [4, 16]. Building upon the insights gained from the existing approaches, we investigate different modeling techniques that have shown promising results in improving the overall performance of multimodal fake news detection. These techniques encompass the effective aggregation of features from different modalities, with a particular focus on cross-modal feature correlations.

Table 1 provides an overview of the most recent cross-modal learning approaches in FND. It is evident from the table that various models have been employed to address the multimodal FND, aiming to aggregate unimodal features and produce comprehensive representations of news. Building upon the insights gained from the existing approaches, we investigate different modeling techniques that have shown promising results in improving the overall performance of multimodal FND. These techniques encompass the effective aggregation of features from different modalities, with a particular focus on cross-modal feature correlations.

2.2.1. Modeling Techniques

Contrastive learning: Aims to learn effective representation by contrasting pairs of data points. In the case of cross-modal contrastive learning, pairs of data points are drawn from different modalities with the objective of mapping them into a shared latent space where similar pairs are brought closer together while dissimilar pairs are pushed apart. One notable advantage of cross-modal contrastive learning is its ability to learn representations without requiring explicit alignment of modalities, making it suitable for cases where data is not perfectly aligned or where certain modalities have missing information. Cross-modal contrastive learning has been successfully applied to a variety of tasks, including image-text retrieval [32], audio-visual speech recognition [33], and multimodal sentiment analysis [34]. This approach, which has also been adapted to vision-language multimodal representation learning, has demonstrated promising results and holds potential for application in various other cross-modal learning tasks.

Contrastive Language-Image Pre-training (CLIP) [35], a simple contrastive learning method, is employed to learn transferable visual representations by maximizing cosine similarity between the image and text modalities. This approach facilitates the zero-shot transfer of the model to various downstream tasks. FND-CLIP [24] utilizes two pre-trained CLIP encoders to extract deep representations from both images and modalities. These extracted CLIP features are then employed as a multimodal representation and aggregated using modality-specific attention mechanisms. In contrast, COOLANT [16] follows a similar dual-stream design as CLIP, incorporating a cross-modal contrastive learning module. However, it distinguishes itself by introducing an auxiliary cross-modal consistency learning task that measures semantic similarity between images and texts. It then provides soft targets for the contrastive learning module, leading to effective FND. Multi-grained Multimodal Fusion Network (MMFN) [4] is another model based on contrastive learning that employs two Transformer-based pre-trained models to encode features. MMFN uses BERT and Swin-T with a Transformer to obtain fine-grained multi-modal text-based and image-based features and uses CLIP [35] to acquire coarse-grained multi-modal features. DGM [21] aims to detect the authenticity of multi-modal media and also ground the manipulated content. To address this problem, a hierarchical multi-modal manipulation reasoning transformer is introduced. The proposed model includes two uni-modal encoders for image and text data, a multi-modal aggregator, and a manipulation detection on top. To improve the performance of the model on exploiting the semantic correlation of images and texts, image and text embeddings are aligned through cross-modal contrastive learning.

Cross-modal knowledge distillation: Knowledge Distillation (KD) [36] is an effective technique in deep learning that facilitates the training of a smaller student network under the supervision of a larger teacher network. It was initially introduced by [37] and further expanded by [38]. KD serves as a promising solution for integrating multimodal data, allowing information transfer between networks when training constructively. Building upon the concept of deep mutual learning, where an ensemble of networks can learn collaboratively and teach each other during training, Wei et al. [23], propose Cross-Modal Knowledge Distillation (CMC), for multi-modal FND. This model involves training two single-modal networks in a mutual manner. By employing a cross-modal distillation objective function, CMC facilitates the learning of feature correlations between different modalities by guiding the single-modal networks with a soft target.

Multimodal progressive fusion: In this method, information from various modalities is gradually fused over multiple

Table 1: *Summary of the most recent works on FND, covering two modalities of text and image.*

Model	Year	Method
COOLANT [16]	2023	Cross-modal contrastive learning
MMFN [4]	2023	Multi-grained information fusion
MPFN [17]	2023	Multimodal progressive fusion
SAMPLE [18]	2023	Similarity-aware multimodal prompt learning
TieFake [19]	2023	Integration of multimodal context and author sentiment: focusing on title-text similarity and emotion awareness
FNR [20]	2023	Similarity and transformer-based learning
DGM [21]	2023	Transformer based on manipulation-aware contrastive learning and modality-aware cross-attention
CAFE [22]	2022	Cross-modal ambiguity learning
CMC [23]	2022	Cross-modal knowledge distillation
FND-CLIP [24]	2022	Contrastive language-image pretraining-guided learning
LIIMR [25]	2022	Leveraging intra and inter modality relationship
FMFN [26]	2022	Fine-grained multimodal fusion network
MCAN [27]	2021	Multimodal co-attention networks
AMFB [28]	2021	Attention-based multimodal factorized bilinear pooling
HMCAN [29]	2021	Hierarchical multi-modal contextual attention network
CARMN [30]	2021	Crossmodal attention residual and multichannel CNN
SAFE [31]	2020	Cross-modal similarity measurement

stages or layers of a neural network. The main idea is to enable the model to progressively integrate information from different modalities, instead of attempting to do so in a single step. For instance, in the context of video analysis, the model initially processes the visual data from individual frames independently and subsequently combines it with the audio data across multiple layers, allowing a step-by-step integration process. In the domain of FND, Jing et. al proposed the Multimodal Progressive Fusion Network (MPFN) [17] which comprises three key components: a text feature extractor utilizing a pretrained BERT model [39], a visual feature extractor combining Swin Transformer [40] and VGG19 [41] for spatial and frequency domain information extraction, and a progressive multimodal feature fusion process.

Knowledge prompt learning: In recent years, prompt learning has emerged as a relatively new approach in cross-modal learning that focuses on acquiring task-specific representations by conditioning the model on a context-providing prompt. This approach draws inspiration from recent advancements in natural language processing, where prompt-based learning has shown enhanced performance across various language tasks [42, 43]. Furthermore, prompt-based models have been employed in FND, demonstrating their potential in this domain as well. For example, in the Similarity-Aware Multimodal Prompt Learning (SAMPLE) framework [18], three popular prompt learning methods (discrete prompting, continuous prompting, and mixed prompting) are integrated into a soft verbalizer. It utilizes CLIP [35] to extract text and image features and generates a multimodal representation. The framework addresses the issue of uncorrelated semantic representation between image and text by calculating semantic similarity and normalizing it to adjust the intensity of the aggregated multimodal representation.

Cross-modal attention mechanisms: Attention mechanisms have demonstrated considerable efficacy in different tasks such as image captioning [44], machine translation [45], and recommendation system [46]. Their ability to capture fine-grained relevance across different modalities has led to significant improvements, particularly with cross-modal attention mechanisms. Recently, attention mechanisms have been integrated into FND methods. This approach leverages attention mechanisms to selectively emphasize specific modalities within a multi-modal dataset, based on their relevance to the specific task. For example, in FND, attention may be directed towards the text and image modalities, while disregarding less relevant modalities.

TieFake [19] introduces a novel method for detecting fake news, which incorporates title-text similarity and the author's subjective emotion. It employs a scaled dot-product attention mechanism to capture the similarity between the title and text effectively. By leveraging the news text, images, subjective emotion of the author, and the similarity between the title and text, this method shows promising results well. AMFB [28] proposes an attention-based stacked bi-directional LSTM network that captures textual information at different levels, attention-based multi-level convolutional neural network–recurrent neural network for visual feature extraction, multimodal factorized bilinear pooling to combine the textual and visual feature representations, and then passes them through a multi-layer perceptron for FND. Wang et al. [26] introduce a Fine-grained Multimodal Fusion Network (FMFN) to fuse textual and visual features fully. The proposed model utilizes scaled dot-product attention mechanisms for the fine-grained fusion of the textual and the visual features, which not only takes into account the correlations between different visual features but also captures the dependencies between textual and visual features. In CARMN [30], a multimodal fake news detection model based on Cross-modal Attention Residual Network (CARN) and Multichannel Convolutional Neural Network (MCN) is proposed. The CARN is utilized to fuse relevant information across modalities while preserving the unique characteristics of each modality. To address the potential noise generated during cross-modal fusion, the MCN is employed to extract feature representations from both the original and fused textual information simultaneously. Qian et. al [29] propose a hierarchical multi-modal contextual attention network (HMCAN). The model utilizes BERT to generate hierarchical semantic representations of text and utilizes ResNet to generate regional representations of images. Then different levels of semantics are fed into co-attention layers with regional image features to achieve hierarchical multi-modal feature fusion. Finally, a hierarchical encoding network captures hierarchical semantics for fake news detection. Multimodal Co-Attention Networks (MCAN) [27] further employ three different sub-networks to extract spatial-domain features and frequency-domain features from images, along with textual features from the text. These three types of features are then fused through stacking co-attention layers to learn inter-modality relations. Finally, the fused representation obtained from the last co-attention layer is used for FND.

Table 2: *Performance comparison on Weibo, Twitter, Gossipcop, and PolitiFact datasets. The best performance is highlighted in bold, and the second rank is indicated with blue color.*

Dataset	Model	Accuracy	Fake News			Real News		
			Precision	Recall	F1-score	Precision	Recall	F1-score
Weibo	COOLANT	**0.923**	0.927	0.923	**0.925**	0.919	0.922	0.920
	MMFN	**0.923**	0.921	**0.926**	0.924	**0.924**	0.920	**0.922**
	CMC	0.908	**0.940**	0.869	0.899	0.876	**0.945**	0.907
	FND-CLIP	0.907	0.914	0.901	0.908	0.914	0.901	0.907
	FNR	0.879	0.87	0.89	0.88	0.88	0.87	0.88
	LIIMR	0.900	0.882	0.823	0.847	-	-	-
	MCAN	0.899	0.913	0.889	0.901	0.884	0.909	0.897
	HMCAN	0.885	0.920	0.845	0.881	0.856	0.926	0.890
	FMFN	0.885	0.878	0.851	0.864	0.874	0.896	0.885
	CARMN	0.853	0.891	0.814	0.851	0.818	0.894	0.854
	CAFE	0.840	0.855	0.830	0.842	0.825	0.851	0.837
	MPFN	0.838	0.857	0.894	0.889	0.873	0.863	0.876
	AMFB	0.829	0.86	0.90	0.88	0.75	0.68	0.71
	MCNN	0.823	0.858	0.801	0.828	0.787	0.848	0.816
	SAFE	0.816	0.818	0.815	0.817	0.816	0.818	0.817
Twitter	COOLANT	0.900	0.879	**0.922**	0.900	0.923	0.880	0.901
	MMFN	**0.935**	0.960	0.856	**0.905**	**0.924**	**0.980**	**0.951**
	FNR	0.789	0.78	0.85	0.82	0.79	0.71	0.75
	LIIMR	0.831	0.836	0.832	0.830	-	-	-
	HMCAN	0.897	**0.971**	0.801	0.878	0.853	0.979	0.912
	CAFE	0.806	0.807	0.799	0.803	0.805	0.813	0.809
	MPFN	0.833	0.846	0.921	0.880	0.809	0.721	0.740
	AMFB	0.749	0.76	0.79	0.78	0.73	0.70	0.71
GossipCop	MMFN	**0.894**	0.799	0.598	0.684	0.910	**0.964**	**0.936**
	TieFake	0.892	**0.887**	**0.902**	**0.894**	-	-	-
	CMC	0.893	0.826	0.657	0.692	**0.920**	0.963	0.935
	FND-CLIP	0.880	0.761	0.549	0.638	0.899	0.959	0.928
	CAFE	0.867	0.732	0.490	0.587	0.887	0.957	0.921
PolitiFact	TieFake	0.912	**0.931**	**0.909**	**0.920**	-	-	-
	CMC	0.894	0.806	0.862	0.833	0.944	0.92	0.932
	FND-CLIP	**0.942**	0.897	0.897	0.897	**0.960**	**0.960**	**0.960**

2.3. Datasets

One of the primary challenges faced when working with diverse modalities such as text and image is the challenge of accessing a suitable dataset. Additionally, it is necessary to integrate data from multiple platforms, such as news, Twitter posts, and social media, as each data source exhibits distinct styles and focuses on diverse topics.

The Weibo dataset [47], derived from China's popular social media platform Weibo, is widely utilized in fake news detection. It consists of a comprehensive collection of real and fake news samples, each accompanied by relevant text, images, and labeling information. This dataset has been extensively employed in recent studies to evaluate the effectiveness of multimodal fake news detection strategies. For the construction of the training set, real news items were gathered from the authoritative Xinhua News Agency, while fake news was acquired from the official fake news debunking system of Weibo, spanning the period from May 2012 to January 2016. The training set comprises a total of 7,532 news, with 3,749 representing fake news and 3,783 non-fake news; the test sets contain 1,996 posts.

The Twitter dataset [48], released for the MediaEval Verifying Multi-media Use 2015 [49], is a well-known multi-modal dataset used for fake news detection. The dataset includes sequential content of tweeted texts and images, with 6,000 rumor posts and 5,000 true posts in the training set. The test set consists of up to 2,000 posts featuring various types of breaking news. Posts containing only images or text are excluded from the training and testing processes.

The Politifact and Gossipcop datasets [50], sourced from the political and entertainment domains of the FakeNewsNet repository, are English datasets commonly used in fake news research. They contain news articles published between May 2002 to July 2018 and July 2000 to December 2018, respectively. The Politifact dataset consists of 244 real news and 135 fake news in the training set and 75 real news, and 29 fake news in the test set. The Gossipcop, on the other hand, contains 10,010 training news, of which 2,036 are fake news, and 7,974 are real news. The test set contains 2,285 real news and 545 fake news.

COOLANT, **MMFN**, and **TieFake** show the best performances for multimodal fake news detection for image and text modalities. However, with the upcoming regulations within the EU, there needs to be a discussion on how to apply such models for real-world use cases and what legal requirements they have to follow.

2.4. Performance Comparison

Table 2 presents the performance comparison of cutting-edge methods and the corresponding outcomes on four widely used datasets: Weibo, Twitter, GossipCop, and PolitiFact. Accuracy,

precision, recall, and F1-score are the evaluation metrics which are typically used to evaluate the performance of a binary classification task. In the following, we analyze the performance of different approaches by highlighting the top-performing models across all datasets.

2.4.1. Most outstanding models

As shown in Table 2, the top-ranked performance models are highlighted in bold for the first rank and in blue color for the second rank. Specifically, MMFN achieves the highest accuracy of scores of 92.3% (tied with COOLANT), 93.5%, and 89.4% on the three datasets, outperforming the state-of-the-art by margins of 1.5%, 3.5%, and 0.1% respectively. Moreover, MMFN demonstrates its effectiveness by securing either the first or second rank in 61.1% of the different metrics, encompassing precision, recall and F1 score, across the majority of evaluations. COOLANT consistently achieves a position within the top two ranks for accuracy and maintains a notable presence in 75% at all remaining metrics. On the GossipCop and PolitiFact datasets, TieFake outperforms all the baselines on precision, recall, and F1-score of fake news.

Table 3 presents the top three models with the best performance across the four datasets. To highlight their respective strengths, we utilize checkmarks to indicate the datasets on which each model excels. It can be observed that **COOLANT**, **MMFN**, and **TieFake** are the best-performing models in multimodal FND processing of both modalities, text and image. Additionally, in terms of modeling techniques, both COOLANT and MMFN demonstrate the utilization of **cross-modal contrastive learning** in their framework, highlighting the efficacy of this approach compared to other cross-modal learning methods.

Table 3: *The top three models with the best performance on the Weibo, Twitter, GossipCop, and PolitiFact datasets.*

Dataset	COOLANT	MMFN	TieFake
Weibo	✓	✓	
Twitter	✓	✓	
GossipCop			✓
PolitiFact			✓

Based on the analysis provided in Section 2, the three best-performing models **COOLANT**, **MMFN**, and **TieFake** can be recommended for the multimodal fake news detection task. These three models will be used in the following to analyze the applicability of the risk-based approach proposed in the AI Act and the obligations which need to be considered when using them in real-world use cases.

3. Multimodal Fake News Detection Aligned with the AI Act

Before analysing the legal requirements for the best performing models mentioned above, the AI Act and related concepts are first introduced. The European Union (EU) is currently engaged in a comprehensive regulatory initiative called the AI Act, which aims to establish a comprehensive framework for the regulation of AI. The European Parliament have given its support to new regulations aimed at promoting transparency and risk management in the development of Artificial Intelligence (AI) systems in Europe, with a strong emphasis on ensuring a human-centric and ethical approach.

3.1. AI Act: Risk-based Approach

The regulatory framework under consideration extends its coverage to AI applications across both public and private sectors. It applies to systems sold within the EU market or having an impact on EU citizens. Its primary objective is to provide comprehensive guidance to AI developers, deployers, and users by specifying precise requirements and obligations for diverse applications of AI systems. The development of a risk-based approach has been facilitated through extensive consultations with key stakeholders, including the High-Level Expert Group on AI. This approach effectively recognizes the inherent advantages and potential of AI while also taking into account the potential hazards and risks associated with emerging AI applications and systems. The regulation provides a comprehensive definition of AI in *Article 3*, encompassing AI systems that produce outputs influencing their environment, such as content, predictions, recommendations, or decisions. This definition includes various components, such as machine learning (supervised, unsupervised, reinforcement, and deep learning), logic- and knowledge-based approaches (inductive logic programming, knowledge representation, inference, and deductive engines), as well as statistical methods like Bayesian estimation and search optimization [51]. Consequently, the concepts of disinformation and hate speech align with the broad definition of AI within the proposed EC regulation.

The risk-based approach within the AI Act includes four different risk categories, where the distinction of AI systems falling into the *high-risk* or *limited risk* category are of most interest for the application of *deep learning* models for the fake news detection domain. In a sensitive domain such as fake news detection often resulting in actions what kind of information might be flagged, removed, or blocked, there is a risk of interference with the freedom of expression, therefore the application and resulting actions based on classification or prediction results need to be carefully evaluated based on the four risk categories, to verify which obligations need to be followed in order to align with the AI Act.

The four risk categories for AI systems and their corresponding obligations entail:

- **Unacceptable Risk:** AI systems presenting evident dangers to individuals' safety, rights, and well-being, such as government social scoring systems or hazardous voice-assisted toys, are prohibited from the European market.
- **High Risk:** AI systems developed for domains critical to human life and well-being, including infrastructure, education, safety components, law enforcement, and public services, fall under this category. Strict obligations, outlined in *Chapter 2 and 3* of the AI Act [12], must be followed before these high-risk AI systems can be put on the EU market. These obligations include high-quality datasets, risk assessment, traceability, accuracy, robustness, security, user information, human oversight, and conformity assessment.
- **Limited Risk:** AI applications requiring transparency obligations to make interactions with AI systems more comprehensible for human users fall into this category. Four obligations mainly focus on informing users about their interaction with an AI system or AI-generated content, such as audio or video content (e.g., deepfakes).
- **Minimal Risk:** This category encompasses AI systems developed for domains such as AI-enabled video gaming, spam filters, and applications that do not harm human users. No further obligations are defined for the minimal risk category.

In response to the evolving nature of technology, authorities have included a provision that emphasizes the need for ongoing assessment of the risk classification of AI systems. The EU is directed to take into account the "intended purpose of the AI system" when assessing the risk category of AI systems [12]. This provision highlights a significant concern that arises from the fact that AI systems might circumvent or avoid the safeguards outlined in the Act. This is due to the intricate relationship between the developers, deployers, tasks performed by the AI systems, and the specific purpose(s) they serve as a product [52].

3.2. General Purpose AI Systems

Generative AI, particularly Large Language Models (LLMs), have garnered significant attention due to notable advancements, with ChatGPT serving as a prominent example [53]. When examining the respective risk category and obligations of AI systems in the disinformation detection domain, it needs to be clarified first, if such AI systems fall under the category of *General Purpose AI Systems* (GPAIS)[1]. GPAIS are not developed for a specific task and can be used for versatile application domains. However, often GPAIS, such as LLMs are often used by fine-tuning the last layer for a specific purpose, which makes it difficult to categorize various types of LLMs used for specific purposes as GPAIS or not [53]. There are various ways how to define GPAIS as they exhibit significant variations in terms of autonomy, agency, modality, and training methods [54]. An amendment to the draft AI Act, specifically Articles 4a-4c, concerning GPAIS, was circulated by the French Council presidency on May 13, 2022, defining GPAIS as systems "intended by the provider to perform generally applicable functions such as image and speech recognition, audio and video generation, pattern detection, question answering, translation and others; a general purpose AI system may be used in a plurality of contexts and be integrated in a plurality of other AI systems" (Article 3(1b) AI Act) [53]. GPAIS are required to adhere to high-risk obligations (such as those outlined in Articles 8 to 15 of the AI Act) if they have the potential to be used as high-risk systems or as components of such systems (as specified in Article 4b(1) and 4b(2) of the AI Act).

The current definition of GPAIS in the AI Act does not align with the requirement for broad generality in abilities, tasks, or outputs beyond integration into different use cases. The existing rule treats every simple image or speech recognition system as qualifying, disregarding their capabilities, which contradicts the technical literature on GPAIS. The issue arises from the language used in Article 3(1b) of the AI Act, where the use in different contexts and AI systems is presented as possible examples rather than necessary conditions. To accurately define truly general-purpose systems, the definition should be revised to make use in different contexts and AI systems necessary conditions, not merely sufficient ones. Additionally, the definition should emphasize the significant generality in ability, task, or output, with relevance as a priority. This would require models with a single set of abilities and tasks to demonstrate highly diverse output to be classified as GPAIS, while multimodal models would generally qualify even if they are limited to a specific task without significant output variation [53].

Consequently, according to the current definition given, the three top-performing models for multimodal fake news detection in Section 2 **COOLANT**, **MMFN**, and **TieFake** would fall under the category of GPAIS, due to their multimodal processing of information and because they include pre-trained LLMs and computer vision models:

- **COOLANT:** the cross-modal contrastive learning framework for multimodal fake news detection extracts the first unimodal features by the modal-specific encoder. For the unimodal encoders the transformer-based LLM BERT [55] is used. To learn image representations the large pre-trained Convolutional Neural Network (CNN) ResNet [56] is used.
- **MMFN:** similar to **COOLANT**, **MMFN** uses BERT to learn unimodal representations from text and deploys Swin-T, a general-purpose model for computer vision [57] for learning unimodal image representations.
- **TieFake:** also deploys BERT for learning textual representations. A new variant of the ResNet model, ResNetSt [58], with an additional *split attention module* is used for learning image representations.

All models (BERT, ResNet, Swin-T, ResNeSt) used for learning textual and image representations in the best performing multimodal fake news detection frameworks, fall under the definition of GPAIS, outlined in Article 3(1b) AI Act [53].

3.3. Risk-Assessment of Multimodal for Fake News Detection Frameworks

In the event that any of the three models, namely **COOLANT**, **MMFN**, and **TieFake**, are utilized (even as components) for high-risk applications, they must be classified as *high-risk* AI systems according to the definition of General Purpose Artificial Intelligence (GPAIS). In order to determine whether the models are employed for *high-risk* applications, it is essential to incorporate the actions resulting from the model outputs into the risk assessment of multimodal fake news detection systems. According to the European Parliament's report by Marsden (2019) [13] on regulating false content, there are several potential actions that can be taken as a result of detecting false content. These actions include **filtering**, **blocking**, **(de)prioritisation**, **flagging**, and **disabling**. These approaches can be employed to address the presence of false information and mitigate its impact. According to Schmitt et al. (2021) [59], fake news detection systems resulting in actions like **filtering**, **(de)prioritization**, and **flagging** can be categorized under the *limited risk* category, as these actions, when carefully designed, do not affect free speech and media pluralism directly. However, categorized as *limited risk* AI systems, the multimodal fake news detection frameworks mentioned above need to follow the *transparency obligation* outlined in *Article* 52 of the AI Act. The *transparency obligation* requires informing natural persons interacting with the system, to inform them about that they are interacting with an AI system and make the reasons for the decisions transparently available and how the respective **filtering**, **(de)prioritization**, or **flagging** mechanisms work.

When the actions **blocking**, or **disabling** results from detecting fake content, the analysis from Schmitt et al. (2021) [59] recommends categorizing fake news detection models into the *high-risk* category, as free speech and media pluralism are affected. In case the *high-risk* category applies, the current proposal requires providers of AI systems to comply with a more extensive list of obligations (outlined in Chapter 2, *Article* 9-15). The following obligations need to be followed from a technical perspective:

1. **Risk-management system** (*Article* 9): providers are obligated to establish a risk management system that includes the

[1] Also referred to as *AGI*, i.e. Artificial General Intelligence.

following components: formalizing potential risks, assessing their impact on users, and implementing strategies to test various countermeasures against identified risks. Within the context of fake news detection, a notable concern arises when the models produce *false positives*, incorrectly identifying information as false when it is, in fact, true. When *false positives* get **blocked**, or spreaders get **disabled**, then the human right of *freedom of expression* granted by all European member states is negatively affected. Thus, the risk-management system needs to address all potential risks which can lead to a negative effect on the human right of *freedom of expression*, which are *inter alia* addressed in the *Articles 10-15*.

2. **Data governance and management** (*Article* 10): **COOLANT**, **MMFN**, and **TieFake** are trained on large datasets which need to be reconsidered carefully for the fake news detection task. *Article 10* requires providers of *high-risk* AI systems to deploy appropriate data governance and management practices. Hereby, the process of data collection, data cleaning, and pre-processing such as annotation, labeling, cleaning, and data enrichment needs to be carefully designed and evaluated. Moreover, a prior assessment of the availability, quantity, and suitability of the data sets needs to be conducted, and relevant assumptions, with respect to the information that the data are supposed to measure and represent, need to be formulated. Data privacy in terms of *Article 9(1)* of Regulation (EU) 2016/679, *Article 10* of Directive (EU) 2016/680 needs to be guaranteed as soon as personal-related data is included in datasets. Lastly, relevant biases in the data should be examined. The data governance and management requirements can be fulfilled for the additional provided datasets, which are used for fine-tuning the models with respect to their intended application, where fake news datasets can be analyzed based on the requirements outlined above. However, in the case of using pre-trained models such as BERT, ResNet, Swin-T, and ResNeSt for multimodal fake news detection, it is very difficult to fulfil the data governance requirements outlined in *Article 10*. First, the providers of the pre-trained LLMs and CNNs need to address these requirements when publishing the models, even though these GPAIS do not directly fall under the *high-risk* AI category. Here, a more fine-grained separation between requirements of pre-trained GPAIS and their context-specific requirements is still missing, to make the application of *Article 10* more applicable to real-world use cases.
3. **Technical documentation** (*Article* 11): it mandates compliance with all obligations specified in Chapter 2 and necessitates the completion of technical documentation before an application is introduced to the EU market. Furthermore, continuous updates to the documentation are required. Providers are also obligated to provide all necessary information to national competent authorities to assess compliance with the AI Act. Similar to the challenges presented in *Article 10*, this documentation can only be accomplished for retraining the pre-trained GPAIS as it is only feasible to document the specific implementation for particular use cases.
4. **Record-keeping** (*Article* 12): *high-risk* AI systems must have logging capabilities that automatically record operating events. These logging capabilities should adhere to recognized standards or common specifications. The logging capabilities must include at least: (a) recording the duration of each system use (start and end date and time); (b) referencing the database used for input data verification; (c) logging the input data that resulted in a match; (d) identifying the individuals involved in result verification, as mentioned in *Article 14(5)*. This can be fulfilled for the data inputs used in any *downstream task* for which, in this case, BERT, ResNet, Swin-T, and ResNeSt are used. However, here similar challenges are faced when it is required to also apply the record-keeping requirement to the pre-trained GPAIS.
5. **Transparency and provision of information to users** (*Article* 13): High-risk AI systems must be designed to ensure sufficient transparency in their operation, enabling users to interpret and utilize the system's output appropriately. User instructions for high-risk AI systems must be concise, comprehensive, accurate, and easily understandable, provided in a suitable digital format or otherwise, ensuring relevance, accessibility, and comprehensibility for users. Hereby, several requirements are listed referring to the identity of the provider, characteristics, capabilities, and limitations of *high-risk* AI systems. Moreover, performance requirements and the level and form of human oversight (defined in*Article 14*) need to be specified by the provider, and the insurance of proper functioning of the AI system needs to be provided. These requirements are vaguely defined, which makes it difficult to define the right level e.g. of human oversight, how deep transparency obligations need to be designed, and about what exactly the user needs to be informed. Human oversight can be designed on various levels. For the data collection and pre-processing task, for the model development and training task, and for the interaction with users part. Also, the level of human involvement is difficult to determine from the descriptions given in *Article 13*. For the multimodal fake news detection use case, human oversight can be designed in various different ways, e.g. a human needs to check each classified fake news item, or only those with low accuracy, or only when a user reports an issue. Thus, first, it needs to be clarified where human oversight is necessary, then how deeply the human should get involved, and lastly, what level of expertise is relevant for the task the human needs to review and verify. This basic conceptualization is necessary to then develop XAI features that allow for transparency for the model outputs on a level that the reviewer and the user understand and can react to in critical cases. For example, when using **COOLANT**, **MMFN**, and **TieFake** for the fake news detection task and *Shapley-based* explanations are generated for the text-based outputs [60], they might be not meaningful for users, who do not have any computer science background. Therefore, transparent design and providing information on deep-learning frameworks is only useful to a certain extent and needs to be specified for the respective context, reviewer, and user of such systems.
6. **Human oversight** (*Article* 14): *high-risk* AI systems should be designed and developed with effective human-machine interface tools to enable oversight by individuals while the system is in use. Human oversight aims to prevent or minimize risks to health, safety, and fundamental rights that may arise from the use of *high-risk* AI systems. This oversight is particularly important when such risks persist despite meeting other requirements outlined in Chapter 2. In the case of **blocking**, or **disabling** fundamental rights are affected which makes it necessary, that for the fact-checking use case resulting in either of the two actions, a human needs to be integrated into the decision-making process when content gets blocked or users are disabled from further distributing information online. The provided measures aim to enable individuals assigned to human oversight to: (a) fully understand the capabilities and limitations of the *high-risk* AI system and ef-

fectively monitor its operation to promptly detect and address any anomalies, dysfunctions, or unexpected performance; (b) be aware of the potential risk of automation bias, which refers to the tendency to automatically rely on the system's output; (c) correctly interpret the system's output, considering its characteristics and the available interpretation tools and methods; (d) disregard, override or reverse *high-risk* AI system outputs; (e) intervene or interrupt the functioning of the *high-risk* AI system. These points can be followed for a human-in-the-loop for collaborative decision-making for the multimodal fake news detection task. However, as outlined in *Article 13* the transparency obligation and integration of human oversight is often not straightforward and needs to be carefully designed to address the actual challenges in preventing the infringements of fundamental human rights, such as *freedom of expression*.

7. **Accuracy, robustness, and cybersecurity** (*Article* 15): *high-risk* AI systems must be designed and developed to achieve an appropriate level of accuracy, robustness, and cybersecurity in line with their intended purpose. These systems should maintain consistent performance in these aspects throughout their lifecycle. Moreover, such a system should be resilient to errors, inconsistencies, bias, and cyberattacks. In the case of multimodal fake news detection, these points can be followed by defining a performance level (e.g. $accuracy > 0.90$, which applies to **COOLANT**, **MMFN**, and **TieFake** only partially) tested on multiple datasets, which models need to fulfill before put on the EU market. It becomes more challenging when developing models which are resilient to errors, inconsistencies, and biases, especially when GPAIS are used as part of the frameworks developed. Here it is also necessary, that clear requirements need to be defined for GPAIS before being published, such that they can be integrated into other frameworks. It makes it infeasible when providers of *high-risk* AI systems need to tackle resilience to errors, inconsistencies, and bias for GPAIS, such as BERT and ResNet by themselves.

These requirements apply to fake news detection frameworks used such, that they result in actions such as **blocking** content, or **disabling** users from further sharing information. For developing such a system it is allowed to use *high-risk* AI systems in a controlled environment called *regulatory sandboxes* for a limited time (*Article 54*, Chapter 3). *Regulatory sandboxes* can be also used by the scientific community to test and validate models and frameworks for different use cases, such as multimodal fake news detection. However, there is no clear guidance of what requirements GPAIS such as BERT and ResNet need to incorporate when being published for scientific purposes, without any clear reference to potential use cases and applications.

Overall, in practical use cases, it appears simpler to utilize multimodal fake news detection frameworks like **COOLANT**, **MMFN**, and **TieFake** in conjunction with **filtering**, **flagging**, or **(de)prioritization** methods. This enables the use of multimodal fake news detection frameworks for real-world use cases with a smaller set of obligations that can be followed more straightforwardly.

4. Discussion and Conclusion

This paper presents a comparative analysis of various models used for multimodal fake news detection, evaluating their performance. The best-performing models have then been categorized based on the risk categories outlined in the AI Act, describing the obligations attached to the respective risk category with respect to the fake news detection domain. The research questions stated in the introduction can be answered as follows:

1. RQ: *What are the essential architectures and evaluation scores of current best-performing multimodal fake news detection AI models?* Based on the analysis presented in Section 2 **COOLANT**, **MMFN**, and **TieFake** can be identified as the best performing models for multimodal fake news detection based on their performance on four different datasets. Unfortunately, not all models have been trained and tested on all four datasets, which leaves room for further comparison by implementing all models with the same pre-processing steps and hyper-parameters on all four datasets. In terms of model architectures, the underlying methods show a wide range of effective strategies, such as contrastive learning with **COOLANT** and **MMFN**, the latter in combination with progressive fusion of features and modalities, as well as similarity and emotion component inclusion with **TieFake**.
2. RQ: *What risk categories do the models fall into and what are the consequences for societies and AI providers, e.g. with respect to risk management, data governance, documentation, transparency, human oversight and required accuracy?* The assessment of the application of the GPAIS definiton and the respective risk category outlined in Section 3 shows, that for all three models **COOLANT**, **MMFN**, and **TieFake**, the GPAIS definition can be applied. The risk categorization can be based on the actions which result from applying the models in real-world use cases. Actions such as **filtering**, **(de)prioritisation**, **flagging** can be mapped to the *limited-risk* category, where mainly transparency obligations need to be followed. Actions, such as **disabling**, and **blocking**, result in the *high-risk* category with a extensive list of obligations, which are very hard to follow, especially when GPAIS are used, such as BERT and ResNet. Moreover, the obligations are described very opaque, that it is challenging to translate them into concrete technical countermeasures and actions for the fake news detection use case. Therefore, the application of the risk-based approach proposed by the EC in the AI Act needs to be specified with respect to the following aspects.

Overall, the AI Act provides guidance in various criteria which need to be considered to provide ethically aligned applications for multimodal fake news detection. Nevertheless, the risk-based approach shows clear limitations in terms of clarity and applicability which creates challenges when trying to follow the proposed obligations and leaves room for interpreting the obligations very differently. Especially, when LLMs or pre-trained deep-learning models for other modalities are used, the obligations cannot be followed for the *limited* and *high-risk* categories, as the transparency obligations cannot be provided for pre-trained deep-learning models. This limitation is prevalent in many machine learning applications today and poses a significant risk, as it renders the AI Act inapplicable to a majority of AI applications in the EU market. Therefore, the definition of GPAIS and the risk categories need to be redefined to make the usage of pre-trained deep-learning models more applicable and more concrete, to ensure that these models can still be used in an ethically and legally compliant way, without having to go through the extensive list of obligations defined for the *high-risk* category. The AI Act provides a general framework for AI applications. However, for specific use cases such as FND further aspects regarding bias in data, representative training samples, transparency of AI systems and security measures are relevant to consider, to increase human trust in the systems an make them usable in practice.

5. References

[1] O. Darwish, Y. Tashtoush, M. Maabreh, R. Al-essa, R. Aln'uman, A. Alqublan, M. Abualkibash, and M. Elkhodr, "Identifying fake news in the russian-ukrainian conflict using machine learning," in *Advanced Information Networking and Applications: Proceedings of the 37th International Conference on Advanced Information Networking and Applications (AINA-2023), Volume 3*. Springer, 2023, pp. 546–557.

[2] V. Balakrishnan, N. W. Zhen, S. M. Chong, G. J. Han, and T. J. Lee, "Infodemic and fake news–a comprehensive overview of its global magnitude during the covid-19 pandemic in 2021: A scoping review," *International Journal of Disaster Risk Reduction*, p. 103144, 2022.

[3] L. Monsees, "Information disorder, fake news and the future of democracy," *Globalizations*, vol. 20, no. 1, pp. 153–168, 2023.

[4] Y. Zhou, Y. Yang, Q. Ying, Z. Qian, and X. Zhang, "Multi-modal fake news detection on social media via multi-grained information fusion," 2023.

[5] J. A. Tucker, A. Guess, P. Barberá, C. Vaccari, A. Siegel, S. Sanovich, D. Stukal, and B. Nyhan, "Social media, political polarization, and political disinformation: A review of the scientific literature," *Political polarization, and political disinformation: a review of the scientific literature (March 19, 2018)*, 2018.

[6] J. Allen, B. Howland, M. Mobius, D. Rothschild, and D. J. Watts, "Evaluating the fake news problem at the scale of the information ecosystem," *Science Advances*, vol. 6, no. 14, p. eaay3539, 2020.

[7] Y. Chen, D. Li, P. Zhang, J. Sui, Q. Lv, L. Tun, and L. Shang, "Cross-modal ambiguity learning for multimodal fake news detection," in *Proceedings of the ACM Web Conference 2022*, 2022, pp. 2897–2905.

[8] X. Zhou and R. Zafarani, "A survey of fake news: Fundamental theories, detection methods, and opportunities," *ACM Computing Surveys (CSUR)*, vol. 53, no. 5, pp. 1–40, 2020.

[9] F. Monti, F. Frasca, D. Eynard, D. Mannion, and M. M. Bronstein, "Fake news detection on social media using geometric deep learning," *arXiv preprint arXiv:1902.06673*, 2019.

[10] K. Shu, S. Wang, and H. Liu, "Beyond news contents: The role of social context for fake news detection," in *Proceedings of the twelfth ACM international conference on web search and data mining*, 2019, pp. 312–320.

[11] D. Shepardson and D. Bartz. (2023) Us begins study of possible rules to regulate ai like chatgpt. [Online]. Available: https://www.reuters.com/technology/us-begins-study-possible-rules-regulate-ai-like-chatgpt-2023-04-11/

[12] "Proposal regulation: laying down harmonised rules artificial intelligence," 2021. [Online]. Available: https://digital-strategy.ec.europa.eu/en/library/proposal-regulation-laying-down-harmonised-rules-artificial-intelligence

[13] C. Marsden and T. Meyer, *Regulating disinformation with artificial intelligence: effects of disinformation initiatives on freedom of expression and media pluralism*. European Parliament, 2019.

[14] C. Wardle and H. Derakhshan, "Information disorder: Toward an interdisciplinary framework for research and policy making," *Council of Europe report*, vol. 27, pp. 1–107, 2017.

[15] H. L. E. G. on Fake News and O. Disinformation, "Report to the european commission on a multi-dimensional approach to disinformation," 2018. [Online]. Available: https://ec.europa.eu/digital-single-market/en/news/final-report-high-level-expert-group-fake-news-and-online-disinformation

[16] L. Wang, C. Zhang, H. Xu, S. Zhang, X. Xu, and S. Wang, "Cross-modal contrastive learning for multimodal fake news detection," 2023.

[17] J. Jing, H. Wu, J. Sun, X. Fang, and H. Zhang, "Multimodal fake news detection via progressive fusion networks," *Information Processing Management*, vol. 60, no. 1, p. 103120, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0306457322002217

[18] Y. Jiang, X. Yu, Y. Wang, X. Xu, X. Song, and D. Maynard, "Similarity-aware multimodal prompt learning for fake news detection," 2023.

[19] Q. Guo, Z. Kang, L. Tian, and Z. Chen, "Tiefake: Title-text similarity and emotion-aware fake news detection," 2023.

[20] F. Ghorbanpour, M. Ramezani, M. A. Fazli, and H. R. Rabiee, "Fnr: A similarity and transformer-based approach to detect multi-modal fake news in social media," 2023.

[21] R. Shao, T. Wu, and Z. Liu, "Detecting and grounding multimodal media manipulation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 6904–6913.

[22] Y. Chen, D. Li, P. Zhang, J. Sui, Q. Lv, L. Tun, and L. Shang, "Cross-modal ambiguity learning for multimodal fake news detection," in *TheWebConf 2022*. ACM, April 2022. [Online]. Available: https://www.microsoft.com/en-us/research/publication/cross-modal-ambiguity-learning-for-multimodal-fake-news-detection/

[23] Z. Wei, H. Pan, L. Qiao, X. Niu, P. Dong, and D. Li, "Cross-modal knowledge distillation in multi-modal fake news detection," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 4733–4737.

[24] Y. Zhou, Q. Ying, Z. Qian, S. Li, and X. Zhang, "Multimodal fake news detection via clip-guided learning," 2022.

[25] S. Singhal, T. Pandey, S. Mrig, R. R. Shah, and P. Kumaraguru, "Leveraging intra and inter modality relationship for multimodal fake news detection," in *Companion Proceedings of the Web Conference 2022*, ser. WWW '22. New York, NY, USA: Association for Computing Machinery, 2022, p. 726–734. [Online]. Available: https://doi.org/10.1145/3487553.3524650

[26] J. Wang, H. Mao, and H. Li, "Fmfn: Fine-grained multimodal fusion networks for fake news detection," *Applied Sciences*, vol. 12, no. 3, 2022. [Online]. Available: https://www.mdpi.com/2076-3417/12/3/1093

[27] Y. Wu, P. Zhan, Y. Zhang, L. Wang, and Z. Xu, "Multimodal fusion with co-attention networks for fake news detection," in *Findings*, 2021.

[28] R. Kumari and A. Ekbal, "Amfb: Attention based multimodal factorized bilinear pooling for multimodal fake news detection," *Expert Systems with Applications*, vol. 184, p. 115412, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417421008320

[29] S. Qian, J. Wang, J. Hu, Q. Fang, and C. Xu, "Hierarchical multimodal contextual attention network for fake news detection," ser. SIGIR '21. New York, NY, USA: Association for Computing Machinery, 2021, p. 153–162. [Online]. Available: https://doi.org/10.1145/3404835.3462871

[30] C. Song, N. Ning, Y. Zhang, and B. Wu, "A multimodal fake news detection model based on crossmodal attention residual and multichannel convolutional neural networks," *Information Processing Management*, vol. 58, no. 1, p. 102437, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0306457320309304

[31] X. Zhou, J. Wu, and R. Zafarani, "Safe: Similarity-aware multimodal fake news detection," 2020.

[32] L. Zhang, M. Yang, C. Li, and R. Xu, "Image-text retrieval via contrastive learning with auxiliary generative features and support-set regularization," ser. SIGIR '22. New York, NY, USA: Association for Computing Machinery, 2022, p. 1938–1943. [Online]. Available: https://doi.org/10.1145/3477495.3531783

[33] Y. Hu, R. Li, C. Chen, H. Zou, Q. Zhu, and E. S. Chng, "Cross-modal global interaction and local alignment for audio-visual speech recognition," 2023.

[34] G. Hu, T.-E. Lin, Y. Zhao, G. Lu, Y. Wu, and Y. Li, "Unimse: Towards unified multimodal sentiment analysis and emotion recognition," 2022.

[35] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, ser. Proceedings of Machine Learning Research, M. Meila and T. Zhang, Eds., vol. 139. PMLR, 2021, pp. 8748–8763.

[36] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," 2015.

[37] C. Buciluundefined, R. Caruana, and A. Niculescu-Mizil, "Model compression," in *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '06. New York, NY, USA: Association for Computing Machinery, 2006, p. 535–541. [Online]. Available: https://doi.org/10.1145/1150402.1150464

[38] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: Lessons learned from the 2015 MSCOCO image captioning challenge," *CoRR*, vol. abs/1609.06647, 2016. [Online]. Available: http://arxiv.org/abs/1609.06647

[39] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," 2019.

[40] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," 2021.

[41] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2015.

[42] Y. Zhu, X. Zhou, J. Qiang, Y. Li, Y. Yuan, and X. Wu, "Prompt-learning for short text classification," 2022.

[43] X. Han, W. Zhao, N. Ding, Z. Liu, and M. Sun, "PTR: prompt tuning with rules for text classification," *CoRR*, vol. abs/2105.11259, 2021. [Online]. Available: https://arxiv.org/abs/2105.11259

[44] L. Ren, G. Duan, T. Huang, and Z. Kang, "Multi-local feature relation network for few-shot learning," *Neural Comput. Appl.*, vol. 34, no. 10, p. 7393–7403, may 2022. [Online]. Available: https://doi.org/10.1007/s00521-021-06840-8

[45] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," 2016.

[46] J. Chen, H. Zhang, X. He, L. Nie, W. Liu, and T.-S. Chua, "Attentive collaborative filtering: Multimedia recommendation with item- and component-level attention," ser. SIGIR '17. New York, NY, USA: Association for Computing Machinery, 2017, p. 335–344. [Online]. Available: https://doi.org/10.1145/3077136.3080797

[47] Z. Jin, J. Cao, H. Guo, Y. Zhang, and J. Luo, "Multimodal fusion with recurrent neural networks for rumor detection on microblogs," in *Proceedings of the 25th ACM International Conference on Multimedia*, ser. MM '17. New York, NY, USA: Association for Computing Machinery, 2017, p. 795–816. [Online]. Available: https://doi.org/10.1145/3123266.3123454

[48] C. Boididou, S. Papadopoulos, M. Zampoglou, L. Apostolidis, O. Papadopoulou, and Y. Kompatsiaris, "Detection and visualization of misleading content on twitter," *International Journal of Multimedia Information Retrieval*, vol. 7, pp. 71–86, 2018.

[49] C. Boididou, K. Andreadou, S. Papadopoulos, D.-T. Dang-Nguyen, G. Boato, M. Riegler, and Y. Kompatsiaris, "Verifying multimedia use at mediaeval 2015." Wurzen, Germany: CEUR Workshop Proceedings, 09/2015 2015.

[50] K. Shu, D. Mahudeswaran, S. Wang, D. Lee, and H. Liu, "Fakenewsnet: A data repository with news content, social context and spatialtemporal information for studying fake news on social media," 2019.

[51] P. Glauner, "An assessment of the ai regulation proposed by the european commission," *arXiv preprint arXiv:2105.15133*, 2021.

[52] C. I. Gutierrez, A. Aguirre, R. Uuk, C. C. Boine, and M. Franklin, "A proposal for a definition of general purpose artificial intelligence systems," *Available at SSRN 4238951*, 2022.

[53] P. Hacker, A. Engel, and M. Mauer, "Regulating chatgpt and other large generative ai models," *arXiv preprint arXiv:2302.02337*, 2023.

[54] J. Wei, Y. Tay, R. Bommasani, C. Raffel, B. Zoph, S. Borgeaud, D. Yogatama, M. Bosma, D. Zhou, D. Metzler *et al.*, "Emergent abilities of large language models," *arXiv preprint arXiv:2206.07682*, 2022.

[55] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[56] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[57] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 012–10 022.

[58] H. Zhang, C. Wu, Z. Zhang, Y. Zhu, H. Lin, Z. Zhang, Y. Sun, T. He, J. Mueller, R. Manmatha *et al.*, "Resnest: Split-attention networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 2736–2746.

[59] V. Schmitt, V. Solopova, V. Woloszyn, and J. d. J. de Pinho Pinhal, "Implications of the new regulation proposed by the european commission on automatic content moderation," in *Proc. 2021 ISCA Symposium on Security and Privacy in Speech Communication*, 2021, pp. 47–51.

[60] D. Fryer, I. Strümke, and H. Nguyen, "Shapley values for feature selection: The good, the bad, and the axioms," *IEEE Access*, vol. 9, pp. 144 352–144 360, 2021.