Integrating Multimodal Deception Detection with Automated Fact Checking

[ Do not insert author name or contact details now - only to be completed after acceptance]

**Abstract:** The fight against misinformation has reached a critical juncture, demanding advanced techniques to expose deception. Combining multimodal detection methods, analysing text, audio, visuals, and context, with automated fact-checking systems powered by reliable sources, creates a comprehensive shield against online falsehoods. Multimodal analysis delves deep into different content formats, unearthing deceptive cues through advancements in machine learning and computer vision. This multidimensional approach captures linguistic, visual, and contextual clues for a more nuanced understanding of deceit. Automated fact-checking systems complement this approach by employing computational muscle to verify claims against established databases. This cross-referencing enhances the accuracy and reliability of assessments, further strengthening the defence against misinformation. However, challenges remain. The vast range of platforms where deceptive content surfaces make it difficult to maintain comprehensive training datasets. The sheer volume of online information poses scalability issues, while the ever-evolving tactics of misinformation creators add another layer of complexity. Despite these challenges, the integration of multimodal detection and automated fact-checking represents a significant step forward in the fight against misinformation. By understanding these challenges and continuously refining our tools, we can build a more resilient information landscape where truth prevails.

**Keywords:** Multilayer perceptron architecture; Natural Language Processing; Deception Detection; Emotion Detection;

1. Introduction

The deluge of information in our world carries an insidious undercurrent: deception. From casual fibs to calculated manipulations, it weaves its way through every facet of life, eroding trust, shattering reputations, and even inflicting real-world harm. To safeguard ourselves in this perilous landscape, the quest for effective deception detection becomes ever more crucial.

Traditionally, we relied on methods like behavioral observation, scrutinizing nonverbal cues in a dance often riddled with misinterpretations. Polygraphs, measuring physiological changes, offered a shaky glimpse, susceptible to external factors. Statement analysis, while valuable, demanded keen linguistic and deception expertise, leaving room for subjectivity. These limitations exposed the urgent need for more robust tools.

Enter the era of cutting-edge techniques. Powered by AI and machine learning, computer-based analysis delves into the minutiae of facial expressions, voice patterns, and textual data, unearthing subtle cues invisible to the naked eye. Neuroscientific frontiers, though in their nascent stages, explore brain imaging as a potential objective measure of deception. And multimodal approaches, weaving together traditional and modern methods, paint a comprehensive picture of deceptive behaviour.

Yet, challenges remain. Individuals' unique deception styles defy a one-size-fits-all approach. Cultural nuances influence how deception manifests, adding another layer of complexity. And some, skilled in the art of the lie, can shroud their intentions with an adeptness that poses a formidable challenge.

The future of deception detection lies in continuous refinement. Research is actively pursuing context-aware systems that consider situational factors and relationship dynamics to sharpen accuracy. Real-time detection systems hold the promise of immediate feedback in critical situations. But amidst these advancements, ethical considerations regarding privacy and responsible data utilization remain paramount.

In conclusion, the battle against deception is a continuous one, demanding unwavering vigilance and innovative tools. As technology evolves and our understanding of human behavior deepens, we inch closer to a future where truth holds firm and deception finds no refuge. The path ahead may be challenging, but the stakes are too high to falter. Let us remain committed to safeguarding trust, security, and ethical conduct in our information-rich world, one unmasked lie at a time.

2 Literature Survey

*2.1 Deception Detection*

The proposed neural model (Mathur L et.al ,2020), MLPU, is designed for deception detection in real-life videos, leveraging a multi-layer perceptron architecture that integrates features from diverse modalities, including visual, textual, audio, and micro-expressions. MLPU outperforms existing methods, though its reliance on a relatively small dataset necessitates further refinement for enhanced generalization across larger datasets and diverse environments.

In the realm of feature extraction, the model (Ahmed et.al, 2021) employs a 3D-CNN for visual modalities, demonstrating superior performance in capturing spatiotemporal features for facial expressions such as smiles, fear, or stress. Textual features are extracted using a CNN and a pretrained Word2Vec model (Krishnamurthy et.al, 2018) while audio features undergo extraction with the OpenSMILE toolkit, including preprocessing steps for background noise removal and voice normalization.

The MLPU model itself is a multi-modal architecture with a hidden layer of size 1024 and a linear output layer. ReLU activation functions introduce non-linearity, and dropout with a keep probability of 0.5 is applied for regularization. Notably, individual evaluations are conducted on each modality without data fusion at this stage.

Subsequently, data fusion techniques in (Zhang, J, et.al 2019) are introduced, including Concatenation (MLPC) and Hadamard + Concatenation (MLPH+C), aiming to combine information from different modalities to enhance overall model performance. These fusion techniques, with carefully specified dimensions, represent an effort to capture complementary information and improve deception detection.

Shifting to a related context, the exploration of deception detection through unsupervised transfer learning, particularly in videos, stands out. This approach, distinct from traditional supervised learning, does not rely on labelled training data, making it suitable for high-stakes scenarios where obtaining labelled data can be challenging. Visual and audio features are aligned using subspace alignment techniques, projecting them into a shared subspace to enhance the detection of deception patterns that span both modalities.

The evaluation involves testing on videos of individuals engaged in high-stakes deception tasks and those expressing deception on YouTube, providing a realistic testbed for real-world scenarios and a naturalistic setting for assessing generalization. Results indicate that the unsupervised transfer learning approach outperforms traditional supervised learning methods, even when the latter are trained on extensive labelled data. This performance suggests the potential of unsupervised transfer learning for deception detection in situations where labelled data is limited.

The significance and implications of these approaches extend across domains such as law enforcement, national security, and business applications. They could be applied in assessing witness statements, identifying potential threats, and detecting fraud or assessing employee honesty. The incorporation of subspace alignment techniques introduces a novel concept for multimodal deception detection, capturing a more comprehensive representation of the deception signal and enhancing overall detection accuracy.

*2.2 Fact Checking System*

According to (Al-Rakhami et.al 2020) the prevalence of misinformation, particularly concerning the COVID-19 pandemic on Online Social Networks (OSNs) like Twitter, underscores the urgent need for reliable fact-checking. A framework employing ensemble learning has been developed to categorize COVID-19 related tweets into credible and non-credible categories based on tweet- and user-level characteristics. The framework exhibits high precision in discerning the credibility of such tweets, crucial in preventing the spread of misleading information.

Misinformation, often disseminated through platforms like Twitter, can exacerbate the spread of the disease as individuals act based on the information they receive. With over 4000 incorrect claims about the pandemic circulating by April 2020, misinformation can induce fear, misguide medical practices, and lead to non-compliance with medical guidelines. Examples include false assertions about COVID-19 remedies and misleading tweets regarding potential treatments, impacting mental and physical well-being.

To address this challenge, the proposed frameworks (Furuta, T, & Suzuki, Y 2021 and Sarr, E. N et.al 2017) combines six machine-learning algorithms through ensemble learning, demonstrating higher accuracy and generalization.

Even though there are many fact-checking systems available on the internet, they are often paid, and some of them are only centred on detecting fact claims surrounding a certain area or topic. Some videos that contain misinformation might be in a language that the viewer does not even understand. It is easy to misguide a viewer by showing a video in a language that they do not understand, with subtitles that are false or misleading, or falsely translated. Therefore, a fact-checking system that is able to translate and perform fact-checking in multiple languages is needed.

3 Proposed Work

Detecting deception and ensuring the accuracy of information in digital content remains a significant challenge. Currently, there is a lack of a unified system capable of seamlessly analyzing audio-visual features for deception detection and scrutinizing spoken dialogue for fact-checking. This gap hinders the provision of a comprehensive tool that can concurrently unveil deceptive practices and assess the factual accuracy of multimedia content, limiting the ability to offer users a trustworthy digital experience.

The proposed system is an advancement of (Rill-García et.al 2021) model which aims to address this challenge by integrating insights from various research papers to develop an advanced multimodal deception detection model. Leveraging both visual and acoustic features, the system targets enhanced accuracy in detecting deception in high-stakes scenarios. The primary objective is to design a comprehensive system that outperforms existing models, and provide multiple features like automated fact-checking.

The proposed approach employs early-fusion and late-fusion techniques to identify effective modalities and fusion approaches. The contribution of each modality, particularly eye gaze and other visual cues, is investigated.

Taking the endeavor further the proposed approach integrates advanced fact-checking systems seamlessly into multimodal deception detection. This integration introduces a layer of scrutiny to evaluate the factual accuracy of information exchanged in dialogues. The scarcity of a fact-checking system that covers multiple areas is very rare, and most of them are not free. So, Google’s Gemini has the clear lead in terms of topics and areas and is also free to use, making it the most suitable choice.

The operational framework of the proposed model, as depicted in the flowchart, seamlessly integrates the Deception Detection model and the fact-checking model. This integration is expected to enhance the overall accuracy of the model, improving its efficiency in discerning deceptive information.



Fig 3.1 The proposed model

Potential for further refinement exists, including exploring the impact of visual content (images, videos) on deception detection and fact-checking, creating a truly comprehensive multimodal system. In conclusion, the integration of fact-checking systems into multimodal deception detection represents a pioneering step toward fostering a more reliable and discerning communication environment. This symbiotic relationship between various detection mechanisms culminates in a system that identifies deceptive cues and evaluates the factual accuracy of information, fostering a trustworthy communication landscape.

4 Implementation

The implementation involves using OpenCV to extract facial expressions, head movements, and eye gaze features for visual analysis. Simultaneously, PyAudio is employed to extract acoustic features like tone, pitch, and jitter for audio analysis. Normalization of these features is performed to account for variations in illumination, pose, facial expressions, and recording conditions. To enhance accuracy, fact-checking information is integrated into the model along with speech-to-text conversion, associating each video with verified factual information. Next, multimodal fusion combines visual and acoustic features using either early or late fusion techniques to establish correlations and capture multimodal context.

The dataset is then partitioned into training and testing subsets, commonly with an 80-20 split ratio, facilitating model development and evaluation. Choosing a suitable machine learning model (SVM, Decision Trees, or advanced Deep Learning Architectures) depends on the deception detection task's complexity. The training data is fed into the selected model, and hyperparameters are fine-tuned using optimization techniques like grid search.

Leverage NLP techniques to comprehensively analyze the speaker's dialogue content, extracting meaningful insights. Integrate fact-checking results into the multimodal model for enhanced accuracy. The model evolves iteratively, adapting to new insights from incoming labelled data and refining based on feedback. Consider transfer learning to harness knowledge from pre-existing models for enhanced capabilities. Validate the trained model across diverse datasets, ensuring robustness in handling varied scenarios through extensive testing that mimics real-world conditions.

* 1. *Software Requirements*

Software requirements for deception detection typically involve a combination of advanced algorithms, natural language processing (NLP) capabilities, machine learning models, and data analytics tools like OpenAi's Whisper AI. The software should be able to analyze various verbal and non-verbal cues, including speech patterns, facial expressions, and physiological signals, to identify potential deception. Additionally, it should support real-time processing to provide immediate feedback or analysis. Furthermore, integration with databases or external sources for contextual information and historical data can enhance the accuracy of deception detection. The software must also prioritize security and privacy measures to ensure the ethical use of sensitive information during the detection process. This system is compatible with both Windows and Linux operating systems. Hardware requirements include a processor with Intel core i5 or above, 64-bit architecture, quad-core with a minimum of 3.0 GHz per core, or an AMD equivalent. Additionally, it necessitates a minimum of 4 GB of VRAM with a minimum of 2 GB Hard disk storage.

* 1. *Dataset*

The dataset for deception detection is crucial for training and evaluating algorithms aimed at discerning truthful from deceptive communication. It typically comprises a diverse range of samples, including transcripts of verbal interactions, videos capturing facial expressions and body language, physiological data such as heart rate and skin conductance, and metadata providing contextual information. The dataset should encompass various scenarios, contexts, and demographics to ensure the robustness and generalizability of the deception detection model. Additionally, annotations indicating the ground truth of deception or truthfulness are essential for supervised learning approaches, enabling the algorithm to learn patterns associated with deceptive behaviour. Careful curation of the dataset is necessary to maintain ethical standards and privacy considerations, ensuring the responsible use of sensitive information in the pursuit of improving deception detection technology.

1. **Models and Algorithms**

The proposed system aims to align audio-visual representations of deception observed in low-stakes scenarios with those in high-stakes situations, circumventing the need for high-stakes labels during training. This alignment employs Principal Component Analysis (PCA) to generate subspace embeddings for both the source domain (representing low-stakes situations) and the target domain (corresponding to high-stakes scenarios).

The unsupervised Subspace Alignment (SA) models are compared with existing supervised models, considering ACC, AUC, and F1-Score metrics to contrast outcomes with prior research models, ensuring a robust evaluation of the proposed approach.

1. **Workflow**

This is where fact-checking comes into play, where human fact-checkers or automated systems cross-reference the flagged content against credible sources, databases, and expert opinions to verify its accuracy. The integration ensures that deception detection algorithms provide an additional layer of scrutiny, assisting fact-checkers in prioritizing their efforts and identifying potentially misleading information more efficiently. This collaborative workflow aims to bolster the reliability of fact-checking processes, ultimately promoting the dissemination of accurate and trustworthy information to the public.



Figure 6.1: Deception Detection model



Figure 6.2: Fact Checking

1. **Evaluation and Results**

A better way to represent the performance of the deception detection model is through f1 Score. The model is tested with a testing dataset that includes 36 videos, 18 truth videos and 18 deceptive videos. The model is tested with the truth videos first and then the deceptive videos are used for testing. Among the 18 truth videos tested 13 of them displayed a deception probability of below 50% and 5 of them were above 50%. In order to find the f1 score, the videos with a deception probability below 50% is considered as Truth Positive (TP) and videos above 50% is considered as False Positive (FP). Then the model is again tested with 18 deceptive videos and 14 of the videos displayed a deception probability above 50% and it is considered as Truth Negative (TN) and the remaining 4 videos displayed a deception probability below 50% and is considered as False Negative (FN).

Precision (P) is the ratio of correctly predicted positive observations to the total predicted positives. From the experimentation results, the number of correctly predicted positive observations is 13. The precision is found to be 0.7222 by substituting the values in equation (1).

Recall (R) is the ratio of correctly predicted positive observations to all actual positives. From the above observation Recall is found to be 0.7647 by substituting values in equation (2).

By substituting the values of precision and recall in equation (3), the f1 score of the model is found to be 0.7429.

The accuracy of the model calculated by substituting the values in equation (4) is 0.75. This means that the model correctly classified 75% of the videos in the dataset.

The constructed model was able to receive a video from the user as input and perform deception detection on the video. The deception detection model was able to detect deception with an accuracy of 75% and the f1 score of the model is found to be 0.743. The program also extracts audio from the video and performs speech-to-text on the extracted audio and performs fact-checking on the generated text using Google’s Gemini API. The speech-to-text function is performed by OpenAI’s Whisper AI. The accuracy of the fact-checking model is also dependent on the transcription of the video.

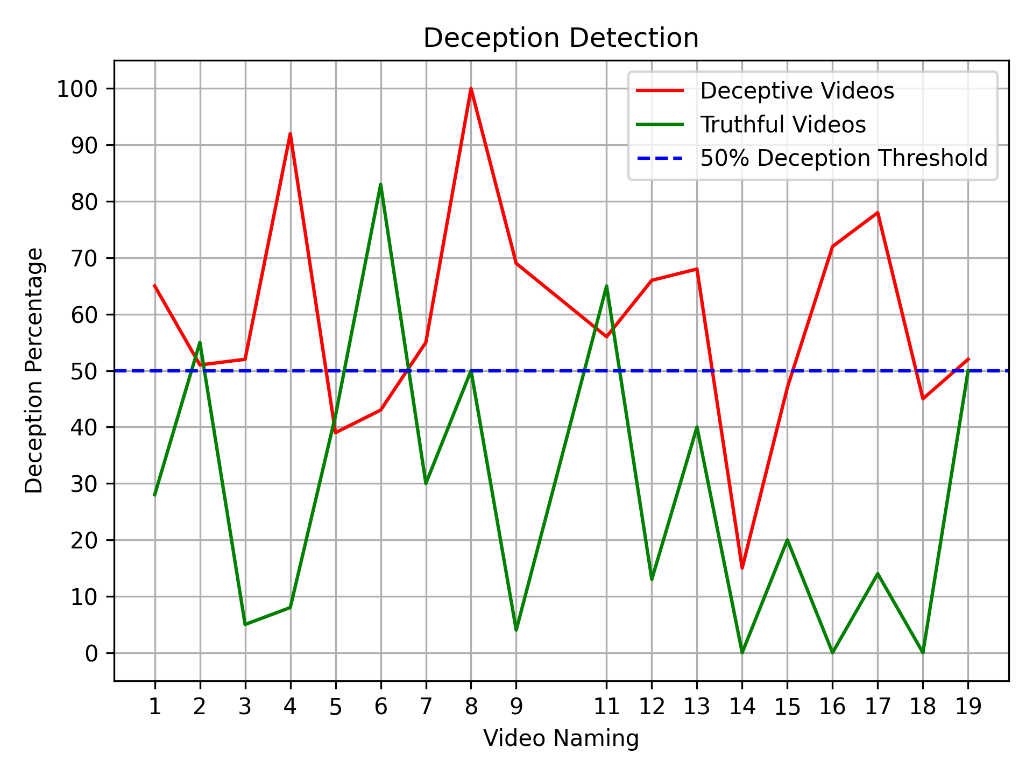


Figure 7.1: Truth Vs Deceptive

The text-to-speech accuracy of the transcript can be increased by using a larger model which also results in a longer processing time, but the transcript is the most accurate. Using the Gemini API to fact-check the transcript generated by Whisper AI is very easier and the Gemini API performs fact-checking in the transcript, it can also identify sarcasm to an extent. Overall, the fact-checking system works fine most of the time. The system built was able to perform deception detection in videos, speech-to-text from the extracted audio, and fact-checking from the video transcript.

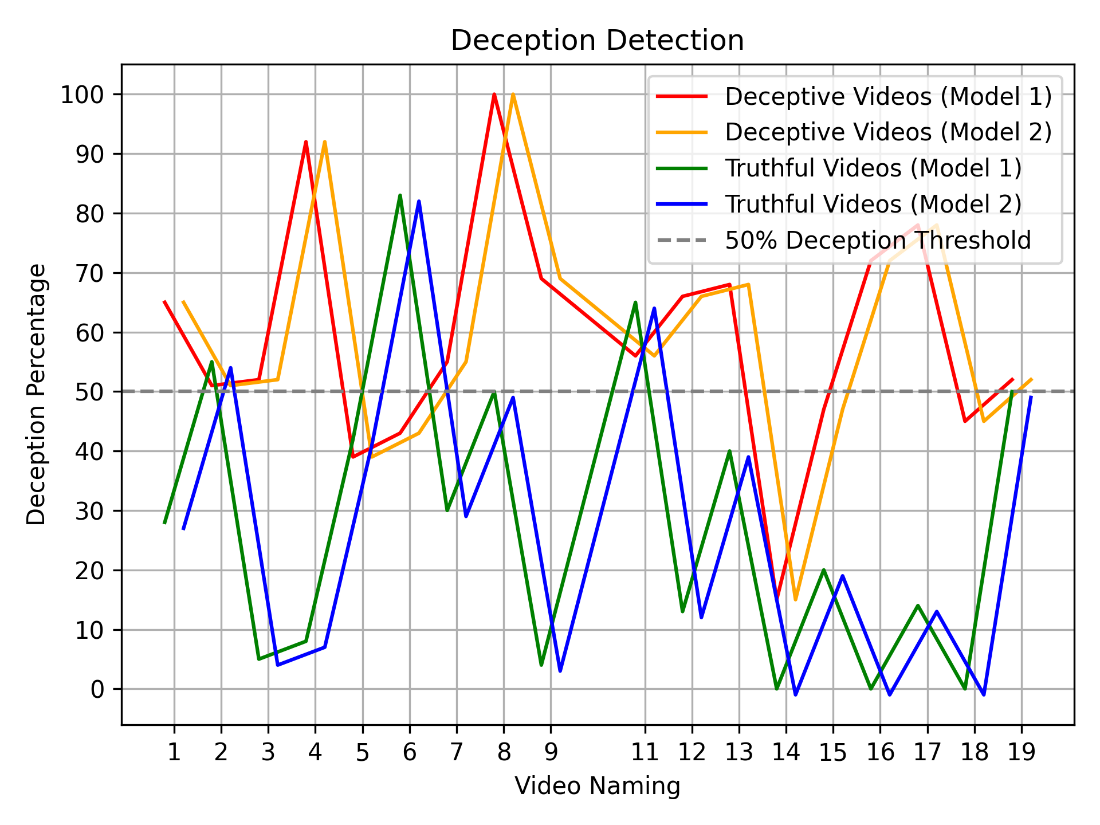


Fig 7.2: Comparison between Existing and Proposed System

The newly implemented system was able to perform deception detection slightly better than the existing systems, thanks to the changes and minor tweaks made in the dataset. The new implemented model was better at classifying videos as deceptive or not.

8 Conclusion

The model is able to predict the probability of deception in videos with an accuracy of 75%, on par with existing models of the same type. The system is also equipped with automated fact-checking, making it easier to validate the information from the speaker’s speech. Additionally, the model can detect sarcasm, jokes, etc., and if the speaker is asking any questions, the Gemini-powered fact-checking model can also answer them easily.

In conclusion, a system where the user can upload a video of a person speaking in any language about anything is created to perform deception detection on the video by extracting features from the speaker's face, performing speech-to-text, and fact-checking, and then displaying the result.

In the future, the model can be expanded or enhanced with new features and functionalities, such as a pre-processed video generator that can generate a video showing the deception detection process with translated subtitles if the speaker is not speaking in English. Additionally, the accuracy can be increased by training this model on a larger and more refined dataset. Furthermore, the functionality of automated fact-checking can be improved by combining the results from multiple sources like Gemini, ChatGPT, etc., thereby enhancing the reliability and comprehensiveness of the fact-checking process.

**References**

Krishnamurthy, G., Majumder, N., Poria, S., & Cambria, E. (2018, March). A deep learning approach for multimodal deception detection. In *International Conference on Computational Linguistics and Intelligent Text Processing* (pp. 87-96). Cham: Springer Nature Switzerland.

Sarr, E. N., Sall, O., & Diagne, A. (2017, July). SenFact Algorithm: Fact-checking by the confrontation of opinions. In *2017 13th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD)* (pp. 2235-2241). IEEE.

Raj, C., & Meel, P. (2021, October). Microblogs Deception Detection using BERT and Multiscale CNNs. In *2021 2nd Global Conference for Advancement in Technology (GCAT)* (pp. 1-6). IEEE.

Zhang, J., Levitan, S. I., & Hirschberg, J. (2020, October). Multimodal Deception Detection Using Automatically Extracted Acoustic, Visual, and Lexical Features. In *INTERSPEECH* (pp. 359-363).

Rill-García, R., Jair Escalante, H., Villasenor-Pineda, L., & Reyes-Meza, V. (2019). High-level features for multimodal deception detection in videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* (pp. 0-0).

Al-Rakhami, M. S., & Al-Amri, A. M. (2020). Lies kill, facts save: Detecting COVID-19 misinformation in twitter. *Ieee Access*, *8*, 155961-155970.

Mathur, L., & Matarić, M. J. (2021, June). Unsupervised audio-visual subspace alignment for high-stakes deception detection. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 2255-2259). IEEE.

Venkatesh, S., Ramachandra, R., & Bours, P. (2019, March). Robust algorithm for multimodal deception detection. In *2019 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)* (pp. 534-537). IEEE.

Furuta, T., & Suzuki, Y. (2021, September). A Fact-checking Assistant System for Textual Documents. In *2021 IEEE 4th International Conference on Multimedia Information Processing and Retrieval (MIPR)* (pp. 243-246). IEEE.

Kumar, S., Bai, C., Subrahmanian, V. S., & Leskovec, J. (2021, May). Deception detection in group video conversations using dynamic interaction networks. In *Proceedings of the International AAAI Conference on Web and Social Media* (Vol. 15, pp. 339-350).

Ahmed, H. U. D., Bajwa, U. I., Zhang, F., & Anwar, M. W. (2021). Deception detection in videos using the facial action coding system. *arXiv preprint arXiv:2105.13659*.