Multimodal Natural Language Processing: A survey on Innovations, Challenges, and Applications Across Domains

Deepika A[1][0000-0001-5765-0870] and Rajeswari M[2][000-0001-8949-549X]

[1] Research Scholar, PSGR Krishnammal College for Women, Coimbatore, India
[2] Assistant Professor, PSGR Krishnammal College for Women, Coimbatore, India
deepuarumugam22@gmail.com

Abstract. In recent years, the Large Language models and Multimodal Artificial Intelligence have become highly important topics in the field of Natural Language Processing. Multimodal Natural Language Processing (NLP) which gives high impact by integrating text with other data types, such as images, audio, and videos which can create models that offer richer and more contextualized understanding. The latest advanced machine learning and deep learning algorithms have made the machines much easier to understand and generate the human language effectively which in turn enables efficient conversation between humans and machines. Techniques such as transfer learning, data augmentation are explored as key methods to overcome the challenges of data scarcity, ultimately promoting linguistic diversity and digital equity. Also in addressing Low-Resource Languages, the paper highlights the growing focus on developing NLP tools for languages with limited data availability. This survey paper provides a deep discussion about the current trends, challenges, and advancements in Natural Language Processing (NLP). As this field evolves continuously, several emerging areas are gaining prominence, including Multimodal NLP, Low-Resource Languages, and Explainability in NLP. The survey also delves into the critical issue of Explainability in NLP, emphasizing the need for transparent and interpretable models, especially for the high-stakes applications. Various methods to enhance model explainability, such as feature importance, attention mechanisms, and model-agnostic techniques, are also discussed. The paper concludes by identifying the ongoing challenges in NLP, including data quality, evaluation metrics, and ethical considerations. Apart from this, this paper explores the future directions where the combination of multimodal approaches and the development of tools for various languages, and the pursuit of model explainability will play crucial roles in shaping the next generation of NLP technologies.

Keywords: Natural Language Processing, LLM, Multimodal NLP, Text Mining, Transfer Learning, Attention Mechanism, Multimodal Generative Models.

1 Introduction

The Concept of Natural Language Processing (NLP) came into existence in the 20th century which initially specifically focused on only text data and the systems were rule based that only relied on handcrafted linguistic rules and language understanding. This kind of system has the inability to adapt to the vast and diverse languages. Later statistical methods were introduced and techniques like Hidden Markov Models (HMMs) and n-grams became standard for tasks such as part-of-speech tagging and for machine language translation. Despite these advancements, the scope remained confined to text, limiting the potential applications and effectiveness of NLP systems.

Over the past years, the developments in this field contribute to the machines' intelligence so that they can perceive and communicate emotions. One of the areas in NLP which is expanding is the computational analysis of Human Multimodal Language. It broadens the scope of NLP by exploring language in both face-to-face communication and online multimedia. [1]

Multimodal Natural Language Processing (NLP) is one of the subfields of Artificial Intelligence that primarily focuses on the integration and processing of multiple types of data or modalities to understand and generate human language. The modalities which include are image data, audio data and video data. The multimodal NLP obtains deeper understanding of the content and context by utilizing the added information from the several modalities. The importance of multimodal NLP lies in its ability to mimic human-like understanding and interaction with the world, which inherently involves multiple senses. Multimodal NLP allows systems to grasp context more accurately by combining information from different sources. For example, understanding a scene described in text alongside an accompanying image can provide a richer context than either modality alone. By using multiple modalities, multimodal systems can cross-validate information, leading to more accurate and reliable outcomes. For instance, in sentiment analysis, combining textual data with facial expressions from images can lead to better sentiment detection.

Early multimodal systems were relatively simple, typically involving the fusion of text and images. For instance, image captioning systems combine visual information from images with textual descriptions, leveraging convolutional neural networks (CNNs) and Recurrent Neural Networks (RNNs) for image processing and text generation respectively. These initial attempts laid the groundwork for more sophisticated multimodal models. L. Ma, et al., proposed multimodal Convolutional Neural Network(m-CNN) for image matching and text matching using two CNN algorithms. Experimental results show that the model outperforms other approaches in bidirectional image-sentence retrieval tasks on the Flickr8K and Flickr30K datasets, making it a valuable contribution to multimodal NLP.[2]

Beyond the text generation, these advancements are increasingly used in the following areas like speech generation, Robotic process control, image search and Human Computer Interaction. In the review article, authors provided a comprehensive review of the evolving field of human-computer interaction (HCI), covering its historical development, key principles, and emerging technologies. It also discusses the improvements in this field from early command-line interfaces to modern systems like touchscreens and voice recognition.[3]

However, extending the capabilities of large language models (LLMs) to handle both multimodal text and images remains an ongoing research challenge. Pure text-based LLMs are typically trained solely on textual data and lack the perceptual ability to interpret visual signals. While there are several reviews on multimodal models, each tends to focus on different aspects of this evolving field. Summaira et al. [4] provided a detailed introduction to the application of different modalities by categorizing them based on modes.

The paper highlights various multimodal products from major technology companies and provides a practical guide on the technical aspects of these models. [5]

This paper seeks to bridge the existing gap by starting with a foundational definition of multimodal systems. It offers an overview of the historical evolution of multimodal algorithms and explores the potential applications and challenges within this field. We begin by defining multimodal models/algorithms and then examine their historical development. Additionally, we discuss the practical guide covering various technical aspects of multimodal models, such as knowledge representation, learning objectives, model construction, information fusion, and the use of prompts. The paper reviews the latest algorithms in multimodal models and commonly used datasets, offering essential resources for future research and evaluation. Lastly, we explore many applications of advanced multimodal models and the key challenges arising from their ongoing development are also discussed.

This article, organized as section 2, discusses the core techniques in Multimodal NLP. In section 3 the deep learning techniques are discussed and section 4 discusses the Attention Mechanism and Transformers. Section 5 discussed the Multimodal Transformers and Multimodal Generative models in Section 6. Section 7 highlighted the importance and advancements in Transfer Learning. Section 8 highlights the importance of Multilingual Multimodal NLP. Further sections 9 and 10 elaborates about the applications and results with Discussions and finally concludes the survey in Section 11.

2 Core techniques in Multimodal NLP: Traditional Machine Learning Approaches

There are many important techniques that involve processing multiple types of data (such as text, images, and audio) to increase the performance of predictive models. These methods involve the pre-processing steps, extracting key features, and combining these features to generate a representation that can be used to perform classification, regression, and clustering.

This survey reviews the growing role of multimodal machine learning (ML) in various fields. It examines key challenges and innovations in multimodal representation, fusion, translation, alignment, and co-learning, emphasizing the potential of these models for improving clinical predictions.[6]

This section explains various key traditional machine learning techniques used for handling multimodal data.

2.1 Feature Extraction

Feature extraction is one of the important steps in processing multimodal data. It is used to transform the raw data into a format acceptable by machine learning algorithms by extracting meaningful and relevant features from the given dataset.

Textual Features. Techniques include BoW (word counts), TF-IDF (weighted importance), and Word Embeddings (dense vectors for semantic relationships).

Visual Features. Methods like HOG (edge detection), SIFT/SURF (local features), and CNNs (hierarchical learning from raw images) extract image data features.

Audio Features. MFCCs (power spectrum), Spectrograms (frequency-time analysis), and Chroma Features (pitch classes) are used for audio data extraction.

2.2 Multimodal Machine Learning Models

Traditional machine learning models have been adapted to handle multimodal data by incorporating techniques for feature extraction and data fusion. Some common models include the following.

Support Vector Machines (SVMs). SVMs can be extended to handle multimodal data by combining features from different modalities into a single input vector. Kernel methods can be used to map the combined features into a higher-dimensional space where a linear separation is possible. In his paper, author proposed hybrid algorithm algorithm[7]

Decision Trees and Random Forests. These models can naturally handle multimodal data by treating each feature independently. Random forests, which are ensembles of decision trees, can improve performance by averaging the predictions of multiple trees, each trained on different subsets of the data. In their paper, authors used Decision Trees and Random Forest Algorithms to Predict and Determine Factors Contributing to First-Year University Students' Learning Performance [8]

Naive Bayes. This probabilistic model can be adapted for multimodal data by assuming independence between features from different modalities. While this assumption may not always hold, Naive Bayes can still perform well in practice, especially with text data. Authors used Multinomial Naive Bayes classification model to perform sentiment analysis on sentiment analysis dataset. [9]

Recently the authors in their work used the Multinomial Naive Bayes Technique to detect the language among the given dataset. [10]

K-Nearest Neighbors (KNN). KNN can handle multimodal data by computing distances between combined feature vectors. The choice of distance metric is crucial, as it needs to account for the different characteristics of each modality.

Principal Component Analysis (PCA) and Canonical Correlation Analysis (CCA). PCA can be used to reduce the dimensionality of combined feature vectors, retaining the most important information. CCA, on the other hand, finds linear transformations of two sets of variables (e.g., text and image features) that maximize their correlation, facilitating multimodal data fusion.

3 Deep Learning Techniques in Multimodal NLP

Deep learning has revolutionized the field of Natural Language Processing (NLP) by introducing advanced techniques capable of handling and integrating multiple modalities, such as text, images, and audio. These techniques leverage neural networks to learn complex patterns and representations, facilitating the development of sophisticated models that can understand and generate multimodal data. This section elaborates on several core deep learning techniques in multimodal NLP.

3.1 Convolutional Neural Networks (CNNs)

Convolutional Neural Networks (CNNs) are widely used for processing visual data, and their principles have been adapted for text and multimodal data processing. The A The mathematical linear operations which are performed in matrices are called Convolution. The CNN consists of the many layers which includes a convolutional layer, pooling layer, non - linearity layer and fully connected layer. The convolutional and fully-connected layers have learnable parameters, while the non-linearity and pooling layers do not have parameters.[11]

Visual Data Processing. CNNs are designed to recognize visual patterns through convolutional layers that apply filters to extract features such as edges, textures, and shapes from images. Popular architectures include AlexNet, VGG, ResNet, and Inception. These networks can extract high-level visual features that can be combined with text features for tasks like image captioning and visual question answering.

Text Data Processing. CNNs can also be applied to text data by treating sentences as sequences of word embeddings. Convolutional layers capture local dependencies and patterns within the text, making CNNs suitable for tasks like text classification and sentiment analysis.

3.2 Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) Networks

Recurrent Neural Networks (RNNs) and their variants, such as Long Short-Term Memory (LSTM) networks, are essential for handling sequential data, particularly in text and audio processing. RNN are used to handle time series data. The learning data have to be with the even timestamp for making learn the RNN based neural Network. So in this, for each run one variable act as the single model and one cell is a gated recurrent unit (GRU) which gives the needed accuracy for even the small environments like mobile devices. [12]

Text Data Processing. RNNs and LSTMs are designed to capture temporal dependencies in sequential data, making them ideal for tasks like language modeling, machine translation, and text generation. These networks can process text one word at a time, maintaining a hidden state that captures contextual information.

Audio Data Processing. LSTMs are also effective for processing audio data, such as speech recognition and synthesis. They can model the temporal dynamics of audio signals, capturing patterns in speech and other audio sequences.

4 Attention Mechanisms and Transformers

Attention mechanisms and Transformer architectures have significantly advanced the state of multimodal NLP by allowing models to focus on relevant parts of the input data dynamically.

4.1 Self-Attention Mechanisms

Self-attention mechanisms, introduced in the Transformer architecture, enable models to weigh the importance of different parts of the input sequence. This allows the model to capture long-range dependencies and contextual relationships more effectively than RNNs or LSTMs. Transformers, such as BERT and GPT, have achieved state-of-the-art performance in various NLP tasks. Chenquan et.al., proposed a Multimodal Fusion Network with a multi- head self-attention mechanism to improve performance of visual – textual sentiment analysis. [13]

4.2 Cross-Attention Mechanisms

Cross-attention mechanisms extend the concept of self-attention to multimodal data by enabling the model to attend to features from one modality based on the context provided by another modality. This approach is used in models like VisualBERT and VilBERT for tasks like visual question answering and image captioning, where textual and visual information need to be integrated. In recent years, multimodal data are increasing in social media so to focus on these multimodal data multimodal sentiment analysis are developed.[14]

5 Multimodal Transformers

Multimodal Transformers build on the success of Transformer architectures by extending them to handle multiple modalities. These models use self-attention and cross-attention mechanisms to integrate and process multimodal data effectively. In handling various machine learning operations these multimodal transformers achieved high success. It has become an important focus in the AI research field due to its increasing efficiency.[15]

5.1 CLIP (Contrastive Language-Image Pre-training)

CLIP is a multimodal Transformer that learns joint representations of text and images by training on a large dataset of image-text pairs. It uses a contrastive learning objective to align textual and visual features, enabling robust performance on tasks like image classification and zero-shot learning. Weixiong Lin et.al., build and released a biomedical dataset which is analyzed by CLIP algorithm. [16]

5.2 VisualBERT and VilBERT

These models extend the BERT architecture to process both text and visual data. VisualBERT and VilBERT use cross-attention mechanisms to align and integrate textual and visual features, making them effective for tasks like visual question answering and image captioning.

6 Multimodal Generative Models

Multimodal generative models aim to generate data in multiple modalities, such as text-to-image synthesis and image captioning. These models learn to generate coherent and contextually relevant data across different modalities. [17]

6.1 Generative Adversarial Networks (GANs)

GANs are composed of a generator and a discriminator, where the generator creates synthetic data and the discriminator evaluates its realism. Multimodal GANs, like Conditional GANs (cGANs) and CycleGANs, can generate data in one modality conditioned on another, such as generating images from textual descriptions.

6.2 Variational Autoencoders (VAEs)

VAEs are probabilistic models that learn latent representations of data and generate new data by sampling from these latent spaces. Multimodal VAEs extend this concept to integrate and generate data from multiple modalities, such as generating images and corresponding text descriptions simultaneously.

Evaluation Metrics. Evaluating multimodal deep learning models requires specialized metrics. BLEU, ROUGE, and METEOR assess the quality of generated text for tasks like image captioning and multimodal translation. Accuracy, Precision, Recall, and F1 Score are used for tasks like visual question answering and sentiment analysis to measure correctness and relevance. Human Evaluation provides additional insights into output quality and coherence that automated metrics might miss.

In summary, deep learning techniques for handling multimodal data in NLP involve advanced neural architectures like CNNs, RNNs, LSTMs, and Transformers, as well as sophisticated methods for feature extraction, data fusion, and generative modeling. These techniques enable the effective integration and processing of multiple modalities, paving the way for more sophisticated and capable AI systems. The continuous advancement and refinement of these techniques are crucial for driving further progress in the field of multimodal NLP.

7 Transfer Learning in Multimodal NLP: Fine-tuning pre-trained models for multimodal tasks

Transfer learning is a powerful technique in deep learning where a model pre-trained on a large dataset for a specific task is fine-tuned on a smaller, task-specific dataset. This approach leverages the knowledge gained from the initial training phase to improve performance on the target task. In the context of multimodal NLP, transfer learning involves adapting pre-trained models to handle tasks that require integrating and processing multiple types of data, such as text, images, and audio. Pre-trained

models have become the cornerstone of modern NLP and computer vision. These models are trained on extensive datasets and capture a wide range of features and representations that can be beneficial for various downstream tasks. Some notable pre-trained models used in multimodal NLP include:

7.1 BERT (Bidirectional Encoder Representations from Transformers)

BERT is a powerful language model pre-trained on vast amounts of text data. It captures contextual relationships between words and can be fine-tuned for various NLP tasks, including those involving multimodal data when combined with other pre-trained models.

7.2 GPT (Generative Pre-trained Transformer)

GPT is another influential language model known for its generative capabilities. It can generate coherent and contextually relevant text and be adapted for multimodal tasks by integrating it with models handling other modalities.

7.3 ResNet (Residual Networks)

ResNet is a widely-used convolutional neural network pre-trained on image datasets like ImageNet. It captures hierarchical visual features that can be combined with textual features for multimodal tasks.

7.4 CLIP (Contrastive Language-Image Pre-training)

CLIP is a multimodal model pre-trained on a large dataset of image-text pairs. It aligns textual and visual features in a shared embedding space, making it highly effective for tasks involving both text and images.

8 Multilingual Multimodal NLP: Handling multiple languages in multimodal contexts

Multilingual multimodal NLP involves the integration of multiple languages and multiple types of data, such as text, images, and audio, to create models that can understand and process information in a multilingual and multimodal context. This field merges the complexities of both multilingual NLP and multimodal NLP, addressing challenges such as language diversity, cultural nuances, and the alignment of different modalities across languages. The goal is to develop models that can operate seamlessly across different languages while effectively integrating information from various modalities.

The strong ability in understanding the texts and generating the accurate texts, the LLM models such as the BERT, GPT, PaLM, LLaMa and PanGu gained more popularity in recent days. Also the Computer Vision models like CLIP and the stable Diffusion have become more powerful. In general, LLMs have made significant breakthroughs and they form general-purpose Artificial General Intelligence (AGI). [13]

8.1 **Challenges in Multilingual Multimodal NLP**

Multimodal NLP faces challenges such as language diversity, requiring models to handle different syntaxes and semantics across languages, and capturing cultural nuances for accurate interpretation. Aligning modalities like text, images, and audio is complex, particularly with cross-lingual data. Data scarcity remains a significant issue, as multimodal datasets in various languages are limited and resource-intensive to create. Additionally, the integration of multiple languages and modalities increases model complexity, necessitating efficient algorithms to maintain performance.

8.2 **Techniques for Handling Multilingual Multimodal Data**

Several techniques and approaches are employed to address these challenges and enable effective multilingual multimodal NLP:

Pre-Trained Multilingual Models. Models like mBERT (multilingual BERT) and XLM-R (Cross-lingual Language Model - RoBERTa) are pre-trained on large multilingual corpora. These models capture linguistic features across multiple languages and can be fine-tuned for specific multimodal tasks, leveraging their multilingual capabilities.

Multimodal Embeddings. Multimodal embeddings represent data from different modalities in a shared space. Techniques like multilingual CLIP (Contrastive Language-Image Pre-training) extend this concept to handle multiple languages, aligning textual and visual features across different languages.

Cross-Lingual and Cross-Modal Attention. Attention mechanisms are adapted to handle both cross-lingual and cross-modal interactions. Cross-lingual attention allows the model to focus on relevant parts of the text in different languages, while cross-modal attention integrates information from various modalities, such as aligning descriptions in different languages with corresponding images.

Translation and Alignment. Machine translation models can translate text from one language to another, facilitating the alignment of multimodal data across languages. This is particularly useful for tasks like image captioning, where captions need to be generated or interpreted in multiple languages.

Multimodal Transformers. Transformers are extended to handle multilingual and multimodal data by incorporating layers that process and integrate features from multiple languages and modalities. These models, such as Multilingual Multimodal

Transformers (MMT), leverage both self-attention and cross-attention mechanisms to capture complex relationships.

Data Augmentation. Techniques like back-translation and synthetic data generation are used to augment multilingual multimodal datasets. This helps address data scarcity by generating additional training examples that capture diverse linguistic and multimodal patterns.

9 Applications of Multimodal NLP

Multimodal NLP has diverse applications across various fields. In Visual Question Answering (VQA), systems answer questions about images by combining CNNs for image processing and RNNs/transformers for text, using attention mechanisms for alignment. Image Captioning similarly blends computer vision and NLP to generate descriptions for images. Speech Recognition and Synthesis involve converting speech to text and vice versa, often using multimodal data like lip movements to improve accuracy. Multimodal Machine Translation incorporates visual or video data for better contextual translations. Multimodal Sentiment Analysis gauges emotions by analyzing text, audio, and images together. AR/VR integrates text, visual, and audio elements for immersive user experiences. Dialogue Systems and Chatbots leverage multiple modalities for more natural, engaging interactions. Autonomous Vehicles process visual, textual, and audio inputs to navigate safely. Healthcare and Assistive Technologies combine multimodal data for diagnostics and support, while Educational Tools use multimodal NLP to create interactive, adaptive learning environments. These applications demonstrate how integrating multiple modalities improves AI system accuracy, understanding, and user engagement.

10 Results and Discussions

Some of the research issues and potential needs in water quality assessment and prediction have been identified based on literature surveys and are highlighted below. The possible solutions have been proposed.

10.1 Challenges in NLP and Proposed Solutions

Data Quality and Availability – Challenge. The large scale datasets with high quality and diverse formats are important for the training of the various NLP models. However, there are various languages and many specific domains which suffer from the scarcity of good quality data which in turn lead to inefficient and inaccurate models.

Solutions.

Data Augmentation. The following techniques such as back-translation, paraphrasing, and synthetic data generation which can help in increasing the size of the training datasets.

Crowdsourcing and Community Initiatives. incorporating the communities and crowdsourcing which can help collecting more data specifically from very uncommon languages.

Unsupervised Learning. These unsupervised learning algorithms which do not rely on labeled data to learn from raw data.

Bias and Fairness – Challenge.

NLP models mostly pick the biases from the training data which will lead them in making the inefficient outcomes specifically towards the minority groups.

Solutions.

Bias Mitigation Techniques. The mitigation techniques like adversarial training, fairness-aware regularization, and debiasing word embeddings which can be used to reduce the bias of the model.

Diverse Data Collection. To reduce the inherent bias, the dataset comprises the different languages and various cultural contexts.

Transparent Model Development. The model development and testing can adopt the transparency practices with including various stakeholders to ensure the accountability and fairness throughout the entire process.

Explainability and Interpretability – Challenge. Based on the deep learning algorithms the complexity of the NLP models are difficult to make decisions.

Solutions.

Explainable AI Techniques. The methods like SHAP, LIME and attention visualization are used in making the model predictions more interpretable

Simpler Models for Critical Applications. Use the interpretable models for the basic understanding of the decisions taken accurately.

Human-in-the-Loop Systems. Include the experts to review and also incorporate the oversight by the human in making the process of making the final decision.

Evaluation Metrics – Challenge.

BLEU or F1 score are the evaluation metrics that are not exactly capturing the practical performance of the developed models. Also these metrics often fail to identify the bias and real world applicability.

Solutions.

Comprehensive Evaluation Frameworks. Consider developing and adapting to the evaluation frameworks to consider the multiple dimensions of the algorithm performance like accuracy, fairness, bias and interpretability.

Task-Specific Metrics. Develop specific metrics for the specific NLP tasks.

Continuous Benchmarking. Benchmark new and diverse datasets to make the model perform well across various contexts and scenarios.

Ethical Considerations – Challenge. Due to the world wide adoption of the NLP technologies there raised the ethical concerns which includes the privacy issues, misinformation through manipulation and the impact on the societal structures.

Solutions.

Ethical AI Guidelines. Ethical AI guidelines should be followed for the responsible development and deployment of the NLP algorithms and ensure that they respect societal values and legal ethical requirements.

Privacy-Preserving Techniques. Incorporate differential privacy or federated learning to protect sensitive information to enable model development.

Stakeholder Engagement. Engage various stakeholders to design and develop in NLP systems which can ensure the ethical consideration that can be addressed from the start and that potential societal impact which are fully understood and mitigated.

Handling Low-Resource Languages – Challenge. Languages which are spoken by the smaller populations which have lack of resources to build robust NLP models.

Solutions.

Cross-Lingual Transfer Learning. Using the transfer learning by using the mBERT and XLM - R to increase the performance for high resource languages to the low resource languages.
Data Sharing and Collaboration. By doing international collaboration, the data sharing can be done. By which we can use various resources and diverse expertise to build NLP models which can be used for the low resource languages.
Language-Specific Initiatives. Promote and support initiatives which focus on collecting and curating the data from low resource languages which can be any funded projects or government initiatives.

By tackling the challenges discussed above the new innovative technical approach, ethical consideration and collaborative efforts can be achieved for the NLP community. By concentrating on the discussed solutions, scientists can improve this field to be responsible for society.

11 Conclusion

Nowadays, the powerful program Multimodal NLP has been developed which enables to develop a system that can understand more modalities like text along with images and audio. Recent progress in multimodal transformers and both supervised pre-training and fine-tuning models show the efficacy of these kinds of models in learning representations from complex, multimodal inputs. Nevertheless, it still faces some challenges like data alignment and fusion strategies, and interpretability. While this interaction paves the way for a whole new set of tasks such as more accurate sentiment analysis, machine translation and visual question answering to be tackled, it also makes necessary the research on how neural models can efficiently cope with such complex and resource-intensive multimodal systems. Future work should emphasize the interpretability of multimodal models and ethical considerations, especially as they find increased utility in high-stakes applications. The latest advancements in Multimodal NLP are opening the door for efficient models and the system that is aware of the context effectively.

12 References

1. Poria, S., Soon, O.Y., Liu, B. *et al.* Affect Recognition for Multimodal Natural Language Processing. *Cogn Comput* **13**, 229–230 (2021). https://doi.org/10.1007/s12559-020-09738-0
2. L. Ma, Z. Lu, L. Shang and H. Li, "Multimodal Convolutional Neural Networks for Matching Image and Sentence," *2015 IEEE International Conference on Computer Vision (ICCV)*, Santiago, Chile, 2015, pp. 2623-2631, doi: 10.1109/ICCV.2015.301.
3. A. L. Kotian, R. Nandipi, U. M, U. R. S, VARSHAUK and V. G. T, "A Systematic Review on Human and Computer Interaction," *2024 2nd International Conference on Intelligent Data Communication Technologies and Internet of Things (IDCIoT)*, Bengaluru, India, 2024, pp. 1214-1218, doi: 10.1109/IDCIoT59759.2024.10467622.
4. arXiv:2105.11087.
5. arXiv:2311.13165.
6. Venugopalan, J., Tong, L., Hassanzadeh, H.R. *et al.* Multimodal deep learning models for early detection of Alzheimer's disease stage. *Sci Rep* **11**, 3254 (2021). https://doi.org/10.1038/s41598-020-74399-w.

7. Kim KH, Sohn SY. Hybrid neural network with cost-sensitive support vector machine for class-imbalanced multimodal data. Neural Netw. 2020 Oct;130:176-184.
8. Huynh-Cam T-T, Chen L-S, Le H. Using Decision Trees and Random Forest Algorithms to Predict and Determine Factors Contributing to First-Year University Students' Learning Performance. *Algorithms*. 2021; 14(11):318. https://doi.org/10.3390/a14110318.
9. Abbas, M., Memon, K. A., Jamali, A. A., Memon, S., & Ahmed, A. (2019). Multinomial Naive Bayes classification model for sentiment analysis. IJCSNS Int. J. Comput. Sci. Netw. Secur, 19(3), 62.
10. Mangla, P., Singh, G., Pathak, N., Chawla, S. (2024). Language Identification Using Multinomial Naive Bayes Technique. In: Swaroop, A., Polkowski, Z., Correia, S.D., Virdee, B. (eds) Proceedings of Data Analytics and Management. ICDAM 2023. Lecture Notes in Networks and Systems, vol 786. Springer, Singapore. https://doi.org/10.1007/978-981-99-6547-2_24.
11. S. Albawi, T. A. Mohammed and S. Al-Zawi, "Understanding of a convolutional neural network," *2017 International Conference on Engineering and Technology (ICET)*, Antalya, Turkey, 2017, pp. 1-6, doi: 10.1109/ICEngTechnol.2017.8308186.
12. Kim J-C, Chung K. Recurrent Neural Network-Based Multimodal Deep Learning for Estimating Missing Values in Healthcare. *Applied Sciences*. 2022; 12(15):7477. https://doi.org/10.3390/app12157477.
13. Huang D, Yan C, Li Q, Peng X. From Large Language Models to Large Multimodal Models: A Literature Review. *Applied Sciences*. 2024; 14(12):5068. https://doi.org/10.3390/app14125068.