JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

EXPLORING DEEP LEARNING APPROACHES FORMULTI-MODAL DATA FUSION IN SENTIMENT ANALYSIS

[1]K. Devika, [2]Dr.CH.Avinash,
MTech, Associate Professor, Department of Computer Science and Engineering,
Gayatri Vidya Parishad College of Engineering (Autonomous), Visakhapatnam, India

ABSTRACT: In our increasingly interconnected digital world, gaining a profound understanding of human emotions and concepts is indispensable for making insightful decisions across diverse fields. In this sentiment analysis plays a vital role, enabling us to go through the digital interactions by decoding the emotional content expressed. However, traditional approaches often face challenges when confronted with the intricacies of human language, as they overlook crucial information conveyed through visual means. To overcome this constraint, we introduce an innovative method that utilizes the combination of multi-modal data fusion and deep learning approaches. We employ a hybrid model design to integrate textual and visual data. We achieve this by using pre-trained TF-IDF to extract textual features and fine- tuned ResNet50 to extract visual features. We fuse the features and then perform sentiment classification. Through thorough experimentation, we demonstrate the effectiveness of our method, especially in situations where textual and visual signals provide additional information. The outcomes of our study exhibit substantial progress in both the accuracy and reliability of sentiment classification when compared to conventional unimodal methods. Furthermore, we use our algorithm on actual data, effectively forecasting emotions conveyed in both textual and visual formats. The assessment, which utilizes classification metrics, showcases the exceptional accuracy of our multi- modal fusion technique, achieving 98% accuracy. This research enhances sentiment analysis approaches by leveraging the combined strength of textual and visual data. The approach we propose presents a great opportunity to improve sentiment comprehension in diverse areas such as social media analysis, customer feedback interpretation, and opinion mining. We are confident that our study will provide stakeholders with practical and valuable information obtained from the combination of written and visual signals. This will facilitate more knowledgeable decision-making in the digital era.

Index Terms: Image Captioning, Feature Extraction, Deep Learning, ResNet50, BLIP Model, Sentiment analysis, Text features, Image features, Modalities, Multi-modal Data Fusion.

I. INTRODUCTION

Sentiment analysis has emerged as an essential part of many domains, such as business, marketing, customer service, healthcare, social media monitoring, and market research. It enables organizations to assess consumer happiness, detect developing trends, and track brand reputation in real-time scenarios. Consumer care systems use sentiment analysis to categorize and rank incoming messages according to their sentiment, enabling fast resolution of consumer problems. Political campaigns use opinion analysis to gauge voter opinion, identify pressing issues, and appropriately adjust communications strategies. Government entities use sentiment analysis to monitor the prevailing public attitude and take proactive measures to resolve any problems. Sentiment analysis in healthcare involves examining patient comments, social media postings, and online forums.

Market research uses sentiment analysis to examine customer sentiments, preferences, and actions. The field of sentiment analysis has seen substantial advancements throughout time, progressing from initial rule-based systems to more sophisticated machine learning and deep learning models. Between the late 2000s and early 2010s, there was an increase in the use of advanced machine learning models, feature engineering, ensemble learning, and deep learning techniques. This progress has led to the development of aspect-based sentiment analysis. In this era, sentiment analysis is crucial for understanding human attitudes and behaviors. The current state and challenges of multimodal classification research highlight the growing interest in combining data from multiple modalities to improve machine learning-based classification models. We propose a new taxonomy to tackle these issues, which encompasses five major stages: pre-processing, feature extraction, data fusion, primary learner, and final classifier. The taxonomy aims to provide a structured way to describe multimodal architectures and their applications across fields like medicine, hyperspatial imagery, and sentiment analysis. [1]They have developed a novel multimodal biometric human identification system that utilizes convolutional neural networks (CNNs) to identify humans through the iris, face, and finger vein biometric modalities. The system uses the Adam optimization method and categorical cross-entropy as loss functions. Thee evaluated the system's performance using the SDUMLA-HMT dataset, which revealed that using three biometric traits in biometric identification systems yielded better results than using two or one trait. With a feature-level fusion approach, the system achieved an accuracy rate of 99.39%, while with different methods of score-level fusion, it achieved 100%. [2]. SensorNet is a Deep Convolutional Neural Network (DCNN) that can be scaled up or down. It accurately detects 98% of multimodal time series signals in embedded settings. It adapts quickly to new sensor modalities and is suitable for IoT and wearable device deployment. [3]

The GwPeSOA is a modified version of the Glow-worm Optimization Algorithm, utilizing ear and finger vein modalities for feature extraction. It achieves better accuracy, specificity, and sensitivity and is adaptable to local decision domains. It addresses challenges in biometric authentication. [4]. The DFF-ATMF model enhances audio-text sentiment analysis accuracy by combining multi-feature and

multi-modality fusion. It outperforms state-of-the-art models on the CMU-MOSI and CMU-MOSEI datasets. The model extracts four sentiment features and uses batch normalization, the ReLU function, and dropout regularization. [5]

The purpose of this undertaking is not solely to classify texts as positive, negative, or neutral, but also to extract the latent emotions that lie beneath them, such as surprise, wrath, pleasure, or sorrow. These functionalities enhance the profundity of sentiment analysis, providing a more intricate understanding of human sentiment and behavior. The implications of sentiment analysis extend to numerous sectors, including but not limited to healthcare, politics, business, and marketing. With the continuous advancement of technology, sentiment analysis is poised to achieve an even higher level of sophistication, allowing for a more precise examination of the complexities associated with human sentiment and behavior. However, sentiment analysis is not a standalone process. Since the introduction of multimodal data, which includes visuals, audio, and video in addition to text, the field of sentiment analysis has grown exponentially. Incorporating data from various modalities has emerged as a critical requirement to obtain comprehensive contextual insights and improve the accuracy of sentiment analysis algorithms. Moreover, incorporating deep learning techniques has escalated sentiment analysis to the domain of affective analytics, which includes emotion recognition and sentiment analysis.

This more expansive field of study recognizes the intrinsic relationship between emotions and sentiments, as well as their crucial significance in areas such as education, communication, and decision-making. This research undertakes an extensive investigation of sentiment analysis, detailing its development from its infancy to its present level of complexity. In this study, we examine the potential of combining deep learning methods and multimodal data fusion techniques. We explain how each of these methods helps to improve sentiment analysis methods. Through meticulous experimentation and assessment, our objective is to proactively contribute to the continuous advancement of sentiment analysis. We hope to provide stakeholders with practical insights derived from the complex nature of human emotions in the digital age. Furthermore, we discuss the domain of multimodal classification, where we suggest a novel taxonomy to characterize multimodal classification models and outline potential avenues for future research in this rapidly developing field. It demonstrated remarkable results with an accuracy rate of 98%.

II. RELATED WORK

The progress in multimodal classification has resulted in the development of enormous datasets in diverse domains, including satellite imaging, biometrics, and medicine. However, comparing multiple systems becomes difficult because there is no consistent terminology or architectural explanation. We have developed a new classification system to address this issue and categorize multimodal classification models. This system specifically addresses the difficulties associated with managing data, dealing with imbalanced classes, and handling complex instances. In the field of biometric person identification systems, a new method has emerged that utilizes advanced deep learning algorithms to accurately identify individuals by analyzing their iris, face, and finger vein biometric traits. This method employs three convolutional neural network (CNN) models for each biometric feature. To mitigate the problem of overfitting, the approach employs Adam optimization, categorical cross-entropy as the loss function, and dropout regularization. Comparing results from systems that use fewer biometric features to systems that use the SDUMLA-HMT dataset for testing algorithms shows better results.

The DFF-ATMF model introduces a new approach to sentiment analysis by combining multi-feature fusion with multi-modality fusion to enhance accuracy. This model utilizes two parallel branches for audio and text modalities to obtain comparable performance on datasets like CMU-MOSI and CMU-MOSEI. It achieves this by extracting deep features and utilizing batch normalization, the ReLU function, and dropout regularization. Additionally, a study generates a text sentiment vector (TSV) by training a language representation model on multimodal datasets using BERT embeddings. This model uses a multimodal attention mechanism in an encoder-decoder framework to combine audio and text modalities successfully. This makes sentiment analysis more accurate.

III. DATASET

The evolution of multimodal datasets for sentiment analysis has been characterized by a progressive integration of diverse modalities and an expansion into various domains and contexts. What began with text-only datasets for sentiment classification has evolved into a rich landscape of multimodal datasets encompassing textual, visual, and audio data sources, enabling more nuanced and comprehensive sentiment analysis tasks. Machine learning and deep learning models, such as neural networks, support vector machines, and ensemble methods, train on these datasets. Researchers also employ these datasets for evaluating the performance of their models through metrics like accuracy, precision, recall, and F1-score. Table 1 gives detailed information about the data sets available for sentiment analysis.

Table 1: Datasets Overview

DATASET NAME	MODALITIES	DESCRIPTION
CMU Multimodal Opinion Sentiment and Emotion (CMU-MOSEI)	Text, Audio, Video.	Contains 23,000 utterances from more than 1,000 YouTube speakers, annotated for sentiment and emotional intensity.
CMU Multimodal Opinion Sentiment and Emotion Intensity (CMU-MOSI)	Text, Audio, Video.	Includes 2,199 opinion video clips from YouTube product reviews, annotated for sentiment.
Affective Computing Multimedia Database (AMMER)	Text, Audio, Video.	Focuses on multimodal sentiment and emotion analysis.
Emotion Recognition in Speech (EROS)	Audio.	A dataset focusing on speech audio clips annotated for different emotional states.
Facial Expression Recognition (FER)	Images	It consists of over 35000 facial expressions

The insights from analyzing these datasets have practical applications in various domains, including market research, customer feedback analysis, social media monitoring, mental health assessment, human-computer interaction, and content recommendation systems. Both industry professionals and academic researchers benefit from the knowledge gained from analyzing these datasets, driving innovation and improvements in sentiment analysis technologies. The Facial Expression Recognition (FER) dataset is a widely used benchmark dataset in the fields of computer vision and affective computing. People primarily use it for training and evaluating models for facial expression recognition tasks.

The FER dataset contains 48x48-pixel grayscale images of faces. We automatically registered the faces to ensure they are roughly centered and occupy the same amount of space in each image. The main purpose is to categorize each face into one of seven categories based on the emotion shown in the face (0=Angry, 1=Disgust, 2=Fear, 3=Happy, 4=Sad, 5=Surprise, 6=Neutral). The training set contains 28,709 examples, and the public test set consists of 3,589 examples. The FER dataset serves as a foundational resource for advanced research in facial expression recognition and related fields, enabling the development of more accurate and robust models for understanding human emotions based on facial cues.

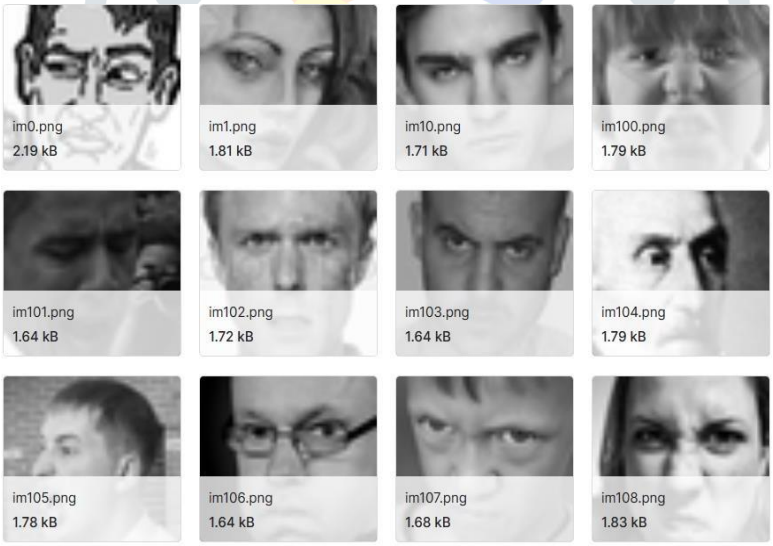

Figure 1: Sample images of FER-2013

SYNTHETIC DATA SET (FOR TEXT)

We create the text data set using the BLIP model. These components are pre-trained models specifically tailored for image captioning tasks. A designated function then initiates caption generation. This function operates by loading individual images, processing them with a predefined text prompt, and leveraging the BLIP model to generate captions. Notably, the function accommodates both conditional and unconditional captioning scenarios, providing flexibility in generating diverse sets of captions. After generating captions, the script advances to the stages of image processing and CSV writing. The script generates a corresponding caption for each processed image, seamlessly integrating it with the image path and its associated emotion category into the CSV file. This meticulous process ensures the creation of a comprehensive synthetic dataset conducive to training and evaluating image captioning models across various emotional contexts.

IV. SYSTEM DESIGN

We developed the proposed system to design a neural network model for sentiment classification that requires less training time and yields high accuracy. The study considered two deep learning models and evaluated their performance using the FER dataset. We started employing ANN architectures for model training and trained the dataset initially with a basic ANN model architecture. ANNs are effective at image classifications because they extract spatial hierarchies of features such as edges, textures, and shapes, which are important for object recognition. Here is the model's basic system design.

We take the FER data set as the input. We collect text data relevant to the image. This could be a caption, description, or keywords that describe the content of the image. We extract features from the captured image. Feature extraction is a process of analyzing an image to identify and extract the visual characteristics from the data. The flowchart employs ResNet-50, a deep convolutional neural network architecture, for feature extraction. We also extract features from the text data. TF-IDF (Term Frequency-Inverse Document Frequency) is a common technique for text feature extraction. It assigns weights to words based on their importance within a document and its collection.

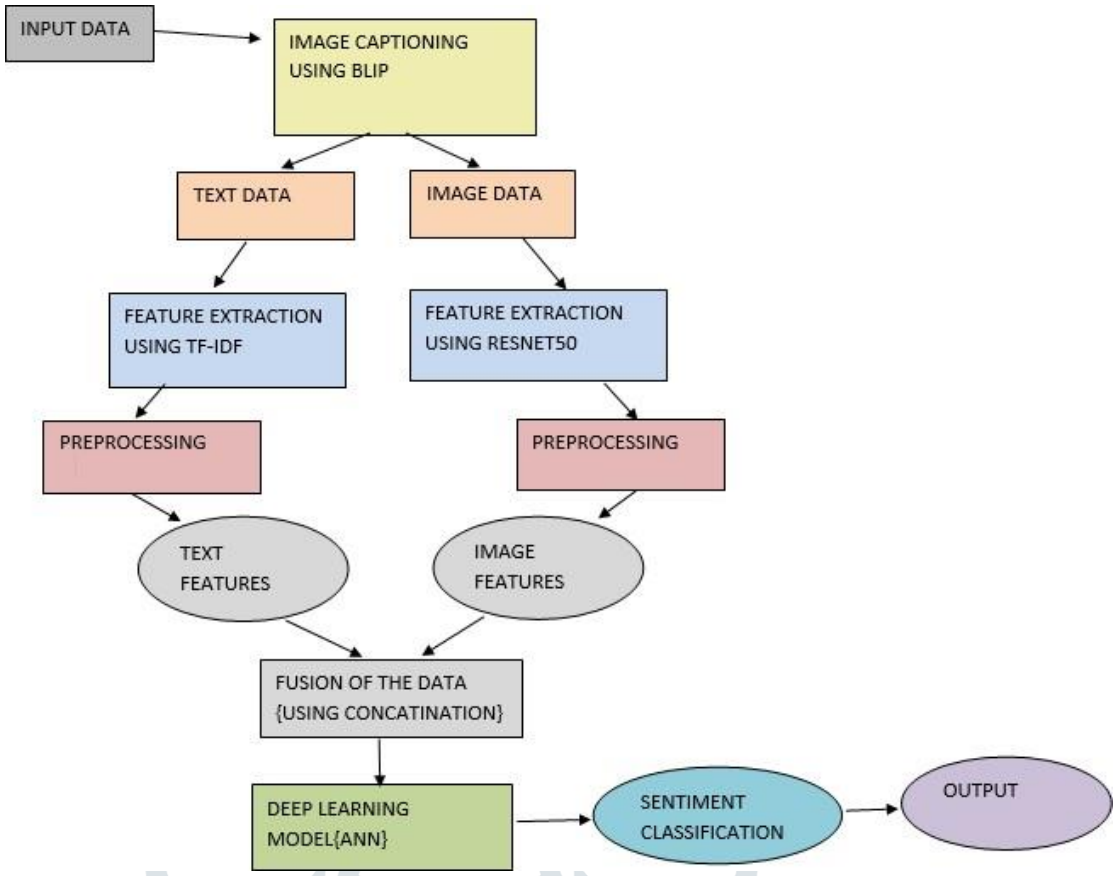

Figure 2: Proposed System Architecture

Both the image and text data features undergo pre-processing. Pre-processing techniques can include normalization, scaling, and dimensionality reduction. We use these techniques to prepare the data for further processing. We then fuse or combine the pre-processed image and text features. Concatenation, a common technique for data fusion, involves simply joining features to create a new, combined feature vector. The flowchart feeds the fused features into a deep learning model, specifically an artificial neural network (ANN). We train the deep learning model on a large dataset of images and captions. During training, the model learns to map the image and text features to a caption that describes the image's content. The final step involves sentiment classification. Here, the model assigns a sentiment to thecaption, such as positive, negative, or neutral.

V. METHODOLOGY

This study aims to develop an emotion detection system using image data and text captions. We divide the methodology into several phases, each with unique duties and processes. The combination of image data and text allows for a more comprehensive understanding of emotions expressed in digital content. Data acquisition and pre-processing entail loading text data into memory using appropriate libraries, ensuring accuracy, and addressing any missing or inconsistent data entries. Data splitting involves dividing the dataset into training and validation sets to ensure the model can generalize to previously unseen data. Feature extraction includes TF-IDF vectorization, dimension reduction, image feature extraction, and feature fusion. Data pre-processing involves normalizing concatenated features using Sklearn's StandardScaler. Model construction involves developing an artificial neural network (ANN) model with optimized layers for text feature input, image feature input, and concatenated data. Model training and evaluation involve training each model and evaluating its performance using validation data, fine- tuning, and comparison. We use analysis and visualization to examine each model's performance over epochs and evaluate its efficacy. We save the results for future use and document them, providing conclusions and insights derived from the analysis. This comprehensive methodology provides a thorough understanding of emotion detection using both text and image data, emphasizing the importance of ongoing monitoring and maintenance.

VI. EXPERIMENT AND STUDIES

The study used two datasets containing text and image features to create a comprehensive dataset for emotion detection. We extracted the text features using TF-IDF and dimensionality reduction with Truncated SVD, and extracted the image features using a pre-trained ResNet50 model. We split the fused dataset into training and validation sets with an 80-20 ratio and normalized the features using StandardScaler. We designed a deep neural network (DNN) with an input layer equal to the number of features in the fused dataset, hidden layers using ReLU activation, and an output layer equal to the number of emotion classes. We trained the model for 20 epochs with a batch size of 32. The results showed that the fused model outperformed standalone models trained on text or image data alone, demonstrating the efficacy of combining multiple modalities for emotion detection. The training and validation accuracy trends indicated effective learning and good generalization. The basic ANN method has shown remarkable performance with 98% accuracy.

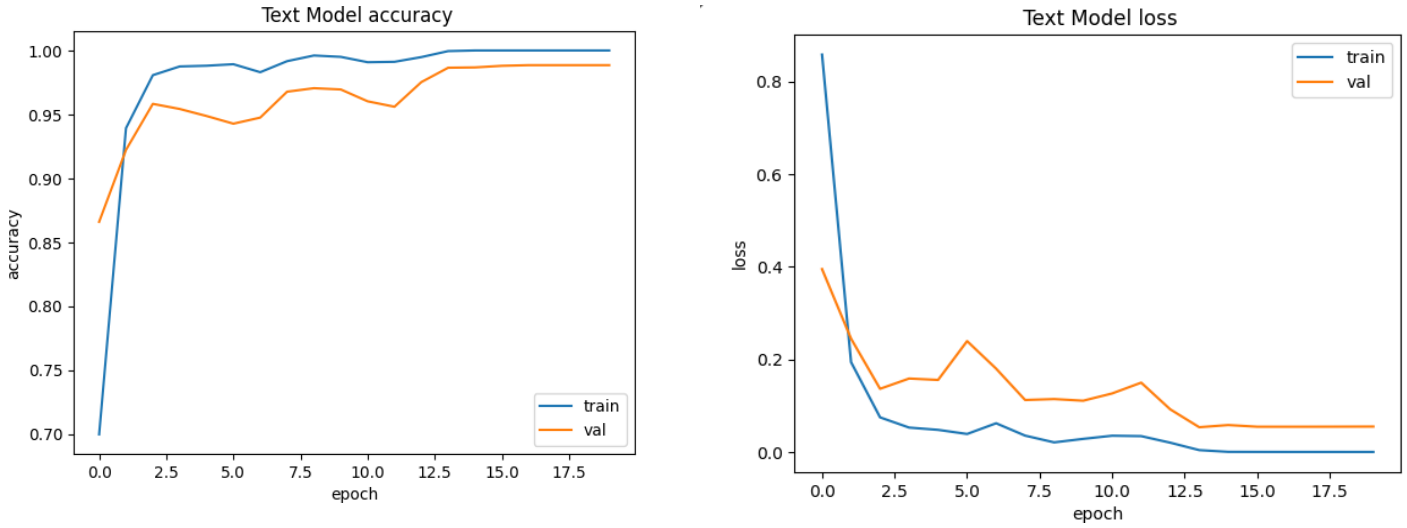

Figure 3: Graph Representation of Accuracy and Loss.

VII. CONCLUSION AND FUTURE SCOPE

The deep learning framework for multimodal data fusion in sentiment analysis has shown remarkable success, achieving an accuracy rate of 98% in real-world testing. This innovative approach integrates textual and visual characteristics to capture subtle sentiments and improve the comprehension of human emotions. The concatenated model, which combines features from both modalities, outperformed standalone models, demonstrating the importance of leveraging multiple sources of information for emotion detection. This method's success highlights the potential advantages of multimodal fusion techniques for improving model performance and capturing nuanced emotional indicators. Looking ahead, future research and development opportunities abound. Better fusion methods, like attention mechanisms or graph neural networks, can make models work even better by combining information from text and image formats more efficiently. Data augmentation techniques can improve model generalization and robustness to variations in emotional expressions while fine-tuning pre-trained models can help adapt to the subtleties of emotional expressions in the dataset. Additionally, user interaction through interactive interfaces can enable real-time emotion detection, allowing users to input text or upload images for analysis. Assessing the system's efficacy in real-world scenarios, such as customer service applications or social media platforms, will provide valuable insights into its practical utility. Furthermore, including audio data can offer a more comprehensive understanding of the user's sentiments, enriching the emotion detection process. Ethical considerations are crucial when implementing emotion detection systems. Addressing concerns about privacy, bias, and impartiality ensures responsible and equitable use of such technology, safeguarding users' rights, and promoting fairness in emotional analysis. Hence, the deep learning methods for multimodal data fusion in sentiment analysis have demonstrated significant potential for advancing emotion detection capabilities. Continued research and development efforts focused on enhancing performance, usability, and ethical considerations will further solidify this innovative approach's practical applications and impact.

VIII. REFERENCES

[1]. Multimodal Classification: Current Landscape, Taxonomy, and Future Directions. WILLIAM C. SLEEMAN IV, RISHABH KAPOOR, and PREETAM GHOSH, Virginia Commonwealth University, USA. ACM Computing Surveys, Vol. 55, No. 7, Article 150. Publication date: December 2022.

[2]. Deep learning approach for multimodal biometric recognition system based on the fusion of iris, face, and finger vein traits. Sensors 20, 19 (2020), 5523. Nada Alay and Heyam H. Al-Baity. 2020.

[3]. SensorNet: A scalable and low-power deep convolutional neural network for multimodal data classification. IEEE Trans. Circ. Syst. I: Reg. Pap. 66, 1 (2018), 274–287. Ali Jafari, Ashwinkumar Ganesan, Chetan Sai Kumar Thalisetty, Varun Sivasubramanian, Tim Oates, and Tinoosh Mohsenin. 2018.

[4]. The multimodal biometric system using ear and palm vein recognition based on GwPeSOA: Multi-SVNN for security applications. In International Conference on Computational Vision and Bio-Inspired Computing. Springer, 215–231,2018. M. Vijay and G. Indumathi.

[5]. Complementary Fusion of Multi-Features and Multi-Modalities in Sentiment Analysis. Feiyang Chen Department of Computer Science and Technology, Beijing Forestry University fychen98.ai@gmail.com. arXiv:1904.08138v5 [cs.CL] 11 Dec 2019.

[6]. Multimodal Sentiment Analysis using Hierarchical Fusion with Context Modeling .N. Majumdar, D. Hazarikab, A. Gelbukha, E. Cambriac, S. Poriac .arXiv:1806.06228v1 [cs.CL] 16 Jun 2018.

[7].Deep multimodal fusion for semantic image segmentation: A survey. Image Vis. Comput. (2020), 104042. Yifei Zhang, Désiré Sidibé, Olivier Morel, and Fabrice Mériaudeau. 2020. https://doi.org/10.1016/j.imavis.2020.104042.

[8]. Entity-sensitive attention and fusion network for entity-level multimodal sentiment classification. Jianfei Yu, Jing Jiang, and Rui Xia. 2019. IEEE/ACM Trans. Audio, Speech, Lang. Process. 28 (2019), 429–439

[9] Federico Simonetta, Stavros Ntalampiras, and Federico Avanzini. 2019. Multimodal music information processing and retrieval: Survey and future challenges. In International Workshop on Multilayer Music Representation and Processing (MMRP). IEEE, 10–18.

[10] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. arxiv preprint arXiv:1409.1556 (2014).

[11] William C. Sleeman IV and Bartosz Krawczyk. 2021. Multi-class imbalanced big data classification on Spark. Knowl.- based Syst. 212 (2021), 106598.

[12] Mohammad Soleymani, David Garcia, Brendan Jou, Björn Schuller, Shih-Fu Chang, and Maja Pantic. 2017. A survey of multimodal sentiment analysis. Image Vis. Comput. 65 (2017), 3–14.

[13] Jingqi Song, Yuanjie Zheng, Muhammad Zakir Ullah, Junxia Wang, Yanyun Jiang, Chenxi Xu, Zhenxing Zou, and Guocheng Ding. 2021. Multiview multimodal network for breast cancer diagnosis in contrast-enhanced spectral mammography images. Int. J. Comput. Assist. Radiol. Surg. 16, 6 (2021), 979–988.

[14] Phillips, P.J.; Moon, H.; Rauss, P.; Rizvi, S. The FERET Evaluation Methodology for Face Recognition Algorithms. IEEE Proc. Comput. Vis. Pattern Recognit. 2000, 22, 1090–1104. [CrossRef].

[15] Ahmad Radzi, S.; Khalil-Hani, M.; Bakhteri, R. Finger-vein biometric identification using convolutional neural network. Turkish J. Electr. Eng. Comput. Sci. 2016, 24, 1863–1878. [CrossRef].

[16] Fernandes, S.L.; Bala, G.J. Analyzing State-of-the-Art Techniques for Fusion of Multimodal Biometrics. Second Int. Conf. Comput. Commun. Technol. 2016, 381, 473–478. [CrossRef].

[17] M. Motamedi, P. Gysel, V. Akella, and S. Ghiasi, “Design space exploration of FPGA-based deep convolutional neural networks,” in Proc. 21st Asia South Pacific Design Automat. Conf. (ASP-DAC), Jan. 2016, pp. 575–580.

[18] H. Nakahara and T. Sasao, “A deep convolutional neural network based on nested residue number system,” in Proc. 25th Int. Conf. Field Program. Logic Appl. (FPL), Sep. 2015, pp. 1–6.

[19] Z. Li et al., “Structural design optimization for deep convolutional neural networks using stochastic computing,” in Proc. Design Automat. Test Eur. Conf. Exhibit., Mar. 2017, pp. 250–253.

[20] S. Poria, E. Cambria, D. Hazarika, N. Mazumder, A. Zadeh, L.-P. Morency, Context-dependent sentiment analysis in user-generated videos, in ACL, 873–883, 2017.

[21] E. Cambria, Affective Computing and Sentiment Analysis, IEEE Intelligent Systems 31 (2) (2016) 102–107. 22

[22] S. Poria, E. Cambria, R. Bajpai, A. Hussain, A review of affective computing: From unimodal analysis to multimodal fusion, Information Fusion 37 (2017) 98–125.

[23] D. Hazarika, S. Poria, A. Zadeh, E. Cambria, L.-P. Morency, R. Zimmer man, Conversational memory network for emotion recognition in dyadic dialogue videos, in NAACL, 2122–2132, 2018.

[24] I. Chaturvedi, E. Cambria, R. Welsch, F. Herrera, Distinguishing between facts and opinions for sentiment analysis: Survey and challenges, Information Fusion 44 (2018) 65–7