Journal of Ambient Intelligence and Humanized Computing (2023) 14:4489–4501
https://doi.org/10.1007/s12652-023-04567-z

ORIGINAL RESEARCH

A multimodal sentiment analysis system for recognizing person aggressiveness in pain based on textual and visual information

Anay Ghosh[1] · Bibhas Chandra Dhara[2] · Chiara Pero[3] · Saiyed Umer[4]

Received: 8 April 2022 / Accepted: 8 February 2023 / Published online: 2 March 2023
© The Author(s) 2023

Abstract

This article proposes a multimodal sentiment analysis system for recognizing a person's aggressiveness in pain. The implementation has been divided into five components. The first three steps are related to a text-based sentiment analysis system to perform classification tasks such as predicting the classes into non-aggressive, covertly aggressive, and overtly aggressive classes. The remaining two components are related to an image-based sentiment analysis system. A deep learning-based approach has been employed to do feature learning and predict the three types of pain classes. An aggression dataset for the text-based system and the UNBC-McMaster database for an image-based system has been employed, respectively. Experimental results have been compared with the state-of-the-art methods, showing the superiority of the proposed approach. Finally, the scores due to text-based and image-based sentiment analysis systems are fused to obtain the performance for the proposed multimodal sentiment analysis system.

Keywords Sentiment analysis · Pain recognition · Person aggressiveness · Text data · Image data

1 Introduction

Sentiment analysis (SA) is realizing various opinions from different entities like events, issues, aggression, anger, attitude, etc. Sentiment analysis tries to categorize the sentiment of people's opinions into three main categories: positive, negative, and neutral (Ghosh et al. 2022). Nowadays, the research activities are not limited only to finding positive, negative, or neutral sentiment but also to finding the amount of positivity and negativity with the help of sentiment scores through natural language processing, images/video, and audio. It has been observed that the various activities of different users on online platforms are growing day by day at a rapid proportion. With the increased amount of interaction and the increased number of people involved in these interactions over the web, various aggression-oriented activities like flaming, trolling, roasting, and cyberbullying have also increased (Ebrahimi et al. 2017) globally. Sentiment Analysis includes applications such as spam email detection, bullying, trolling, suicidal tendency, aggression, fraud messages, roasting, etc. Several problems were solved using text, audio, image, or video-based sentiment analysis methods (Li and Xu 2019). For example, if we have a detailed analysis of the aggressive text, we will be able to measure the intensity of the aggression. We can analyze various images to detect pain, happiness, and other human emotions. Any particular video can easily track the changes in emotions with the change in the person's activity. There are several cases like aggressive detection that are possible with text, image, and video, but there will be some situations where sentiment analysis based on a single data will not be enough to get

✉ Chiara Pero
cpero@unisa.it

Anay Ghosh
anay.ghosh1@gmail.com

Bibhas Chandra Dhara
bibhas@it.jusl.ac.in

Saiyed Umer
saiyedumer@gmail.com

[1] Department of Computer Science and Engineering, University of Engineering and Management, Newtown, Kolkata, India

[2] Department of Information Technology, Jadavpur University, Kolkata, India

[3] Department of Computer Science, University of Salerno, Fisciano, Italy

[4] Department of Computer Science and Engineering, Aliah University, Kolkata, India

the correct result, as explained in detail in the following subsections.

1.1 Text-based sentiment analysis

Sentiment analysis using text is being performed using text data. This approach goes through a specific set of activities like preprocessing of text data (removing stop words, articles, prepositions, pronouns, be verbs, and other words), tokenization of words, feature extraction (using a bag of words, aspect-based features, context-based features, etc.), and finally, the classification tasks using well-known classifiers such as random forest (Mursalin et al. 2017), support vector machine (Noble 2006), logistic regression (Hosmer Jr et al. 2013), Naïve Bayes (Malmasi and Zampieri 2017) and Neural Network based approaches (Sundermeyer et al. 2012; Zaremba et al. 2014; Gambäck and Sikdar 2017). Text-based sentiment analysis solves the identification problems of bullying, roasting, aggression, trolling, and suicidal tendencies. In contrast, feedback analysis provides better customer service, such as tracking the performance of employees of any organization. One of the important issues is that we cannot correctly judge the actual feelings of the person, as, during the analysis of the text, the prediction of emotions is absent. Text-based sentiment analysis is based on the type of content most often characterized by a lack of labeled data, an inability to handle complex sentences, and a misunderstanding of context in specific conversations, making this task particularly demanding. Multimodal sentiment analysis, as opposed to the traditional single modality, considers diverse manifestation patterns. Therefore, the sentiment analysis method must be effective in bridging the gap between different modalities. The semantic information covered by the text description and the visual content may differ. It is necessary to extract comprehensive and discriminative data from each modality most related to sentiment classification. Finally, the absence of one modality from the multimodal data is a common phenomenon. Dealing with incomplete multimodal data for sentiment analysis is still a challenging issue.

1.2 Image/video-based sentiment analysis

Analyzing information from images is vital in understanding human behavior by capturing their activities. The image-based SA better handles human attitudes and emotions. Nowadays, people use social media to share various images with friends, relatives, and near or dear ones. The description or caption for most shared images is unavailable (You et al. 2015). Various video-sharing network applications, websites, and other multimedia platforms help researchers work in multimodal sentiment analysis (Li and Xu 2019). The information gathered from natural language processing is not enough since humans communicate and express their emotions and sentiments through different channels. The simultaneous and cognitive analysis of text, audio, and visual modalities enables the effective extraction of semantic and affective information. In this direction, visual information is essential as it contains significant sentiment characteristics in the speaker's gestures and facial expressions. Accepting the various challenging issues, we have proposed a sentiment-based analysis system to identify the types of sentiments among the non-aggressive (NAG, no-pain), covertly aggressive (CAG, low-pain), and overtly aggressive (OAG, high-pain) classes in human behaviors. Covert-aggressive behaviors (due to pain in the human body) are standard and relatively low compared to OAG behavior, which has a very high intensity of aggressiveness. Finally, the NAG class has no aggressive behavior. Therefore, the contributions of this work are as follows:

- A sentiment analysis system using both text and image data of a person such that the proposed framework will predict the level of the sentiment of a person, such as pain among NAG, CAG, and OAG human behaviors.
- A text-based sentiment analysis using both conventional and deep learning approaches has been proposed [(long short term memory (LSTM) architecture].
- An image-based sentiment analysis adopts a deep learning approach, i.e., the convolutional neural network (CNN) architecture.
- Fusion of scores due to text and image-based sentiment analysis systems have been adopted to derive the prediction for the proposed multimodal sentiment system.

This paper's organization is as follows: Sect. 2 describes the related works for the proposed system. Methodology for implementing the proposed sentiment analysis system using text and image data has been discussed in Sect. 3. The experimental results and discussion have been performed in Sect. 4. Finally, Sect. 5 concludes this paper.

2 Related work

The online platform is not considered just a matter of nuisance but has been marked as a significant criminal activity that can be dangerous for many people (Kumar et al. 2018b). So, it is essential to take some preventive action to provide a safeguard to the people of the web. Thus, analyzing various texts, images, or videos using natural language processing, image processing, pattern recognition, computer vision techniques, and algorithms will be highly effective in detecting aggression-related issues. Several deep learning-based sentiment analysis methods have recently drawn more attention to text detection techniques and word embedding algorithms (Chen and Zhang 2018).

2.1 Text-based sentiment analysis

From a current observation, there are some works of detecting hate speech vs. vulgarity in Kumar et al. (2018b), which leads to the scope to differentiate speech vs. vulgarity into covert and overt aggression. Dinakar et al. (2011) had completed the work relating to cyber-bullying. In contrast, Dadvar et al. (2013), Dadvar et al. (2014) and Van Hee et al. (2015) have implemented the cyber-bullying detection system with some improved performance. Trolling is another kind of human behavior on which we have found some initial research activities by Cambria et al. (2010) and, after that, Kumar et al. (2014). Mihaylov et al. (2015) and Mojica (2016) have also published their work on trolling identification with improved performance. Human behavior like racism is another important factor for sentiment analysis, and for this, Greevy and Smeaton (2004) had derived the solutions for analyzing racism as sentiment analysis. The next human behavior is hate speech identification, extensively analyzed by Burnap and Williams (2015), Djuric et al. (2015), Gitari et al. (2015) and Badjatiya et al. (2017). Davidson et al. (2017) used a multiclass classification problem, while Del Vigna12 et al. (2017) employed neural networks classifier. An abusive language detection system has been proposed by Chen et al. (2012) using Lexical Syntactic Feature. In contrast, Nobata et al. (2016) have adopted machine learning-based algorithms for detecting abusive languages on the web.

2.2 Image/video-based sentiment analysis

In the present scenario, due to the availability of upgraded communication systems such as smartphones, plenty of data is uploaded in video (Cambria et al. 2017). Due to the prosperity of research activity in sentiment analysis, it is not sufficient to analyze the aggression from text-only; thus, it is imperative to analyze other data types to get the most relevant result (You et al. 2015). A survey on deep learning approaches to medical images with a systematic look up into object detection tasks has been demonstrated in Kaur et al. (2021). As concerned with image or video-based sentiment analysis, Arya et al. (2021) had served some multidisciplinary domains contributing to affective computing for emotion recognition. The image sentiment analysis using deep learning approaches had been discussed in Mittal et al. (2018). Emotions using facial expressions (Neth 2007) and body language of any person remain within the visual information. In multimodal sentiment analysis (MSA), facial expression recognition using video-based data plays an important role (Li and Xu 2019). There exist two types of data for facial expressions recognition: (Butler et al. 2009) one is spatial, and the other is Spatio-temporal. In the case of spatial, the image sequences are encoded frame-by-frame; on the other hand, the neighboring frames are considered in Spatio-temporal representation (Sariyanidi et al. 2014). Ekman and Keltner (1970) described a thorough investigation into facial expression recognition. Chen et al. (1998) and De Silva et al. (1997) presented their early work on emotion detection fusing visual and audio modalities. Werner et al. (2019) demonstrated a survey on automatic recognition methods supporting pain.

2.3 Multimodal sentiment analysis

Ullah et al. (2017) had served several different types of difficulties faced during the implementation of multimodal sentiment analysis (MSA) using text, image, audio, and video posted regularly on social media. And also, the survey reports a list of existing and upcoming difficulties and opportunities for MSA research. Another paper on a survey on MSA (Soleymani et al. 2017) where also, problems of MSA in different domains such as spoken reviews, video blogs, images, human–machine, and human-interaction systems have been discussed with their opportunities and challenges. Rao et al. (2021) have employed speech-based LSTM features, k-nearest neighbors (kNNs), Bayesian networks, hidden Markov models (HMMs), and artificial neural networks (ANN) based features from facial expressions for acoustic features (Gaussian mixture model, Mel frequency cepstral coefficients) using RAVDESS (Ryerson audio-visual database of emotional speech and song) audio dataset. Emotion classification from speech and text in videos using a multimodal approach has been performed in Caschera et al. (2022) where an automatic extraction of emotional information from a variety of data provided by different interaction modalities and from different domains has been demonstrated. A survey on multimodal video sentiment analysis using deep learning approaches has been reported in Abdu et al. (2021) where multimodal sentiment analysis systems with the Multimodal Multi-Utterance based architecture have been discussed. Based on these studies, a text-based and an image-based sentiment analysis system have been developed in this work using machine learning and computer vision techniques. Using categorically based learning, the proposed system can detect aggressiveness in the pain of human behavior. The classes are NAG, CAG, and OAG. For the text-based sentiment analysis, only the text data is used. In contrast, for image-based sentiment analysis, static images extracted from the facial region are employed to analyze emotions and thus predict the level of aggression.

3 Proposed method

The proposed approach mainly deals with two different types of data. One is text, and the other is images or sequences of frames from a video to perform the recognition task for the type of sentiment when a person feels pain due to any mis-shaping in their body. The block diagram of the proposed system is shown in Fig. 1.

3.1 Text preprocessing

During sentiment analysis, the employed text datasets have several noises due to the diverse domains. So, to make these databases usable, some text preprocessing techniques are required. In this work, the text-based sentiment analysis system has been performed in three components: text preprocessing, feature extraction, and classification. During text preprocessing, the particular document $\mathcal{D}$ has been considered in the same domain for which the problem is to be solved. Let's assume that there are $\mathcal{N}$ comments in the document $\mathcal{D}$, where N is the number of comments. Here each comment $\mathcal{C}_i$ may contain several stop words such as 'is', 'are', 'i', 'am', 'would', 'will', 'what is, 'more', 'such', 'has', 'have', etc. During processing, these stop words are removed from $\mathcal{C}_i$ such that after preprocessing $\mathcal{C}_i$ is transformed to $\mathcal{C}'_i$ (Fig. 2). This $\mathcal{C}'_i$ now undergoes to feature extraction component in which two different algorithms, Scheme_1 and Scheme_2, have been adopted for feature computation, respectively. Scheme_1 derives f^{TEXT} while Scheme_2 provides g^{TEXT} feature vectors. The classification tasks have been performed on the computed feature vectors. A detailed description of the two schemes is proposed below.

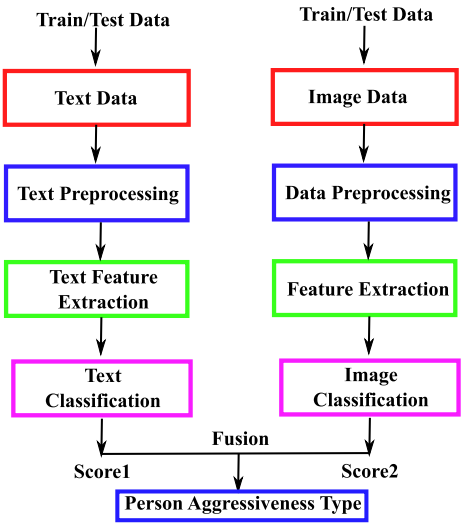

Fig. 1 Block diagram of the proposed system

3.1.1 Scheme_1

The preprocessed comments $\mathcal{C}' = \{\mathcal{C}'_1, \ldots, \mathcal{C}'_N\}$ contains several words which are related to measuring the aggressiveness label of a person. Since a particular word may have several meanings, the participation of each word is important. So, the conversion of words into numeric form not only removes the redundancy between the words but also introduces the distinctive properties between those words (Boulis and Ostendorf 2005). Moreover, the numeric transformation of these words reduces the dimension and helps the classifier to drive better predictions for the proposed system. In Scheme_1, the sentences in each comment $\mathcal{C}'_i$ has been tokenized into the words i.e. sentence $\mathcal{S}_j \in \mathcal{C}'_i = \{w_1, w_2, \cdot, w_K\}$. During feature extraction, we extract features from each $\mathcal{C}'_i$ in the form of feature vector (f^{TEXT}) such that for each word $w_K \in \mathcal{C}'_i$, two values: (a) term frequency (TF), and (b) inverse document frequency (IDF) have been computed. In this work, TF is defined as $a = \frac{n_{w_i}}{|\mathcal{C}'_i|}$ whereas inverse document frequency (IDF) is defined as $b = log(\frac{N}{M_{w_i}})^2$, where n_{w_i} denotes the number of occurrences of word w_i in the particular $\mathcal{C}'_i$, N be the total number of comments i.e. $|\mathcal{C}'| = N$ and M_{w_i} be the number of $\mathcal{C}'_i$s in which w_i appears. The counting of words in each comment and also in the entire comments are handled using the Bag-of-Words (Boulis and Ostendorf 2005) technique where a list of unique words (let there are $\mathcal{K}$ unique words in the dictionary) that has been considered that defines '1' for presenting while '0' for absent of the word in the comment $\mathcal{C}'_i$. The final feature

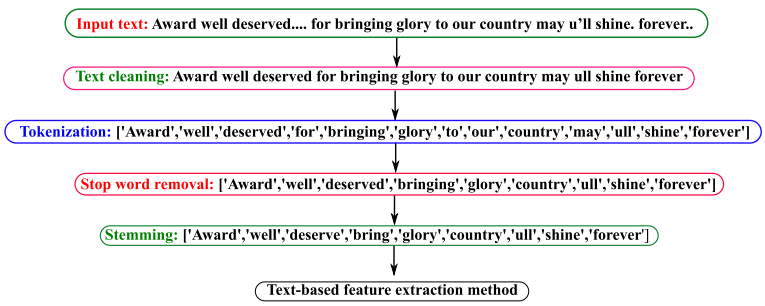

Fig. 2 Text pre-processing steps for the proposed text-based sentiment analysis system

value for each word w_i is given by $c = a * b$. So, the feature vector for each comment $\mathcal{C}'_i$ is given by $f^{TEXT} \in \mathbb{R}^{1\times\mathcal{K}}$ that is sparse in nature with some zeros values. The value of K varies, i.e., each comment $\mathcal{C}'_i$ may not have the same number of words. The block diagram for extracting the features for the text-based sentiment analysis system has been shown in Fig. 3. Now, the features extracted from the text documents undergo the classification task. Several classifiers such as LR, SVM, classification and regression tree (CART), and kNN have been employed. During classification, 50% of samples from each class have been used to form a training set, while the remaining 50% are used for testing purposes. This partitioning of training and testing has been performed ten times, and since then, average performance has been reported for the text-based sentiment analysis system. Here the performances have been reported for the testing set in terms of both F1-Score and correct recognition rate [accuracy (%)].

3.1.2 Scheme$_2$

In the second scheme, we have employed LSTM based feature extraction followed by a classification technique. LSTM (Hochreiter and Schmidhuber 1997) is a specific type of Recurrent Neural Network (RNN) applied for sequence labeling and prediction tasks. The feature extraction using this scheme is as follows: each preprocessed comment $\mathcal{C}'_i$ undergoes space-separated sequences of words which are further split into a list of tokens, and then these tokens are vectorized with some data structure technique. This list of tokens is finally input to the LSTM-based architecture that performs feature learning of tokens from the separate comment and classifies the comment into NAG, CAG, and OAG classes, respectively. The LSTM based architecture has been shown in Fig. 4 while parameters are shown in Table 1. Here also, 50% of samples from each class have been randomly used to form a training set. In contrast, the remaining 50% are used for testing purposes, and average performance has been reported over ten times the partitioning of training–testing samples of the employed dataset. Here, the Scheme$_1$ is based on TF and inverse document frequency methods of NLP-based handcrafted feature representation techniques, where the contributions of word occurrences are used for feature computation at the lexical level. This scheme gives more importance to the word frequency in the comments, so the emotional information retrieval is easier using this scheme if the significance of the words is learned to the system, making search engines faster to identify emotions in the given content. On the other hand, this scheme is based on the bag-of-words (BoW) model. Therefore, this scheme does not capture the word's position, co-occurrences, and semantics in the other comments; hence, this scheme cannot introduce the concept of word embeddings and topic modeling during information extraction. So, the Scheme$_2$ feature representation technique has been incorporated here also. In this scheme, LSTM based technique has been employed to better memorize specific patterns of words in the comments. This

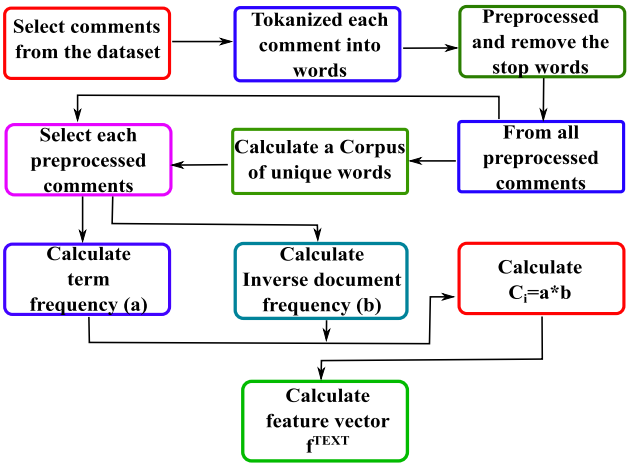

Fig. 3 Block diagram of Scheme$_1$ for feature extraction from text data for text-based sentiment analysis

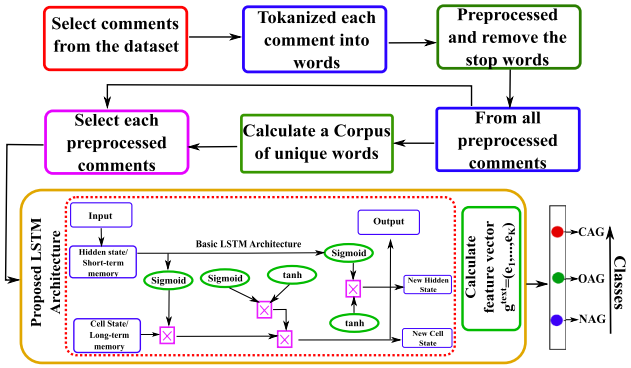

Fig. 4 Block diagram of Scheme$_2$ for feature learning with classification from text data for text-based sentiment analysis

Table 1 LSTM architecture for the proposed system

Layer	Output shape	Parameter
Input	200	0
Embedding	(200,100)	1,000,000
LSTM	100	80,400
Dense	128	12,928
Activation (ReLu)	3	0
Dense	3	387
Activation (sigmoid)	3	0
Total	1,093,715	

technique can also support extracting the semantic information from the given content. Hence, these two schemes are adopted for the proposed text-based sentiment analysis to extract more distinct and discriminant features from texts and incorporate both lexical and semantic-based information from the comments.

3.2 Image preprocessing

The image-based sentiment analysis has two components: (i) image preprocessing for extracting the facial region as a region of interest and (ii) feature learning with classification for predicting the aggressiveness level in human behavior. During an unconstrained imaging environment, noise, illuminations, variations in poses, and background represent irrelevant features (Umer et al. 2019). So, to extract more relevant and valuable characteristics, the face region detection as a region of interest from the input image has been extracted and normalized to obtain the same dimensional feature vector from each extracted face region. During the preprocessing image phase, a tree-structured part model (Zhu and Ramanan 2012) has been employed, which works for all variants of face poses. This technique computes sixty-eight facial landmark points for the frontal face, while thirty-nine landmarks have been extracted for the profile face. The bilinear image interpolation method has been employed for normalization purposes on each extracted face region. The face detection process for the proposed image-based sentiment analysis system has been shown in Fig. 5.

3.2.1 Feature learning with classification

A deep learning-based approach such as CNNs for feature learning with the classification of images into three aggressiveness classes has been employed. With CNN-based models, various research-oriented problems like object detection, texture classification, face recognition, object recognition, scene understanding, and many more applications from the computer vision field can be analyzed and solved (Szegedy et al. 2015; Saxena 2016). The CNN-based approaches extract shape and texture information using machine learning optimizing algorithms. The training of CNN architectures has been performed with a bulky database, and according to the number of classes for the given problem, the weights in the network are adjusted. The CNN architecture has two parts: (i) feature learning and (ii) classification (Zhang et al. 2018).

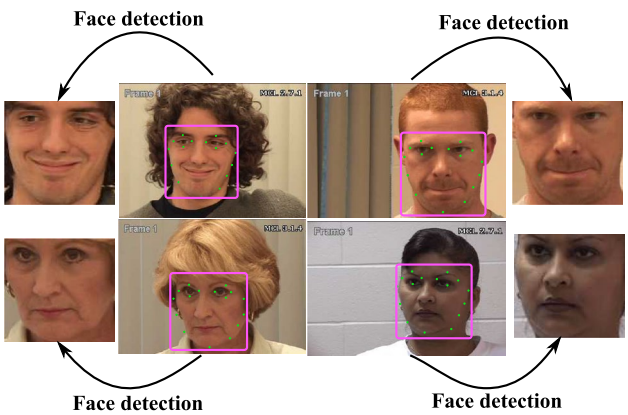

Fig. 5 Image preprocessing task for the proposed image-based sentiment analysis system

In the convolution layer, the input layer always accepts the image. In contrast, the convolution operations are performed with various unique kernels (filter banks) to get a convoluted image (feature map) against each filter. Here, the parameters are considered the adjusted weights in the filter sets. The max-pooling layers (Liu et al. 2013) reduce the computational barriers by decreasing the number of parameters within the network. In the max-pooling layer, the 2×2 filter has been employed on each feature map (retrieved from the preceding layer). Then a step of a 2-down-sample with the minimum or maximum or average values is computed among the 4-numbers towards horizontal and then in the vertical direction. A fully connected layer transforms all the features from the previous layer into a 1-dimensional vector. The dense layer is another fully connected layer that performs linear operations in the dense layer. In the case of linear operations, each input is connected with the output and the probability scores are generated as the outcomes with the help of an activation function such as Softmax. Hence, using the concepts and theories of the CNN layers, in this work, we have proposed a CNN architecture. The architectural design of the proposed CNN has been shown in Fig. 6. It can be seen that the architecture has mainly six blocks (each block is composed of convolution, batch normalization, activation, max-pooling, and dropout layers) followed by two fully connected layers with three dense layers, out of which the probability values are obtained from the last dense layer for the three aggressiveness classes, i.e. NAG, CAG, and OAG. The detailed description of the proposed CNN architecture along with adopted layers, the output shape of feature maps at each layer, and the parameters involved at each layer have been demonstrated in Table 2.

During learning the parameters, the data from the preceding layer is normalized by the Batch Normalization technique (Ding et al. 2018). This normalization technique processes the batch of data by subtracting it from the batch mean and dividing it by the batch standard deviation. Then the batch mean (γ) and standard deviation (β) are two trainable parameters added to the batch normalized data. The used activation function is the Rectified Linear Unit (ReLU) (Ding et al. 2018) function. The Dropout method (Wang et al. 2017) ignores arbitrarily selected neurons during

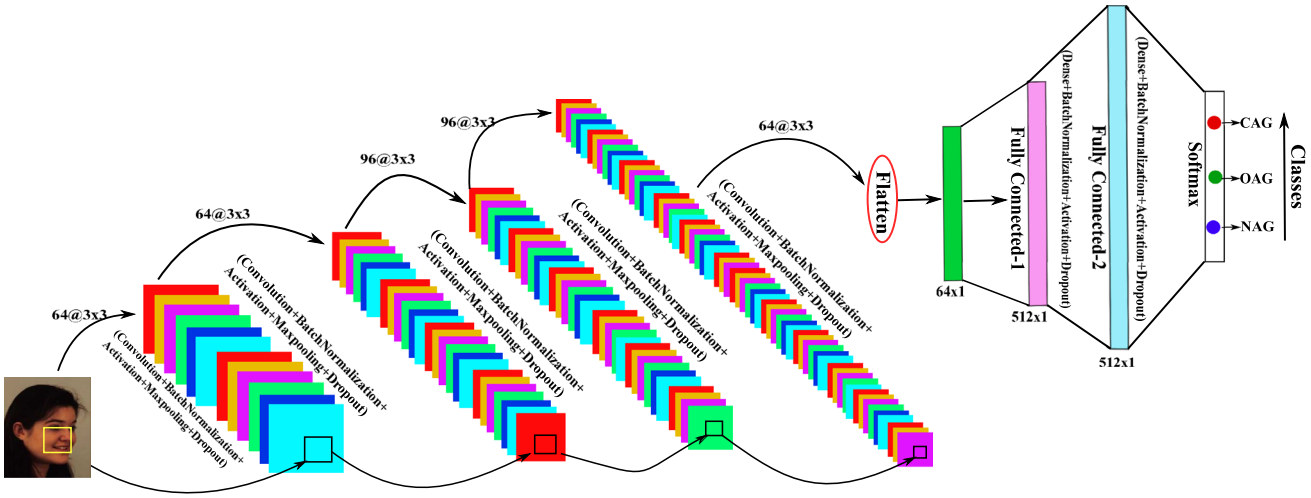

Fig. 6 The employed CNN architecture for the proposed system

training in the forward pass because the weights on those ignored neurons are not updated in the backward pass. The use of the Dropout method reduces the overfitting problem in the network. Also, it provides a way to combine the predictions made by the different neural networks efficiently. The softmax activation function predicts the outcomes at the last dense layer in the form of probability scores concerning the intensity of pain as aggressiveness classes.

4 Experimental results

4.1 Datasets

An Aggression dataset (Kumar et al. 2018b) has been used for text, which is divided into three classes: non-aggressive (NAG, no-pain), covertly aggressive (CAG, Low-pain), and overtly aggressive (OAG, high-pain). As mentioned in Sect. 2, the OAG class of comments basically represents comments where the user's aggression is expressed with great intensity against any particular topic. Both the external and internal statements are highly aggressive for these types of comments. For the CAG class of comments, the intensity of the overall aggressiveness is quite low compared to OAG comments. Moreover, if we notice the external statement of the comments, it may not look aggressive. Still, if we notice the internal statement of the comments, the clear aggressiveness will be identified distinctly. In the case of the NAG class of comments, no aggressiveness can be identified from both the external and internal statements of the comments. NAG is $\mathcal{C}_1$ class, CAG is $\mathcal{C}_2$ class, and OAG is $\mathcal{C}_3$ class. The comments in this dataset are composed of two different languages: 'Hindi', and 'English'. Some examples of these comments in 'English', and 'Hindi' (words are in the English Alphabets) with respect to NAG, CAG, and OAG classes have been shown in Table 3 whereas Table 4 shows the description of this used text database.

Similarly, for the image-based sentiment analysis, we have employed the UNBC-McMaster (Lucey et al. 2011) shoulder pain expression archive database. This dataset is composed of 129 subjects (63 male and 66 female). The participants have shoulder pain, three physiotherapy clinics have identified their problems, and the videos were captured on the campus of McMaster University. The subjects have suffered from arthritis, bursitis, tendinitis, subluxation, rotator cuff injuries, impingement syndromes, bone spurs, capsulitis, and dislocation. The frames have been extracted from each video, and the images are labeled from 'No pain' to 'High-intensity pain' classes. These images are classified as non-aggressive (NAG, no-pain) ($\mathcal{C}_1$), covertly aggressive (CAG, low-pain) ($\mathcal{C}_2$), and overtly aggressive (OAG, high-pain) ($\mathcal{C}_3$), with descriptions of the samples in Table 5. Some images from this database have been shown in Fig. 7.

4.2 Results and discussion

The proposed sentiment analysis system has been implemented in Python on Ubuntu O/S with 32GB RAM and an Intel Core i7 processor of 3.20 GHz. Several Python packages have been employed during implementation with Theano (Bergstra et al. 2010) and Keras (Gulli and Pal 2017) special packages. The sentiment of a person has been analyzed using their text information (written by them) and the image data (that has pain emotion on their

Table 2 Description of employed CNN architecture about the number of layers, output shapes and parameters at each layer where the size of input image is 48×48

Layer	Output shape	Image size	Parameters
Block -1			
Convolution2D (3×3@64) (Activation: Relu)	$(n, n, 64)$	(48,48,64)	$((3 \times 3 \times 3) + 1) \times 64$=1792
Maxpooling2D (2×2)	$(n_1, n_1, 64)$	(24,24,64)	0
Batch normalization	$(n_1, n_1, 32)$	(24,24,64)	$4 \times 64 = 256$
Block-2			
Convolution2D (3×3@64) (Activation: Relu)	$(n_1, n_1, 64)$	(24,24,64)	$((3 \times 3 \times 64)$+1)x64= 36928
Maxpooling2D (2×2)	$(n_2, n_2, 64)$	(12,12,64)	0
Batch normalization	$(n_2, n_2, 64)$	(12,12,64)	$4 \times 64 = 256$
Block-3			
Convolution2D (3×3@96) (Activation: Relu)	$(n_2, n_2, 96)$	(12,12,96)	$((3 \times 3 \times 64) + 1) \times 96 = 55{,}392$
Maxpooling2D (2×2)	$(n_3, n_3, 96)$	(6,6,96)	0
Batch normalization	$(n_3, n_3, 96)$	(6,6,96)	$4 \times 96 = 384$
Block-4			
Convolution2D (3×3@96) (Activation: Relu)	$(n_3, n_3, 96)$	(6,6,96)	$((3 \times 3 \times 96)$+1)x96=83040
Maxpooling2D (2×2)	$(n_4, n_4, 96)$	(3,3,96)	0
Batch normalization	$(n_4, n_4, 96)$	(6,6,96)	$4 \times 96 = 384$
Block-5			
Convolution2D (3×3@64) (Activation: Relu)	$(n_4, n_4, 64)$	(3,3,64)	$((3 \times 3 \times 96) + 1) \times 64 = 55{,}360$
Maxpooling2D (2×2)	$(n_5, n_5, 64)$	(1,1,64)	0
Batch normalization	$(n_5, n_5, 64)$	(1,1,64)	$4 \times 64 = 256$
Fully connected			
Flatten	$1 \times 1 \times 64 = 64$		0
Dense + dropout	512		$(64 + 1) \times 512 =$ 33,280
Dense + dropout	512		$(512 + 1) \times 512 = 262{,}656$
Dense + softmax	3		$(512 + 1) \times 3 = 1539$
Total parameters			475,875

face). Here, text and image-based sentiment analysis have been performed individually, and the results are reported accordingly. Finally, to improve the performance of the proposed system, the results from text and image-based sentiment analysis systems are fused at the post-classification level such that the unconstrained environments can be handled with improved performance. In the below sections, text-based and image-based sentiment analyses have been discussed accordingly.

4.2.1 Results on text

During text-based sentiment analysis, each dataset's comment is considered and classified into three classes. Here at first each comment $\mathcal{C}_i$ has been preprocessed to $\mathcal{C}'_i$ using the technique discussed in Sect. 3.1. Now using Scheme_1 for each comment $\mathcal{C}'_i$, $f^{TEXT} \in \mathbb{R}^{1\times(\mathcal{W}=1000)}$ dimensional feature vector has been obtained and hence for all comments $F^{TEXT} \in \mathbb{R}^{12000\times(\mathcal{W}=1000)}$ feature matrix has been obtained. This feature matrix has been randomly partitioned with 50% of its data as a training set while 50% as a testing set.

The training set undergoes classification tasks using LR, kNN, CART, and SVM classifiers, respectively. Each classifier results in a model used to obtain the performance of the proposed text-based sentiment analysis system using the testing set. The performance of proposed text-based sentiment analysis using Scheme_1 and Scheme_2 methods has been demonstrated in Table 6. In this table, we show the performance for both 2-class problems (where samples from $\mathcal{C}_2$ and $\mathcal{C}_3$ are considered to be from the same class, i.e., the aggressive class, and samples from $\mathcal{C}_1$ are considered to be a NAG class).

From Table 6 it has been observed that the proposed system has attained better performance using SVM and LSTM classifiers for both *2-class* and *3-class* problems. It has also been observed that the proposed system has achieved better performance for *2-class* problem than *3-class* problem, and it is because the samples are better distributed in *2-class* than *3-class* problem. The performance of the proposed text-based sentiment analysis due to both Scheme_1 and Scheme_2 methods has been compared with the existing state-of-the-art methods in Table 7 where we have noted the performance from Samghabadi et al. (2018), Kumar et al. (2018a), Modha et al. (2018), and Constantin Orasan (Orăsan 2018) methods respectively in terms of both F1-Score and Accuracy (%). This comparison shows that the proposed text-based sentiment analysis has obtained better performance using both the Scheme_1 and Scheme_2 methods. For performance improvement in the proposed text-based sentiment analysis system, the scores from the SVM classifier using Scheme_1 features and the scores from the LSTM classifier using Scheme_2 features are fused together using sum, product and weighted-sum rule-based score level fusion techniques. Let s_1 and s_2 be the scores after post-classification of SVM and LSTM classifiers; (i) sum rule-based technique is defined as $s = s_1 + s_2$, (ii) product rule-based technique is defined $s = s_1 \times s_2$, and (iii) weighted-sum rule-based technique is defined as $s = w_1 \times s_1 + w_2 \times s_2$, where s is the fused score while w_1 and w_2 be the corresponding weights such that $w = w_1 + w_2$. The fused performance of text-based sentiment analysis due to Scheme_1 and Scheme_2 methods has been shown in Table 8.

4.2.2 Results on image

In the image-based sentiment analysis, the implementation of the proposed system has been divided into (i) image preprocessing and (ii) feature learning with classification. For this system, during face preprocessing, the face region $\mathcal{F}$ has been extracted using the TSPM model. Then the extracted face region $\mathcal{F}$ is normalized to $\mathcal{N} \times \mathcal{N}$ fixed image size. Further, the extracted facial region from the training samples undergoes the proposed CNN architecture (Fig. 6). Here the size of the face region $\mathcal{N} \times \mathcal{N}$ is 48×48 while the batch size and the number of epochs vary. During experimentation, it has been observed that the performance of the proposed system improves due to the batch sizes $\{30, 40, 50\}$ with epochs such as $\{50, 100, 200\}$. Figure 8 demonstrates the effectiveness of batch sizes with the number of epochs over the performance of the proposed image-based sentiment analysis system for the UNBC-McMaster shoulder pain database.

From Fig. 8, it has been observed that the performance improves with increasing the epochs, and for batch size 30, the performance is better. For further experiments, we have employed batch size 30 on training samples with 200 epochs for learning the parameters of the proposed CNN architecture. So, the top-2 performance (in terms of accuracy and f1-score) of the proposed image-based sentiment analysis system has been shown in Table 9. Here also, the performance of the proposed system has been shown for both *2-class* and *3-class* problems. For *2-class* problem the image samples of $\mathcal{C}_2$ and $\mathcal{C}_3$ are considered to be from 'Pain' class while the image samples of $\mathcal{C}_1$ belongs to 'Non-Pain' class. Hence, from the performance, it has been observed that the proposed system has achieved better performance for both *2-class* and *3-class* problems. We have compared the performance of this proposed system with some existing state-of-the-art methods for the UNBC-McMaster shoulder pain database in Table 10. The performance reported for the competing methods adopted the same training–testing protocols. These results show that the proposed system performs better than the other competing methods for the UNBC-McMaster shoulder pain database. Tables 7 and 10 shows that the proposed system has achieved outstanding performance for both the employed text and image dataset. So, the performances of text-based and image-based sentiment analysis systems are fused for the multimodal sentiment analysis system. Both datasets have different samples for NAG, CAG, and OAG classes. Consequently, an equal number of samples have been considered for the multimodal sentiment analysis system, selecting 5000 samples for NAG, 4000 samples for CAG, and 2700 samples for OAG classes. Here the datasets have been partitioned with 50% data as training while 50% as testing. The performance of the proposed multimodal sentiment analysis system has been shown in Table 11. According to the reported performance, the data in the multimodal system is challenging. Due to a lack of data in training set for both datasets, the individual performance is somewhat similar. The performance is much better after fusion. It is also shown that the multimodal system has obtained better performance for the product-rule-based fusion technique than other fusion techniques.

Table 3 Some examples of texts for the proposed text-based sentiment analysis system

Comments in English	
NAG	Award well deserved for bringing glory to our country may u'll shine forever
CAG	This is absolutely right everyone has the right to choose what they want to eat or not
OAG	It is right. These students are doing all nuisance leaving their studies. Stupids. Beat them well and put them in lockup for a week
Commentrs in Hindi	
NAG	Hum kyu le bhai Pehle hi Afghani,Balochi, Bangladeshi,rohingya rehte hai
CAG	Bhaiya sabse pahle ye hai kon... mahangaayi badhi to banwaas pr gaye hue the, Note ban pr ese ghayab hue jaise gadhe k sir se seeng, Patrol and eatables price hike pr to ye mr india ban gaye the k kahin dikhe hi nahi. Or ab aagaye neend se jaag kar k chalo thoda publicity b lelun
OAG	sale sab neta o ko boder pe bhejdo use ke bad dekho jaan ki kimat kya hoti hai hai aur hamare javano

Table 4 Description of text database for the proposed system

Class	Sample
Non-aggressive (NAG) (No-pain) $\mathcal{C}_1$	5052
covertly aggressive (CAG) (Low-pain) $\mathcal{C}_2$	4240
Overtly aggressive (OAG) (High-pain) $\mathcal{C}_3$	2708

Table 5 Description of UNBC Image database for the proposed system

Class	Sample
Non-aggressive (NAG) (No-pain) $\mathcal{C}_1$	40,029
Covertly aggressive (CAG) (Low-pain) $\mathcal{C}_2$	2909
Overtly aggressive (OAG) (High-pain) $\mathcal{C}_3$	5460

5 Conclusions

This paper presents a multimodal sentiment analysis system for recognizing people's aggression in pain using textual and visual information from a person. The implementation of the proposed system has five components: text preprocessing, feature extraction, and classification. These are the components of text-based sentiment analysis, processing the person's textual information to predict the label of

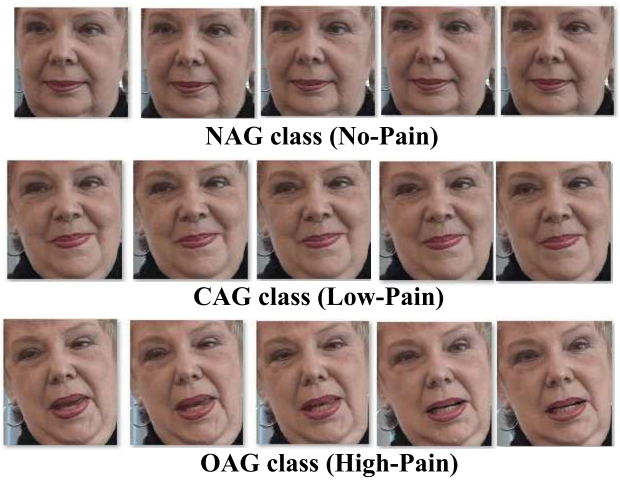

Fig. 7 Some samples of the employed UNBC database

Table 6 Performance of the proposed text-based sentiment analysis system with training and testing times in Sec. for both the schemes

2-Class problem		
Scheme$_1$ (*Sect.* 3.1.1)		
Classifier	F1-Score	Accuracy (%)
LR	0.6543	64.11
KNN	0.6194	60.66
CART	0.6371	62.85
SVM	0.6845	66.25
Training-time	5.00 sec.	
Testing-time	0.02 sec.	
Scheme$_2$ (*Sect.* 3.1.2)		
LSTM	0.6705	65.32
Training-time	7.03 sec.	
Testing-time	0.02 sec	
3-Class problem		
Scheme$_1$ (*Sect.* 3.1.1)		
Classifier	F1-Score	Accuracy (%)
LR	0.3535	32.78
KNN	0.5717	52.95
CART	0.5672	52.51
SVM	0.5967	58.56
Training-time	5.89 sec.	
Testing-time	0.02 sec.	
Scheme$_2$ (*Sect.* 3.1.2)		
LSTM	0.5803	57.91
Training-time	7.71 sec.	
Testing-time	0.03 sec.	

Table 7 Performance comparison of the proposed text-based Sentiment analysis system with the other competing methods for the 3-class problem

Method	F1-Score	Accuracy (%)
Samghabadi et al. (2018)	0.5875	56.17
Kumar et al. (2018a)	0.3572	33.72
Modha et al. (2018)	0.5580	53.76
Kumar et al. (2014)	0.5229	54.78
Cambria et al. (2017)	0.53.71	56.89
Constantin Orasan (Orăsan 2018)	0.5830	55.79
Proposed Scheme$_1$	**0.5967**	**58.56**
Proposed Scheme$_2$	0.5872	57.91

Bold values indicate our results

Table 8 Fused Performance of the proposed text-based Sentiment analysis system for the 3-class problem

Method	Accuracy (%)
Scheme$_1$	58.56
Scheme$_2$	57.91
Sum rule	59.72
Product rule	**61.29**
Weighted-sum rule	60.33

Bold values indicate our results

Table 9 Performance of the proposed image-based sentiment analysis system

2-Class problem			
Top-1		Top-2	
Accuracy(%)	F1-Score	Accuracy(%)	F1-Score
84.59	0.8391	87.46	0.8679
Training-time		39.00 sec.	
Testing-time		0.03 sec.	
3-Class problem			
Top-1		Top-2	
Accuracy(%)	F1-Score	Accuracy(%)	F1-Score
82.35	0.8193	84.78	0.8367
Training-time		41.29 sec.	
Testing-time		0.03 sec.	

aggressiveness among the NAG, Covertly Aggressive, and OAG classes when they have pain or no pain.

Similarly, image preprocessing and feature learning with classification are the components of an image-based sentiment analysis system, where the visual intensity of pain emotion on the facial region of a person has been employed to predict the classes of NAG, CAG, and OAG. Both these

Table 10 Performance comparison of the proposed image-based Sentiment analysis system with the other competing methods using UNBC-McMaster shoulder pain database for the 3-class problem

Method	Accuracy (%)
Vgg16 (Simonyan and Zisserman 2014)	76.84
ResNet50 (Szegedy et al. 2016)	79.32
Inception-v3 (McNeely-White et al. 2020)	79.64
Werner et al. (2016)	75.50
Li and Xu (2019)	76.89
Cambria et al. (2017)	79.71
Lucey et al. (2011)	81.80
Proposed	82.35

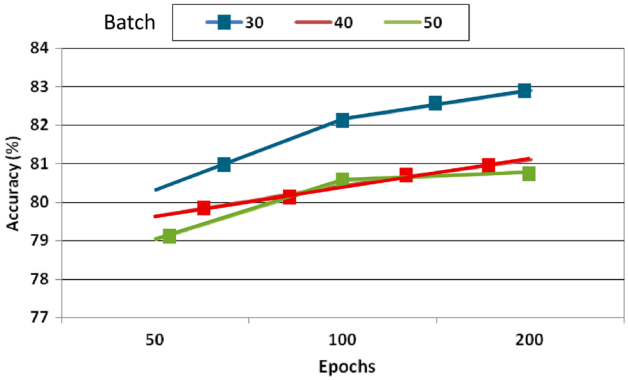

Fig. 8 Effectiveness of batch vs epochs for the proposed image-based sentiment analysis system

Table 11 Performance of the proposed multimodal Sentiment analysis system

Method	Acc. (%)	F1-Score
Image-based	52.54	0.5145
Text-based (*Scheme*$_1$)	51.92	0.5039
Sum-rule	53.56	0.5301
Product-rule	**55.85**	**0.5384**
Weighted sum rule	54.34	0.5319
Method	Acc. (%)	F1-Score
Image-based	52.54	0.5162
Text-based (*Scheme*$_2$)	50.62	0.4897
Sum rule	53.41	0.5237
Product rule	**55.35**	**0.5441**
Weighted sum rule	55.08	0.5428

Bold values indicate our results

systems have been implemented individually and experimented with using the respective databases. The performance has been compared with the state-of-the-art methods,

showing the superiority of the proposed system. Finally, the scores due to both these systems have been fused to derive the performance of the proposed multimodal sentiment analysis system.

Funding Open access funding provided by Università degli Studi di Salerno within the CRUI-CARE Agreement.

Declarations

Conflict of interest The authors declare no conflict of interest. The funding agency had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

Abdu Sarah A, Yousef Ahmed H, Ashraf S (2021) Multimodal video sentiment analysis using deep learning approaches, a survey. Inf Fus 76:204–226

Arya R, Singh J, Kumar A (2021) A survey of multidisciplinary domains contributing to affective computing. Comput Sci Rev 40:100399

Badjatiya P, Gupta S, Gupta M, Varma V (2017) Deep learning for hate speech detection in tweets. In: Proceedings of the 26th international conference on world wide web companion. International World Wide Web Conferences Steering Committee, pp 759–760

Bergstra J, Breuleux O, Bastien F, Lamblin P, Pascanu R, Desjardins G, Turian J, Warde-Farley D, Bengio Y (2010) Theano: a cpu and gpu math expression compiler. In Proceedings of the Python for scientific computing conference (SciPy), vol 4. Austin, TX, pp 1–7

Boulis C, Ostendorf M (2005) Text classification by augmenting the bag-of-words representation with redundancy-compensated bigrams. In: Proceedings of the international workshop in feature selection in data mining. Citeseer, pp 9–16

Burnap P, Williams ML (2015) Cyber hate speech on twitter: an application of machine classification and statistical modeling for policy and decision making. Policy Internet 7(2):223–242

Butler S, Tanaka J, Kaiser M, Le Grand R (2009) Mixed emotions: Holistic and analytic perception of facial expressions. J Vis 9(8):496

Cambria E, Chandra P, Sharma A, Hussain A (2010) Do not feel the trolls. ISWC, Shanghai

Cambria E, Hazarika D, Poria S, Hussain A, Subramanyam RBV (2017) Benchmarking multimodal sentiment analysis. In: International conference on computational linguistics and intelligent text processing. Springer, pp 166–179

Caschera MC, Grifoni P, Ferri F (2022) Emotion classification from speech and text in videos using a multimodal approach. Multimodal Technol Interact 6(4):28

Chen Y, Zhang Z (2018) Research on text sentiment analysis based on cnns and svm. In: 2018 13th IEEE conference on industrial electronics and applications (ICIEA). IEEE, pp 2731–2734

Chen Lawrence S, Huang Thomas S, Miyasato T, Nakatsu R (1998) Multimodal human emotion/expression recognition. In: Proceedings third IEEE international conference on automatic face and gesture recognition. IEEE, pp 366–371

Chen Y, Zhou Y, Zhu S, Xu H (2012) Detecting offensive language in social media to protect adolescent online safety. In: 2012 international conference on privacy, security, risk and trust and 2012 international conference on social computing. IEEE, pp 71–80

Dadvar M, Trieschnigg D, Ordelman R, de Jong F (2013) Improving cyberbullying detection with user context. In: European conference on information retrieval. Springer, pp 693–696

Dadvar M, Trieschnigg D, de Jong F (2014) Experts and machines against bullies: a hybrid approach to detect cyberbullies. In: Canadian conference on artificial intelligence. Springer, pp 275–281

Davidson T, Warmsley D, Macy M, Weber I (2017) Automated hate speech detection and the problem of offensive language. In: Eleventh international AAAI conference on web and social media

De Silva Liyanage C, Miyasato T, Nakatsu R (1997) Facial emotion recognition using multi-modal information. In: Proceedings of ICICS, 1997 international conference on information, communications and signal processing. Theme: trends in information systems engineering and wireless multimedia communications (Cat.), vol 1. IEEE, pp 397–401

Del Vigna F, Cimino A, Dell'Orletta F, Petrocchi M, Tesconi M (2017) Hate me, hate me not: hate speech detection on facebook. In: Proceedings of the first Italian conference on cybersecurity (ITASEC17), pp 86–95

Dinakar K, Reichart R, Lieberman H (2011) Modeling the detection of textual cyberbullying. In: Fifth international AAAI conference on weblogs and social media

Ding Z, Zhu M, Tam VWY, Yi G, Tran CNN (2018) A system dynamics-based environmental benefit assessment model of construction waste reduction management at the design and construction stages. J Clean Prod 176:676–692

Djuric N, Zhou J, Morris R, Grbovic M, Radosavljevic V, Bhamidipati N (2015) Hate speech detection with comment embeddings. In: Proceedings of the 24th international conference on world wide web. ACM, pp 29–30

Ebrahimi M, Yazdavar AH, Sheth A (2017) Challenges of sentiment analysis for dynamic events. IEEE Intell Syst 32(5):70–75

Ekman P, Keltner D (1970) Universal facial expressions of emotion. Calif Ment Health Res Dig 8(4):151–158

Gambäck B, Sikdar Utpal K (2017) Using convolutional neural networks to classify hate-speech. In: Proceedings of the first workshop on abusive language online, pp 85–90

Ghosh A, Umer S, Khan Muhammad K, Rout Ranjeet K, Dhara Bibhas C (2022) Smart sentiment analysis system for pain detection using cutting edge techniques in a smart healthcare framework. Cluster Comput 1–17

Gitari ND, Zuping Z, Damien H, Long J (2015) A lexicon-based approach for hate speech detection. Int J Multim Ubiquitous Eng 10(4):215–230

Greevy E, Smeaton AF (2004) Classifying racist texts using a support vector machine. In: Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, pp 468–469

Gulli A, Pal S (2017) Deep learning with Keras. Packt Publishing Ltd, Birminghan

Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural Comput 9(8):1735–1780
Hosmer DW Jr, Lemeshow S, Sturdivant RX (2013) Applied logistic regression, vol 398. Wiley, New York
Kaur A, Singh Y, Neeru N, Kaur L, Singh A (2021) A survey on deep learning approaches to medical images and a systematic look up into real-time object detection. Arch Comput Methods Eng 1–41
Kumar S, Spezzano F, Subrahmanian VS (2014) Accurately detecting trolls in slashdot zoo via decluttering. In: Proceedings of the 2014 IEEE/ACM international conference on advances in social networks analysis and mining. IEEE Press, pp 188–195
Kumar R, Bhanodai G, Pamula R, Chennuru MR (2018a) Trac-1 shared task on aggression identification: Iit (ism)@ coling'18. In: Proceedings of the first workshop on trolling, aggression and cyberbullying (TRAC-2018), pp 58–65
Kumar R, Ojha AK, Malmasi S, Zampieri M (2018b) Benchmarking aggression identification in social media. In: Proceedings of the first workshop on trolling, aggression and cyberbullying (TRAC-2018), pp 1–11
Li H, Xu H (2019) Video-based sentiment analysis with hvnlbp-top feature and bi-lstm. In: Proceedings of the AAAI conference on artificial intelligence, vol 33, pp 9963–9964
Liu M, Li S, Shan S, Chen X (2013) Au-aware deep networks for facial expression recognition. In: 2013 10th IEEE international conference and workshops on automatic face and gesture recognition (FG). IEEE, pp 1–6
Lucey P, Cohn JF, Prkachin KM, Solomon PE, Matthews I (2011) Painful data: the unbc-mcmaster shoulder pain expression archive database. In: Face and gesture 2011. IEEE, pp 57–64
Malmasi S, Zampieri M (2017) Detecting hate speech in social media. arXiv:1712.06427
McNeely-White D, Beveridge JR, Draper BA (2020) Inception and resnet features are (almost) equivalent. Cogn Syst Res 59:312–318
Mihaylov T, Georgiev G, Nakov P (2015) Finding opinion manipulation trolls in news community forums. In: Proceedings of the nineteenth conference on computational natural language learning, pp 310–314
Mittal N, Sharma D, Joshi ML (2018) Image sentiment analysis using deep learning. In: 2018 IEEE/WIC/ACM international conference on web intelligence (WI). IEEE, pp 684–687
Modha S, Majumder P, Mandl T (2018) Filtering aggression from the multilingual social media feed. In: Proceedings of the first workshop on trolling, aggression and cyberbullying (TRAC-2018), pp 199–207
Mojica LG (2016) Modeling trolling in social media conversations. arXiv:1612.05310
Mursalin Md, Zhang Y, Chen Y, Chawla NV (2017) Automated epileptic seizure detection using improved correlation-based feature selection with random forest classifier. Neurocomputing 241:204–214
Neth DC (2007) Facial configuration and the perception of facial expression. Ph.D. thesis, The Ohio State University
Nobata C, Tetreault J, Thomas A, Mehdad Y, Chang Y (2016) Abusive language detection in online user content. In: Proceedings of the 25th international conference on world wide web. International World Wide Web Conferences Steering Committee, pp 145–153
Noble WS (2006) What is a support vector machine? Nat Biotechnol 24(12):1565–1567
Orăsan C (2018) Aggressive language identification using word embeddings and sentiment features. In: Proceedings of the first workshop on trolling, aggression and cyberbullying (TRAC-2018), pp 113–119
Rao A, Ahuja A, Kansara S, Patel V (2021) Sentiment analysis on user-generated video, audio and text. In: 2021 international conference on computing, communication, and intelligent systems (ICCCIS). IEEE, pp 24–28
Samghabadi NS, Mave D, Kar S, Solorio Thamar (2018) Ritual-uh at trac 2018 shared task: aggression identification. arXiv:1807.11712
Sariyanidi E, Gunes H, Cavallaro A (2014) Automatic analysis of facial affect: A survey of registration, representation, and recognition. IEEE Trans Pattern Anal Mach Intell 37(6):1113–1133
Saxena A (2016) Convolutional neural networks: an illustration in tensorflow. XRDS: crossroads. ACM Mag Stud 22(4):56–58
Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556
Soleymani M, Garcia D, Jou B, Schuller B, Chang S-F, Pantic M (2017) A survey of multimodal sentiment analysis. Image Vis Comput 65:3–14
Sundermeyer M, Schlüter R, Ney H (2012) Lstm neural networks for language modeling. In: Thirteenth annual conference of the international speech communication association
Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A (2015) Going deeper with convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1–9
Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z (2016) Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2818–2826
Ullah MA, Islam MM, Azman NB, Zaki ZM (2017) An overview of multimodal sentiment analysis research: Opportunities and difficulties. In: 2017 IEEE international conference on imaging, vision & pattern recognition (icIVPR). IEEE, pp 1–6
Umer S, Dhara BC, Chanda B (2019) Face recognition using fusion of feature learning techniques. Measurement 146:43–54
Van Hee C, Lefever E, Verhoeven B, Mennes J, Desmet B, De Pauw G, Daelemans W, Hoste V (2015) Detection and fine-grained classification of cyberbullying events. In: Proceedings of the international conference recent advances in natural language processing, pp 672–680
Wang P-S, Liu Y, Guo Y-X, Sun C-Y, Tong X (2017) O-cnn: Octree-based convolutional neural networks for 3d shape analysis. ACM Trans Graphics (TOG) 36(4):1–11
Werner P, Al-Hamadi A, Limbrecht-Ecklundt K, Walter S, Gruss S, Traue HC (2016) Automatic pain assessment with facial activity descriptors. IEEE Trans Affect Comput 8(3):286–299
Werner P, Lopez-Martinez D, Walter S, Al-Hamadi A, Gruss S, Picard R (2019) Automatic recognition methods supporting pain assessment: a survey. IEEE Trans Affect Comput
You Q, Luo J, Jin H, Yang J (2015) Robust image sentiment analysis using progressively trained and domain transferred deep networks. arXiv:1509.06041
Zaremba W, Sutskever I, Vinyals O (2014) Recurrent neural network regularization. arXiv:1409.2329
Zhang Y, Tian Y, Kong Y, Zhong B, Fu Y (2018) Residual dense network for image super-resolution. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2472–2481
Zhu X, Ramanan D (2012) Face detection, pose estimation, and landmark localization in the wild. In: 2012 IEEE conference on computer vision and pattern recognition. IEEE, pp 2879–2886

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.