تطوير نظام اوتوماتيك قائم على تقنيات التعرف على الكائنات للتنبؤ بالمشاعر المرئية من صور الشبكات الاجتماعية

Develop an Automatic System based on Object Recognition Techniques for Predicting Visual Sentiments from Social Network Images

Hend A. ElLaban
Computer Teacher Preparation Department
Faculty of Specific Education, Damietta University

Prof. Dr. El Saeed E. AbdElrazek
Computer Teacher Preparation Department
Faculty of Specific Education, Damietta University

Prof. Dr. A.A. ElHarby
Computer Science Department,
Faculty of Computers and Artificial Intelligence, Damietta University

Dr. Doaa M.Hawa
Computer Teacher Preparation Department
Faculty of Specific Education, Damietta University

المجلة العلمية لكلية التربية النوعية – جامعة دمياط
عدد (٨) – ديسمبر ٢٠٢٣

Develop an Automatic System based on Object Recognition Techniques for Predicting Visual Sentiments from Social Network Images

Hend A. ElLaban[1], Elsaeed E. AbdElrazek[2], A.A. El-Harby[3] and Doaa M. Hawa[4]

[1,2,4] Computer Teacher Preparation, Faculty of Specific Education, Damietta University

[3] Computer Science Department, Faculty of Computers and Artificial Intelligence, Damietta University

Abstract

Social networks have become a vital part of everybody's life, as users on popular social networking platforms share millions of images to express their opinions and personal emotions. Therefore, images have emerged as one of the most effective methods for transmitting sentiments on social networks. This has resulted in a solid vision to analyze social network images to predict positive and negative sentiments from these images. In this paper, an automatic system based on object recognition is developed by combining InceptionV3 and Long Short-Term Memory networks for predicting visual sentiments. This system aims to recognize the salient objects from social network images and predict their sentiments. Firstly, the InceptionV3 pre-trained CNN network is fine-tuned to recognize objects from images. After that, the object features are extracted using the trained network. Finally, a Long Short-Term Memory network is used to learn sentiments from object features to predict visual sentiment. The experiment results showed that the proposed system is a more powerful system for predicting visual sentiments by combining Inception V3 and Long Short-Term Memory networks. The proposed system achieved 98.2% for predicting visual sentiments.

Keywords: Visual sentiments, InceptionV3, LSTM, Social Network images, Object recognition.

1. Introduction

Nowadays, social network platforms like Instagram, Flickr, Twitter and Facebook play a crucial role in allowing people to share information about significant events, news and topics related to various domains like business, entertainment, crisis management, politics and education. These platforms offer various tools for expression that enable people to easily share their opinions through various forms of user-generated content like text, images, audio and videos [1]. Images are a more convenient way for users to express their opinions and emotions. The emotions that users feel when looking at an image are commonly referred to as the image's visual sentiment [2].

Visual sentiment analysis is the study of human emotional responses to visual stimuli such as images and videos [3]. It is one of the computer vision tasks. It aims to analyze images and enable computers to understand what is happening in those images to detect and express the positive and negative sentiments and opinions conveyed [4]. It entails being able to recognize the objects and scenes in images, as well as their emotional context from images. Furthermore, Visual sentiment prediction is a branch of image understanding that involves a higher level of abstraction regarding the affects that an image will convey [5]. Automated sentiment prediction is difficult since there is an affective gap between low-level visual features and high-level sentiments [6].

The development of computer vision techniques has improved the process of understanding visual sentiment by moving from low-level to high-level features. Several techniques have been used in previous studies to predict visual sentiments from images. These techniques include low-level visual feature-based, semantic-level feature-based and deep-learning architecture-based techniques [7]. Convolutional Neural Networks (CNNs) are used in deep learning architecture-based techniques to automatically extract high-level sentiment features. Deep features outperformed hand-tuned features for sentiment prediction when using a CNN network for visual sentiment analysis [8]. Furthermore, psychology studies proved that the images' sentiments maybe come from a variety of salient objects in the images [9]. For this reason, recognizing the objects in the image can help to predict important features that influence the sentiments. Therefore, this thesis aims to develop an automatic system based on object recognition techniques for predicting visual sentiments from social network images. The proposed system is designed by combining InceptionV3 and LSTM networks to recognize objects and predict their visual sentiments.

2. Motivation and Problem Definition

Sharing visual media content, such as images and videos, has become easier than ever thanks to smartphone cameras and social networks. Research shows that images are particularly popular on platforms like Tumblr, Instagram, and Flickr, as they provide a natural and rich way to communicate sentiments [10]. Images have a greater impact than text because of their ties to our memories and emotions. Moreover, the human brain is wired to process visual information more quickly, retain it for a longer period and react to it more emotionally. This is why sociologists, psychiatrists and marketers have long recognized the power of images to convey a mixture of different sentiments[11].

Different types of images can evoke a range of emotions in humans, and they can be classified into three categories. The first type is called visual information images, which elicit sentimental responses through their colors and texture features. The second type is called semantic object images, which elicit sentimental responses through the semantic features derived from object detection. The last type is called semantic relation images, which elicit sentimental responses through the relationship between objects [12]. Analyzing these images is crucial to understanding the positive and negative sentiments conveyed within them. Predicting visual sentiments from social network images can help determine how users feel about the image and why they choose to share it online.

Previous studies on social network sentiment analysis showed that predicting sentiments from images is a major challenge. One of these challenges is generating hand-crafted features from images for sentiment prediction, which requires a substantial amount of human effort and time. Another challenge is the affective gap between low-level visual features and high-level sentiments. Recent visual sentiment analysis studies have focused on learning sentimental features such as texture, brightness, color and so on from the entire image [11]. Whereas psychology studies [13] have proved that emotional stimuli within images can elicit visual sentiments. These emotional stimuli are color, salient objects, facial expressions or any other attributes. Social network images typically contain diverse categories of objects: living (flowers, birds, animals, insects) and non-living (vehicles, buildings). According to psychologists, the objects of the same class can express a variety of sentiments. Furthermore, diverse objects express the same sentiment (positive or negative) [14].

Therefore, this study aims to design a system to automatically predict visual sentiments from social network images based on object recognition techniques.

The problem can be determined in the following question: -

How to develop an automatic system for predicting visual sentiments from social network images using object recognition techniques?

3. The Objectives

The present study seeks to reach:

1) Determining the phases of developing the automatic system for visual sentiments prediction.
2) Developing a proposed automated system based on object recognition techniques for predicting visual sentiments from social network images.
3) Evaluating the effectiveness of the proposed system in predicting visual sentiments.

4. The Importance

The importance of the current study:

1) By utilizing computer vision techniques, particularly object recognition and implementing deep learning, it becomes possible to recognize and interpret the context, meaning and sentiment of images shared on social networks.
2) Use the proposed system as a novel technique for determining if an image will elicit positive or negative sentiments in a human viewer.
3) The current study adds to the realm of computer vision and is used for sentiment analysis.
4) In the education field, the proposed system can be utilized to ascertain the purpose behind students sharing images on social networking sites and their opinions on a specific lesson or subject.

5. Background

5.1.Related works

This section discusses previous studies on visual sentiment prediction. These studies can be divided into three aspects: predicting visual sentiment from whole image features, combining entire images with local regions and predicting visual sentiment from salient objects.

Deep CNNs were used by researchers to automatically learn sentiment features for image sentiment prediction. Building end-to-end deep neural network models is developed to extract image features for predicting visual sentiments [5]. As an early effort to construct an emotion-specific network, Peng et al. [15] presented a multi-level deep network to extract sentiment features from multi-scale patches (i.e., pixel-level feature, aesthetic feature and semantic feature). To more efficiently mix the features acquired from different levels, Rao et al.[16] also have presented a multi-level deep network that would predict visual sentiments using both low-level and high-level features.

Recently, researchers have concentrated on predicting visual sentiment using features taken from the image's salient objects. Fan et al. [2] published the first study to demonstrate the relationship between an image's emotional characteristics and visual attention. They created deep convolutional neural networks with additional channels for encoding local information. Yang et al. [8] used image regions to predict visual sentiment for the first time in their study. They proposed ARconcatenation, a framework for leveraging affective regions based on the object recognition method EdgeBoxes. A bounding box was drawn around each image to calculate the objectness score. They used CNN to learn the sentiments by extracting both global and local features from the image. The affective regions were created by combining the objectness and sentiment scores. Experiment results showed that visual sentiment analysis performance was improved using local information.

Xiong et al. [10] developed a framework that utilized the pre-trained VGGNet model and group sparse regularization to generate an initial sentiment prediction. By combining the sentimental features with the entire image, the sentiment regions were automatically identified to calculate the final sentiment score of the image. Zhang et al. [7] proposed a Bayesian network-based model to predict sentiments from object semantics in images.

A proposed system detailed in [14] introduced a deep learning network that effectively predicts visual sentiments using residual attention. The CNN architecture was utilized to learn the image feature spatial hierarchies. To concentrate on the local regions of the image that have rich sentiment, the residual attention model was employed. Additionally, this study explored the influence of fine-tuning on seven CNN-based architectures, such as InceptionResnet-V2, Inception-V3, NASNet, ResNet-50, Xception, VGG-16 and VGG-19 in the visual sentiment analysis field.

A scheme for recognizing emotion tags from images based on object-related features was introduced in [17]. They created a dataset of emotion-tagged images using subject evaluation from 212 users. The ResNet-50 model was utilized to extract deep features from object images for recognizing nine emotion categories, including amusement, awe, anger, boredom, contentment, disgust, excitement, fear, and sadness. The proposed scheme employed a two-level approach for emotion tagging (top-level and bottom-level). Results showed approximately 85% accuracy for top-level and 79% accuracy for bottom-level emotion tag recognition. A framework that utilizes a Graph Convolutional Network (GCN) was introduced in [18] to extract sentimental interaction features between objects. This framework consists of two branches, one that extracts visual and emotional features from images using a deep network, and the other that extracts emotional interaction features from objects using GCN. However, the experimental results have shown that effectively utilizing object interaction information remains a difficult task. In [13], a stimuli-aware VEA network was developed that can simulate how humans experience emotions when stimulated. The network consisted of three stages: stimuli selection, feature extraction and emotion prediction. The network uses off-the-shelf tools to select specific emotional stimuli from images, such as color, object and face. Three separate networks - Global-Net, Semantic-Net and

Expression-Net - extracted emotional features from multiple stimuli at the same time. They proposed a hierarchical cross-entropy loss method to help the network distinguish between hard false examples and easy ones and to learn in a way that is specific to each emotion.

Comment on the related works:

According to these previous studies, it is found that visual sentiment analysis studies have focused on the entire image to learn sentimental features such as color, brightness, texture, and so on. Whereas studies that focus on predicting visual sentiments based on object recognition are still in their elementary stage. Furthermore, it was shown that deep learning architecture-based techniques employing CNNs for deep feature extraction outperformed hand-crafted feature-based techniques for sentiment prediction. Therefore, this study introduces an automatic system based on object recognition techniques. The proposed system combines the InceptionV3 pre-trained model and a Long Short-Term Memory (LSTM) network to predict visual sentiment from social network images. This system aims to analyze images to predict visual sentiment by recognizing the salient objects from social network images and predicting their positive and negative sentiments.

5.2.Inception-V3 Model

The Inception V3 model is an upgraded version of Google's pre-trained model. It was trained on over 1.4 million images and 1000 classes. The architecture of an Inception v3 network is designed to has a depth of 42 layers over 20 million parameters as shown in Figure 1 [19].

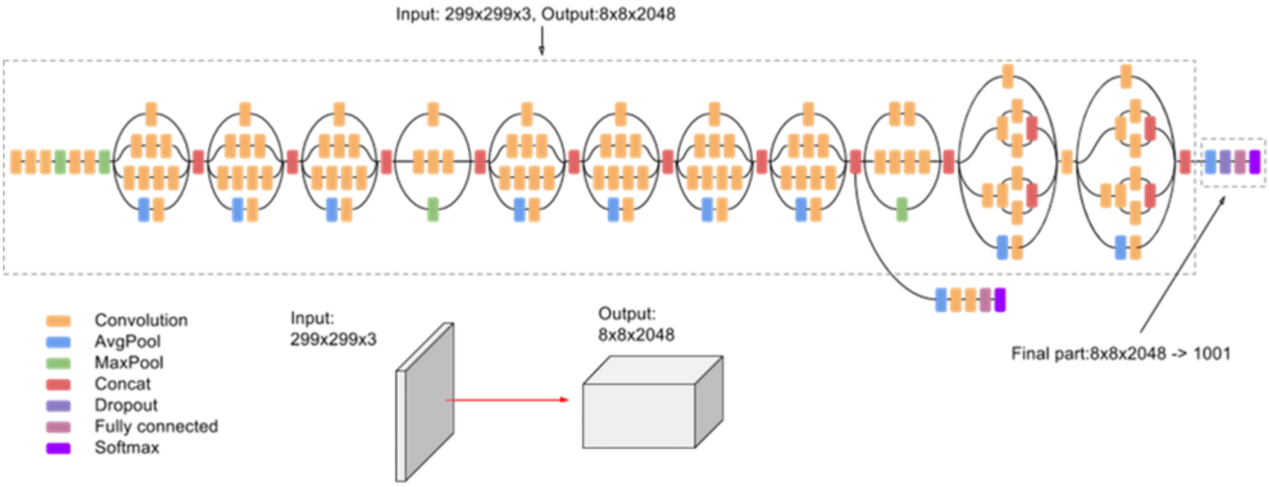
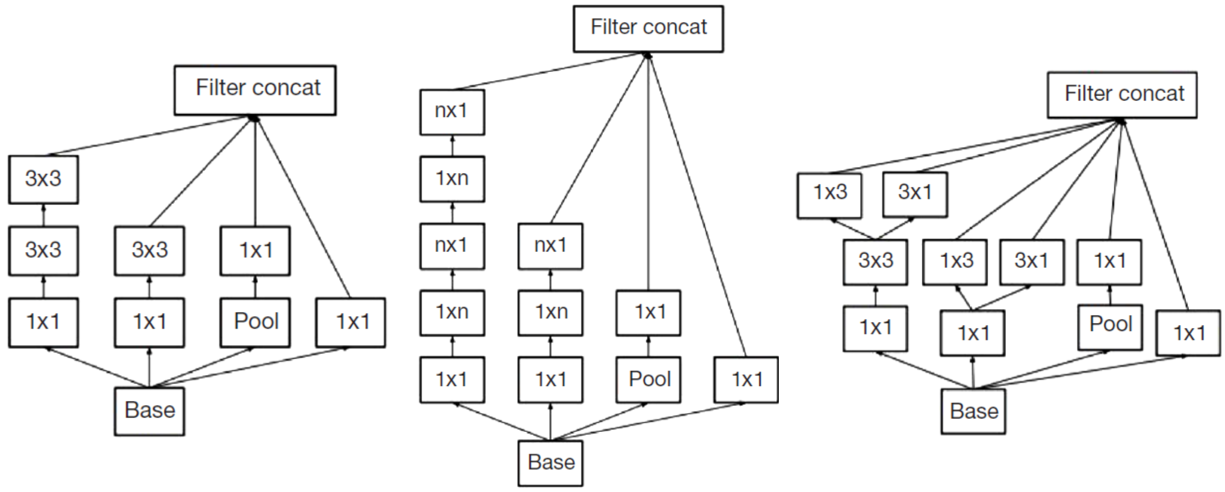


Figure 1: Inception V3 network architecture.

This network contains 94 convolution layers with different filter sizes. The first layer is input layer with size $(299 \times 299 \times 3)$. Furthermore, the network consists of symmetrical and asymmetrical building blocks referred to as inception modules. These convolution modules are designed to produce discriminating features and reduce the number of parameters. There are three types of Inception modules: InceptionA, InceptionB and InceptionC stacked in series as shown in Figure 2 [20].



(a) InceptionA (b) InceptionB (C) InceptionC

Figure2: The Inception modules of Inception-v3.

Each module contains various layers, including convolutional layers and pooling layers. In addition, a batch normalization layer is inserted into this network as a regularizer between the auxiliary classifier and the fully connected layer [21]. This layer is carried out to the activation layer for the network [22]. The batch normalization method ensures the output activations' performance. The output is normalized by subtracting the mean and dividing it by the standard deviation. Its purpose is to reduce the "internal covariance shift" of the activation layers [23].

5.3.Long Short-Term Memory (LSTM) Model

LSTM is a recurrent neural network (RNN) that can learn long-term dependencies between sequence data time steps. It is utilized to classify time series or sequential data in applications such image classification [24, 25], activity recognition and video description [26]. LSTM proposes memory blocks instead of RNN units to address the vanishing and exploding gradient problems. A memory block is a processing unit consisting of one or multiple memory cells. The operation of a memory block is controlled by a group of adaptive multiplicative gates [27]. The architecture of the LSTM network includes a memory cell and three gates: an input gate, an output gate and a forget gate as shown in Figure 3 [28, 26] .

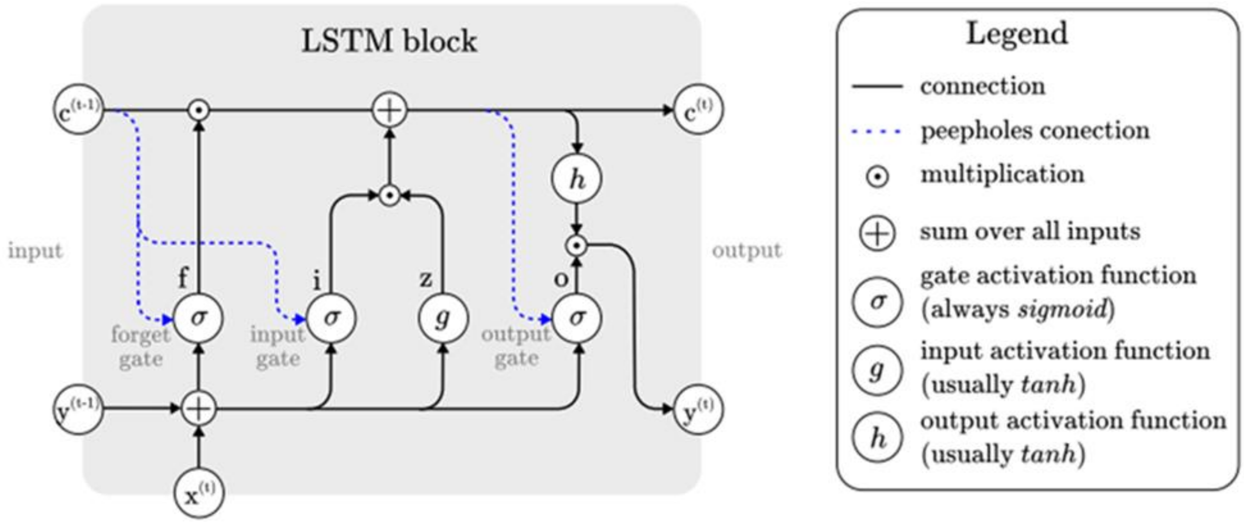


Figure 3 : The architecture of LSTM network [26].

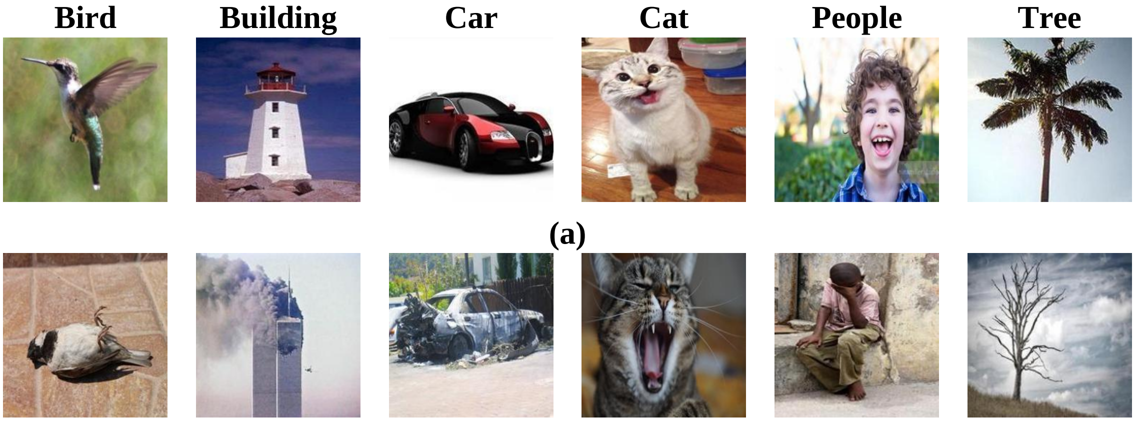
A memory cell is referred to as a 'cell state' because it keeps its state over time by remembering values at arbitrary time intervals. The three gates are used to control the memory cell's output and internal state [29]. These gates allow information to flow into and out of the cell. The mechanism uses a sigmoid neural net layer and pointwise multiplication. The sigmoid layer outputs values between 0 and 1. Zero means nothing should pass, while one means everything should pass. An LSTM has a hidden state: $H_{(t-1)}$ for the previous timestamp and H_t for the current timestamp. It also includes a cell state denoted by $C_{(t-1)}$ and C_t for the previous and current timestamps, respectively. In this context, the short term memory refers to the hidden state, while the long term memory refers to the cell state [26].

6. Dataset Preparation

The dataset utilized comprises images sourced from social networks. These images have been selected to be useful for object recognition and sentiment prediction. The dataset consists of 550 images of nine real-life objects. These objects are a bird, building, car, cat, dog, flower, knife, people and Tree. Each object is represented by 50 images, except for the "people" object which has 150 images as shown in Table 1. As shown in Figure 4, the object images display positive and negative visual sentiments, with each object displaying a unique range of visual sentiments. The prepared datasets are split into two divisions in the applied experiments: one for training and the other for testing. The most effective split is 70% for training and 30% for testing [30, 31].

Table 1: The prepared dataset information.

Content	Sentiments Category		Number
	Positive	Negative	
Bird	43	7	50
Building	15	35	50
Car	27	23	50
Cat	36	14	50
Dog	37	13	50
Flower	37	13	50
Knife	5	45	50
People	80	70	150
Tree	30	20	50
Total	310	240	550



(a)

(b)

Figure 4: Selected six images from the test dataset.
(a) Positive sentiment, (b) Negative sentiment.

7. Proposed System Stages

This system is developed to work in an automated way to predict the visual sentiments from social network images by recognizing the proposed objects and learning their visual sentiment features. The proposed system is built by combining the InceptionV3 pre-trained model with LSTM networks. Figure 5 shows the detailed processes of the system. The visual sentiments prediction system works in two stages:

Stage1: Object Recognition

This stage aims to recognize the proposed objects from the input image. In this stage, the Inception V3 pre-trained CNN model is used and fine-tuned on the prepared training dataset to adjust the deep network's parameters. The object recognition model is created using the following procedures. Firstly, image augmentation and preprocessing operations are applied to all input images. They are automatically resized to 299 × 299 and converted into RGB. They are also randomly flipped vertically, translated up to 30 pixels, and scaled up to 10% horizontally and vertically. These images serve as the training network's data supply. Secondly, the Inception V3 pre-trained network is used to retrain the network to recognize objects by extracting their features. This involves fine-tuning the structure of the Inception V3 model by replacing the last fully connected layer with a new layer containing nine outputs equal to the number of objects in the dataset. In addition, the weights of the first ten layers are frozen by setting their learning rates to zero.

For training the network, the stochastic gradient descent with momentum is used to optimize the network and the cross-entropy is used as the loss function. Furthermore, the different hyperparameters are adjusted. The corresponding experiments are applied to determine the optimal values for *InitialLearnRate*, *MiniBatchSize* and *MaxEpochs*. From the experimental results, the optimal values of these three parameters are listed in Table 2. Finally, the trained network is used to recognize objects from test images. The softmax activation function calculates the probabilities of objects based on Equation 1 [32]. The softmax classifier takes as input a vector of features derived from the learning process and returns the probability for each object class. The classification layer outputs a single label for each object based on this probability.

$$y_i = \frac{exp^{T_i}}{\sum_{n=1}^{i} exp^{T_n}} \tag{1}$$

where y_i: the probabilities for each object, exp^{T_i}: the standard exponential function applied on T vector of i dimension. i: number of objects.

Table 2: The hyperparameters values of the Inception V3 Network.

Parameters	**Values**
InitialLearnRate	0.0003
MiniBatchSize	10
MaxEpochs	12
Momentum	0.9000
L2Regularization	0.0001
ValidationFrequency	42

Stage2: Sentiment Prediction

This stage aims to predict the visual sentiments from the recognized object's features by developing a sentiment prediction model. Three steps are applied to build the sentiment prediction model.

The first step is object feature extraction. In this step, the object's features are extracted using the Inception V3 trained network from stage 1. These features are extracted from the final pooling layer using the activation function and obtained as a 2048-dimensional vector. The features vector is saved to be used as input to the sentiment prediction model.

In the second step, the LSTM network is built to handle sentiment features related to objects as sequential data and represent their joint probability distribution. The network contains a sequence input layer with a 2048 input size, an LSTM layer with 20 hidden units and a dropout layer. Additionally, the network includes three layers for visual sentiment prediction: a fullyconnected layer with two output categories, a softmax layer and a classification layer. The optimizer used to train the network is adaptive moment estimation. Furthermore, the different hyperparameters are adjusted for training the LSTM network to learn the object's visual sentiments. These hyperparameters values are given in Table 3. The corresponding experiments are applied to determine the best values for the *InitialLearnRate* and *MiniBatchSize* parameters. From the experimental results, the best values of these parameters are 0.0001 for *InitialLearnRate* and 12 for *MiniBatchSize* as shown in Table3.

The third step is to classify sentiment. The LSTM trained network is used to predict the visual sentiments from the social network images. The trained network uses the object features vector to map sentiment polarity. The softmax activation function is used to convert the joint distribution of features into sentiment scores. The cross-entropy loss function measures the loss of predicted probability, defined as:

$$L(\acute{y}, y) = -\frac{1}{N}\sum_{i=0}^{N}[y_i \log(\acute{y}_i) + (1 - y_i)\log(1 - \acute{y}_i)] \tag{2}$$

Where $L(\acute{y}, y)$: the loss function value, N represents the total number of images per mini-batch, y_i and $\acute{y}_i$ denote target and predicted labels, respectively.

Table3: The hyperparameters values for LSTM network.

Parameters	**Values**
InitialLearnRate	0.0003
MiniBatchSize	12
GradientDecayFactor	0.9000
SquaredGradientDecayFactor	0.9990
Epsilon	10^ (-8)
L2Regularization	0.0001
GradientThresholdMethod	L2norm

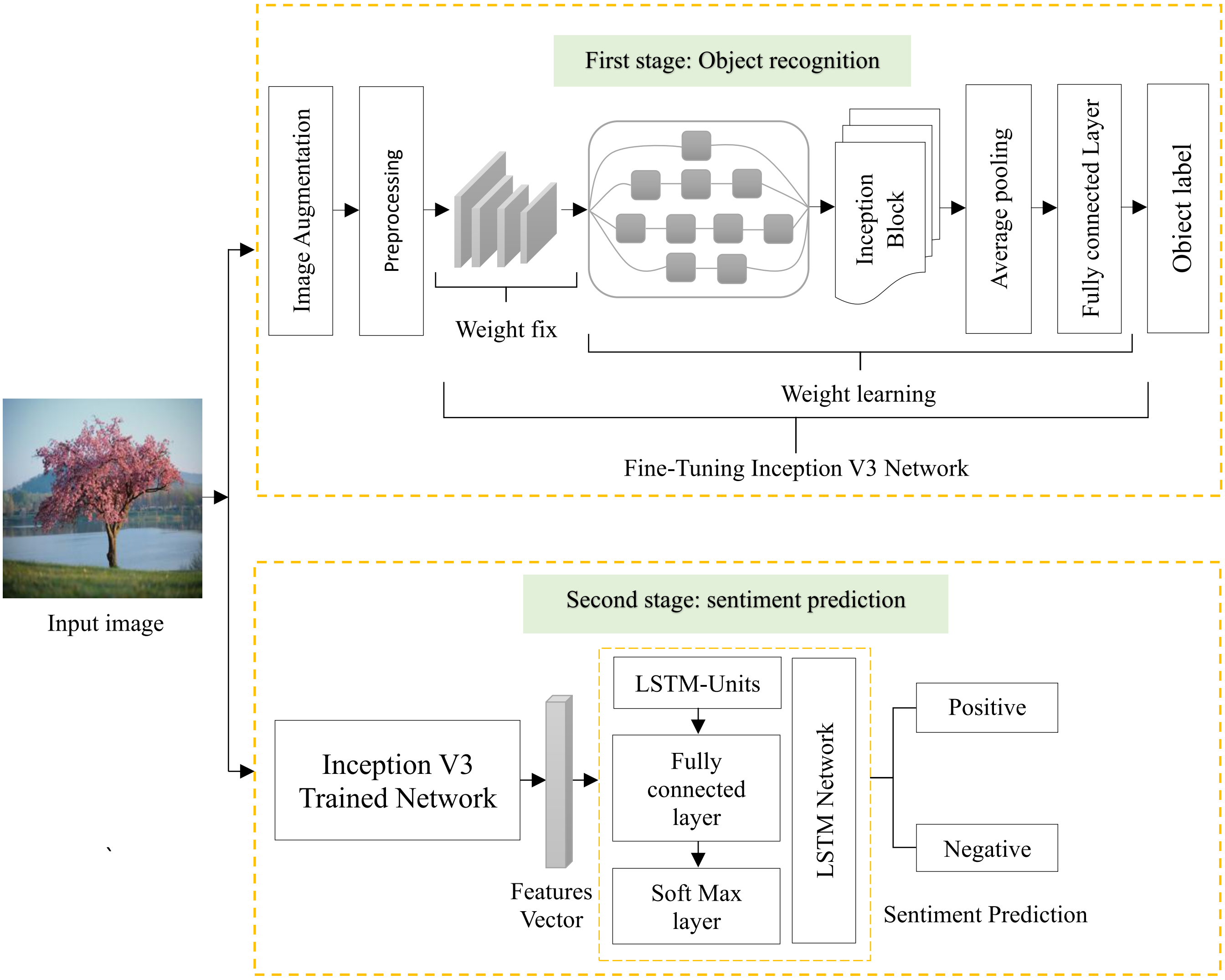


Figure 5: The architecture of the proposed system.

8. Experiments

These experiments are designed to calculate and assess the performance of the proposed system for recognizing objects and predicting their visual sentiments from social network images. All experiments have been performed using MATLAB 2021a on a Windows X86-64 machine with Intel(R) Core (TM) i7-6700HQ CPU, 16 GB RAM and NVIDIA GeForce GTX 950M. The proposed system is written using MATLAB toolboxes. They are Deep Learning, image processing and Parallel Computing toolboxes. The dataset prepared in Section 6 is used in all experiments. The experiments are divided into two groups.

The first group of experiments is designed to determine the optimal trained network of inceptionV3 for object recognition. Therefore, 27 experiments are applied to determine the best values of three hyperparameters for the training network. These hyperparameters are *InitialLearnRate*, *MaxEpochs* and *MiniBatchSize*. Each hyperparameter has three values. The network is trained using all combinations of these values to specify the best as shown in Table 4. After the training, the best-trained network is then used to recognize objects from the test dataset and extract visual sentiment features from them. These extracted features are used as the input to the LSTM network in the next experimental group.

Table 4: The hyperparameters values of Inception V3 used in the training experiments.

ExpNo.	***InitialLearnRate***	***MaxEpochs***	***MiniBatchSize***
1	0.0001	4	8
2	0.0001	4	10
3	0.0001	4	12
4	0.0001	8	8
5	0.0001	8	10
6	0.0001	8	12
7	0.0001	12	8
8	0.0001	12	10
9	0.0001	12	12
10	0.0003	4	8
11	0.0003	4	10
12	0.0003	4	12
13	0.0003	8	8
14	0.0003	8	10
15	0.0003	8	12
16	0.0003	12	8
17	0.0003	12	10
18	0.0003	12	12
19	0.0005	4	8
20	0.0005	4	10
21	0.0005	4	12
22	0.0005	8	8
23	0.0005	8	10
24	0.0005	8	12
25	0.0005	12	8
26	0.0005	12	10
27	0.0005	12	12

The second group of experiments is designed to determine the optimal LSTM network architecture and evaluate the proposed system performance for predicting visual sentiments. In these experiments, the best-trained inceptionV3 network is used to extract visual sentiment features of objects that are used as input to the LSTM network. 27 experiments are applied by varying the number of hidden units in the LSTM network layer. Furthermore, the various *InitialLearningRate* and *MiniBatchSize* parameter values are tested to specify the best values for training the network to learn the visual sentiment features of objects. Table 5 shows the three values specified for each parameter. The LSTM network is trained using all combinations of these parameters. The proposed system that uses a combination of a well-trained InceptionV3 network and optimal LSTM network architecture is applied to the testing dataset to predict visual sentiments from images.

Table5: The proposed parameter values of LSTM network.

Parameters	**Values**		
Num of Hidden Units	10	20	30
Initial Learning Rate	0.0001	0.0003	0.0005
MiniBatchSize	10	12	14

9. Results and discussion

This section discusses the experimental results obtained from the proposed experiments. There are two groups of experiments as mentioned above. The first group consists of 27 experiments that are applied to determine the best hyperparameters values for fine-tuning the InceptionV3 pre-trained network. These hyperparameters are *InitialLearnRate*, *MaxEpochs* and *MiniBatchSize*. Table 6 shows the best results of only nine experiments. These experiments' networks are well-trained and achieved high performance when applied to the test dataset. Based on the results of these nine experiments, the best hyperparameters values are determined to be 0.0003, 12 and 10 respectively. After that, the best-trained network is applied to recognize objects. The confusion matrix shows that all objects are accurately recognized as shown in Figure 6.

Table 6: The best hyperparameters values of the InceptionV3 network with only 9 experiments.

ExpNo.	***InitialLearnRate***	***MaxEpochs***	***MiniBatchSize***	**Accuracy %**
10	0.0003	4	8	100
11	0.0003	4	10	100
12	0.0003	4	12	99.4
13	0.0003	8	8	100
14	0.0003	8	10	98.8
15	0.0003	8	12	99.4
16	0.0003	12	8	100
17	0.0003	12	10	100
18	0.0003	12	12	100

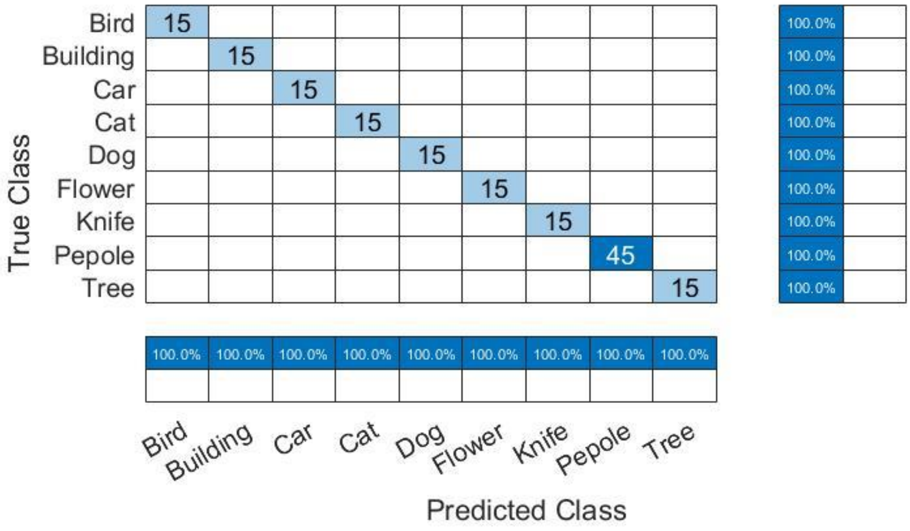


Figure 6: The confusion matrix of object recognition.

In the second group of experiments, 27 experiments are applied to determine the optimal LSTM network architecture and hyperparameter values. The best results are obtained in the nine experiments shown in Table 7. It is found that specifying 20 hidden units yields the optimal LSTM network architecture. Furthermore, the best values for *InitialLearnRate* and *MiniBatchSize* parameters are 0.0003 and 12, respectively.

Table 7: The best parameters values of the LSTM network with only 9 experiments.

ExpNo.	**Hidden Units**	***InitialLearnRate***	***MiniBatchSize***	**Accuracy%**	
				Training	**Test**
10	20	0.0001	10	100%	97.6%
11	20	0.0001	12	100%	98.2%
12	20	0.0001	14	99.7%	97.6%
13	20	0.0003	10	100%	98.2%
14	20	0.0003	12	100%	98.2%
15	20	0.0003	14	100%	98.2%
16	20	0.0005	10	100%	98.2%
17	20	0.0005	12	100%	98.2%
18	20	0.0005	14	100%	98.2%

After that, the proposed system that combined the InceptionV3-LSTM networks is applied on the test dataset images to evaluate the performance of predicting visual sentiments. The results of the proposed system are listed in Table 8. This table shows the proposed system's accuracy in predicting visual sentiments for each recognized object. The second column shows the correctly recognized objects. The next two columns show the visual sentiments correctly predicted for positive and negative images. The results show that the proposed system correctly recognized all objects. Furthermore, it is found that the proposed system's accuracy for predicting visual sentiments is 98.2% with three visual sentiments misclassified; as shown in Figure 7. The images of these sentiments are shown in Figure 8.

Table 8: The proposed system accuracy.

Objects	**Recognized Object**	**Correct Visual Sentiment Prediction**	
		Positive	**Negative**
Bird	15	12	1
Cat	15	11	3
Building	15	3	12
Car	15	8	7
Dog	15	12	3
Flower	15	12	3
People	45	25	20
Knife	15	1	14
Tree	15	10	5
Total	165	94	68
Accuracy	100%	100%	95.8
Average accuracy		98.2%	

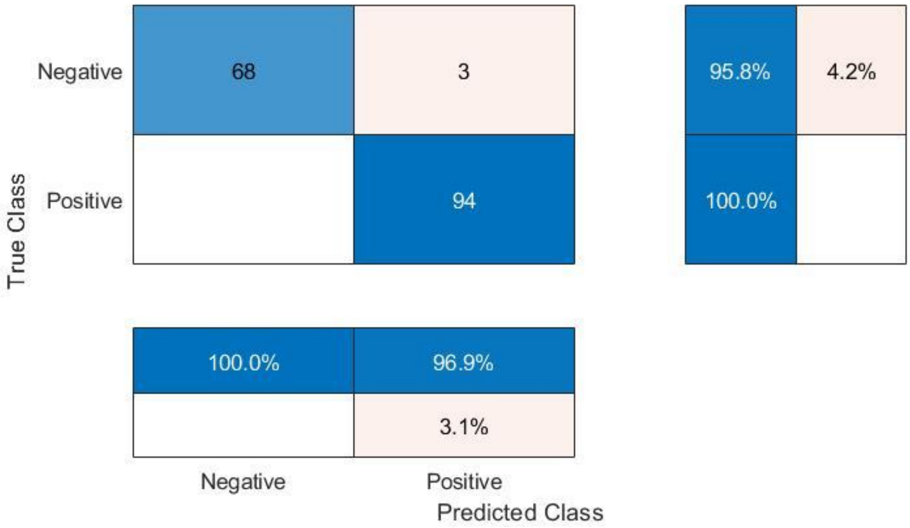


Figure 7: Performance of visual sentiments prediction.

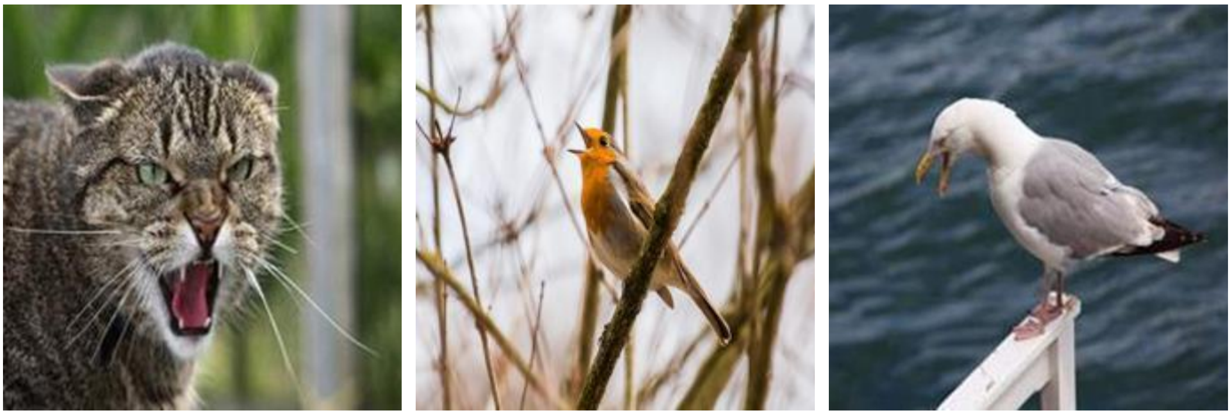


Figure 8: The mispredicted sentiments.

On analyzing the results, it can easily be observed that the visual sentiment prediction accuracy achieved high performance when using the proposed InceptionV3-LSTM network. Moreover, the positive sentiment achieved the highest accuracy compared to the negative sentiment. The positive sentiment accuracy is 100% whereas the negative sentiment accuracy is 95.8%. As shown in Figure 8, there are three images for negative sentiments incorrectly predicted. Incorrect predictions can occur due to the similarities in features between images that convey positive emotions. For example, the first image displays an angry cat baring its teeth, while some other images of cats that convey positive emotions may also show the same feature. This can make it understandable why such mistakes occur. Learning prior knowledge through simple training may be challenging for deep networks dealing with complex scenarios. As a result, designing emotion-specific networks for each object becomes necessary.

10. Conclusion

This work has presented an automatic system for visual sentiment prediction based on object recognition from social network images. The proposed system is designed by combining InceptionV3 and LSTM networks. Firstly, the pre-trained Inception V3 network is fine-tuned to recognize the proposed objects in an image. Then the visual sentiment features of the recognized object are extracted as a 2048-dimensional vector. The features vector is given to the LSTM network to learn the distribution of features. Finally, these features are classified into positive and negative sentiments by using the softmax activation function. Two groups of experiments are applied to evaluate the system's performance. The first one is used to specify the best values of hyperparameters for the Inception V3 network. The second group is used to find the optimal LSTM network architecture and evaluate the proposed system performance for predicting visual sentiments. The results show that the proposed system which combines the Pre-trained Inception V3 network with the LSTM network is a more powerful system for predicting visual sentiments. The proposed system's accuracy for predicting visual sentiments is 98.2%.

In the future, the system can detect multiple objects in images by designing an emotion-specific network to predict sentiment based on object detection.

References

[1]Stieglitz, S., Mirbabaie, M., Ross, B., & Neuberger, C. (2018). Social media analytics–Challenges in topic discovery, data collection, and data preparation. International journal of information management, 39, 156-168.

[2]Fan, S., Jiang, M., Shen, Z., Koenig, B. L., Kankanhalli, M. S., & Zhao, Q. (2017, October). The role of visual attention in sentiment prediction. In Proceedings of the 25th ACM international conference on Multimedia (pp. 217-225).

[3]You, Q., Jin, H., & Luo, J. (2017, February). Visual sentiment analysis by attending on local image regions. In Proceedings of the AAAI conference on artificial intelligence (Vol. 31, No. 1).

[4]Chen, J., Mao, Q., & Xue, L. (2020). Visual sentiment analysis with active learning. IEEE Access, 8, 185899-185908.

[5]Zhang, J., Liu, X., Chen, M., Ye, Q., & Wang, Z. (2022). Image sentiment classification via multi-level sentiment region correlation analysis. Neurocomputing, 469, 221-233.

[6]Song, K., Yao, T., Ling, Q., & Mei, T. (2018). Boosting image sentiment analysis with visual attention. Neurocomputing, 312, 218-228.

[7]Zhang, J., Chen, M., Sun, H., Li, D., & Wang, Z. (2020). Object semantics sentiment correlation analysis enhanced image sentiment classification. Knowledge-Based Systems, 191, 105245.

[8]Yang, J., She, D., Sun, M., Cheng, M. M., Rosin, P. L., & Wang, L. (2018). Visual sentiment prediction based on automatic discovery of affective regions. IEEE Transactions on Multimedia, 20(9), 2513-2525.

[9]Wu, L., Qi, M., Jian, M., & Zhang, H. (2020). Visual sentiment analysis by combining global and local information. Neural Processing Letters, 51, 2063-2075.

[10]Xiong, H., Liu, Q., Song, S., & Cai, Y. (2019). Region-based convolutional neural network using group sparse regularization for image sentiment classification. EURASIP Journal on Image and Video Processing, 2019(1), 1-9.

[11]Campos, V., Jou, B., & Giro-i-Nieto, X. (2017). From pixels to sentiment: Fine-tuning CNNs for visual sentiment prediction. Image and Vision Computing, 65, 15-22.

[12]Yamamoto, T., Takeuchi, S., & Nakazawa, A. (2021). Image emotion recognition using visual and semantic features reflecting emotional and similar objects. IEICE TRANSACTIONS on Information and Systems, 104(10), 1691-1701.

[13]Yang, J., Li, J., Wang, X., Ding, Y., & Gao, X. (2021). Stimuli-aware visual emotion analysis. IEEE Transactions on Image Processing, 30, 7432-7445.

[14]Yadav, A., & Vishwakarma, D. K. (2020). A deep learning architecture of RA-DLNet for visual sentiment analysis. Multimedia Systems, 26(4), 431-451.

[15]Rao, T., Li, X., & Xu, M. (2020). Learning multi-level deep representations for image emotion classification. Neural processing letters, 51, 2043-2061.

[16]Rao, T., Li, X., Zhang, H., & Xu, M. (2019). Multi-level region-based convolutional neural network for image emotion classification. Neurocomputing, 333, 429-439.

[17]Manzoor, A., Ahmad, W., Ehatisham-ul-Haq, M., Hannan, A., Khan, M. A., Ashraf, M. U., ... & Alfakeeh, A. S. (2020). Inferring Emotion Tags from Object Images Using Convolutional Neural Network. Applied Sciences, 10(15), 5333.

[18]Wu, L., Zhang, H., Deng, S., Shi, G., & Liu, X. (2021). Discovering sentimental interaction via graph convolutional network for visual sentiment prediction. Applied Sciences, 11(4), 1404.

[19]Kurama, V. (2020). A review of popular deep learning architectures: Resnet, inceptionv3, and squeezenet. Consult. August, 30.

[20]Guan, Q., Wan, X., Lu, H., Ping, B., Li, D., Wang, L., ... & Xiang, J. (2019). Deep convolutional neural network Inception-v3 model for differential diagnosing of lymph node in cytological images: a pilot study. Annals of translational medicine, 7(14).

[21]Cao, J., Yan, M., Jia, Y., Tian, X., & Zhang, Z. (2021). Application of a modified Inception-v3 model in the dynasty-based classification of ancient murals. EURASIP Journal on Advances in Signal Processing, 2021(1), 1-25.

[22]CA, G. S., Bhowmik, N., & Breckon, T. P. (2019, December). Experimental exploration of compact convolutional neural network architectures for non-temporal real-time fire detection. In 2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA) (pp. 653-658). IEEE.

[23]Alzubaidi, L., Zhang, J., Humaidi, A. J., Al-Dujaili, A., Duan, Y., Al-Shamma, O., ... & Farhan, L. (2021). Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions. Journal of big Data, 8, 1-74.

[24]Aditi, M. K., & Poovammal, E. (2019). Image classification using a hybrid LSTM-CNN deep neural network. Int. J. Eng. Adv. Technol, 8(6), 1342-1348.

[25]Li, P., Tang, H., Yu, J., & Song, W. (2021). LSTM and multiple CNNs based event image classification. Multimedia Tools and Applications, 80, 30743-30760.

[26]Van Houdt, G., Mosquera, C., & Nápoles, G. (2020). A review on the long short-term memory model. Artificial Intelligence Review, 53, 5929-5955.

[27]Swapna, G., Kp, S., & Vinayakumar, R. (2018). Automated detection of diabetes using CNN and CNN-LSTM network and heart rate signals. Procedia computer science, 132, 1253-1262.

[28]Tatsunami, Y., & Taki, M. (2022). Sequencer: Deep lstm for image classification. Advances in Neural Information Processing Systems, 35, 38204-38217.

[29]Bawa, V. S., & Kumar, V. (2019). Emotional sentiment analysis for a group of people based on transfer learning with a multi-modal system. Neural Computing and Applications, 31, 9061-9072.

[30]Vrigazova, B. (2021). The proportion for splitting data into training and test set for the bootstrap in classification problems. Business Systems Research: International Journal of the Society for Advancing Innovation and Research in Economy, 12(1), 228-242.

[31]Nguyen, Q. H., Ly, H. B., Ho, L. S., Al-Ansari, N., Le, H. V., Tran, V. Q., ... & Pham, B. T. (2021). Influence of data splitting on performance of machine learning models in prediction of shear strength of soil. Mathematical Problems in Engineering, 2021, 1-15.

[32]Rangarajan Aravind, K., & Raja, P. (2020). Automated disease classification in (Selected) agricultural crops using transfer learning. Automatika: časopis za automatiku, mjerenje, elektroniku, računarstvo i komunikacije, 61(2), 260-272.

تطوير نظام اوتوماتيك قائم على تقنيات التعرف على الكائنات للتنبؤ بالمشاعر المرئية من صور الشبكات الاجتماعية

الملخص:

أصبحت الشبكات الاجتماعية جزءًا حيويًا من حياة الجميع، حيث يشارك المستخدمون على منصات الشبكات الاجتماعية الشهيرة ملايين الصور للتعبير عن آرائهم ومشاعرهم الشخصية. لذلك، ظهرت الصور كواحدة من أكثر الطرق فعالية لنقل المشاعر على الشبكات الاجتماعية. وقد أدى ذلك إلى رؤية قوية لتحليل صور الشبكات الاجتماعية للتنبؤ بالمشاعر الإيجابية والسلبية من هذه الصور. في هذا البحث، تم تطوير نظام اوتوماتيك قائم على التعرف على الكائنات من خلال الدمج بين الشبكات العصبية لتقنية InceptionV3 Network وتقنية Long Short-Term Memory Network للتنبؤ بالمشاعر المرئية. يهدف هذا النظام إلى التعرف على الكائنات من صور الشبكات الاجتماعية والتنبؤ بالمشاعر بتلك الصور. أولا، تم ضبط الشبكة العصبية CNN المدربة مسبقا InceptionV3 pre-trained CNN network للتعرف على الكائنات من الصور. بعد ذلك، تم استخدام الشبكة المدربة فى استخراج مميزات الكائن. أخيراً، تم استخدام الشبكة العصبية Long Short-Term Memory Network لتعلم المشاعر من ميزات الكائن للتنبؤ بالمشاعر المرئية. أظهرت النتائج أن النظام المقترح هو نظام أكثر قوة للتنبؤ بالمشاعر المرئية من خلال الدمج بين الشبكات العصبية لتقنية InceptionV3 Network وتقنية Long Short-Term Memory Network حيث حقق النظام المقترح نسبة ٩٨.٢٪ للتنبؤ بالمشاعر المرئية.

الكلمات المفتاحية :

معالجة الصور – التعلم العميق – التعرف على الكائنات – التنبؤ بالمشاعر المرئية – صور الشبكات الاجتماعية