IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

SPECIAL SECTION ON ADVANCED DATA MINING METHODS FOR SOCIAL COMPUTING

Received December 28, 2019, accepted February 6, 2020, date of publication February 27, 2020, date of current version March 13, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2976530

Detection of Possible Illicit Messages Using Natural Language Processing and Computer Vision on Twitter and Linked Websites

SERGIO L. GRANIZO, ÁNGEL LEONARDO VALDIVIESO CARAGUAY, LORENA ISABEL BARONA LÓPEZ, AND MYRIAM HERNÁNDEZ-ÁLVAREZ

Department of Informatics and Computer Science (DICC), Escuela Politécnica Nacional, Quito 170517, Ecuador

Corresponding author: Lorena Isabel Barona López (lorena.barona@epn.edu.ec)

This work has been partially supported by Escuela Politécnica Nacional under Grant PIS-17-10 (Research and Social Projection Management Unit).

ABSTRACT Human trafficking is a global problem that strips away the dignity of millions of victims. Currently, social networks are used to spread this crime through the online environment by using covert messages that serve to promote these illegal services. In this context, since law enforcement resources are limited, it is vital to automatically detect messages that may be related to this crime and could also serve as clues. In this paper, we identify Twitter messages that could promote these illegal services and exploit minors by using natural language processing. The images and the URLs found in suspicious messages were processed and classified by gender and age group, so it is possible to detect photographs of people under 14 years of age. The method that we used is as follows. First, tweets with hashtags related to minors are mined in real-time. These tweets are preprocessed to eliminate noise and misspelled words, and then the tweets are classified as suspicious or not. Moreover, geometric features of the face and torso are selected using Haar models. By applying Support Vector Machine (SVM) and Convolutional Neural Network (CNN), we are able to recognize gender and age group, taking into account torso information and its proportional relationship with the head, or even when the face details are blurred. As a result, using the SVM model with only torso features has a higher performance than CNN.

INDEX TERMS CNN, features detection, image classification, natural language processing, SVM.

I. INTRODUCTION

Initially the websites were isolated and just placed for reading since the user could not truly interact with the web. However, from the innovation and arrival of web 2.0, there was a revolutionary and radical change since the user stopped being a simple spectator and became an active individual in social networks such as Facebook, Twitter, Instagram, among others [1].

Unfortunately, a door has also been opened for illegal businesses such as human trafficking [2]–[5], where some countries, such as Latin American countries, have the highest rates of smuggling of people, especially children and adolescents under 14 years old. It is important to note that the average age of consent is 14 years old in Latin American countries, so if underage people are used for illicit services are directly considered victims of human trafficking. Currently, in Twitter [6], it is possible to find websites that offer escort or similar services where young girls are promoted for the consumption of ''customers.'' These girls are generally abused physically [7], psychologically, and sexually [8]–[11].

In recent years many criminal organizations advertise these ''sexual services'' using social networks hiding their illegal activity with seemingly innocuous terms such as ''chicken soup'' to refer to child pornography. Websites and social networks are used to extend this crime to the online environment, where covert advertising and messages are used to promote illegal services to exploit people who are victims of this crime, mainly minors.

Although there are previous tweet filtering and image classification works to detect illicit messages, most of them use natural language processing methods or computer vision techniques separately. However, a different treatment of text and images is shown in [12]. In this paper, the authors focus

The associate editor coordinating the review of this manuscript and approving it for publication was Huizhi Liang.

their efforts on the analysis of advertisement published on the web for automatic detection of suspected messages. They use 10,000 ads manually annotated for this task. This work labels advertising that has text and images, and the analysis combines both types of information. They use a deep multimodal model called Human Trafficking Deep Network, and they obtained an F1 value of 75.3% with a recall of 70.9%.

On the other hand, the current image classification models use only facial information without taking into account that most of the images have the face blurred. In [12], the authors use computer vision algorithms to predict age with an approximate accuracy of 86.64%. In [13], SVM and CNN classification models are used to define the gender of a person. To the best of our knowledge, there are no works that consider characteristics of the upper body (upper torso) in the images to classify age groups.

The present work has two phases. In the first stage, natural language processing techniques are used in order to identify messages on Twitter that promote illicit services provided by minors. In the second phase, from the websites categorized as suspects, images are extracted in order to perform image processing and gender recognition of two age groups: over 14 years and under or equal to 14 years old. For this recognition, not only the characteristics of the torso but also the facial features were used. It is worth to mention that several images are often blurred and pixelated.

This document is organized into eight sections, the first one being the present introduction. Section 2 presents related works. Then, Section 3 shows our system proposal to detect possible human trafficking based on the analysis of the upper body (torso). Section 4 describes the first phase of our project, which is the extraction and processing of tweets that can be related to human trafficking. Section 5 details the image extraction, processing, and classification by gender and age. Section 6 describes the machine learning algorithms used, SVM and CNN, and how they work. Then, the experimental results are described in Section 7. Finally, the conclusions and future works are presented in Section 8.

II. RELATED WORK

There are research papers related to deception using social networks. In [10], the authors analyze how cheating techniques are used for the manipulation of content, information falsification, and handling of images and videos.These approaches are used in blogs, collaborative projects, microblogging, news sites or social networks, content communities, virtual social worlds, or virtual games. These deception techniques can have a high probability of success, depending on the skill of the attacker.

Some analyses have been performed using natural language processing, and their findings contribute to the combat of this crime. For instance, in [14], a study that analyzes suspicious tweets to detect illicit advertisements is presented. In [15], the authors use a semi-supervised learning approach to discern potential patterns of human trafficking to identify related ads. Moreover, they use non-parametric learning methods to implement text analysis. In the area of computer vision, some works classify images by age groups such as children, adolescents, and adults, focusing only on the face of the person [16].

In the same way, in [11], the weaknesses of social networks, particularly Facebook, are analyzed since this social network does not apply security filters. Similarly, there is a text manipulation [17] through the use of false surveys and opinions of products, which are sent to victims as spam in order to deceive unsuspecting users, especially those who are looking for a job or academic opportunities.

In [18], it is revealed that attackers use sophisticated software and techniques such as encrypted communication, or strongly protected online servers, to avoid tracking and remain an unknown status. Some authors [19], recommend analyzing: inconsistency of the age, variance in the alias, frequency of content, shared management, race, nationality, and third-party publications, in order to detect anomalies in the profiles of alleged followers which are online attackers. Some indicators of deception in social networks have been discovered. However, there are still limitations related to the quality of the information processed.

In [20], it is stated that in many countries, there are no real data about migrants due to the amount of illegal and undocumented people, or because many children are not registered or enrolled in school. Additionally, there are no complaints from the victim's kin, incompetence, or indifference of the authorities. On the other hand, some dysfunctional families obtain money through the prostitution of their children.

There are many challenges to fulfill in the area, so enhanced techniques must be developed in order to contribute in the detection of human trafficking indicators to social networks and linked websites.

III. SYSTEM PROPOSAL

Our proposal for the detection of suspicious websites is divided into two phases: i) Treatment, analysis, and classification of tweets using natural language processing and ii) Processing and classification of images hosted on websites classified as suspicious. For the first phase, some search criteria related to possible human trafficking were applied [8], especially with girls underage.

Figure 1 shows the whole process from the tweet searching related to human traffic [17] or slavery [21] of people; download and processing of this information [18], [22]; until the extraction of characteristics and their classification. The main objective of this phase is to obtain a blacklist of suspicious websites related to tweets. The second phase deals with the classification of images downloaded from the blacklist. Using predictive models, such as Vector Support Machine (SVM) and Convolutional Neural Networks (CNN), the image classification process is done through a training phase and a testing phase.

An overview of the second phase of our proposal is shown in Figure 2. Sections IV and V describe both phases of the work.

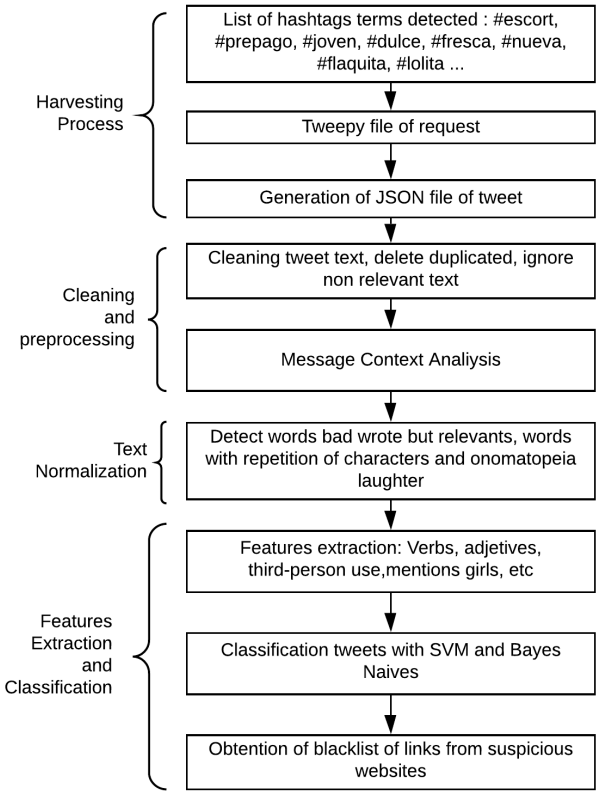

FIGURE 1. Tweet classification based on natural language processing.

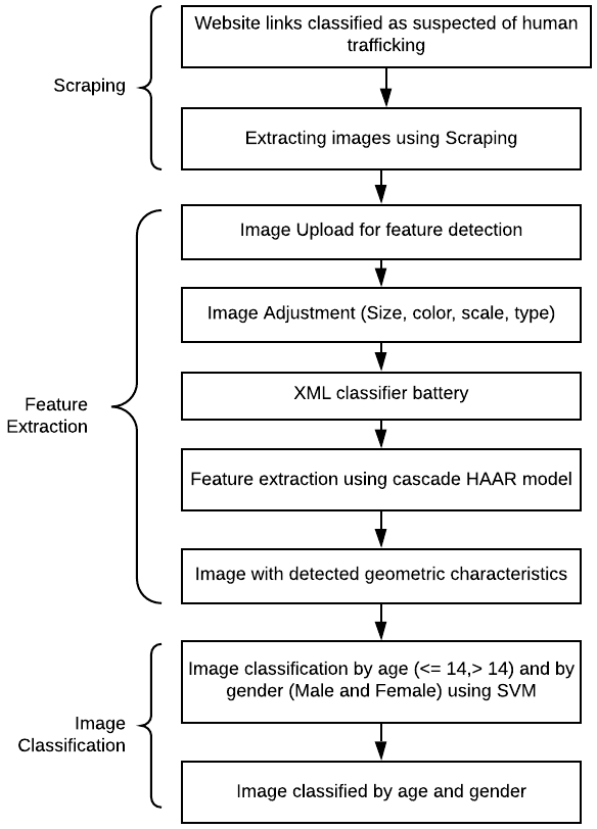

FIGURE 2. System overview of image processing.

IV. TWEET EXTRACTION AND PROCESSING

This section describes the process of tweets extraction, processing, and classification to determine if there are signs of human trafficking.

A. HARVESTING PROCESS

Initially, the data for each harvested day is stored on a JSON file that has information regarding the tweet post. The most relevant data includes the text of the tweet, user information, user mentions, associated URLs, and posted time. The capture process is shown in Figure 3.

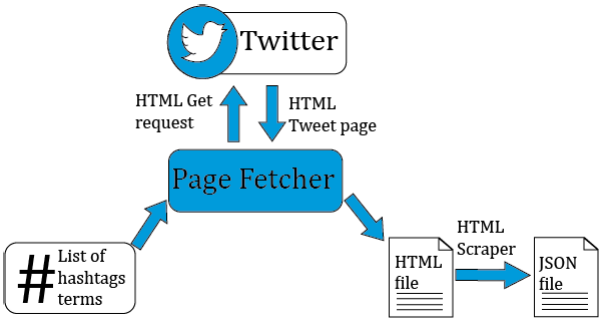

FIGURE 3. Tweets crawling process [14].

Spanish data is collected by executing a search request with the following hashtags: #escort, #prepago, #joven, #Dulce, #Fresca, #nueva, #lolita, and #flaquita. Hashtags were chosen as indicators of underage criteria. Tweets were used mined using the following criteria: mention of people from other countries if the tweets are written in the third person that shows that the Twitter user promotes the services of another person, or if the same user promotes the services of several people. To detect age, terms that indicate that people are underage victims were applied, such as the mention of a skinny young person or words from the jargon of pedophilia.

A preliminary analysis of tweets and Facebook messages that were denounced as guilty of sex trafficking was conducted. The following words, in Spanish, are frequently used: joven (young), dulce (sweet), fresca (fresh), nueva (new), Lolita, and flaquita (skinny). Other words like Caldo de Pollo, club penguin, and cp are used for criminals as an abbreviation of child pornography, the hashtags #escort, #prepago (prepaid), and the words mentioned before are chosen for this analysis. In Table 1, the number of tweet posts for each hashtag is summarized. For testing purposes, 100000 tweets were mined following the chosen words.

B. CLEANING AND PREPROCESSING

All downloaded messages always contain at least one of the hashtags mentioned in Table 1, and these messages are stored locally in a JSON file. The information is processed and cleaned using a Python application, and tweets are deleted according to the following criteria:

TABLE 1. Total tweets post by hashtag.

Hashtag	Number of occurrences
#escort	45604
#prepago	15890
#joven	3456
#dulce	1256
#fresca	1456
#nueva	5743
#flaquita	6580
#lolita	867
#penguin	23980
#caldodepollo	45990
#cp	34562
Twits with a URL link	1765

- Tweets with certain characters and not standardized were removed in order to build a readable and more precise text.
- Repetitive tweets, a user may post the same tweet many times. Then, eliminating repetitions avoids redundant information; otherwise, we will create bias in the subsequent analysis.
- Tweet that does not contribute to the project. For example, one of the words for the tweet filtering is ''young'' referring to children and adolescents, but if a user writes: ''Long live Quito! Young city!'', this message has another context. Therefore, it must be discarded.

The cleaning of the data is essential to perform the classification, so digits, stop words, and special characters are removed. Moreover, duplicated advertisements or tweets out of context were eliminated, automatically. If a URL link was matched to a night club website or massage therapy site, the tweet was manually tagged.

C. TEXT NORMALIZATION

Twitter messages usually have much noise due to the shortness of the texts and because they are mostly generated using mobile devices. Besides, many tweets have incompleted, misspelled, or distorted words, so the performance of natural language processing is degraded. Consequently, in the preprocessing, it is necessary to apply methods of tweet standardization written in Spanish. The text of Tweet messages was processing with a Spanish Spell-checker for lexical normalization algorithm in order to detect words OOV (Out of Vocabulary) using the following criteria:

- Detect incorrectly spelled words that are relevant (e.g., adolescents, young lady).
- Detect words with repetition of characters (p.e: besooooooo, siiiiiiiii).
- Detect the correct context of the words and foreign terms (e.g., xq, q +, plis).
- Detect onomatopoeias for laughter (e,g,, hahaha).

This corpus classified the words in three OOV categories: i) 0 = Variant (show suggested word), ii) 1 = Correct, or iii) 2 = Not processed (e.g., XD, xq). Some examples of Spanish tweets applying these normalization techniques are shown below:

Example 1: joder mis vecinos estan peleando a chillio limpio dioss pero en el fondo son una familia #dulce y #fresca xd

Example 2: @CazaEscorias d miiii? Perdona #lolita pero no sé d q m hablas mira por aquí ni m hables o por mensajes o por tuenti pero no m lies

Example 3: *@anngeleescastro graciias #flaquita ..... Jajajajajaj claroo q te invitooo.. Y tuu pagas vale?? Jajajaj tee quierooo graciias!!*

Example 4: *@Jurgi1998 no te vayas, que haces #flaquita gaupasa with my o algo asin*

Example 5: *Sale el anuncio en la #nueva radio del centro comercial de la Zenia y dicen: DE LA Z A LA A. Y dice mi madre: o sea que Zara sí que está jajajajaja*

Then, applying these three normalization criteria in these examples, the text normalization of these examples is shown in Table 2.

TABLE 2. Text normalizer [23].

	VOO	criteria	output
Example 1	estan	0	están
	chillio	0	chillido
	dioss	0	Dios
	xd	2	-
Example 2	miiii	0	mí
	tuenti	0	Tuenti
	lies	0	líes
Example 3	graciias	0	gracias
	Jajajajajaj	0	Ja
	claroo	0	claro
	invitooo	0	invito
	tuu	0	tú
	Jajajaj	0	Ja
	tee	0	te
	quieroooo	0	quiero
	graciias	0	gracias
Example 4	gaupasa	2	-
	with	2	-
	asin	0	asín
Example 5	Zenia	1	-
	jajajajaja	1	Ja
	Zara	0	-

Once the information is loaded and cleaned, the JSON format is converted into data frames corresponding to the features. Then, data can be filtered according to particular criteria. Additionally, some columns were created to perform a more efficient analysis. Then, using a Python program, the tweets were automatically tokenized; that is, all words of a text were split into individual words and converted to lower case.

D. FEATURES EXTRACTION AND CLASSIFICATION

Features are defined based on some criteria related to the deception and cybercrime [24], [25]. These criteria consider young age as an indicator for the detection of victims. For the input characteristics, the reason why considered each one is explained in Table 3.

With a Python program, syntax analysis is done in order to evaluate the higher frequency of adjectives and verbs,

TABLE 3. Considered features.

Features	*Reasons to considered each characteristic*
Quantity of words	Deceptive messages have more words to make them forgettable
URL links that are related to night club websites or massage therapy sites	It serves to detect the use of Twitter to publicize these sites
Third-person use	Deceptive messages have few self-references to avoid accountability. On the other hand, other people advertise the victims' services.
Same Twitter user talking about more than one victim	Covert publicity of illicit activities
The number of hashtags considered to harvest the data	Confirmation of Twitter relevance
A user account from one country that mentions girls from another country	it is a well-known signal of sex trafficking
The number of adjectives and verbs is an indicator of a possible deceptive message	This number is higher than a standard message because deceptive messages are usually very expressive
Similar advertising from the same account promoting different women	It is a well-known sign of sex trade
Weight of women	Less than 100 pounds for very young girls
One account is promoting more than two different women	it is a well-known sign of sex trade

which is valuable information since the deceptive message usually is very expressive. This information was entered as a new feature to be considered in the classification. With this program, we also detected other features as 1) quantity of words, the maximum possible, 2) recognition of third-person speech, 3) same Twitter user talking about more than one victim, 4) number of defined hashtags present in the message, 5) mentions of girls from one country in messages from another origin country, 6) number of adjectives and verbs, 7) similar advertising with the same words promoting different women, 8) mentions of lightweight that could correspond to very young girls, and 9) one account promoting several women.

The analysis of URL links to detect if they are night club or massage therapy site is made manually from a report produced by a program with a list of URLs, and the other features are obtained processing the corpus with some Python programs. Then, the data is loaded in the input feature file in order to feed the classifier. Moreover, the corpus is collected using the hashtags presented in Table 1. The corpus was filtered in order to obtain the most relevant tweets according to the presence of more than one defined hashtag. This corpus was automatically classified in suspicious or not suspicious tweets, to prove the validity of the features (Table 3).

The automatic classification performance was evaluated against the annotated corpus that is considered the ground truth. Approximately 10% of the annotated original corpus was used as ground truth. This part of the corpus was randomly selected and labeled. The characteristics of the twitter account and the nature of their messages were analyzed. Additionally, the tweets were tagged as ''suspicious'' when a message comes from accounts that were closed by Twitter due to complaints of child pornography content.

A semi-supervised learning technique with Naïve Bayes and SVM algorithms was used in order to classify the tweets as ''suspicious'' or ''not suspicious'' of being related to sex trafficking. The performance of each classifier was evaluated based on average Precision (P), average Recall (R), and average F-Measure (F). Some algorithms were tested, and because SVM and Naïve Bayes presented a good performance and processing speed, they were chosen to classify the data using a semi-supervised approach.

10-fold cross-validation was used in order to evaluate the classifiers. This cross-validation divided the data randomly into ten sets. Each one was tested against the rest of the sets (9). The performance result was the average of all tests. As it is mentioned above, Precision, Recall, and F-Measure were used in order to evaluate classifiers' performance against evaluation using the ground truth established from the previous annotation. The performance for Naïve Bayes and SVM was presented in Tables 4 and 5, respectively.

TABLE 4. Naïve bayes performance.

	Actual Class	Suspected Tweet	Non-suspected
Predicted Class	Suspicious Tweet	25890	6570
	Not-suspicious	4389	18274

TABLE 5. SVM performance.

	Actual Class	Suspected Tweet	Non-suspected
Predicted Class	Suspicious Tweet	32453	3678
	Not-suspicious	3320	15672

As shown in Table 6, SVM has a better performance than Naïve Bayes, as is expected. SVM has a bit more Precision, Recall, and F -Measure than Naïve Bayes. Besides, both classifiers have a good performance measure.

The URLs from webs that had a relation with suspected tweets were collected, and then they were saved in an individual file called ''Black-List.'' Once this blacklist with suspected web sites was generated, we cleaned the data following the next steps:

- Eliminate duplicated results.
- Eliminate links that do not exist or unavailable web sites.
- Eliminate links that belong to withdraw or deactivated web sites.

The next phase uses this refined Blacklist as input data (Extraction and Processing).

V. IMAGEN EXTRACTION AND PROCESSING

This section describes the image extraction and processing to determine if the image is a person under 14 years old or not.

TABLE 6. Comparative of classifier performance.

Classifier	Precision	Recall	F-measures
Naïve Bayes	85,5%	0,80,1%	82,7%
SVM	90,7%	87,3%	88,9%

A. SCRAPING

There are web sites that generated massive concurrence of social network users where deliberately tweets that mention sexual services are linked. Some of these sites that repeatedly appear in suspected blacklist tweets are Infoscort, Punterking, and Shinagawaesthe. The web sites were scraped to obtain links to the images that are shown in their domains. For this purpose, the Html and CSS code of these suspected web sites are analyzed following two steps:

1) Massive download: The manual download of suspicious images is a time-consuming task because we have to do one by one, so this process is done by means of automatic techniques of scrapping. In order to scrape a web site, its link or URL is needed as an input. As a result, a plain text file was obtained, and all links to the images of web sites are saved. Then, the images were downloaded through an automatic process using the URLs of this file. It is essential to note that if the web sites require a paid subscription, it is not possible to download the internal images because the subscription restricts access to the subsequent pages freely. Therefore, scraping is limited to open access pages. Moreover, it is important to highlight that although security techniques have evolved, the use of digital certificate [23] does not guarantee that the activities of some web sites are legal. This kind of illegal businesses use seemingly harmless phrases such as ''caldo de pollo'' or ''club penguin'' to interchange the digital material (photos and videos). For instance, a threat in Twitter, written in Spanish, which exposes the use of this kind of words, is shown in https://twitter.com/MyLifeAsThunder/status/1173267980811194368.

2) Image Preprocessing: Once the set of possible images are downloaded, the data cleaning is done through the following process:

- Discard all not relevant pictures, for instance, icons, labels, among others.
- Discard duplicated images: For a better analysis, it is vital to avoid duplicated data.
- Discard images in greyscale because the classification models need an RGB color composition for the images (red, green, and blue).
- Discard images with unknown formats: the present project uses.jpg format for the classification model. If there are images in a different format, these are converted to.jpg in the case that the image has relevant information.
- Resize the images because the classification models need that all images have the same dimensions. Therefore, the image size was standardized to 150×150 pixels.

The image bank was formed by all images of suspicious web sites for testing data. On the other hand, for training data, we used the following open databases:

- Flickr (https://www.flickr.com).
- UKBench (https://archive.org/details/ukbench).
- Deep learning (http://deeplearning.net/datasets/).

B. GEOMETRIC FEATURE EXTRACTION

Once the data is ready, the feature extraction process is started, as is detailed in Figure 4.

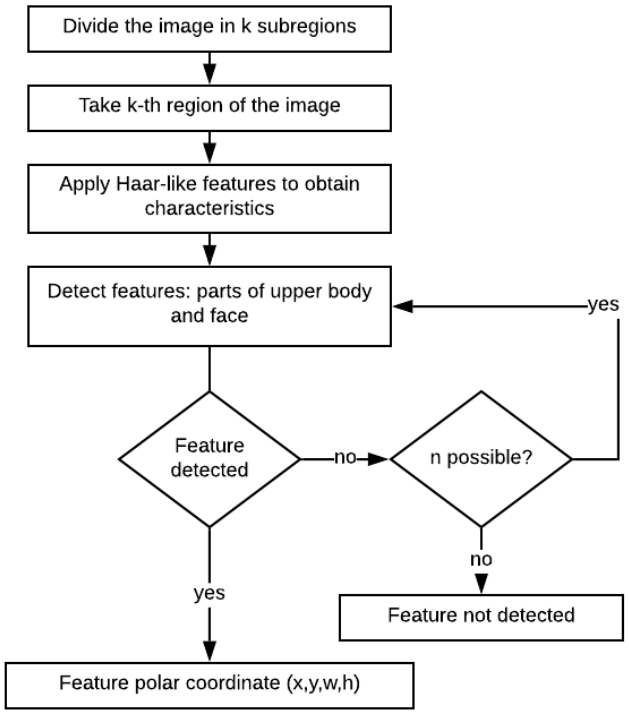

FIGURE 4. Feature extraction process.

Initially, the loaded image is divided into k different local regions; then, a HAAR cascade classifier was applied to each area. It uses the Viola-Jones algorithm to detect some image patterns analyzing geometric properties, such as Euclidian distance from the pixels. Besides, these patterns are handled as specific physic features (eyes, face, and upper body). In this paper, only the geometric features were taken into account in order to detect faces and upper body of the images collected from suspicious web sites. Three geometric features are needed (eyes, nose, and mouth) because these are joined in order to detect a face successfully. Figure 5 shows as different filters are joined in order to recognize a face in an image.

When images of minors in suspicious web sites are searched, most of these images have the face blurred, covered, or there is not a photo. For this reason, some classifiers are used in order to detect the upper body in each image (Figure 6).

In Figure 6, the integration of different Haar filters to detect the upper body of a person is shown. In some images, it is not possible to identify all individual geometric features because the picture has lousy quality or is blurred. The majority voting method is used to define if the face or the upper body detection is predominant. It takes into account the most

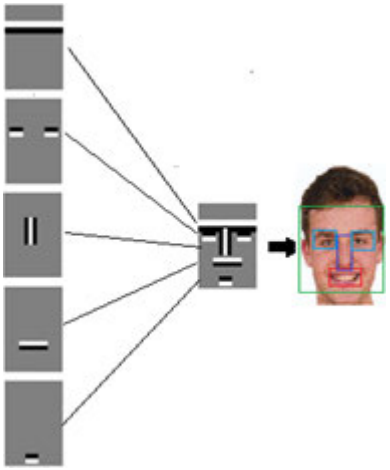

FIGURE 5. Face detection using Haar model.

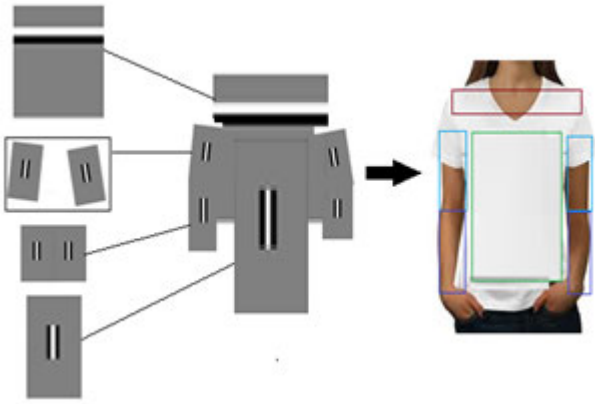

FIGURE 6. Upper body detection.

significant number of individual features detected for each case.

VI. IMAGEN CLASSIFICATION

Two algorithms were used in order to evaluate and compare the image classification process: i) SVM using Haar features and ii) CNN. Then for each algorithm, the classification by age group (under fourteen years old) or by gender (women or men) was carried out.

A. SUPPORT VECTOR MACHINE - SVM

SVM is a supervised machine learning model that can be used to predict two classes, like yes or no (binary classifier) [8]. In this work, a linear kernel was used because we have binary variables. Figure 7 shows the SVM prediction process.

The extracted features were used as input for a SVM algorithm. SVM classifier uses a linear kernel function to construct the boundary function $f(x)$, defined by (1), where b is the bias value, and y_i, α_i are the Lagrange optimization

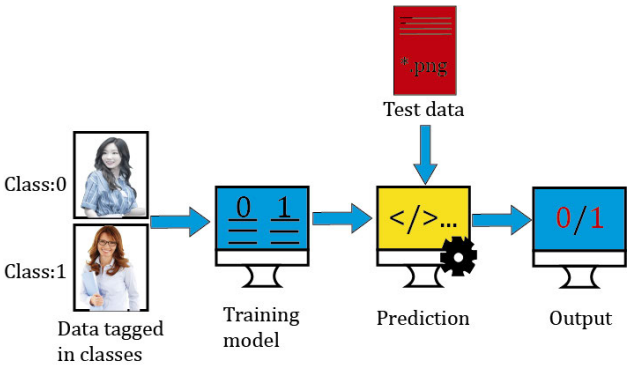

FIGURE 7. SVM model architecture.

parameters.

$$f(x) = \sum_{i=1}^{m} y_i \alpha_i K(x, x_i) \tag{1}$$

Furthermore, the SVM is a binary classifier. On the one hand, for age classification, two classes were defined: i) Class 0 for people under 14 years old and ii) Class 1 for people over 14 years old. On the other hand, for gender classification, two classes were defined too: i) class 0 corresponds to men and ii) Class 1 for women. A SVM classifier for each feature was applied, and the final decision was calculated by majority voting. Moreover, a Convolutional Neural Network with 16 layers was used in order to compare the classification performance between these two algorithms.

B. CONVOLUTIONAL NEURAL NETWORK - CNN

CNN is a supervised machine learning model that requires a big image dataset to build a classification model after some iterations. Many images and some iterations are needed to obtain the right prediction, so the main disadvantage of CNN is that it requires many computational resources, and it is a time-consuming model. In this work, for testing purposes, a computer i7-3770 with 8 GB RAM was used. For each iteration, 55 minutes were required, so for binary classification with ten iterations will take around 9 to 10 hours. The performance of CNN versus SVM using Haar-like features was compared. For this purpose, two leading indicators are taken into account: i) Accuracy to measure the number of correct predictions, and ii) Mean Square Error (MSE) that is defined in (2).

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (y_i - y_i') \tag{2}$$

where: y_i is the predicted value, y_i' is the observed value, and n is the number of data samples.

VII. TESTING RESULTS

In this section, the testing results for both SVM and CNN were described. Two experiments were carried out: 1) image classification using faces, and 2) image classification using

the upper body. Then, a comparison between these experiments was made. It is important to note that each test classifies by gender and age group.

A. SVM RESULTS

SVM algorithm can use Haar-like features, so it is possible to detect the age using the upper-body characteristics. This combination allows classifying based on the upper-body because, in many cases, we do not have face information (only upper-body). Moreover, the training process must categorize and label the data in two different classes correctly, and then its performance is measured through mean square error and accuracy indicator. For training data, a labeled dataset was created with 4096 images from public repositories. Most of them were frontal or slightly profile; however, side view pictures or others with faces, hats, caps, or glasses were also included. For testing data, 820 images obtained from the scraping of the suspicious sites were added. For this purpose, segments that contain the face and upper body were extracted, and they were classified using their Haar-like features.

1) EXPERIMENT 1: IMAGE CLASSIFICATION WITH FACE

Firstly, the classification was done using a set of images where the face can be detected and applying Haar filters. The confusion matrix of gender classification is shown in Table 7.

TABLE 7. Confusion matrix for gender(face).

	Man	Woman
Man	160	20
Woman	40	100

Then, the performance indicators MSE and accuracy were obtained using the analysis of the confusion matrix. As is shown in Figure 8, the accuracy and MSE are 81,2% and 3,5%, respectively. Then, the confusion matrix of the classification by age group is represented in Table 8.

The MSE and accuracy for age group classification are shown in Figure 9. The results presented show that our model has a classification accuracy of 80,6 % and a Mean Square Error of 3,7%. The obtained values in Tables 7 and 8 show that experiment 1 classifies not only by gender but also by age group.

2) EXPERIMENT 2: IMAGE CLASSIFICATION USING UPPER BODY FEATURES

The second experiment considers the result that HAAR filters detected as upper body features. Firstly, the data were classified by gender (men and women), as is shown in Table 9.

Based on the confusion matrix of the gender classification, the performance indicators of accuracy and MSE were calculated, as are shown in Figure 10.

As is shown in Figure 10, accuracy and SME values are 81,63% and 3,37%, respectively. Then, the images were classified by age group, as is shown in Table 10.

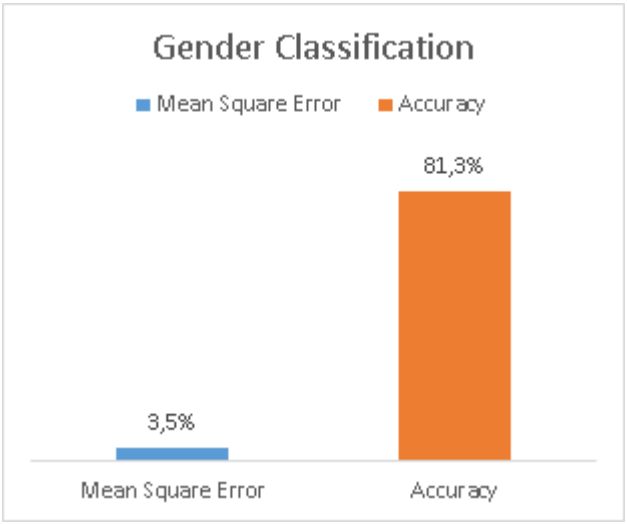

FIGURE 8. SVM gender classification (face).

TABLE 8. Confusion matrix for age group classification.

	Over14	Under14
Over14	160	20
Under14	40	100

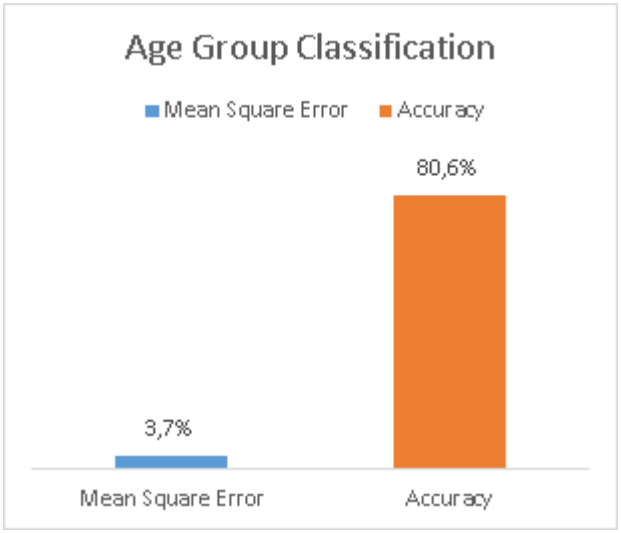

FIGURE 9. SVM age group classification (face).

TABLE 9. Confusion matrix of the gender classification (upper body).

	Man	Woman14
Man14	220	40
Woman	50	180

Based on this confusion matrix (Table 10), the accuracy and SME values were calculated 82,14% and 3,19% respectively, as is shown in Figure 11.

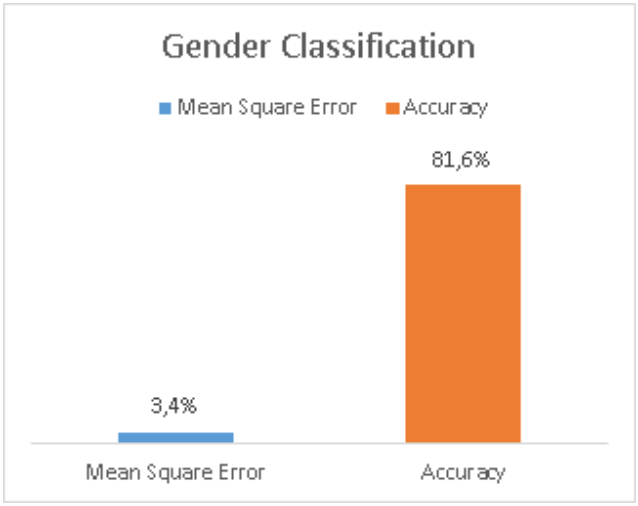

FIGURE 10. SVM gender classification (upper body).

TABLE 10. Confusion matrix age group classification.

	Over14	Under14
Over14	110	30
Under14	20	120

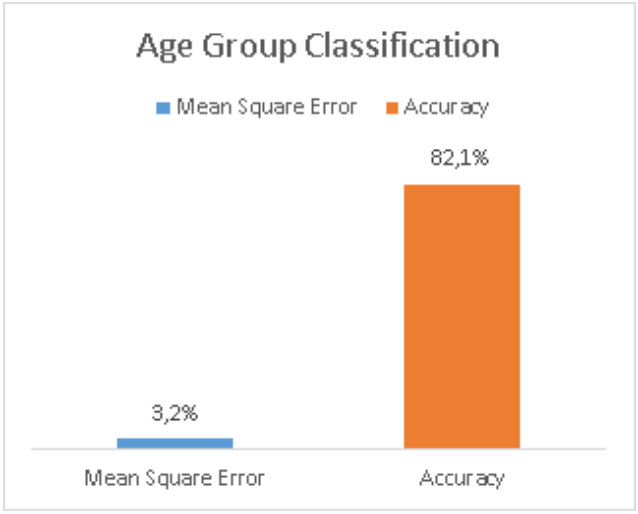

FIGURE 11. SVM age group classification (upper body).

3) SVM COMPARISON

The accuracy comparison between face classification and upper body classification results are shown in Figure 12.

Figure 12 shows the accuracy values for upper body experiment, 81,6 % for gender classification and 82,1% for age group. The accuracy values for face classification are 81,2% (gender) and 80,6% (age group). These results evidence that the classification accuracy using the upper body is higher than the accuracy using faces. To demonstrate the performance of the SVM model, Tables 11 and 12 show the obtained indicators: Accuracy, Precision, Recall, and F-measures.

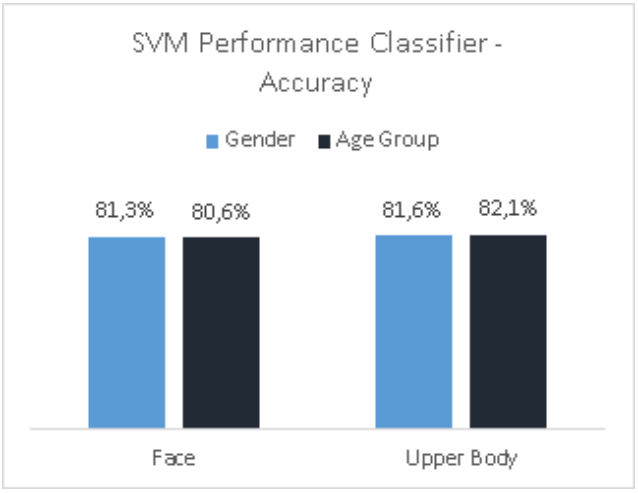

FIGURE 12. SVM accuracy face and upper body.

TABLE 11. Performance SVM (face).

Face	Accuracy	Mean Square Error	Precision	Recall	F-measures
Gender	81,3%	3,5%	83,3%	71,4%	76,9%
Age Group	80,0%	3,75%	80,9%	66,6%	73,1%

TABLE 12. Performance SVM (upper body).

Upper Body	Accuracy	Mean Square Error	Precision	Recall	F-measures
Gender	81,6%	18,4%	81,82%	78,26%	80,0%
Age Group	80,0%	17,9%	84,6%	85,71%	85,16%

B. CNN RESULTS

This model requires a little user intervention, and it is essential to choose an appropriate number of iterations to avoid overfitting in classification results [26]. The image dataset was divided into two groups: i) images with faces and ii) pictures with the upper body.

1) EXPERIMENT 1 IMAGE CLASSIFICATION WITH FACES

The classification is done based on the face features for both categorizations by gender and by age group. Figure 13 shows the results of gender classification.

As is shown in Figure 13, the accuracy values of the first and tenth iteration are 85,5% and 98,5% respectively, and the last iteration has an MSE value of 1,2%. The accuracy value has grown progressively during all iterations. Then, the classification by age group is done, and its result is shown in Figure 14.

The accuracy values are 80,6% and 97,3% in the first and last iteration, respectively, and the MSE value of the last iteration is 97,3%. These results evidence that the classification accuracy using faces has a good result for both, gender and group of age.

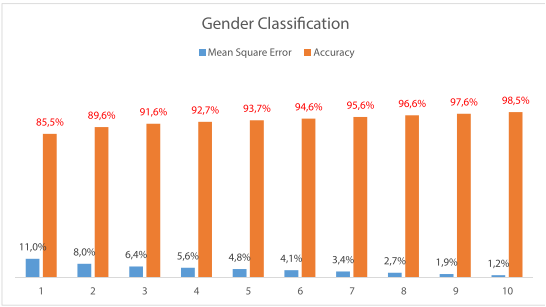

FIGURE 13. CNN gender classification (face).

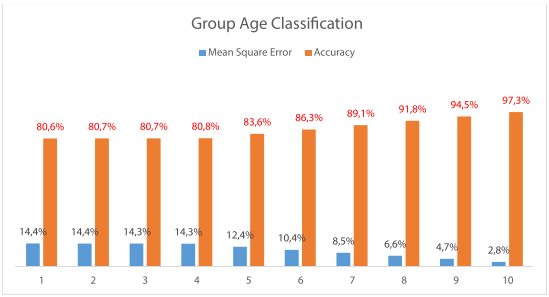

FIGURE 14. CNN age group classification (face).

2) EXPERIMENT 2 IMAGE CLASSIFICATION WITH THE UPPER BODY

In this experiment, the upper body was used in order to classify images where the faces are blurred, covered, or it is not showing. It happens when the face of minors are pixelated to hide their identity. The result of this experiment for gender classification is shown in Figure 15.

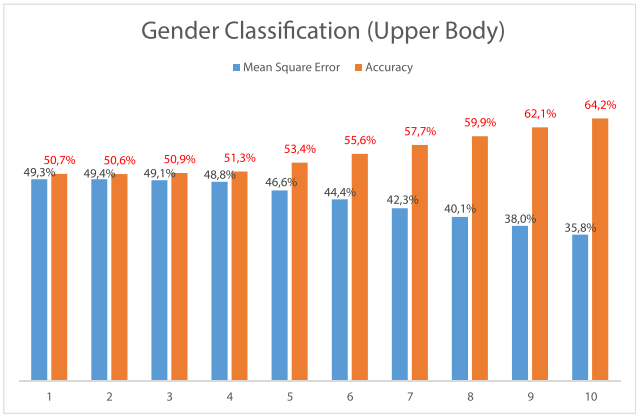

FIGURE 15. CNN gender classification (upper body).

In the gender classification, the accuracy values for the first and the tenth iterations are 50,7% and 64,2%, respectively. Moreover, the MSE value of the last iteration is 3,58%. In the same way, the age group classification using upper body characteristics is depicted in Figure 16.

The accuracy value in the first iteration is 51,1%. This value remains constant during the tenth iteration for this kind

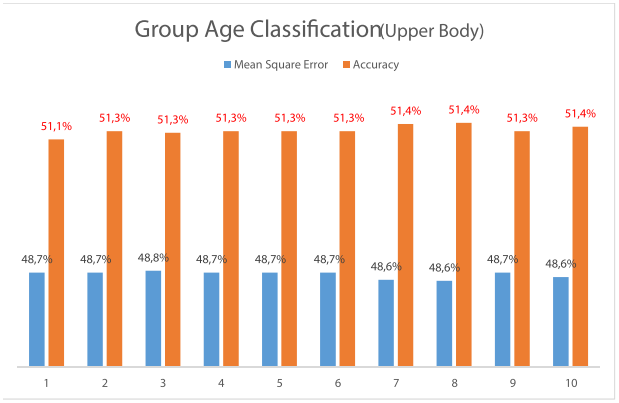

FIGURE 16. CNN age group classification (upper body).

of classification. In fact, the accuracy and SME values for the last iteration are 51,4% and 48,6%, respectively, so this experiment has poor performance because the test classifies correctly one out of 2 cases. This result is similar to select a random image by the toss of a coin. The accuracy values obtained in the last iteration for both gender and age group were 64,2% and 51,4%, respectively (Figure 15 and Figure 16). This experiment takes only upper body features into account, so this model has a poor performance. This outcome happens because CNN has a reasonable prediction rate classifying faces but not the upper torso, like the SVM algorithm.

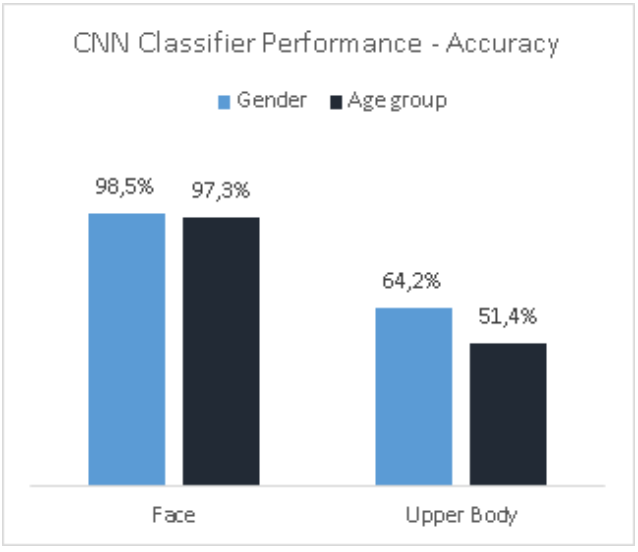

FIGURE 17. CNN Face and UpperBody.

3) CNN COMPARISON

The comparison between experiment 1 and 2 using CNN is shown in Figure 17. This figure shows that under any classification criteria (gender or age group), the CNN algorithm has a better performance when it analyses images with faces (facial features). However, CNN is not recommended when images

only have upper-body features (Experiment 2). For this experiment, the accuracy values for gender and age group are 64,2% and 51,3%, respectively. Consequently, if the images are analyzed, taking into account only upper body features, the SVM is a better option than the CNN model.

TABLE 13. CNN performance (face).

FACE	*Accuracy*	*Mean Square Error*
Gender	98,5%	1,2%
Age Group	97,3%	2,8%

TABLE 14. CNN performance (upper body).

UPPER BODY	*Accuracy*	*Mean Square Error*
Gender	64,2%	35,8%
Age Group	51,4%	48,6%

It is important to note that the main problem using only face features is that the images can be blurred, or they do not exist, especially in web sites that promote minor trafficking. As a result, the classification using CNN, under this condition, does not provide sufficient accuracy. A summary of accuracy and SME values for CNN experiments are shown in Tables 13 and 14.

C. COMPARATIVE ANALYSIS

In this section, a comparison between SVM and CNN performance results, taking into account accuracy value, is presented. Firstly, the results obtained when face features can be detected from an image are analyzed. The accuracy values for both algorithms are shown in Figure 18.

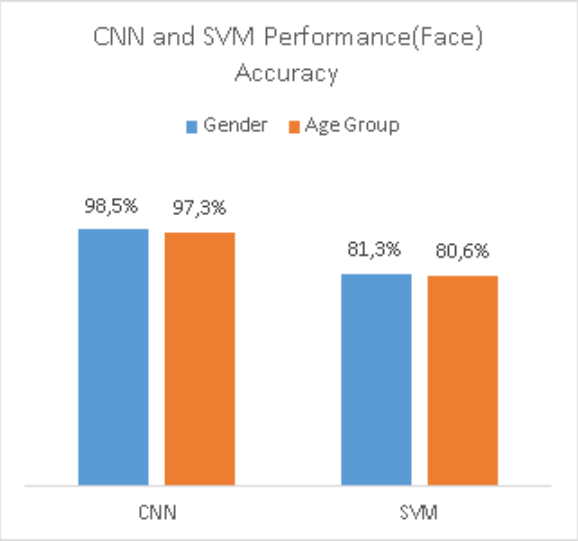

FIGURE 18. Performance classifier models (face).

The results presented (Figure 18) shows that the accuracy of CNN is higher than the accuracy of the SVM model for both gender and age group classification. On the one hand,

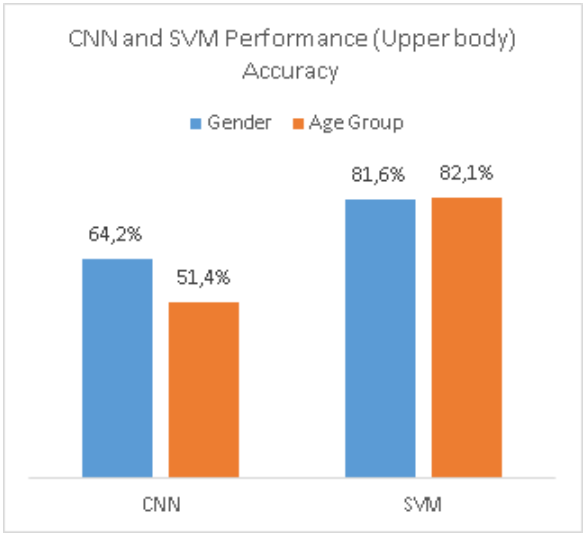

FIGURE 19. Performance classifier models (upper body).

Image Number	Gender	Group Age	Image
1	Man	Under 14 years old	
2	Woman	Over 14 years old	
3	Man	Over 14 years old	
4	Woman	Under 14 years old	
5	Woman	Over 14 years old	
6	Woman	Under 14 years old	
7	Woman	Under 14 years old	
8	Woman	Over 14 years old	
9	Man	Under 14 years old	
10	Man	Over 14 years old	

FIGURE 20. Images classified.

the accuracy values of CNN for gender and age group classification are 98,5% and 97,2%, respectively. On the other hand, the accuracy values of SVM for gender and age group classification are 81,2% and 80,65%, respectively. These two

models have good accuracy values, so both can be used to classify images with facial features. Then, the results of experiment 2 were compared, that it consists of the image classification taking into account upper body features, as is shown in Figure 19.

The results presented in Figure 19 evidence that the classification accuracy of SVM is higher than the accuracy of the CNN model when only upper body features in the images are analyzed. On the one hand, the accuracy values of SVM for gender and age group classification are 81,6% and 82,1%, respectively. On the other hand, the accuracy values of CNN for gender and age group classification are 64,2% and 51,3% respectively.

These results evidence that the SVM accuracy in experiment 1 (face) is similar to experiment 2 (upper body) not only for gender classification but also for age group classification. Moreover, the CNN performance is lower than not only SVM performance but also CNN results obtained in experiment 1. Therefore, the best option to detect a possible case of human trafficking of minors is using the SVM algorithm. As mentioned above, trafficking web sites usually use blurred or pixelated images, or there are no facial features in the image. Figure 20 shows some images classified in this research.

The images with blurred or pixelated faces were classified using the SVM algorithm. Moreover, CNN is commonly used in face detection.

VIII. CONCLUSIONS AND FUTURE WORK

Face recognition algorithms and machine learning models have been improved during the last years. For example, in the ILSVRC competition, an accuracy value of 90% +− 5% was obtained. In these conditions, machine learning recognition can be similar to visual object recognition used by human beings. Many factors have a direct impact on image recognition, such as size, color, opacity, resolution, kind of image format, among others. Therefore, the results of image recognition and classification depend on the dataset quality.

In this work, we probed that satisfactory performance can be obtained using just geometric features of the torso and not only facial characteristics. For this paper, Haar filters combined with an SVM classifier were used for the extraction process of features, and then we classified the age group and gender with an SVM classifier. The obtained results were compared with the outcomes of a CNN algorithm.

SVM is a model widely accepted, and in this work, we obtained a classification accuracy higher than 80% for both experiments (face and upper body), not only for gender classification but also for age group classification. In this paper, our main contribution is the image classification based on the upper body to predict the age group to detect human trafficking.

To the best of our knowledge, this work is the first approach related to image classification without facial features but just the upper-body geometric characteristics. Currently, there is no similar research that takes into account only the upper body features of minors. Thus, the results of this paper can be applied to human trafficking, disappearance, kidnapping, among others. Moreover, the obtained information can be used by the police or other security institutions.

Finally, future work includes: 1) the study of some characteristics related to ethnic and racial features, 2) to extend the proposal to extract geometric features of the entire body, another kind of images, or inclusive videos in different formats, 3) detection of medical issues by means the analysis of features extracted from torso images, legs, back, among other characteristics, and 4) the use of other algorithms or the applicability in other networks like Instagram.

ACKNOWLEDGMENT

The authors would like to thank the Escuela Politécnica Nacional for developing Seed Research Project PIS-17-10 ''Data Mining, Feature Vector Extraction, and Modeling with Pattern Recognition and Machine Learning to Detect Scenarios related to the Crime of Human Trafficking''.

REFERENCES

[1] B. Bangerter, S. Talwar, R. Arefi, and K. Stewart, ''Networks and devices for the 5G era,'' *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 90–96, Feb. 2014.
[2] F. Laczko, ''Data and research on human trafficking,'' *Int. Migration*, vol. 43, nos. 1–2, pp. 5–16, Jan. 2005.
[3] M. Lee, ''Human trafficking and border control in the global south,'' in *The Borders of Punishment: Migration, Citizenship, and Social Exclusion*. Oxford, U.K.: Oxford Univ. Press, 2013, pp. 128–149.
[4] E. Cockbain and E. R. Kleemans, ''Innovations in empirical research into human trafficking: Introduction to the special edition,'' *Crime, Law Social Change*, vol. 72, no. 1, pp. 1–7, Jul. 2019.
[5] R. Weitzer, ''Human trafficking and contemporary slavery,'' *Annu. Rev. Sociol.*, vol. 41, pp. 223–242, Aug. 2015.
[6] T. S. Portal. (2018). *Twitter: Number of Monthly Active Users 2010-2018*. [Online]. Available: https://www.statista.com
[7] M. R. Candes, ''The victims of trafficking and violence protection act of 2000: Will it become the thirteenth amendment of the twenty-first century,'' *U. Miami Inter-Amer. L. Rev.*, vol. 32, p. 571, Jun. 2001.
[8] D. Hughes, *Wilberforce Can be Free Again: Protecting Trafficking Victims*. New York, NY, USA: National Review, 2008.
[9] A. Sultan, ''Countering crime trafcking in persons smuggling migrants Ethiopia: The Law practice,'' Ph.D. dissertation, School Law, Addis Ababa Univ., Ababa, Ethiopia, 2018, pp. 1–72.
[10] M. Tsikerdekis and S. Zeadally, ''Online deception in social media,'' *Commun. ACM*, vol. 57, no. 9, pp. 72–80, Sep. 2014.
[11] A. Vishwanath, ''Diffusion of deception in social media: Social contagion effects and its antecedents,'' *Inf. Syst. Frontiers*, vol. 17, no. 6, pp. 1353–1367, Jun. 2014.
[12] E. Tong, A. Zadeh, C. Jones, and L.-P. Morency, ''Combating human trafficking with deep multimodal models,'' 2017, *arXiv:1705.02735*. [Online]. Available: http://arxiv.org/abs/1705.02735
[13] J. V. D. Wolfshaar, M. F. Karaaba, and M. A. Wiering, ''Deep convolutional neural networks and support vector machines for gender recognition,'' in *Proc. IEEE Symp. Ser. Comput. Intell.*, Dec. 2015, pp. 188–195.
[14] M. Hernandez-Alvarez,''Detection of possible human trafficking in Twitter,'' in *Proc. Int. Conf. Inf. Syst. Softw. Technol. (ICIST)*, Nov. 2019, pp. 187–191.
[15] H. Alvari, P. Shakarian, and J. E. K. Snyder, ''A non-parametric learning approach to identify online human trafficking,'' in *Proc. IEEE Conf. Intell. Secur. Informat. (ISI)*, Sep. 2016, pp. 133–138.
[16] M. M. Dehshibi and A. Bastanfard, ''A new algorithm for age recognition from facial images,'' *Signal Process.*, vol. 90, no. 8, pp. 2431–2444, Aug. 2010.
[17] F. Salvetti, ''Detecting deception in text: A corpus-driven approach,'' Ph.D. dissertation, Comput. Sci. Graduate, Univ. Colorado Boulder, Boulder, CO, USA, 2012, pp. 1–206.
[18] S. Sarkar, ''Use of technology in human trafficking networks and sexual exploitation: A cross-sectional multi-country study,'' *Trans. Social Rev.*, vol. 5, no. 1, pp. 55–68, Jan. 2015.

[19] M. Ibanez and D. D. Suthers, "Detection of domestic human trafficking indicators and movement trends using content available on open Internet sources," in *Proc. 47th Hawaii Int. Conf. Syst. Sci.*, Jan. 2014, pp. 1556–1565.
[20] G. Tyldum, "Limitations in research on human trafficking," *Int. Migration*, vol. 48, no. 5, pp. 1–13, 2010.
[21] J. Quirk, *The Anti-Slavery Project: From Slave Trade to Human Trafficking*. Philadelphia, PA, USA: University of Pennsylvania Press, 2011.
[22] E. Kennedy, "Predictive patterns of sex traffcking online," in *Dietrich College Honors Theses*. Pittsburgh, PA, USA: Carnegie Mellon Univ., 2012, pp. 1–45.
[23] I. Alegria, N. Aranberri, P. R. Comas, V. Fresno, P. Gamallo, L. Padró, I. San Vicente, J. Turmo, and A. Zubiaga, "TweetNorm: A benchmark for lexical normalization of spanish tweets," *Lang. Resour. Eval.*, vol. 49, no. 4, pp. 883–905, Aug. 2015.
[24] A. Mbaziira and J. Jones, "A text-based deception detection model for cybercrime," in *Proc. Int. Conf. Technol. Manag*, Jul. 2016, pp. 1–8.
[25] E. Cockbain and K. Olver, "Child trafficking: Characteristics, complexities, and challenges," in *Child Abuse Neglect*. Amsterdam, The Netherlands: Elsevier, 2019, pp. 95–116.
[26] R. Janani. (2018). *Extracting Structured Data From the Web Usingscrapy Pluralsight*. [Online]. Available: https://app.pluralsight.com

SERGIO L. GRANIZO was born in Quito, Ecuador, in 1988. He received the degree in computer engineering from the Central University of Ecuador and the master's degree in software from the National Polytechnic School. He is currently works as a Software Developer and a Software Analyst at Kruger Corporation and LLACSAA Company. His professional experiences include participation in several machine learning, data mining, image processing, and Python programming projects, in addition to consulting and IT in several Ecuadorian companies.

ÁNGEL LEONARDO VALDIVIESO CARAGUAY received the electronics and telecommunication engineering degree from the Escuela Politécnica Nacional, Quito, Ecuador, in 2009, the M.Sc. degree in information technology from the Hochschule Mannheim University of Applied Sciences, Mannheim, Germany, in 2012, and the Ph.D. degree in computer science from the Universidad Complutense de Madrid, Madrid, Spain, in 2017. He was a Visiting Researcher with the University of Aveiro and Altice Labs, Aveiro, Portugal. He is currently a Lecturer with the Escuela Politécnica Nacional. His professional experience includes research projects with EU H2020 (Selfnet 5G). He is the author of scientific articles on international journals and served on the technical program committees for several leading international conferences. His research interests include the Internet of Things (IoT), SDN, NFV, mobile networks, and information security.

LORENA ISABEL BARONA LÓPEZ received the B.S. degree in electronic and information network engineering from Escuela Politécnica Nacional (EPN), Ecuador, in 2010, and the M.Sc. degree in telecommunications engineering from the Universidad Politécnica de Madrid, Spain, in 2013, and the Ph.D. degree in computer engineering from the Universidad Complutense de Madrid, Spain, in 2017. She was a Visiting Researcher with the University of Aveiro and Altice Labs, Aveiro, Portugal. She is currently a Professor with the Informatics and Computer Science Department, EPN. Her professional experience includes research projects EU H2020. She is author and coauthor of several articles. Her research interests include the Internet of Things (IoT), mobile networks, software defined networking, network function virtualization, and cybersecurity.

MYRIAM HERNÁNDEZ-ÁLVAREZ received the electronics and telecommunication engineering degree from the Escuela Politécnica Nacional, Quito, Ecuador, the M.Sc. degree in computer science from Ohio University, Athens, OH, USA, and the Ph.D. degree in computer applications from the University of Alicante, Spain. She was the Dean of the System Engineering School, from 2014 to 2019, and the Director of the Doctoral Program in informatics of the Escuela Politécnica Nacional, from 2016 to 2019. Her researches include the areas of machine learning, artificial intelligence, computer vision, cognitive security, and natural language processing.