machine learning & knowledge extraction

MDPI

Article

A Multi-Input Machine Learning Approach to Classifying Sex Trafficking from Online Escort Advertisements

Lucia Summers [1], Alyssa N. Shallenberger [1], John Cruz [2] and Lawrence V. Fulton [3,*]

[1] School of Criminal Justice and Criminology, Texas State University, 601 University Drive, San Marcos, TX 78666, USA
[2] Department of Health Administration, Texas State University, San Marcos, TX 78666, USA
[3] Applied Analytics & Economics Programs, Boston College, Chestnut Hill, MA 02467, USA
* Correspondence: lawrence.fulton@bc.edu; Tel.: +1-210-837-9977

Abstract: Sex trafficking victims are often advertised through online escort sites. These ads can be publicly accessed, but law enforcement lacks the resources to comb through hundreds of ads to identify those that may feature sex-trafficked individuals. The purpose of this study was to implement and test multi-input, deep learning (DL) binary classification models to predict the probability of an online escort ad being associated with sex trafficking (ST) activity and aid in the detection and investigation of ST. Data from 12,350 scraped and classified ads were split into training and test sets (80% and 20%, respectively). Multi-input models that included recurrent neural networks (RNN) for text classification, convolutional neural networks (CNN, specifically EfficientNetB6 or ENET) for image/emoji classification, and neural networks (NN) for feature classification were trained and used to classify the 20% test set. The best-performing DL model included text and imagery inputs, resulting in an accuracy of 0.82 and an F1 score of 0.70. More importantly, the best classifier (RNN + ENET) correctly identified 14 of 14 sites that had classification probability estimates of 0.845 or greater (1.0 precision); precision was 96% for the multi-input model (NN + RNN + ENET) when only the ads associated with the highest positive classification probabilities (>0.90) were considered (n = 202 ads). The models developed could be productionalized and piloted with criminal investigators, as they could potentially increase their efficiency in identifying potential ST victims.

Keywords: sex trafficking; deep learning; multi-input models; natural language processing; (convolutional) neural networks

Citation: Summers, L.; Shallenberger, A.N.; Cruz, J.; Fulton, L.V. A Multi-Input Machine Learning Approach to Classifying Sex Trafficking from Online Escort Advertisements. *Mach. Learn. Knowl. Extr.* **2023**, *5*, 460–472. https://doi.org/10.3390/make5020028

Academic Editor: Nicholas Ampazis

Received: 4 March 2023
Revised: 26 April 2023
Accepted: 4 May 2023
Published: 10 May 2023

1. Introduction

Trafficking in persons is one of the most harmful criminal industries internationally. Its prevalence continues to rise each year and it is currently identified as the second-most profitable illegal trade, after drug trafficking [1]. According to the U.S. Department of State's 2022 Trafficking in Persons Report [2], 1111 federal (or joint federal-local/state) investigations of human trafficking were opened during the fiscal year 2021, with the Department of Justice initiating prosecution in 228 cases, the majority of which (221) concerned sex—as opposed to labor—trafficking. At the local/state level, 2203 human trafficking offenses were reported by participating jurisdictions (U.S. Department of State 2022). These figures, however, are thought to grossly underestimate the true extent of the problem. For instance, 11,500 cases of human trafficking were reported to the National Human Trafficking Hotline during 2019 [3]. These calls led to the identification of 22,326 victims and survivors, of whom 14,597 (65%) had been sex trafficked, with an additional 1048 (5%) having been subjected to both sex and labor trafficking.

Federal law defines sex trafficking as "the recruitment, harboring, transporting, provision, obtaining, patronizing or soliciting of a person for the purposes of a commercial sex act, in which the commercial sex act is induced, through the use of force, fraud, or coercion, or in which the person induced to perform such an act has not attained 18 years

of age". This means that, if the victim is under the age of 18, force, fraud, or coercion are not required to prove trafficking. Human trafficking is often conflated with human smuggling, but smuggling requires the movement of an individual across international borders. "Trafficking" in this sense refers to the illegal commodification of an individual and their entrance into the stream of commerce, akin to weapons or drug trafficking, as opposed to the movement of a commodity.

Sex trafficking (ST) can be promulgated through information and communications technologies (ICT) such as the Internet and, more specifically, online escort advertisements [4–6]. These ads are typically very brief and contain a provocative photo, a description of the seller/victim, and language to describe or indicate the advertisement is for commercial sex [7]. However, thinly coded terms and/or other characteristics may be indicative of commercial exploitation. For example, a phrase such as "new in town" may indicate victims being moved around (often an indication of sex trafficking), a crown emoji or image may signify services managed by a pimp, and emojis displaying a growing heart, a cherry, or a cherry blossom emoji may signify a minor in the ad [8,9]. The ads usually appear to be posted by the individual in the ad, but traffickers often post and pay for the ads themselves [7].

While easily accessible to both the public and law enforcement, the sheer volume of ads, the frequency with which the posting location changes, along with the use of obfuscation tactics by those posting and hosting the advertisements, make it difficult for law enforcement to identify, react, and respond [8]. Traffickers and buyers, along with the technology used itself, evolve at a rapid rate, and this demands that law enforcement, prosecutors, and legislatures come up with creative responses to combat the advances in technology (which is inherent to the covert nature of the crime itself). Thus, these stakeholders require accurate, productionalized, and cost-free methods for classifying websites likely engaged in illicit ST to target enforcement [10].

Machine learning (ML) has shown promise in the automated classification of websites—both escort and third-party review sites—that facilitate sex trafficking. Tong and colleagues (2017) were one of the first to apply these classification-supervised learning methods to online escort ads, with the goal of identifying likely sex trafficking activity [11]. They collaborated with law enforcement officials to annotate more than 10,000 escort ads from Backpage (a dominant escort advertisement site that is no longer active) along a seven-point Likert scale to indicate their likely involvement in sex trafficking activity. Based on these, the authors developed a deep model that considered the text, emojis, and images in the ads. This model outperformed all baselines considered (e.g., keywords, bag of words, random forest, logistic regression, and linear support vector machine, or SVM, models) in identifying those ads suspected of being associated with sex trafficking, highlighting the value of the methodology.

Since then, other studies have followed that have employed the Trafficking-10k and other datasets. For example, Alvari et al. [12] also used law enforcement experts to manually label a large portion of crawled data from Backpage. They extended the existing Laplacian SVM model by adding a regularization term to the optimization equation and ultimately reported that their approach had the highest F1 scores (91). In a separate study, an ordinal regression neural network approach yielded a model that outperformed previous conventional regression models [13].

Esfahani and colleagues [14] developed a centralized, semi-automatic tool that utilized natural language processing (NLP) techniques, among others, to identify trafficking ads—the classifiers developed had a significantly better performance than any single feature/variable set alone. The full model utilizing the full feature set (under U-BERT) provided 26% recall improvement over the three individual ones (e.g., 69% vs. 28–42%; recall, or sensitivity, is the model's true positive rate) when precision (positive predictive value) was set to 85%. Zhu et al. [15], using the Trafficking-10k dataset, developed a language selection model and showed improvement against Tong et al.'s human trafficking deep network (HTDN) model [11], with a precision of 66.2% and recall of 73.4%. The application

of the model went further than prior research by using the model to identify unknown trafficking organizations and assign a risk score. However, this model only examined the text of the ads (and not the emojis or images).

Convolutional neural networks (CNN) have become a commonplace approach for image (and other) analyses. Granizo et al. compared SVM and CNN models to estimate the gender and age of individuals on a known public repository where sex services were advertised [16]. The training data set consisted of labeled images (n = 4096 posts). Accuracy rates for age classification were 80.6% (SVM) vs. 97.3% (CNN) for faces and 82.1% (SVM) vs. 51.4% (CNN) for upper body images.

More recently, Wiriyakun and Kurutach [17,18] utilized a feature selection approach and compared three updated ML models against the original work of Tong et al. [11] with the Trafficking-10k dataset, namely random forest, logistic regression, and linear SVM. These models significantly outperformed Tong et al.'s [11] bag of words approach, with F1 scores of 63.3%, 64.8%, and 61.3%, respectively, as compared to 24.5%. These results are relevant because the authors dichotomized the labels in the Trafficking-10k dataset, which is the approach adopted in the present study. Table 1 provides a summary of the results from related work.

Table 1. Summary of results from related work.

Source	Data	Analysis	Findings
Tong et al. (2017) [11]	*Trafficking-10k* dataset (12,350 escort ads from *Backpage* rated by subject experts on 7-point scale)	Human trafficking deep network (HTDN), a deep multimodal model applied to both text and images	HTDN achieved 0.80 accuracy, 0.71 precision, and an F1 score of 0.67, and outperformed several baselines with different inputs.
Alvari et al. (2017) [12]	20,000+ scraped Backpage escort ads; final unfiltered sample was 3543 (of these, 200 were rated as likely associated with ST by expert)	Semi-supervised Laplacian support vector machine (SVM) of unfiltered online escort ads (ads were filtered out if none of 12 ST indicators were present)	Extended S^3VM—R model yielded better precision (0.91 vs. 0.86 for positive cases, 0.92 vs. 0.89 for negative cases), recall (0.91 vs. 0.88 for positive cases, 0.93 vs. 0.90 for negative cases), and F1 scores (0.91 vs. 0.87 for positive cases, 0.92 vs. 0.88 for negative cases) than the original Laplacian SVM.
Wang et al. (2019) [13]	*Trafficking-10k* dataset	Ordinal regression neural network	Accuracy of 0.82, as compared to HTDN's 0.80. No recall or F1 scores reported.
Esfahani et al. (2019) [14]	10,000 online escort ads from *Backpage* and *Eroticmugshots*, with 4385 ads flagged as positive by cross-referencing with list of ST-related phone numbers	Deep learning latent Dirichlet allocation (LDA) model with average word vector (AWV) and bidirectional encoder representation from transformers (BERT)	The LDA+AWD+BERT model outperformed simpler variations of the model (e.g., recall was 0.69 vs. 0.28–0.42) at 85% precision.
Zhu et al. (2019) [15]	*Trafficking-10k* dataset	Language selection model	Precision was 0.66 and recall 0.73.
Granizo et al. (2020) [16]	*Twitter* posts with hashtags related to minors and potential illicit (sexual) activity (n = 4096)	SVM vs. convolutional neural networks (CNN)	Detection of age in posts advertising sex yielded accuracy of 80.6% (SVM) vs. 97.3% (CNN) for faces, and 82.1% (SVM) vs. 51.4% (CNN) for upper body images.
Wiriyakun and Kurutach (2021, 2022) [17]	*Trafficking-10k* dataset with outcome in binary form (collapsed original 7 labels)	Feature selection approach	Reported 0.77 accuracy in best-performing model and F1 scores of 63.3% for random forest model, 64.8% logistic regression, and 61.3% linear SVM.

As ML methods advance and classification modeling improves, the need to determine the extent to which these higher-performance models apply in different contexts arises. This study attempts to improve on earlier methods via the provision of a high-performance deep learning model that includes text and imagery inputs. This has the potential to alleviate resource constraints placed on law enforcement by creating a model that can identify sex trafficking-related online escort ads with both high accuracy and precision, thus maximizing the efficiency of criminal investigators.

2. Materials and Methods

2.1. Data, Software, and Hardware

For this study, 12,350 escort ads were obtained from a user agreement with Marinus Analytics. This agreement provided access to the Trafficking-10k dataset [11]. The data included an identification, a classification label (more on this later), a title text, and a body text. The title and body were further concatenated for use. Observations included emojis (images, e.g., 😊) and emoticons (text viewed as images, e.g., :/). Figure 1 shows the first five observations of the original dataset.

	id	label	title	body	both
0	2632347	4	Ms Chrissy - 22	Tall Sexy Slim 100% real Reas...	Ms Chrissy - 22 Tall Sexy Slim 10...
1	11645643	1	Everybody loves a Filipina Girl - 29	Hi there Boys it's Miss Sweet 1OO% Filipina gi...	Everybody loves a Filipina Girl - 29 Hi there ...
2	15199346	4	NEW Sexy Asian girl⛔⛔⛔⛔⛔SUPER HOOOT⛔⛔⛔...	Hello Gentlemen I am Mimi Asian Girl <br...	NEW Sexy Asian girl⛔⛔⛔⛔⛔SUPER HOOOT⛔⛔⛔...
3	14477454	0	Freaky . TS 🍆🍑 TOP AND BOTTOM (9INCH) fun pack...	• So Treat Yourself to Heaven on Earth... 🌎 YOU...	Freaky . TS 🍆🍑 TOP AND BOTTOM (9INCH) fun pack...
4	11266475	2	❤Kinky Kaylani 👑 Rear Access &VIP visits 👑 Great R...	Aren't you tired of the old bait and switch or...	❤Kinky Kaylani 👑 Rear Access &VIP visits 👑 Great R...

Figure 1. The first five observations of the data.

The Anaconda distribution of Python 3.8 [19] was used for all models. The TensorFlow library [20] provided support for the machine learning algorithms. Programming code is freely available in Supplementary Materials. For this preliminary study, processing was performed on a high-end computer with a 14-core Intel i9-12900K Central Processing Unit (CPU) operating at 2.5 GHz, 64 GB of random access memory (RAM), and a single NVIDIA GeForce RTX 3080 Super Graphical Processing Unit (GPU).

2.2. Training, Validation, & Test Sets

Data were split randomly into an 80% training set with 9880 observations and a 20% test set with 2470 observations. The training set was further split for hyperparameter tuning, with 80% (7904) used for training and 20% (1976) used for validation. After estimating optimal hyperparameters, the entire training set was used for model fitting.

2.3. Image Creation

Marinus Analytics was unable to provide the original images used in the initial study. To investigate the utility of multi-input models that include image components, Unicode emojis from the concatenated title and body were converted to color images for use in investigating CNNs using spaCy [21]. The extracted emojis for each observation were arranged into a square grid just large enough to contain each sequential image. For example, if there were 100 emojis in the combined text, then the grid was 10 by 10. This grid was then converted to a 224 by 224 image using OpenCV [22] and the Segoe UI Emoji True Type Font. Figure 2 shows the image generation for a single observation.

Figure 2. Example of image generation.

2.4. Text Preprocessing

The title and body text were concatenated for use in natural language processing. Text was converted to lowercase, numbers were converted to text, and additional. html tags remaining in the dataset were removed using the "re" library. Emoticons and emojis were mapped to text using the "emot" library. Punctuation was identified by the "string" library in Python and removed. Common English stop words were removed, and text was lemmatized (to account for parts of speech), tokenized (to convert from text to numbers), and padded to the maximum sentence length using the Natural Language Tool Kit (NLTK) [23]. Figure 3 shows four examples of pre-processed vs. original images before lemmatization, tokenization, and padding.

	id	label	title	body	both
0	2632347	4	ms chrissy twenty-two	tall sexy slim one hundred real reasonable cle...	Ms Chrissy - 22 Tall Sexy Slim 10...
1	11645643	1	everybody love filipina girl twenty-nine	hi boys s miss sweet oneoo filipina girl kai l...	Everybody loves a Filipina Girl - 29 Hi there ...
2	15199346	4	new sexy asian girl noentry noentry noentr...	hello gentleman fleurdelis fleurdelis mimi asi...	NEW Sexy Asian girl ⛔ ⛔ ⛔ ⛔ ⛔ SUPER HOOOT ⛔ ⛔ ⛔ ...
3	14477454	0	freaky t eggplant peach top bottom nineinch fu...	• treat heaven earth globeshowingEuropeAfrica ...	Freaky . TS 🍆 🍑 TOP AND BOTTOM (9INCH) fun pack...

Figure 3. Example of title and body conversion.

In Figure 3, "both" reflects the original concatenated title and body. In observation 0, there are numbers and. html tags that are removed (along with punctuation and capitalization) in the transformed columns "title" and "body". In observations 3 and 4, the emojis are converted to text. Doing so supports text-based modeling.

2.5. Feature Creation

Eight features were extracted from the data. Title and body emojis were counted and used as separate variables. Derivations of the word "sex" were counted separately for the body and text as well. Proportions of emojis and derivations of the word "sex" were also calculated for the title and the body. These eight variables were included in a four-layer neural network model (64 neurons, 50% dropout, 32 neurons, and 20% dropout with linear activation functions) after min-max scaling. Min-max scaling uses the range of the variable as the denominator and the difference between each observation and the minimum of the variable as the numerator, effectively providing each observation a location within the distribution as a percentage. Neural networks (NN) require scaling to facilitate convergence and improve computational performance as well as accuracy [24].

2.6. Label Definition & Processing

The dependent variable for this study was based on the original question from the Tong et al. study [11]. "In your opinion, would you consider this advertisement suspicious of human trafficking?" The possible responses were "Certainly no", "Likely no", "Weakly no", "Unsure", "Weakly yes", "Likely yes", and "Certainly yes". Classification decisions were made by three raters with many years of combined experience in the field of human trafficking. Pairwise agreement among the raters was 83%. The original dataset included images; however, these are no longer available. The best F1 score from the original study was 66.5% with a deep network approach [11].

For this study, the label was collapsed from seven levels to binary, as follows: "Certainly no", "Likely no", "Weakly no", and "Unsure" were combined into the category "Not likely or unsure". The remaining categories were collapsed into "More likely than not". This type of binary coding allows for probabilistic mapping between the classification algorithm and the raters' assessments. Dichotomization of the labels in the Trafficking-10k dataset has been proven to be adequate by previous studies [17,18].

2.7. Models

Seven models were evaluated to assess the value of the engineered features, the title and body content, the generated images, and combinations of the three as a design of experiments (see [25]), as it seeks to identify the marginal and additive contributions towards the classification metrics by modifying the architecture. Classification based on engineered features leverages neural network architecture, a common machine learning approach. Text classification (natural language processing, or NLP) uses stacked bidirectional long short-term memory (LSTM) [26] recurrent neural networks (RNN) [27]. Bidirectional LSTMs have proven to be powerful in classifying textual content [28]. Classification models leveraging imagery were based on convolutional neural network (CNN) architecture, specifically the EfficientNetB6 or ENET [29]. Efficient-Net has proven to be a highly accurate architecture for handling image classification tasks [30]. Multi-input models were combined with the same neural network scaffolding. These types of models have proven to be highly effective in prediction [30,31]. The seven models of this study include (see Table 2): Model 1, a neural network with decreasing nodes and dropout (64 nodes → 50% dropout → 32 nodes → 20% dropout, linear activation functions); Model 2, a CNN (EfficientNetB6 or ENET) flattened with decreasing nodes and dropout (128 nodes → 50% dropout → 64 nodes → 20% dropout); Model 3, an RNN with a size 512 embedding layer and two-layer bidirectional LSTMs; and Models 4 through 7, every combination of these three layers. Table 2 provides the definitions and abbreviations for all seven models. All these individual models were joined (or concatenated for multi-input models) with the final architecture, a neural network (16 nodes (relu activation function) → 10% dropout → 8 nodes (relu activation function) → 1 sigmoidal node). Figure 4 is the TensorFlow architecture display for the full model: NN, RNN, and ENET.

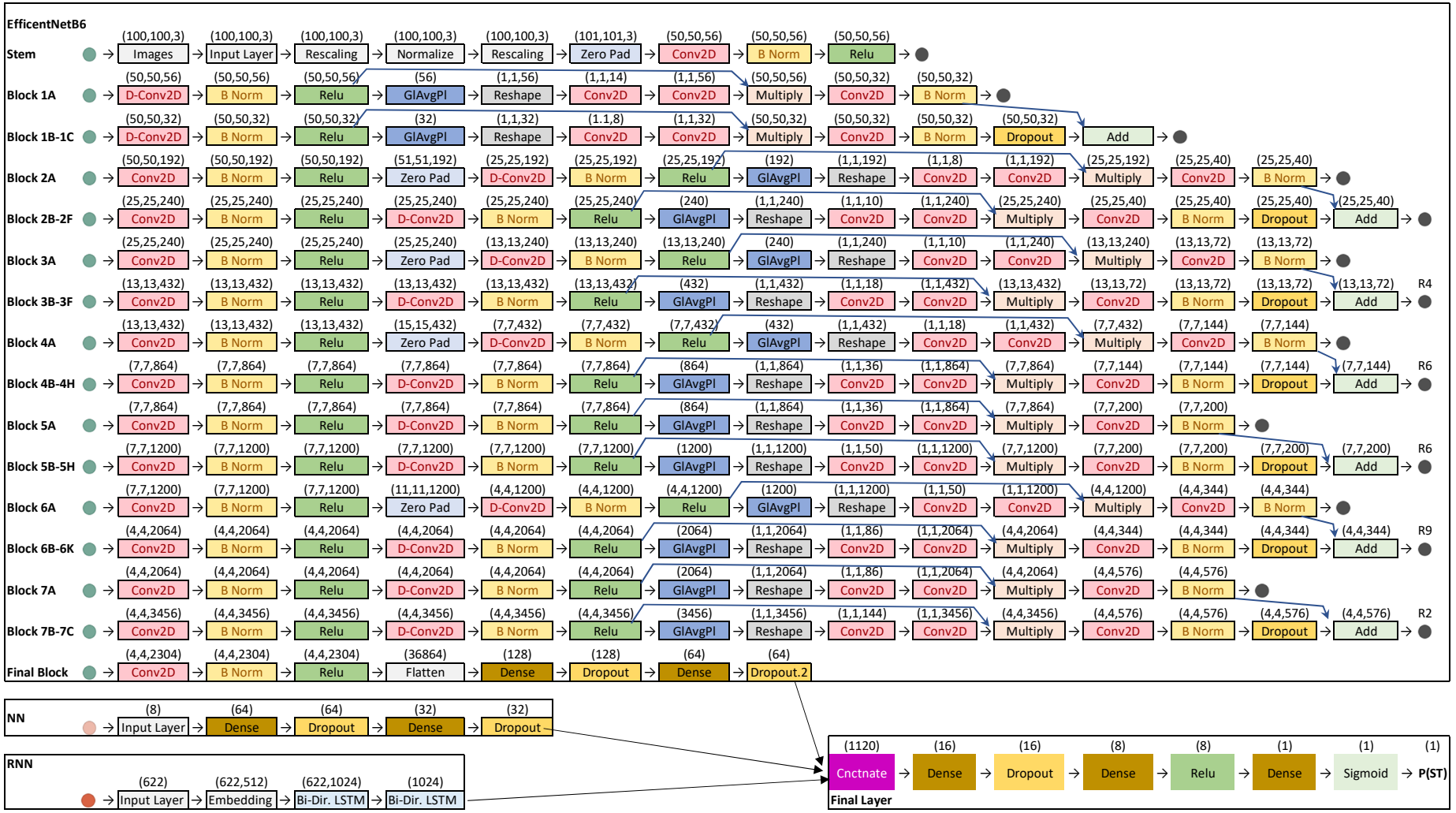

Figure 4. Architecture of the complete multi-input model. Green circle: EfficientNetB6 block start. Black circle: EfficientNetB6 block end. Numbers above blocks: dimension. R #: block repeats # of additional times. Zero Pad: zero padding. Conv2D: 2D convolutional layer. D-Conv2D: Depthwise Conv2D. GlAvgPl: global average pooling. Bi-Dir LSTM: Bi-Directional LSTM. Cnctnate: concatenation. Relu/Sigmoid: activation functions. P(ST): probability of sex trafficking.

Table 2. Models and abbreviations.

Model #	Model Type	Abbreviation
1	Neural Network (NN)	NN
2	EfficientNetB6 (ENET)	ENET
3	Recurrent Neural Network (RNN)	RNN
4	Multi-Input: ENET and NN	ENET + NN
5	Multi-Input: ENET and RNN	ENET + RNN
6	Multi-Input: NN and RNN	NN + RNN
7	Multi-Input: NN, ENET, and RNN	ENET + RNN + NN

2.8. Hyperparameter Tuning and Loss Metric

The number of epochs for each separate model was set based on training/validation split performance. The optimizer used was "adam" [32], although many others were investigated and often switched during performance testing. Batch size for mini-batch metrics was 32. The loss metric for optimizing model weights (and filters) was binary cross-entropy or log loss, appropriate for binary classification.

3. Results

3.1. Descriptive Statistics: Dependent Variable (Label)

The distribution of online escort ads across the classification label (dependent variable) categories is shown in Figure 5. After recoding, 4054 observations (32.8%) were coded as likely sex trafficking, with the remaining 8296 (67.2%) coded as not likely or unknown.

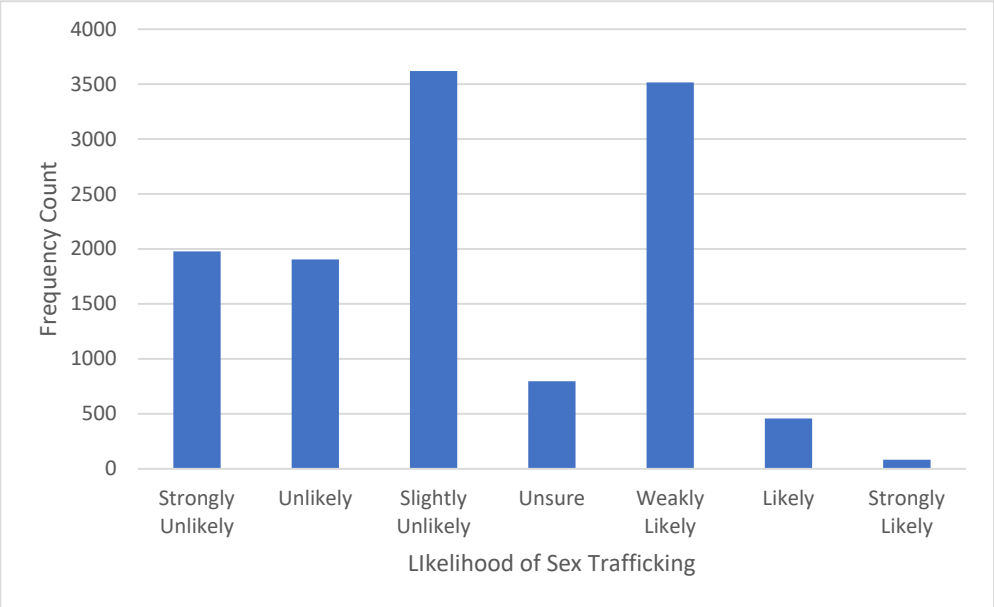

Figure 5. Distribution of the original dependent variable (label) for the full sample (n = 12,350).

3.2. Descriptive Statistics: Features

Eight features were generated for inclusion in the classification models. Table 3 provides descriptive statistics for those features. The "average" observation had 20.3 emojis in the title (30.8% of the words) and 113.7 emojis in the body (24.1% of the words). This "average" observation also had 0.2 derivatives of the word "sex" in the title (0.3% of the words) with an additional 0.7 in the body (0.1% of the words). The distribution of these variables is right-skewed, as they are left-censored at zero. The standard deviation of the emojis in the title and the body show high variability (standard deviation and range), something that ostensibly might be valuable for algorithmic learning of the classification status.

Table 3. Feature descriptive statistics.

Variable	Mean	Median	SD	Skew	Range	Min	Max
# Title Emojis	20.3	18.0	13.3	1.5	162.0	1.0	163.0
# Body Emojis	113.7	96.0	155.8	8.6	2373.0	1.0	2374.0
# "Sex" in Title	0.2	0.0	0.4	2.2	4.0	0.0	4.0
# "Sex" in Body	0.7	0.0	3.0	10.9	71.0	0.0	71.0
% Title Emojis	0.3	0.3	0.1	1.2	0.9	<0.1	0.9
% Body Emojis	0.2	0.23	0.1	2.7	0.9	<0.1	1.0
% "Sex" in Title	<0.1	0.0	<0.1	3.6	0.1	0.0	0.1
% "Sex" in Body	<0.1	0.0	<0.1	4.8	<0.1	0.0	<0.1

3.3. Model Comparison

The focal point of this study is the ability to provide law enforcement tools to efficiently identify those advertisements that are likely to be associated with sex trafficking (ST). To evaluate model performance, five measures were used: sensitivity (also known as recall), specificity, accuracy, precision (positive predictive value), and the F1 score. In terms of probability, these measures provide different types of information. Sensitivity or recall, P(identified as positive | truly ST), tells us the ability of the classifier to identify sex trafficking (i.e., true positive rate), but if the classifier suggested that all observations are positive, then sensitivity would be a perfect 1.0, while the specificity, P(identified as negative | truly not ST) (i.e., true negative rate), would be a worst-possible 0.0, so the model would not be informative. Accuracy combines both in its calculation and equals the proportion of cases that were accurately classified (i.e., true positives + true negatives) over the full sample (i.e., true positives + true negatives + false positives + false negatives). A model that correctly classifies most of the true positives and true negatives may still be problematic in terms of precision, which is the positive predictive value, or P(truly ST | identified as positive), as lower precision means that many identified positives are likely not to represent ST (i.e., there would be too many false positives). Finally, the F1 score provides a mixture of precision and recall (sensitivity) scaled between 0 and 1 and is calculated as 2 × Precision × Recall/(Precision + Recall).

In the original model from Tong et al. [11], the best model achieved 80% accuracy with an F1 score of 0.67. Comparative metrics for all models from the present study, as well as the Tong et al. model [11], are shown in Table 4. The best-performing models in terms of accuracy are Models 5, 6, and 7. For recall (sensitivity), the NN + RNN (Model 6) slightly outperforms Model 5 (ENET+RNN). The most precise model is Model 3, the RNN model. Model 1 (NN) and Model 2 (ENET) are not robust enough for classification, as both have recall measures below 0.50 and F1 scores below 0.51. Building on the work of Tong et al. [11], we were able to improve accuracy (by 2 percentage points), precision (7 points), and the F1 score (3 points) by using the ENET-RNN model alone without having the original imagery. The ENET-RNN-NN model improved accuracy by 2 percentage points and precision by 6. Model estimation time ranged from 1.333 min (NN) to 32.95 min (ENET + RNN + NN).

Table 5 provides the confusion matrix for the complete model (Model 7). The precision, recall, and accuracy metrics suggest a reasonable model.

The precision of models is important for law enforcement, as a lack of precision is associated with wasted effort since precision is calculated as the proportion of cases identified as positive that truly are associated with ST. Thus, we sorted the model predictions by probability (rounded to three decimal places) in descending order, so that the highest estimated probability of ST was ranked first, the second highest was ranked second, etc. We then evaluated the performance of all models at these two hinge points: the highest-ranked probabilities for at least 10 observations and then again at the highest-ranked probabilities for at least 50 observations. (Many classification probabilities occur more than once, so the exact number of samples within each group differs.) For law enforcement officers,

time is limited. Targeting all potential sites simultaneously is infeasible, so some sort of prioritization is required.

Table 4. Comparison of all models against the base.

Model	Abbreviation	Epochs	Minutes per Epoch	Total Minutes	Accuracy	Recall	Specificity	Precision	F1 Score
1	NN	20	0.067	1.333	0.73	0.35	0.91	0.67	0.46
2	ENET	10	2.933	29.333	0.74	0.42	0.89	0.66	0.51
3	RNN	2	3.000	6.000	0.81	0.57	0.93	0.80	0.67
4	ENET + NN	10	2.700	27.000	0.73	0.44	0.88	0.63	0.52
5	ENET + RNN	2	10.000	20.000	0.82	0.63	0.91	0.78	0.70
6	NN + RNN	4	3.250	13.000	0.82	0.65	0.90	0.76	0.70
7	ENET + RNN + NN	3	10.983	32.950	0.82	0.62	0.91	0.77	0.69
Base	Tong et al. (2017) [11]	NR	NR	NR	0.80	0.62	NR	0.71	0.67

NN = neural network; RNN = recurrent neural network; ENET = EfficientNetB6; NR = not reported.

Table 5. Confusion matrix for Model 7 (ENET + RNN + NN).

		Predicted		
		Not ST	**ST**	**Total**
Actual	Not ST	1509	149	1658
	ST	298	514	812
	Total	1807	663	2470
	Precision	78%	Specificity	91%
	Recall	63%	Accuracy	82%

NN = neural network; RNN = recurrent neural network; ENET = EfficientNetB6; ST = sex trafficking.

The results in Table 6 show that the best model, the ENET + RNN multi-input model, was able to correctly identify all 14 of the observations associated with the estimated probability of 0.845 or greater (1.00 precision). The second-best model, the NN + RNN, achieved 96% precision on 23 observations based on a model probability estimate of 0.976 or higher. The complete model (NN + RNN + ENET) achieved 0.9554 precision with 202 observations (a model classification probability estimate of 0.90 or better).

Table 6. Model metrics.

Model	Hinge	Classification Probability	N	True Positives	False Positives	Precision
NN	10	0.770	11	9	2	0.8182
NN	50	0.700	102	84	18	0.8235
RNN	10	0.960	10	9	1	0.9000
RNN	50	0.900	99	94	5	0.9495
ENET	10	0.990	11	8	3	0.7273
ENET	50	0.950	98	82	16	0.8367
NN + RNN	10	0.976	23	22	1	0.9565
NN + RNN	50	0.950	275	255	20	0.9273
NN + ENET	10	0.998	13	11	2	0.8462
NN + ENET	50	0.950	99	79	20	0.7980
ENET + RNN	10	0.845	14	14	0	1.0000
ENET + RNN	50	0.800	54	50	4	0.9259
NN + RNN +ENET	10	0.960	11	9	2	0.8182
NN + RNN + ENET	50	0.900	202	193	9	0.9554

NN = neural network; RNN = recurrent neural network; ENET = EfficientNetB6.

4. Discussion

This study has shown that recent advances in deep learning for classification allow us to more accurately and precisely identify online escort ads that may be associated with sex trafficking activity. High-precision models are particularly favored in that wasted effort by investigators with limited time resources should be avoided; the complete multi-input model (NN + RNN + ENET) developed here achieved 77% precision (as compared to the original 71% precision reported by Tong et al. [11]), and this increased to almost 96% when only the ads associated with the highest positive classification probabilities (>0.90) were considered. Other model metrics for this complete model were comparable to Tong et al.'s [11], demonstrating the increased precision was not associated with a trade-off deterioration in other metrics.

These results are based on the analysis of texts, emoticons, and emojis. Unfortunately, the advertisements' photographic images could not be accessed and incorporated into the model. It is, therefore, possible (if not likely) that even better results could be obtained if the images of the ads were available for analysis.

As with any other research study, this one suffered from certain shortcomings. The Trafficking-10k dataset is aging, so the results reported here would need to be replicated using newly harvested data. While manually labeling ads can be time consuming and expensive, automatic classification based on widely accepted indicators of sex trafficking activity (e.g., movement of sex providers, apparently minor providers) may be performed in its place (see [9]). Further, the multi-input model developed was complex, which yielded greater accuracy but obscured its theoretical underpinnings. Future research could develop and test a more theoretically sound model, then fit neural nets to the residuals to increase the model's accuracy, using expert ratings to annotate the ads.

Although the binary reclassification of the original outcome labels in Tong et al.'s [11] Trafficking-10k dataset could be perceived as a disadvantage, as it would arguably lead to a loss in granularity, our best accuracy score was comparable to those reported by [13], who applied ordinal regression neural network (ORNN) models to this same dataset. The advantage of our binary outcome model is that it estimates the probability of a given online escort ad being associated with ST, which is much easier to interpret than coefficients or estimates from an ordinal model. Such functionality could then be productionalized to allow criminal investigators to identify the ads with the highest probability values. This would allow law enforcement to prioritize such ads, which we have shown to have precision scores as high as 96–100%.

5. Conclusions

Deep learning binary classification models hold much promise in increasing the efficacy with which law enforcement could identify online escort ads that are potentially associated with ST. Any increase in model performance can translate into a more efficient use of limited public safety resources. By optimizing identification and investigation efforts and integrating a low-cost strategy approach, increasing productionalized tool accessibility can be achieved. Multi-input models benefit from the collective strength of each respective model from which they are composed while mitigating individual weaknesses.

Supplementary Materials: The supporting information can be downloaded at: https://github.comdustoff06/ST (accessed on 3 May 2023).

Author Contributions: Conceptualization, L.S., A.N.S. and L.V.F.; methodology, L.V.F.; formal analysis, L.S.; resources, L.S.; writing—original draft preparation, J.C., writing—review and editing, L.S., A.N.S., L.V.F. and J.C.; funding acquisition, L.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by a gift from NEC Corporation of America provided in 2022.

Data Availability Statement: Data are available by request and data use agreement through Marinus Analytics.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. UN Office of Drugs and Crime. *Global Report on Trafficking in Persons 2020*; UNODC: Vienna, Austria, 2021.
2. NA Trafficking in Persons Report July 2022. U.S. Department of State (2022) 2022 Trafficking in persons report. U.S. Department of State, Office to Monitor and Combat Trafficking in Persons. Available online: https://www.state.gov/wp-content/uploads/2022/10/20221020-2022-TIP-Report.pdf (accessed on 22 February 2023).
3. Tillyer, M.S.; Smith, M.R.; Tillyer, R. Findings from the US National Human Trafficking Hotline. *J. Hum. Traffick.* **2021**, 1–10. [CrossRef]
4. Dank, M.L.; Khan, B.; Downey, P.M.; Kotonias, C.; Mayer, D.; Owens, C.; Pacifici, L.; Yu, L. *Estimating the Size and Structure of the Underground Commercial Sex Economy in Eight Major US Cities*; The Urban Institute: Washington, DC, USA, 2014. Available online: https://www.urban.org/sites/default/files/publication/22376/413047-estimating-the-size-and-structure-of-the-underground-commercial-sex-economy-in-eight-major-us-cities_0_1.pdf (accessed on 22 February 2023).
5. L'Hoiry, X.; Moretti, A.; Antonopoulos, G.A. Identifying Sex Trafficking in Adult Services Websites: An Exploratory Study with a British Police Force. *Trends Organ. Crime* **2021**, 1–22. [CrossRef]
6. Martin, C.; Curtis, B.; Fraser, C.; Sharp, B. The Use of a GIS-Based Malaria Information System for Malaria Research and Control in South Africa. *Health Place* **2002**, *8*, 227–236. [CrossRef] [PubMed]
7. Martin, L.; Melander, C.; Karnik, H.; Nakamura, C. *Mapping the Demand: Sex Buyers in the State of Minnesota*; Jones Urban Research and Outreach: Minneapolis, MN, USA, 2017. Available online: https://conservancy.umn.edu/handle/11299/226520. (accessed on 22 February 2023).
8. Keskin, B.B.; Bott, G.J.; Freeman, N.K. Cracking Sex Trafficking: Data Analysis, Pattern Recognition, and Path Prediction. *Prod. Oper. Manag.* **2021**, *30*, 1110–1135.
9. Whitney, J.; Jennex, M.; Elkins, A.; Frost, E. Don't Want to Get Caught? Don't Say It: The Use of Emojis in Online Human Sex Trafficking Ads. In Proceedings of the International Conference on System Sciences, Hawaii, HI, USA, 3–6 January 2018; pp. 4273–4283. [CrossRef]
10. Nodeland, B.; Belshaw, S. Establishing a Criminal Justice Cyber Lab to Develop and Enhance Professional and Educational Opportunities. *Secur. Priv.* **2020**, *3*, e123. [CrossRef]
11. Tong, E.; Zadeh, A.; Jones, C.; Morency, L.-P. Combating Human Trafficking with Deep Multimodal Models. *arXiv* **2017**, arXiv:170502735.
12. Alvari, H.; Shakarian, P.; Snyder, J.K. Semi-Supervised Learning for Detecting Human Trafficking. *Secur. Inform.* **2017**, *6*, 1. [CrossRef]
13. Wang, L.; Laber, E.; Saanchi, Y.; Caltagirone, S. Sex Trafficking Detection with Ordinal Regression Neural Networks. *arXiv* **2019**, arXiv:190805434.
14. Esfahani, S.S.; Cafarella, M.J.; Pouyan, M.B.; DeAngelo, G.; Eneva, E.; Fano, A.E. Context-Specific Language Modeling for Human Trafficking Detection from Online Advertisements. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 28 July–2 August 2019; pp. 1180–1184.
15. Zhu, J.; Li, L.; Jones, C. Identification and Detection of Human Trafficking Using Language Models. In Proceedings of the 2019 European Intelligence and Security Informatics Conference (EISIC), Oulu, Finland, 26–27 November 2019; pp. 24–31.
16. Granizo, S.L.; Caraguay, Á.L.V.; López, L.I.B.; Hernández-Álvarez, M. Detection of Possible Illicit Messages Using Natural Language Processing and Computer Vision on Twitter and Linked Websites. *IEEE Access* **2020**, *8*, 44534–44546. [CrossRef]
17. Wiriyakun, C.; Kurutach, W. Feature Selection for Human Trafficking Detection Models. In Proceedings of the 2021 IEEE/ACIS 20th International Fall Conference on Computer and Information Science (ICIS Fall), Xi'an, China, 13–15 October 2021; pp. 131–135.
18. Wiriyakun, C.; Kurutach, W. Extracting Co-Occurrences of Emojis and Words as Important Features for Human Trafficking Detection Models. *J. Intell. Inform. SMART Technol.* **2022**, *7*, 12.1–12.5.
19. Van Rossum, G. The Python Library Reference, Release 3.8.2. *Python Softw. Found.* **2020**, 32. Available online: https://www.python.org/downloads/release/python-382/ (accessed on 24 February 2020).
20. Developers, T. TensorFlow. Available online: https://zenodo.org/record/6574269#.ZFscCXZByUk (accessed on 23 May 2022).
21. Honnibal, M.; Montani, I. SpaCy 2: Natural Language Understanding with Bloom Embeddings, Convolutional Neural Networks and Incremental Parsing. *Appear* **2017**, *7*, 411–420.
22. Bradski, G. The OpenCV Library. *Dr. Dobbs J. Softw. Tools Prof. Program.* **2000**, *25*, 120–123.
23. Bird, S.; Klein, E.; Loper, E. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*; O'Reilly Media, Inc.: Sebastopol, CA, USA, 2009; ISBN 0-596-55571-7.
24. Nawi, N.M.; Atomi, W.H.; Rehman, M.Z. The Effect of Data Pre-Processing on Optimized Training of Artificial Neural Networks. *Procedia Technol.* **2013**, *11*, 32–39. [CrossRef]
25. Alpaydin, E. Design and analysis of machine learning experiments. In *Introduction to Machine Learning*, 4th ed.; MIT Press: Cambridge, MA, USA, 2010; pp. 475–515.
26. Graves, A. Long Short-Term Memory. In *Supervised Sequence Labelling with Recurrent Neural Networks*; Springer: Berlin/Heidelberg, Germany, 2012; pp. 37–45.

27. Yu, Y.; Si, X.; Hu, C.; Zhang, J. A Review of Recurrent Neural Networks: LSTM Cells and Network Architectures. *Neural Comput.* **2019**, *31*, 1235–1270. [CrossRef] [PubMed]
28. Nowak, J.; Taspinar, A.; Scherer, R. *LSTM Recurrent Neural Networks for Short Text and Sentiment Classification*; Springer: Berlin/Heidelberg, Germany, 2017; pp. 553–562.
29. Tan, M.; Le, Q. Efficientnet: Rethinking Model Scaling for Convolutional Neural Networks. In Proceedings of the Machine Learning Research, Long Beach, CA, USA, 9–15 June 2019; pp. 6105–6114.
30. Apostolopoulos, I.D.; Apostolopoulos, D.I.; Spyridonidis, T.I.; Papathanasiou, N.D.; Panayiotakis, G.S. Multi-Input Deep Learning Approach for Cardiovascular Disease Diagnosis Using Myocardial Perfusion Imaging and Clinical Data. *Phys. Med.* **2021**, *84*, 168–177. [CrossRef] [PubMed]
31. Dua, N.; Singh, S.N.; Semwal, V.B. Multi-Input CNN-GRU Based Human Activity Recognition Using Wearable Sensors. *Computing* **2021**, *103*, 1461–1478. [CrossRef]
32. Defazio, A.; Jelassi, S. Adaptivity without Compromise: A Momentumized, Adaptive, Dual Averaged Gradient Method for Stochastic Optimization. *J. Mach. Learn. Res.* **2022**, *23*, 1–34.