CRNN based Violence Detection using Text and Videos

By Premanand Ghadekar

WORD COUNT	2921	TIME SUBMITTED	27-JAN-2023 01:48PM
		PAPER ID	95958393

CRNN based Violence Detection using Text and Videos

Premanand Ghadekar[1], Adwait Bhosale[2], Dhananjay Deore[3], Kunjal Agrawal[4], Tejas Gadi[5],
Rehanuddin Qazi[6]

[1,2]Vishwakarma Institute of Technology, Pune, India
premanand.ghadekar@vit.edu,
adwaitbhosale04@gmail.com, dhananjaybdeore@gmail.com,
kunjalagrawal2002@gmail.com, tvgadi2003@gmail.com,
rehanqazi02@gmail.com

Abstract. Prevention of violence is quite necessary so that they do not affect the society mentally in general and present an incorrect meaning in front of the young generation. As a consequence of this, a system is proposed which is an AI based application that would detect and identify three major types of violence's namely: Physical violence, Sexual violence and Verbal violence in text and video data. The dataset is created by collecting text data related to violence through articles and video clips consisting of any type of harassment or violence. A hybrid model is created which combines CNN and Bi - directional LSTM (combined architecture enhances the performance of the model). The major purpose of the system is to help the government respond to these violence's before time and accordingly take action so that they can be prevented and avoided before any major harm.

Keywords: Violence, RNN, CRNN, pre-processing, CNN-LSTM, Bidirectional-LSTM,Text Augmentation, Video Augmentation.

1 Introduction

Violence is a social phenomenon that has become quite common nowadays. Every now and then, in lot of scenarios happen rapes and child abuse. Societies and the government are making efforts to minimize the effects of violence against women and men.

A common type of violence, Gender-based violence, includes intimate partner violence, non-partner sexual assault, female genital mutilation, sexual exploitation and abuse, child abuse, emotional violence, verbal abuse, female infanticide and forced child marriage. Gender based violence is a violence that is directed towards a person because of his/her gender. Both women and men are equally the victims of GBV. The United Nations has identified gender-based violence against women as a global health and development issue, and aimed at reducing gender-based violence have been undertaken around the world. GBV does have a lasting effect on the minds of the victims and so it is considered as one of the serious issues on a global level. It has been observed that exposure to violence can increase the risk of mental illness and suicidality; chronic diseases like heart disease.

2 Literature Review

In [1], the authors have presented the techniques of violence detection. Violence detection will be carried out in three categories : 1. Traditional Machine Learning approach, 2. Using Support Vector Machine(SVM) and 3. Deep Learning. A review on the methods for violence detection and datasets like Movies(consisting of 200 clips), Hockey(1000 clips), Media(10,000 clips) etc. In [2], the authors have proposed a model that detects the violence in videos captured from the video surveillance cameras. The proposed model has a UNET-like network model using mobileNet as an encoder. This is followed by a LSTM network for temporal feature extraction. They made use of three datasets : 1. RWF200, 2.Movie fights and 3. Hockey fights dataset. In [3], the authors have made an analysis about recognition of actions like jumping, clapping etc. but not fights or aggressive behaviour. In the paper they have proposed a dataset for violence detection in videos. The dataset consists of 1000 clips divided into two categories: fights and non - fights. The authors have made an attempt to emphasize the recognition of fight detection by methods that detect various actions in the existing systems. In [4], the authors have analyzed the rise of school bullying amongst the teenagers and tried to assess this issue. They propose a system that detects the actions of the students. Actions like hitting, pushing, beating, kicking can be easily distinguished from mundane activities like walking, running and jogging. They have used the Fuzzy Multi Threshold Classifier that detects physical bullying behaviour. In [5], the authors have proposed a system that uses Convolutional Neural Network(CNN) to recognize physical violence actions. The model detects various bullying actions like kicking and punching. In [6], the authors have proposed a model that uses convolutional 3D networks for feature extraction and classification. The paper focuses on campus violence detection. The authors gathered the data for the creation of campus data by performing certain violent actions and daily - activity videos that would help in the classification of violence. The network consists of convolutional and the max pooling layers. The hidden layer makes use of the ReLU activation function to get values as either 0 and 1. They achieved an accuracy of 92.00% for their model. In [7], the authors have created a dataset consisting of violent and non-violent videos. They made use of the CNN model to classify content as violent or non-violent. Extracted features were then fed into the LSTM network. In [8], the authors made use of the fusion technique of two significantly different convolutional neural networks (CNNs) i.e., AlexNet and the SqueezeNet networks. Each network is followed by a separate Convolution Long Short Term memory(ConvLSTM) to extract robust and richer features from a video in the final hidden state. Finally, features were classified using a series of fully connected layers and softmax classifier. In [9], the authors made a comparison on different video classification approaches(like audio, text and video) based on the features. They presented different performance metrics like accuracy, f1 score for video classification and discussed its applications as well. In [10], the authors made a study on the use of CNN architecture in video classification. They conclude that CNN extracts rich features from the video frames and were generic and generalized which enhanced the model accuracy. They trained their model on the UCF-101 ACTION RECOGNITION dataset and improved accuracy from

43.9% to 63.3%. In [11], the authors have proposed a combination of the CNN and LSTM. They have proposed two text classification models called NA-CNN-LSTM and NA-CNN-COIF-LSTM. Through comparative experiments, it is proved that the combination of CNN without activation function and LSTM has better performance. In [12], the authors have proposed a hybrid model of LSTM and CNN, construct CNN model on the top of LSTM, the text feature vector output from LSTM is further extracted by CNN structure. The performance of the hybrid model was compared with that of other models in the experiment.The authors have used LSTM to store the information received and to resolve the vanishing gradient uses CNN to further extract the local features. In [13], the authors have proposed a system to detect gender based violations on twitter messages generated in Mexico. They downloaded 1,857,450 twitter messages for the creation of the dataset and were manually labeled as positive, negative or neutral messages. The authors performed minimal preprocessing on the dataset and thus the original messages were converted to a vector in numeric format. They also studied different feature extraction methods like CountVectorizer, TfidfVectorizer and Hashing Vectorizer. In [14], the authors have proposed a women abuse detection method using CNN where they detect the male and the female present in the location. In [15], the authors have proposed a system to detect violence and non - violence using deep learning. They used CNN for feature extraction from each frame of the video, which were then accumulated and passed to a LSTM network for further extraction and analysis. In [16], the authors have presented various techniques to detect violence. They have categorized these techniques as: using Machine Learning, Support Vector Machine(SVM) and Deep Learning. In [17], the authors proposed a VGG19 Convolutional Neural Network, where they extract the frames from the input videos and label the objects in the frame that show an abnormal behavior. The system is used for the detection of crimes. In [18], the authors have proposed a weak supervised method to detect spatial and temporal actions(that are violent) in the videos. They have used the Fast - RCNN architecture that extracts the spatiotemporal information.

3 Methodology

First of all, the text based classification process and the proposed algorithm is explained in detail. The dataset which is prepared contains 850 records while each record consists of text and its corresponding violence type (label). For videos, there are around 280 videos, approximately 80-100 videos for each class.

3.1 Data Augmentation

The text augmentation such as WordNetAugmenter and CLARE Augmenter is performed by the NLP Alumentations library to increase the text dataset size.

3.2 Data Preprocessing

a) The text preprocessing techniques such as stemming, lemmatization and Tokenization are applied on the sentences and also used by LabelEncoder to encode the class labels.

b) Mapped the vocabulary to the integer value by making use of the StringLookup functionality of keras, which will not perform any splitting or transformation on the input string.

Later a InceptionV3 feature extractor with weights set to imagenet was being applied to the data.

3.3 Splitting the Dataset into training and testing sets

a) For text data, the split ratio is 80:20 so that 80% of the data i.e., 680 records are used for training and 20% i.e 170 records for the testing set.

b) Shuffled the entire dataset wherein allotted 200 videos for training and 100 videos for testing purpose.

3.4 Model building and Compilation

a) For textual data, the various models such as LSTMs, Bidirectional LSTMs, CNN model are used generally but the proposed algorithm makes use of CNN+Bidirectional-LSTMs combined architecture which gives more desirable and accurate results. The model comparison is given in table no. 1. The CNN+ Bidirectional LSTMs is constructed in the following manner: a series of convolutional 2d layers and max pooling 2d layers and then concatenate layer for combining all the maxpooling 2d layer outputs. Now the text features which are extracted by CNN architecture are given to further Bi-LSTMs layers and finally an output Dense layer with neurons equal to the number of classes. Later the model is compiled using Adam optimizer and loss of sparse categorical cross entropy. The mathematical equations for the cell state, candidate cell state and the final output are given in equation 1, 2 and equation 3. The CNN+ Bi-LSTM architecture is shown in figure no. 1.

$$c{\sim}t = tanh(wc[ht\text{-}1, xt] + bc) \quad (1)$$

$$ct = ft * ct\text{-}1 + it * c{\sim}t \quad (2)$$

$$ht = ot * tanh(ct) \quad (3)$$

where ct -> cell state(memory) at timestamp(t).

c~t -> represents candidate for cell state at timestamp(t).

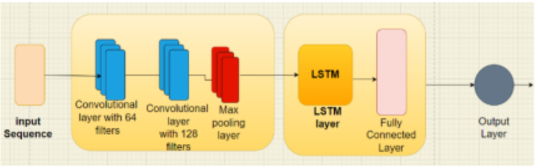


Fig. 1 : Proposed model CNN+Bidirectional LSTM architecture

b) For videos, there is an extraction of vocabulary for every input sentence by using the label processor. Built a CRNN model wherein passed the input vector initially to the GRU, Dense and Dropout layer. Wherein Dropout helped to reduce the overfitting of the model. Later compiled the model with the loss of sparse categorical crossentropy which OneHotEncoded the vectors made use of adam optimizer. Applied the model on the training data and evaluated the model's F1 score and accuracy on the testing data.

4 Flowchart

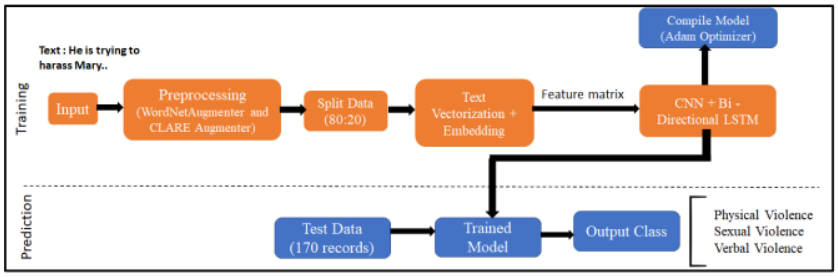

Text : He is trying to harass Mary..
Training
Input
Preprocessing (WordNetAugmenter and CLARE Augmenter)
Split Data (80:20)
Text Vectorization + Embedding
Feature matrix
CNN + Bi - Directional LSTM
Compile Model (Adam Optimizer)
Prediction
Test Data (170 records)
Trained Model
Output Class
Physical Violence
Sexual Violence
Verbal Violence

Fig. 2 : Proposed project flow diagram for Text Data

5 Experimentation

a) First for the text based classification, the base models such as LSTM model is used which gives an F1 score of 0.80 and CNN based text classification with a F1 score of 0.82. Then the combination of CNN + Bidirectional LSTM combined model is used which gives a F1 score of 0.90(figure 4). The model uses Adam optimizer for model compilation and loss function of sparse categorical crossentropy.

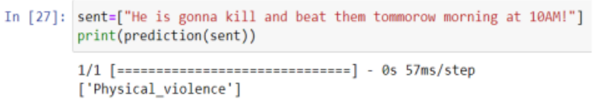
Table No. 1 : Model comparison for text data

Model	F1 Score
LSTM model	0.80
Bi-LSTM model	0.86
CNN model	0.82
CNN + Bi-LSTM model	0.90

b) Video classification is done using the CRNN model, with an Adam optimizer for model compilation and loss function of categorical crossentropy that gives an accuracy of 85%(figure 4).

6 Result and Discussion

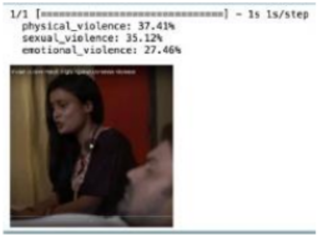
The results show that the proposed approach reaches high accuracy with precise output. For the text classification, CNN+Bidirectional LSTM(figure 2) is used. Here, the features from both the directions are combined and considered for further analysis. The model trains the input text data twice with the help of forward and backward directions. The accuracy on the test data for text as the input was estimated to be 93%. In order to extract high-level information from videos, CNN models are fed the video's pictures. The RNN layer's output is connected to a fully connected layer to produce the classification output after the features have been provided to it. The Convolution Recurrent Neural Network(CRNN) performs better for motion based activities. It extracts the correlation between the images by keeping in mind the past frames and their features. The accuracy for the training dataset was 88% whereas on the test dataset it was found to be 77%. The violence which has a higher percentage amongst the others is predicted(based on the input features).



```
In [27]: sent=["He is gonna kill and beat them tommorow morning at 10AM!"]
         print(prediction(sent))

         1/1 [==============================] - 0s 57ms/step
         ['Physical_violence']
```

Fig.3 : Output image of CNN+Bi-LTSM text classification



```
1/1 [==============================] - 1s 1s/step
  physical_violence: 37.41%
  sexual_violence: 35.12%
  emotional_violence: 27.46%
```

Fig. 4 : Output image of CRNN video classification

CRNN provides better results when compared to transfer learning. The major reason being the number of layers and classes the latter has. The model is trained on IMAGENET which has 1000 classes and layers more than 500. This becomes computationally expensive and increases the training time.

7 Conclusion

The paper highlights the detection of violence in text, images and videos with the help of various deep learning algorithms. CNN+Bidirectional LSTM (combined architecture) is used so that Bi-LSTM can utilize the information from both sides for better understanding. Recurrent neural networks is used for text classification which initially starts with preprocessing like removing of punctuations followed by feature extraction. These features help the model identify and understand the violence. The proposed system provides good accuracy with no overfitting. The created system is beneficial to be used in surveillance systems and social media applications which are prone to violence and harassment.

8 Future Scope

The proposed system involves detection of violence in text and videos. The dataset which is being created consists of various videos depicting the different types of violence like physical, sexual or emotional. Though the results that are drawn from the proposed approach are quite precise and beneficial to be used in surveillance systems, there is always a need for improvement. A dataset consisting of audios which depict any kind of violence based on the phonic information. These audios will help in detecting violences or harassment for example: if someone is trying to emotionally blackmail a person or abuse him, it will automatically be detected and appropriate action will then be taken. Thus, addition of audios to the dataset will make an appropriate system that can be further used in various applications to avoid various kinds of violences before any major mishap.

References

1. M. Ramzan et al., "A Review on State-of-the-Art Violence Detection Techniques," in IEEE Access, vol. 7, pp. 107560-107575, 2019, doi: 10.1109/ACCESS.2019.2932114.

2. Vijeikis, Romas, Vidas Raudonis, and Gintaras Dervinis. 2022. "Efficient Violence Detection in Surveillance" Sensors 22, no. 6: 2216

3. Enrique Bermejo Nievas, Oscar Deniz Suarez, Gloria Bueno García, Rahul Sukthankar, Violence Detection in Video Using Computer Vision Techniques, Computer Analysis of Images and Patterns, 2011, Volume 6855, ISBN : 978-3-642-23677-8

4. Ye, Liang & Ferdinando, Hany & Seppänen, Tapio & Alasaarela, Esko. (2014). Physical Violence Detection for Preventing School Bullying. Advances in Artificial Intelligence. 2014. 1-9. 10.1155/2014/740358.

5. John Clement Suladay Escobanez and Benilda Eleonor Comendador. 2022. Student Physical Violence Detection using Convolutional Neural Networks. In Proceedings of the 12th International Conference on Information Communication and Management (ICICM '22). Association for Computing Machinery, New York, NY, USA, 34–38.

6. Ye, Liang, Tong Liu, Tian Han, Hany Ferdinando, Tapio Seppänen, and Esko Alasaarela. 2021. "Campus Violence Detection Based on Artificial Intelligent Interpretation of Surveillance Video Sequences" Remote Sensing 13, no. 4: 628.

7. Sumon, Shakil & Goni, Raihan & Hashem, Niyaz & Shahria, Md Tanzil & Rahman, Mohammad. (2019). Violence Detection by Pretrained Modules with Different Deep Learning Approaches. Vietnam Journal of Computer Science. 7. 10.1142/S2196888820500013.

8. Mohammed, Heyam & Elrefaei, Lamiaa. (2022). Detecting Violence in Video Based on Deep Features Fusion Technique.

9. Islam, Md & Sultana, Shanjida & Roy, Uttam & Al, Jubayer. (2020). A review on Video Classification with Methods, Findings, Performance, Challenges, Limitations and Future Work. Jurnal Ilmiah Teknik Elektro Komputer dan Informatika. Vol 6, No 2 (2020). 47-57. 10.26555/jiteki.v6i2.18978.

10. A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar and L. Fei-Fei, "Large-Scale Video Classification with Convolutional Neural Networks," 2014 IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 1725-1732.

11. Y. Luan and S. Lin, "Research on Text Classification Based on CNN and LSTM," 2019 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA), 2019, pp. 352-355.

12. J. Zhang, Y. Li, J. Tian and T. Li, "LSTM-CNN Hybrid Model for Text Classification," 2018 IEEE 3rd Advanced Information Technology, Electronic and Automation Control Conference (IAEAC), 2018, pp. 1675-1680.

13. Castorena CM, Abundez IM, Alejo R, Granda-Gutiérrez EE, Rendón E, Villegas O. Deep Neural Network for Gender-Based Violence Detection on Twitter Messages. Mathematics. 2021; 9(8):807

14. Sandhiya, R., & Prassad, A.R. (2020). Women Abuse Detection in Video Surveillance using Deep Learning.

15. Dandage, V., Gautam, H., Ghavale, A., Mahore, R., & Sonewar, P.A. (2019). Review of Violence Detection System using Deep Learning.

16. Milon Biswas, Afjal Hossain Jibon, Mim Kabir, Khandokar Mohima, Rahman Sinthy, Md. Shamsul Islam a and Monowara Siddique. State-of-the-Art Violence Detection Techniques: A review, Asian Journal of Research in Computer Science 13(1): 29-42, 2022; Article no.AJRCOS.79063 ISSN: 2581-8260

17. Umair Muneer Butt, Sukumar Letchmunan, Fadratul Hafinaz Hassan, Sultan Zia and Anees Baqir, "Detecting Video Surveillance Using VGG19 Convolutional Neural Networks" International Journal of Advanced Computer Science and Applications(IJACSA), 11(2), 2020

18. Choqueluque-Roman D, Camara-Chavez G. Weakly Supervised Violence Detection in Surveillance Video. Sensors. 2022; 22(12):

CRNN based Violence Detection using Text and Videos

ORIGINALITY REPORT

12%
SIMILARITY INDEX

PRIMARY SOURCES

1	arxiv.org Internet	54 words — 2%
2	Jiarui Zhang, Yingxiang Li, Juan Tian, Tongyan Li. "LSTM-CNN Hybrid Model for Text Classification", 2018 IEEE 3rd Advanced Information Technology, Electronic and Automation Control Conference (IAEAC), 2018 Crossref	44 words — 2%
3	Yuandong Luan, Shaofu Lin. "Research on Text Classification Based on CNN and LSTM", 2019 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA), 2019 Crossref	32 words — 1%
4	www.turkiyeklinikleri.com Internet	30 words — 1%
5	downloads.hindawi.com Internet	27 words — 1%
6	www.researchgate.net Internet	22 words — 1%
7	Ajiboye, Florence Ibikun. "Teachers' Perception of the Causes and Prevalence of Gender Based Violence among Primary School Pupils in Ifelodun Local Government Area of Kwara State", Kwara State University (Nigeria), 2022 ProQuest	20 words — 1%

8	www.mdpi.com Internet	18 words — 1%
9	Shahin, M.A.. "Settlement prediction of shallow foundations on granular soils using B-spline neurofuzzy models", Computers and Geotechnics, 200312 Crossref	16 words — 1%
10	ojs.unm.ac.id Internet	16 words — 1%
11	openaccess.hacettepe.edu.tr:8080 Internet	15 words — 1%

EXCLUDE QUOTES	ON	EXCLUDE SOURCES	< 14 WORDS
EXCLUDE BIBLIOGRAPHY	ON	EXCLUDE MATCHES	< 9 WORDS