Dissertation on

"Detection Of Cyber Bullying In Images"

Submitted in partial fulfilment of the requirements for the award of degree of

Bachelor of Technology
in
Computer Science & Engineering

UE17CS490A – Capstone Project Phase - 1

Submitted by:

Sruthi Madineni	**PES1201700051**
Meghana Nayak	**PES1201701339**
Rachana H S	**PES1201701726**

Under the guidance of

Prof. Aruna S
Assistant Professor
PES University

August - December 2020

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
FACULTY OF ENGINEERING
PES UNIVERSITY
(Established under Karnataka Act No. 16 of 2013)
100ft Ring Road, Bengaluru – 560 085, Karnataka, India

PES UNIVERSITY

(Established under Karnataka Act No. 16 of 2013)
100ft Ring Road, Bengaluru – 560 085, Karnataka, India

FACULTY OF ENGINEERING

CERTIFICATE

This is to certify that the dissertation entitled

'Detection of Cyber Bullying In Images'

is a bonafide work carried out by

Sruthi Madineni	**PES1201700051**
Meghana Nayak	**PES1201701339**
Rachana H S	**PES1201701726**

in partial fulf ilment for the completion of seventh semester Capstone Project Phase - 1 (UE17CS490A) in the Program of Study - Bachelor of Technology in Computer Science and Engineering under rules and regulations of PES University, Bengaluru during the period Aug. 2020 – Dec. 2020. It is certified that all corrections / suggestions indicated for internal assessment have been incorporated in the report. The dissertation has been approved as it satisfies the 7th semester academic requirements in respect of project work.

Signature **Prof. Aruna S** Assistant Professor	Signature Dr. ShylajaS S Chairperson	Signature Dr. B K Keshavan Dean of Faculty

External Viva

Name of the Examiners	**Signature with Date**
1. _____	_____
2. _____	_____

DECLARATION

We hereby declare that the Capstone Project Phase - 1 entitled **"Detection Of Cyber Bullying In Images"** has been carried out by us under the guidance of Prof. Aruna S, Assistant Professor and submitted in partial fulfilment of the course requirements for the award of degree of **Bachelor of Technology** in **Computer Science and Engineering** of **PES University, Bengaluru** during the academic semester August – December 2020. The matter embodied in this report has not been submitted to any other university or institution for the award of any degree.

PES1201700051	**Sruthi Madineni**	
PES1201701339	**Meghana Nayak**	
PES1201701726	**Rachana H S**	

ACKNOWLEDGEMENT

I would like to express my gratitude to Prof. Aruna S, Assistant Professor, Department of Computer Science and Engineering, PES University, for her continuous guidance, assistance, and encouragement throughout the development of this UE17CS490A - Capstone Project Phase – 1.

I take this opportunity to thank Dr. Suryaprasad J, Vice-Chancellor, PES University, Dr. Shylaja S, Chairperson, Department of Computer Science and Engineering, PES University, Project Coordinators Prof. Silviya Nancy J and Prof. Sunitha R. Finally, this project could not have been completed without the continual support and encouragement we have received from our family and friends.

ABSTRACT

Cyber bullying is a form of bullying which is administered using some electronic means. It has become very common with the advance of Internet and usage of social media sites. This can happen by various means. Posting rumours, sexual remarks, threats, disclosing a person's information without his consent being few among them.

Most of the previous studies are done on texts using some text classification methods. Recently there have been works even on images. Lot of researchers are working on coming with a better classification algorithm on cyberbullyingdetection on images.

The purpose of this project is to build a design methodology to detect cyberbullying by using Convolution Neural Network.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Cyber bullying is a form of bullying which is administered using some electronic means. It has become very common with the advance of Internet and usage of social media sites. This can happen by various means. Posting rumors, sexual remarks, threats, disclosing a person's information without his consent being few among them. This not only affect the victim but it also affects the people close to him. This might lead to a person's low self-esteem, increase suicidal thoughts, being scared, frustrated, angry or depressed.

Detecting cyber bullying and preventing it from affecting the masses has been the concerns of many. There have been many works on the same using different modern computer techniques.

Nowadays Cyberbullying is a common phenomenon in many of the social media sites like Instagram, Twitter, Facebook and many other. According to the statistics for the year 2018-2020,60% of the parents reported that their kids were bullied. There are also other platforms like mails, messaging apps and certain websites where they target certain individuals or even a group of individuals. It has become very difficult to stop cyber bullying because parents and the relatives of the victims are most of the times not aware about the same.
Bullying content can be in the form of texts, images and videos. Most of the previous studies are done on texts using some text classif ication methods. Recently there have been works even on images. Lot of researchers are working on coming with a better classification algorithm on cyberbullying detection on images.

The purpose of this project is to build a design methodology to detect cyberbullying by using Convolution Neural Networks.

CHAPTER 2

PROBLEM STATEMENT

Cyberbullying is utilizing electronic means to embarrass, threaten, harass and targeting another person. Cyberbullying could be in the form of rude or aggressive texts ,abusive or inappropriate images, inappropriate videos etc. Cyberbullying focuses on things like race, person's gender, sexual orientation, religion which violates the law. Apart from violating the law it also causes victim to be in constant state of fear/upset, it can also lead them to fall into depression, anxiety and other mood related disorders. According to recent survey it has been found that overall 36.5% percent of people feel that they have been cyberbullied in their lifetime and 17.4% have reported it has happened at some point in last 30 days. 87% of young people have seen cyberbullying occurring online.

As mentioned above cyberbullying can also happen in the form of text and it has become serious problem among teenagers and adolescents. Works on detection of cyberbullying in text which is also called as sentiment analysis has been done in the past and it is still ongoing to find maximum accuracy.

But our focus is on detecting cyberbullying in images, since less work has been done on this. Cyberbullying using images includes the usage of inappropriate and embarrassing images. Examples of cyberbullying using images include threatening to share inappropriate images of the victim in order to control or blackmail them, taking degrading pictures of the victim and sharing them without his/her permission etc.

Few works on cyberbullying in images have been done using the captions and comments underneath the image but this is not sufficient because sometimes the image may not contain any caption and it still could be bullying or abusive so, the caption or comments are not the only factors that determine the nature of the image. So, our goal is to focus solely on the content of the image irrespective of caption or comments

Models to Detect Cyberbullying content in images have been designed in the past using feature extractors like SIFT(Scale invariant feature transform) and HUE, color histograms with the combination of Bag of visual words and using deep learning algorithms like SVM(Support Vector Machines),LLM(Log Linear model).But this still yields low accuracy rate and our goal is to design a model which yields better accuracy and better feature extraction using deep neural networks(CNN).

CHAPTER 3

LITERATURE SURVEY

3.1 A Framework for Cyberbullying Detection in Social Network

3.1.1 Author and publication details

Authors: Krishna B. Kansara and Narendra M. Shekokar

Date: 5th Feb 2019

3.1.2 Summary

- This article talks about how Social Media could pose threats like cyberbullying by allowing users to post or share some kinds of abusive or offensive content and and the proposed method will be able to detect if it is abusive or not based on the content and features of images.
- Abusive Image Detection:

Step 1: Feature extraction from abusive or inappropriate images and using Local Binary Pattern algorithm. LBP works by extracting interesting points from an image called as local descriptors
Step 2: Using extracted features and forming a visual vocabulary or codebook using K-means algorithm and the obtained clusters or visual words together is called as bag of visual words. Bag of visual words has the features which represents the whole dataset of abusive images.
Step 3: Here SVM is used for classification of given input image as abusive or not. SVM algorithm builds a model that assigns new image into either class based on whether the image contains abusive or inappropriate contents or not. They have used Boolean system to express the likelihood of the cyberbullying content by analyzing image features of the given image.

3.1.3 Results

The proposed model provides recall rate of 66% and since LBP is used as feature extractor here invariant to rotation and Increase in neighbors will lead to more response time and increase of computational complexity. Only pixel difference is used by LBP which is not sufficient.

3.2 LLM and BOVW Models for Adult Image Classification and Filtering

3.2.1 Author and publication details

Authors: Thomas Deselaers and Lexi Pimenidis and Hermann Ney

Date : Sept 2013

3.2.2 Summary

- This article talks about how inappropriate images could be detected by using PCA,GMM(Guassian mixture model) and BOVW(Bag of Visual words).
- Step 1: Image patches are extracted as local features around difference-of-Gaussian interest points and then they are scaled to a common size and then PCA to reduce dimensionality.
- Step 2: Then they have created a bag of visual words using training algorithm for unsupervised training of GMM(Gaussian Mixture Model) and visual vocabulary that has been learned will be able to recognize frequent patterns in training images.
- Step 3: Here classification to determine the class of images is represented by histograms. Using Log-linear models and SVM ,the classification has been done.There are 5 classes here,
  class 0: inoffensive images,
  class 1: If people are lightly dressed,could be offensive in strict environment,
  class 2: partially nude people,could be objected in schools,
  class 3: completely nude people,objectionable in many environments,
  class 4: pornographic images,offensive in almost every environments

3.2.3 Results

Here there was more confusion between class 2 and class 3 and that affected the accuracy. Even though the color information was included it didn't do much improvement in the accuracy and also since LLM is used here it is difficult to capture complete relationships using this model, algorithms which are more powerful and complex such as Neural Networks can easily outperform this algorithm.

3.3 Image Retrieval by Bag of Visual Words and Color Information

3.3.1 Author and publication details

Authors: N. Mansoori, M. Nejati, P. Razzaghi and S. Samavi

Date : May 2013

3.3.2 Summary

- This article talks about how content based image classification using features of an image like color, texture and shape. Here they have used two feature extractors like SIFT (Scale Invariant Feature Transform) and HUE to get more accuracy.
- Training Stage
- Feature Extraction: Here SIFT is used to extract the interesting points from an image which are called as keypoints and each keypoint has a descriptor.
- Feature Description: Here the extracted key points are described.Two descriptors are used here SIFT and HUE descriptors.SIFT detects local patches and HUE produces the hue descriptor which is invariant to both lighting geometry and specularities.
- Quantization: The descriptors obtained using SIFT and HUE are quantized to generate a codebook and this is done using K-Means algorithm,Quantization has a good impact on the final result.
- Training: Here for each image in the dataset,SIFT features are extracted and then these features are described by SIFT and HUE descriptors.Then codebook is obtained using k-means which needs the number of clusters.For each image according to the codebook BOW model are constructed.
- Testing Stage: Retreiving-here images are retrieved or matched according to the similarity values.

3.3.3 Results

Using Both SIFT and HUE descriptors to classify an image into its class gave an average precision of 63.39%.The accuracy is less because during codebook generation it resulted in loss in quantization that can degrade the retrieval or matching performance.

3.4 Document Embedding Generation for Cyber-Aggressive Comment Detection using Supervised Machine Learning Approach

3.4.1 Author and publication details

Authors: Shylaja S S, Abhishek Narayanan , Abhijith Venugopal , Abhishek Prasad

Year:2019

3.4.2 Summary

Introduction:

The main purpose of this paper is to detect cyber bullying comments on social media despite of challenging task in this paper is to detect the sender's tone through his text message.For this one need to obtain manually labelled data. From various sources comments can be collected and can be labelled based on the cyber aggressiveness.These collected comments from various sources will be grouped into datasets ,and document embeddings will be generated which,then will be fed to the supervised machine learning algorithms for the training and prediction of test labels. Classification can been done by following 5 machine learning classification algorithms:

1. Support Vector Machines(SVM)
2. Logistic Regression
3. Bernoulli Naive Bayes Algorithm
4. Decision Trees
5. Random Forest Classifier:Evaluationmetrics

Various factors are considered as evaluation metrics such as Accuracy score, K-Fold Cross Validation, Confusion matrices, Precision, Recall, F-Beta-score, and Area under ROC Curve.

3.4.3 Results and conclusion:

Using all the algorithms for finding out the comments to be aggressive or non aggressive,we found that doc2vec approach along with svm classifier with rbf kernel has given a increased accuracy of 88.465% .Future work of this paper may include application of deep learning algorithms to the existing one for further optimization of the results.

3.5 Introduction to the special issue on deep learning for real-time information hiding and forensics

3.5.1 Authors and Publication details

Authors: Zhili Zhou1, Ching-Nung Yang2, Cheonshik Kim3, Stelvio Cimato4

Date:28 January,2020

3.5.2 Summary

Technology is evolving and growing rapidly. At the same time, the threat to multimedia data is also increasing due to various powerful multimedia processing tools. Due to which multimedia data can be forged and illegally copied. In order to prevent the above from unauthorized use, two approaches are proposed. They are, Information hiding digital forensics which compliment each other. Now let's look into Image forensics, Image steganography and steganalysis

Image forensics

Detection and classification of images can be done in 4 ways:

- Non-aligned double JPEG (NA-DJPEG) compression
- Real-time estimation for the parameters of Guassian filtering via deep learning
- A real-time image forensics scheme based on multi-domain learning
- Image steganography and steganalysis

a. Image steganography:The improvement of multimedia and deep learning technology carries new difficulties to steganography and steganalysis methods. To improve the correspondence dependability and proficiency for current ongoing vigorous steganography strategies, a connected code, made out of syndrome-trellis codes (STC) and cyclic redundancy check (CRC) codes, is proposed here.

b. Image steganalysis:Steganalysis is the contrary technique of steganography. Fundamentally, we attempt to distinguish the presence of steganographic content in a computerized gadget and furthermore find the hidden message.Steganography is a technique that transmits secret data or messages in an appropriate multimedia carrier, e.g., image, audio, and video files.

3.6 Image analysis of cyberbullying using machine learning techniques

3.6.1 Authors and Publication details

Author-Hao Li

Publication year-2016

3.6.2 Summary

- The intent of this paper is to find a prediction method to advise social media site users about what and what not to post in their accounts in order to avoid being bullied. Using ML algorithms they train and test models on data crawled from Instagram.
- Classification

  The image data is crawled from Instagram along with the metadata and other contextual information and also few of the related comments. They used Instagram API to achieve the same. According to a survey by Pew Research Center, Instagram is used by 52% teens and 40% claim it's the most used social media website.

- Labelling Instagram images-They had a two question questionnaire for labelers. They were shown the photos and set of comments. Following that two questions were asked.

  a) Are the comments has any bullying content or not?

  b) Does the image contain any bullying content. If yes is it because of the comments?

- The feature extraction techniques used in this paper include color histogram, edge detection coherence vector, scale invariant feature transform(SIFT),face features to describe the images and caption and user info were also considered as a auxiliary feature. All these features discussed above helped in detection of cyberbullying content when used along with the Machine Learning models.

- The supervised classification methods used are k-nearest neighbors, binary decision tree and SVM. Besides using different models they have also analyzed the feature vectors and their — combinations to get a better classification model and high accuracy and it was observed that SVM gave the best result. They were able to get an overall accuracy of 68.84% .

3.6.4 Research gaps

- Users post photos with different purposes. Bullying might occur because of the user who posted the photo or the content of the images that the commenters have strong sentiment about. This will increase the difficulty of this classification problem.
- Here they have asked the labelers to label the bullied/non-bullied images by going through the given set of images and comments. This would be affected by the commenter's preferences. If the labelers label the photos based only on the subjective opinions, the classification would become more accurate.

CHAPTER 4

PROJECT REQUIREMENTS SPECIFICATION

The high level design for our project includes the following

1)Preprocessing of the bullied and non-bullied images(Specifically resizing).

2)Augmentation of the images.

3) Training a CNN model to do the classification.

4) Building UI to show the above classification implemented.

Current System

A large portion of the current examinations have regular AI models and most of the created models in these investigations are versatile to a solitary informal community at once. Profound learning based models have discovered their way in the identification of digital harassing episodes, asserting that they can defeat the confinements of the traditional models, and improve the discovery execution. However, numerous oldschool models are accessible to control the incident, the need to viably arrange the tormenting is as yet weak. To successfully screen the harassing in the virtual space and to stop the dangerous consequence with the execution of Machine learning and Language handling. The parameters required are muddled to adjust and can be sometimes difficult to understand when bullied confusingly.

Literature survey on the topic was carried out and it was observed that most of them are text classification. Various ML algorithms where used to classify the same and a good accuracy has been achieved.In recent years there have been many researchers who are interested in cyberbullying detection on images. Here few works have been done using available feature extraction techniques and classification is done using ML and DL algorithms.

Design Considerations

1. Design Goals

Our technique utilizes an inventive idea of CNN for content examination. Anyway the current strategies utilize a guileless way to deal with furnish the arrangement with less precision. The dataset is utilized and our system is compared with other existing methods and is found to give better precision and grouping.

2. Architecture Choices

An alternate choice of feature extraction was considered but we feel our approach is better because

a) Selection of the type of feature extraction is very important and selected feature extractor may not work well with the all the images.

b) We need to do various trial and error method to get a good feature extractor to get a better accuracy. This might also involve combining different feature extraction technique and varifying the result.

CHAPTER 5
SYSTEM REQUIREMENTS SPECIFICATION

System Constraints Requires a high GPU memory since a lot of images have to be trained to have a better classification model.A high GPU memory affects the easy run on the local system and requires dependency on platforms like Google Colab.

Software Dependencies

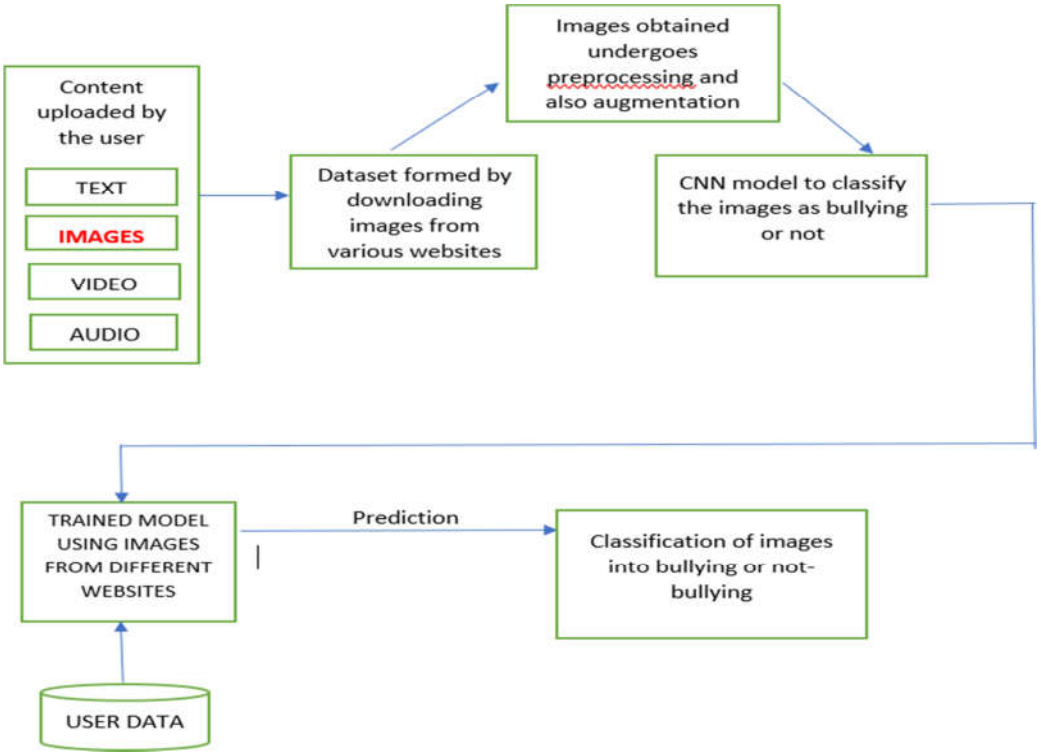
- Technologies: Python ,TensorFlow, scikit learn
- Tools: GPU by Google Colab , AnacondaNavigator
- Database: SQLite • Google Drive , GoogleColab
- Web Development: HTML, JavaScript

Hardware Dependencies

- Working PC
- RAM size > 8GB
- Internet connection

CHAPTER 6

SYSTEM DESIGN



CHAPTER 7

IMPLEMENTATION AND PSEUDOCODE

Crop_Images:

- Convert the image to grayscale, and blur it slightly
- Threshold the image, by erosions+dilations to remove noise in the image
- Find contours in threshold image, then grab the largest one
- Find the extreme points,crop new image out of the original image using the four extreme points (left, right, top, bottom)

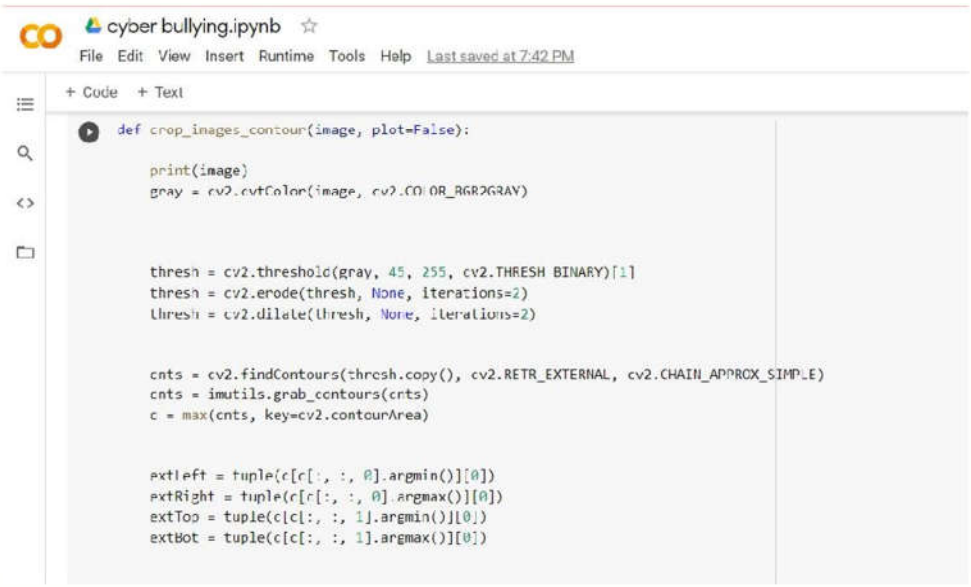


Fig.1

Load _Data:

Read images, resize and normalize them.

Arguments:

dir_list: list of strings representing file directories.

Returns:

X: A numpy array with shape = (#_examples, image_width, image_height, #_channels)

y: A numpy array with shape = (#_examples, 1)

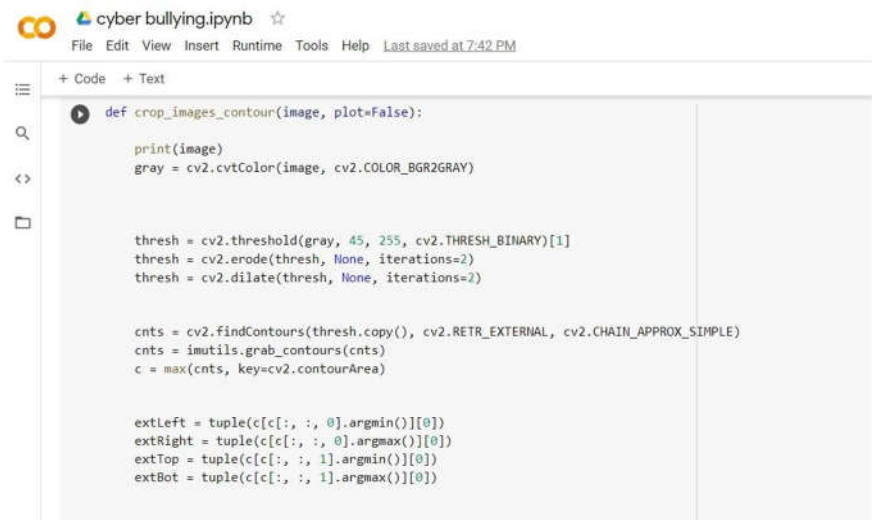


Fig.2

Augmentation:

Augmentation is used to artificially expand the dataset by loading the data this is done by using load_data function.

```
augmented_path = ''

augmented_yes = augmented_path + 'bullying images'
augmented_no = augmented_path + 'Non-bullying images'

IMG_WIDTH, IMG_HEIGHT = (240, 240)

X, y = load_data([augmented_yes, augmented_no], (IMG_WIDTH, IMG_HEIGHT))
```

Fig.3

CHAPTER 8
CONCLUSION OF CAPSTONE PROJECT PHASE-1

As a part of this semester, we explored the various available models for cyberbullying and the previous work done in this field,by doing thorough literature survey.Various ML and DL models are used for cyber bullying detection.Finally the CNN model is chosen after comparing with other models.

CNN is an Artificial neural network model which uses features of visual cortex for classification of images.CNN uses different layers of neural network which increases the accuracy of the model by inputting the output of one layer to the next layer of the network

Google colab was very convenient to use and install all the python deep learning libraries.

CHAPTER 9
PLAN OF WORK FOR CAPSTONE PROJECT PHASE-2

Our plan is to build a convolutional neural networks model from scratch for classification of images with more accuracy than available models.This would be done by doing many variations such as

a)Choosing the initial CNN layer and varying it accordingly.

b)Non-linearity introduction and checking the effect.

c)Usage of several pooling techniques

d) Variations in depth of the network layer

e) Variations of activation units and checking the effect.

Using the above variations,we check the accuracy and predict which is the one with best accuracy and improving the accuracy.

UI is made which when given any image tells whether its bullying or not using the underlying CNN model.

Once the model is done,we will try integrating with the public website and checking our model for any bullying content in that website.

REFERENCE / BIBLIOGRAPHY

[1] Despoina Chatzakou,Ilias Leontiadis,Jeremy Blackburn, Emiliano, De Cristofaro,Gianluca Stringhini,Athena Vakali,Nicolas Kourtellis-"Detecting Cyberbullying and Cyberaggression in Social Media"-

[2] Squicciarini,Rajtmajer,Caragea,HaotiZhong,HaoLi,Christopher Griffin,DavidMiller - "Content-Driven Detection of Cyberbullying on the Instagram Social Network"-

[3] Lu Cheng,Jundong Li,Yasin N,Silva Deborah,L. Hall Huan Liu-"XBully: Cyberbullying Detection within a Multi-Modal Context"-2019

[4] Zhili Zhou1 · Ching-Nung Yang2 · Cheonshik Kim3 · Stelvio Cimato- "Introduction to the special issue on deep learning for real-time information hiding and forensics" - 28 January 2020

[5] Hao Li-"Image analysis of cyberbullying using machine learning techniques"-2016

[6] Belhassen Bayar and Matthew C. Stamm-"Design Principles of Convolutional Neural Networks for MultimediaForensics"-2017

[7] Shylaja S S, Abhishek Narayanan , Abhijith Venugopal ,Abhishek Prasad-"Document Embedding Generation for Cyber-Aggressive Comment Detection using Supervised Machine Learning Approach"-2019

[8] Krishna B. Kansara and Narendra M. Shekokar-" A Framework for Cyberbullying Detection in Social Network"-2019

[9] Thomas Deselaers and Lexi Pimenidis and Hermann Ney-" LLM and BOVW Models for Adult Image Classification and Filtering"-2013

[10] N. Mansoori, M. Nejati, P. Razzaghi and S. Samavi-"Image Retrieval by Bag of Visual Words and Color Information"-2013

APPENDIX A

DEFINITIONS, ACRONYMS AND ABBREVATION

Abbreviations:

1) SIFT-Scale Invariant Feature Transform, The scale-invariant feature transform is a feature detection algorithm in computer vision to detect and describe local features in images.

2) BOVW-Bag Of Visual Words, the bag-of-words model sometimes called bag-of-visual-words model can be applied to image classification, by treating image features as words.

3) SVM-Support Vector Machine, a Machine learning algorithm.

4)CNN-Convolutional Neural Networks, a Deep learning algorithm.