

Portable Learning Approach towards Capturing Social Intimidating Activities using Big Data and Deep Learning Technologies

Mansi Mahendru^a and Sanjay Kumar Dubey^{b,*}

^aABES Engineering College, Ghaziabad, 201009, India

^bAmity University, Noida, 201301, India

Abstract

With the rise in the usage of different social media platforms, social intimidation has increasingly spread into these forums as it has given us endless chances to post anything for anyone. Previous studies have confirmed that exposure to this online social intimidation can have very serious offline consequences. With the growth of these multimodal social media platforms, there is an urgent requirement of some device methods for social intimidation detection and prevention. However, most of the prior research has focused on only textual posts for one or two topics of intimidation, namely sexism and racism. The principal objective of this research is to recognize social intimidation from multimodal posts such as text, memes, videos and audio and to target various social media networks such as Instagram, Twitter, and Facebook for several topics of harassment, namely religion based, personal attack, racism, sexism, physical appearance, etc. Previous research has stopped at detection, but this research has taken one step ahead to test the severity based on hate prediction score. The research study is performed using a combination of big data technology, namely Apache Spark, and several deep learning methods which are described below. The system is validated on five public datasets i.e., MLMA Hate Speech Dataset, MMHS150K Dataset, Hateful Memes Dataset, Instagram, Vine Dataset and measured on the basis of precision, recall and f1-score. Performance of the system has been inspected individually for every category of post under three subsections. The results attained specify that the proposed approach gives more feasible solution for social intimidation detection and its severity in online social networking platforms.

Keywords: social intimidation; multimodal posts; multiple topics; Apache Spark; deep learning; severity detection

(Submitted on June 19, 2022; Revised on July 15, 2022; Accepted on August 10, 2022)

© 2022 Totem Publisher, Inc. All rights reserved.

1. Introduction

Social media platform such as Facebook, Instagram, Twitter, Whatsapp, Reddit, etc. is in a great boom today. There is no confusion in saying that social media has made our lives easier. We can easily connect with anyone worldwide in no time. These social networking forums has completely changed our way of living and thinking about life. Everyone from a small child to the youth and old people are actively participating in the social media world. Apart from these advantages, several negative consequences are also noted in precedent years. One of the worst negative utilizations of these forums is social intimidation where any individual can harass, bully another person just for fun or by intention of taking revenge. Some scandalous persons use these platforms to embarrass other peoples in front of everyone without revealing their true identity. With the rise in the usage of social media and digital forums, comments, posts, and personal information shared by any individual can also be viewed by any outsider. The details that are shared online either it is any kind of personal information, or any pessimistic, evil content is forever saved in the records of social media sites and can be easily accessible by anyone. The online record saved in the database can be thought of as a permanent online reputation, persona of a person which can be smoothly accessible by any schools, organization, analyst, and others who want to search about the person whom they have to work with in the upcoming time. Past work on social intimidation detection has at least one of the following things missing. First, plenty of earlier research focuses on textual comments of the posts and proposed textual solutions only with limited accuracy [1-4]. There are less attempts made that use visual posts and none use audio and video posts for analysis of social intimidation. But it has been seen that modern social media platform is widespread with visual, audio and video posts. Although we cannot expect that video and audio features can replace textual features, it can add some complementary characteristics which can improve the overall achievement of the model [5]. Second, they address only one or two topics of

* Corresponding author.

E-mail address: skdubey1@amity.edu

social intimidation. Third, they target only one social media forum. Fourth they have to stop their research after detection of social intimidating posts but still there are more steps that are important to stop this crime. This research takes one move ahead and tries to overcome all the bottlenecks of the existing research. The contributions of this research are as follows:

- 1) To detect social intimidating content from multiple posts such as textual, memes, videos, and audio posts.
- 2) To target different types of social media platforms, namely Twitter, Facebook, and Instagram for multiple topics i.e., racism, sexism, physical appearance, personal attack, and religion-based etc.
- 3) To classify every incoming post using several big data technologies and to check severity of every post under three categories, namely less severe, moderately severe, and highly severe.

The rest of the study is arranged as follows. Section 2 related work, Section 3 methodology, Section 4 experimental work and analysis, and Section 5 conclusion and future scope.

2. Related Work

The learning of social intimidation detection has expanded promptly in the prior years. In this related work section, the primary goal is to examine various approaches adopted by past researchers for detecting social intimidation. Aluru et al [6] proposed a hate speech detection model using several deep learning algorithms for nine languages dataset. In order to validate the system for multiple languages they have used two types of embedding i.e., LASER and MUSE. Experimental outcomes show that in low resource settings, LASER embeddings with logistic regression performs better and with high resources while BERT performs better. Mossie and Wang [7] developed the online hate detection model for the vulnerable minority groups on social media networks. They have used Amharic dataset from Facebook and compared random forest, gradient boosted decision trees with RNN-GRU and RNN-LSTM algorithm. From the experiments RNN-GRU achieved the best accuracy was 97.85% as compared to other approaches. Florio et al [8] proposed an online hate speech detection model on TWITA dataset. During training they conducted two tests. As per the analysis, ALBERTO performed better than SVM on all datasets.

Rajamanickam et al [9] developed the abusive language detection model using emotion identification to have to have more auxiliary knowledge from the tweets. They have used two categories of datasets. First is the abuse detection category which covers threats, swear words, insults and swear words classification and another one contains tweets regarding sexism and racism. The second category is emotion detection dataset, which has 11 types of emotion classification. In their methodology they have differentiated two types of learning, namely single transfer learning (STL) and multi transfer learning (MTL) on different models. Experimental performance indicates the superiority of MTL over STL techniques. Samghabadi et al [10] advances their research by recognizing the connection between aggression detection and misogyny identification. The figures applied for this research covers three different languages i.e., Hindi, English and Bengali. The research is categorized into two sections i.e., subtask-A classes containing covertly aggressive, overtly aggressive, and non-aggressive. Subtask-B classes cover gendered and non-gendered. They use the BERT algorithm and achieve 0.8579 f-score.

Kumar and Sachdeva [11] proposed the cyberbully detection system using caps net DNN with convolutional neural network. The proposed methodology achieves 0.98 AUC on mix modal datasets having 10,000 comments and posts scraped from Twitter, Instagram and Facebook. Adikara et al [12] developed the model that will extract the harassing content from Instagram into two categories, cyberbullying and non-cyberbullying. They have used Indonesian language dataset of Instagram to train and test the model. In their research they have used naïve bayes along with lexicon-based features and BoW and achieved 0.872 accuracy, 0.948 precision, 0.824 recall, and 0.874 f-measure while testing. Kumari et al [13] developed an approach of social harassment detection using one layer of convolutional layer on both text and images. They have used three social media platforms. The proposed approach correctly identifies 74 % of bullying cases and achieves a 68.5% f-score.

Agrawal & Awaker [14] developed the cyberbully detection model using DNN for multiple social media forums, namely Wikipedia, Form Spring, and Twitter. They advance their research on three topics of cyberbullying, racism, sexism, and personal attack. They have also applied the plan of transfer learning in order to check whether performance of the model on one dataset can be used to strengthen the performance on another dataset. They achieved an f1-score of 0.94 for Twitter and Wikipedia dataset and 0.95 for Form Spring dataset using DNN and feature level transfer learning. Kumari & Singh [15] proposed a cyberbullying detection model that will extract features from both visuals as well as from text to identify various cases of harassment. They have used pre-trained VGG-16 network and CNN for feature collection and genetic algorithm for categorization. They achieve 0.80 precision which is nearly a 9% improvement over earlier research on the same dataset. Rezvani et al [16] develop the cyberbullying detection framework that will extract features from textual comments and from images using neural networks. They adopted this framework for analysing cyberbullying contents from Instagram and Twitter. They achieved 0.85 accuracy measure by applying LSTM+ Context 2 vec features.

Zinovyeva et al [17] proposed an antisocial behaviour detection model to take out hateful textual features. In their research they applied toxic comment classification dataset and compares several ML classifiers such as random forest, logistic regression, SVM, and deep learning classifiers namely CNN, LSTM, GRU. Based on the performance metrics deep learning classifier, LSTM and GRU achieves highest accuracy of 0.86. When tested on bi-directionality, BGRU got the highest accuracy of 0.861. Mohaouchane et al [18] uses four different neural networks for extracting offensive contents from social media. The system is trained and tested on an Arabic dataset of YouTube comments and achieved the highest recall value of 83.46 % using CNN-LSTM followed by 82.24% using CNN, Bi-LSTM with attention (81.51%) and Bi-LSTM (80.97%). Das et al [19] tries to solve the cyberbullying issue by predicting whether the meme is hateful or not. From the above review of past research papers, a conclusion has been drawn that most of the prior studies focus mainly on textual posts for one social media platform, covering two topics of harassment, racism and sexism, and stop their research after detection. All existing research are not flexible enough for the current situation of social intimidation. However, this research takes the work a step ahead by also checking for severity of posts in order to take actions in future

State-of-Art: This segment outlines the comparative analysis of the proposed approach of this work with the prior researchers who have used similar datasets used in this research work and target multimedia posts as shown in Table 1. From the above related work and state-of-art comparison it is clear that all existing studies have only focus on textual or image posts for two or three topics of harassment and no earlier studies have fulfilled objectives of this research using a combination of big data and deep learning technologies and achieve satisfactory results for future advancements.

Table 1. Comparison of proposed approach with other techniques

| Method /Year | Dataset Used | Technique | Objective | F-measure/ Accuracy |
|---------------------------------|--|---|---|---------------------|
| Proposed Approach | MMHS150K Dataset, Hateful Memes Dataset, Instagram Dataset, Vine Dataset | Apache spark, CNN-BERT, VGG-16, CNN-Bi-LSTM | a) Social Intimidation Detection from different categories of posts i.e., textual, memes, videos and audio post. b) To target different social media platforms such as Twitter, Facebook, Instagram and Wikipedia. c) To cover multiple topics of harassment i.e., racism, sexism, personal attack, religion-based and physical appearance. d) To check the severity of the posts on the basis of hate prediction score. | 95.24% |
| Aluru <i>et al.</i> 2020 | Twitter dataset for nine languages | CNN-GRU, BERT, mBERT | To develop hate speech detection model for nine languages using sixteen data sources. | 0.8365 |
| Mossie and Wang 2020 | Amharic text dataset from facebook | Random forest, Gradient boosted decision trees, RNN-GRU, RNN-LSTM algorithm | Hate speech detection model for minority groups on social media platforms like Facebook. | 0.9785% |
| Florio <i>et al.</i> 2020 | TWITA dataset i.e. twitter dataset having Italian tweets | SVM and ALBERTo classifiers | Hate speech prediction system w.r.t different size of training dataset and to analyse the performance of classifier when language and topic changes over time. | 0.694 |
| Rajamanickam <i>et al.</i> 2020 | OffensEval dataset and Waseem & Hovy dataset | MTL and STL on Bi-LSTM classifier | Online hate speech detection model that will check the relation between online abuse comments and the emotion behind it for racism and sexism category. | 81.28% |
| Samghabadi <i>et al.</i> 2020 | English, Hindi, and Bengali Dataset having comments on facebook | BERT algorithm | To develop hate speech detection model that will identify the relation between aggression and misogyny for three languages dataset and annotate each comment into one of three aggression categories i.e., NA, CA, and OA and in one of two categories of misogyny namely gendered and non-gendered. | 0.8579 |
| Kumar, & Sachdeva. 2021 | Facebook, Instagram, YouTube and Twitter | caps net DNN + CNN | Cyberbullying detection system from textual and visual posts. | 98.0 % |
| Adikara <i>et al.</i> 2020 | Indonesian language dataset of instagram | Naïve bayes along with lexicon based features and bag of words | Cyberbully detection from comments posted on Instagram and classify in two categories, cyberbullying or non-cyberbullying | 87.2 % |
| Kumari <i>et al.</i> 2020 | Facebook, twitter and instagram | Single Layer CNN | Social harassment Detection using text and visual categories of post from three social media platform. | 68.5% |
| Agrawal & Awaker 2018 | Wikipedia, formspring and twitter Dataset | DNN algorithm | Cyberbully detection from textual comments for three topics of harassment i.e., racism, sexism, and personal attack | 0.95 |

3. Proposed Methodology

This research aims to identify social intimidating content from the posts using several deep learning and big data technologies. The entire system expects multimodal posts as input and calculate hate prediction scores as output to check the severity of the

input post. In this segment the all-inclusive description of the dataset and models used in the research is presented. The overall architecture of the proposed method is shown in Figure.1. The detailed explanation of every step is given below.

3.1. Data Collection

The training dataset plays a very important role in the entire performance of the model. There are many datasets available from several social media forums [20]. The main concern of this research is to identify social intimidating content from multimodal posts i.e., textual, memes, videos, and audio posts and to target several social media networks for various topics of intimidation. Based on the above listed objectives following datasets are chosen for further processing.

3.1.1. MLMA Hate Speech Dataset

Identification of offensive words is the initial step to start with the social intimidation detection from multimodal posts. This is a publicly available dataset containing 5,647 English tweets, 4,014 French tweets, and 3,353 Arabic tweets. As per the research objectives, only English tweets are used to train the model on multiple topics of intimidation i.e., sexism, racism, disability, appearance related, religion based, and combined [21].

3.1.2. MMHS150K Dataset

MMHS150K is a publicly available multimodal dataset having both text and images collected from Twitter having 1, 50,000 tweets [22]. Out of all tweets, 36,978 were hate tweets and 112,845 were non hate tweets. Then, 11,925 were divided in a racist category, 3,495 under sexist category, and 3,870 homophobic, 163 religion-based, and 5,811 attack to other communities. The main benefit of using this dataset is that this dataset is larger in size and mainly focuses on six categories of harassment. This dataset includes memes that help in training of every potential variation of visual and text in the memes that could reflect in the overall performance of the model.

3.1.3. Hateful Memes Dataset

Facebook AI developed the multimodal dataset for detection of online abuse containing images and text. The prime goal behind the creation of this dataset is to reduce the biases that the model can simply pick on for e.g., black and white, together written easily predict that hateful content is present. To mitigate this issue, memes are reconstructed by using different text without loss of information. The new underlying memes were prepared in cooperation with Getty visuals to allow redistribution of research purposes. It contains 10k memes with 10% test set and 5% dev. [23].

3.1.4. Instagram Dataset

Instagram is an extremely famous social media forum that is well suited for calculating the severity of posts as it is one of the platforms where there are early postings of images with captions followed by later text comments that form the basics of social intimidation. Instagram dataset is collected by [24,25] originally contains 2,000 media sessions including the posted meme with the caption, comments from users. They have used a snowball sampling method to identify the id of Instagram users. In this work, 699 media sessions are considered as their images are only accessible by Instagram URL.

3.1.5. Vine Dataset

Vine [26,27] is a mobile application that gives you permission to upload six second videos, like, and comment on other user videos. For each video, the dataset also consists of time, content, user id, profile description, the ids of their followers, and followings.

3.1.6. Proposed Dataset

From all above datasets used, the machine is validated on various topics and for multiple platforms. To examine the performance of the system more effectively, we construct a dataset for social intimidation detection in memes with 6,000 images using Google images download tool.

3.2. Pre-Processing

As per the stated objectives, the model must be trained with sizeable amounts of unstructured data i.e., visuals, audio, videos or structured data with a numerous of attributes. One of the prime concerns of this work is to handle multimodal posts that

are actively posted on any social media platform, therefore pre-processing varies for structured such as textual posts and unstructured such as visual, video, audio posts.

3.2.1. Processing of Structured Posts

Most of the information present on every social media platform is in the form of text for comments on posts, captions on memes, etc. Nowadays everyone uses unstructured phrases over social media which are not exactly words but short abbreviations such as How r u? , Gr8, f 9, etc. and is very well liked in all platforms. To manage this sort of data, pre-processing has to be done effectively in order to remove all noisy words, missing words, biasness, etc. without any loss of actual information. To resolve the problem while training with large dataset, there are many pre-processing techniques available [28,29], although these techniques are not beneficial and efficient when dealing huge data. There are many deep learning distributed frameworks like tensor flow, hovord [30,31]. These approaches may not give the best pre-processing results when data comes in large volumes of batches. This problem can be solved using apache spark as it can handle any type of vague data and allows for reusing and caching in memory for numerous clusters [32]. Using Apache Spark platform pre-processing of textual contents is performed as follows:

- A. Initially all the raw text, alphanumeric characters, white spaces etc. which bare no meaning for analysis are cleaned using Apache Spark SQL functions named as column-based user defined function (UDF).
- B. The words appearing very frequently in the comments (for e.g. a, at, is,) are excluded by using (spark.ml.feature.StopWordsRemoverTransformer).
- C. From step A and B, stop words and noisy words are excluded, then the whole text is converted into small tokens in order to isolate as much sentiment information as possible. Tokenization was carried out using the Apache Spark Tokenizer and regex pattern matching which comes in package with word tokenization feature (spark.ml.feature.Tokenizer).
- D. After tokenization every word is individually processed but some other challenges are still left as on social media it is in fashion to use abbreviated words like how r u, thnx, think's, etc. hence requires text normalization which will replace the short abbreviations with the actual words they represent. This step is carried out using Apache Spark text normalizer feature transformer to normalize each vector.
- E. After text normalization every token is assigned with the grammatical category i.e., noun, adjective, verb, conjunction etc. in order to recognize the root of every word. This step helps to identify important words which in later stage reduce the time for processing while doing feature extraction. POS tagging and lemmatization is performed using Apache Spark NLP libraries namely open NLP, NLTK, WordNet.

3.2.2. Processing of Un-Structured Posts

Social media data is growing in mass by a second into huge and therefore makes it very challenging to analyse. Recently visual communication using images and video has increased tremendously on social media platforms such as Instagram, Facebook, Twitter, etc. Unstructured data namely images, audio, and video data are particularly more powerful as they deliver sentiments and emotions better [33]. If a single node architecture is used to handle such complex and huge data, then it takes extremely long time to finish pre-processing and make it ready for further training. There are many pre-processing techniques used in past research to handle unstructured data [34-36]. But these technologies are inefficient in handling big data. To handle this issue, Spark shines as an excellent choice which not only works with structured data but can also deal with any type of data like jpeg, png, bmp, tiff, mp4, mp3, etc. by distributing operations in parallel on all nodes of the cluster. Spark framework is a flexible platform that supports every function of python by making use of UDF (User defined Function). Initially all the data in spark breaks in spark data frame, then spark breaks this data frame into multiple batches afterwards every batch can be passed to a python in the form of python UDF. From this step of python UDF any operation can be performed in a distributed manner across all the nodes [37]. Using Apache Spark platform pre-processing of unstructured posts is done as follows:

- A. **Image posts:** Varied memes with captions keeps the audience interested and helps to remember the post for an extended time period. There are many types of noise present in the visual posts i.e., distortion of text and visual, haziness, poor resolution, etc. [38]. For pre-processing many techniques are analysed and after examining all OpenCV is finalized [39]. To pre-process a huge number of images in a distributed manner, Apache Spark is integrated with computer vision using mmlspark.opencv.ImageTransformer [40].
- B. **Audio Posts and Video Posts:** Video posts on social media platforms comprise of running visual and audio which act as an effective way to highlight the visual content with audio speech. Images in video posts are pre-processed in a similar manner as described above for image posts. Apart from the speech behind the running visual in video posts, there are many audio clips that are sent on social media. To reduce noise from audio files, python package librosa is used to extract features and spectrograms for reducing the background noise.

3.3. Feature Collection

The output produced from the pre-processing module will be the input for the feature collection module where the important contents from the input posts are collected and sent for next processing steps. In this section the description of how important characteristics are extracted from all categories of input posts is explained, which can be used afterwards to perform classification to check its hate prediction score and severity of input post. The detailed explanation of feature collection from multi modal input posts is given below.

3.3.1. Feature Collection from Textual Posts

Textual posts are a kind of data that has not been structured in a pre-defined manner and is typically text heavy with the number of features. To extract important features from huge amount of textual conversation is challenging task without artificial intelligence involvement. There are many feature collection techniques used in past research such as TF-IDF, N-gram, bag-of-words, word 2 vec, glove, etc. [41-44]. But all the existing approaches consider every word individually therefore create a single vector for the words with the same name but different sentiments. All these approaches generate the word vector by considering a single word instead of the whole sentence due to which there is loss of semantics. To overcome these challenges and based on the research objectives of handling huge datasets, BERT with CNN is used for feature collection. BERT has the capability to capture long term relation and semantics behind the sentences, and it can enhance the generalization of word vector model so which can more accurately highlight the important features from the sentences [45]. CNN has made remarkable performance in natural language processing as it can generate features automatically and can be easily combined with any classifier [46]. BERT-CNN model mainly consists of two parts. The first part is BERT base model which consists of input sequence and 12-layer transformers, each transformer consisting of a self-attention sub layer with multiple attention heads [47]. The other part is CNN which is applied as a feature extractor to generate feature maps. Once the text is pre-processed from above steps it was set to have a maximum length of tokens to 64, then the text was sent as input to BERT. The last four layers of BERT were concatenated to get word vector representations. Next these vectors are passed to three-layer CNN which is responsible to process the text and to collect features from them. Three convolutional layers of the network were followed by a single max pooling layer. Filters of 8, 16, and 64 each of size 4 were set on three convolutional layers respectively. After the third convolutional layer, a max pooling layer of size 4 is applied to extract important features from it. Then at last one flattens layer followed by two convolutional layers of size 256 and 2 is applied in order to convert pooled map to a single column that can be proceed through a dense layer and sigmoid activation function to get a final output layer which can be passed to any classifier for further classification.

3.3.2. Feature Collection from Memes

Nowadays memes are spreading very rapidly and causing a huge effect on society. Memes consist of two elements i.e., text present in the caption and the image. In order to detect social intimidation from memes we have to compute hate prediction scores of both text and visuals individually, as well as in a combined manner to test every possible contradiction between text and visual. First component consists of the process for extracting the text from the visual using some optical scanner technique. Prior to this work we have developed the real time intelligent optical scanner that can extract all possible text from input visual. The approach used in the research has given better and satisfactory results as compared to other methods used in past research for text extraction. Text extraction from input visual involves a sequence of steps such as text detection and text recognition. Text detection is the operation of extracting regions of interest i.e., regions where there is a possibility of presence of text. For text detection EAST algorithm is used, which also works in a similar manner as CNN but leaves some intermediate steps and makes detection faster and more accurate. The output of the text detection step is the creation of bounding boxes over the region where text is present. Once location of text is detected then the next step is text recognition. Text recognition is performed using CNN-BiLSTM as directed by [48,49]. CNN is applied to learn features from different words according to the location and also used in text classification. But with images as input, some challenges are still left after pre-processing as it may be possible that some characters of the text are partially visible or missing. In that case, CNN will only label the text detectable but not able to predict the missing characters. To conquer this challenge the classified text from CNN is passed as input to Bi-LSTM layer in order to check the long-term dependency between the texts and to check the correctness of the recognized text from a forward as well as from a backward direction. For a detailed description of every step, refer [50]. Once complete, text is recognized from the input meme then it is passed to the classification module. Second component of the meme is the visual. There are many components linked with social intimidation in images. Recently very slighter attempts have been made by the researchers in order to estimate the visual features for identification of social harassment [51]. There might be many potential variations between text and visual in the memes posted on social media that can affect the overall performance of the model. Possible variations are given below:

- A. Text is non-harassing while visual is harassing but whole meme is harassing
- B. Text is non-harassing while visual is harassing but whole meme is non-harassing, etc.

To tackle all these challenges some factors related to the visual such as facial emotions, hand gesture, body posture should be considered. As a baseline model CNN is used to extract low level features from visuals as it is still a very successful model for image-based tasks. There are many convolutional networks that can be applied for extracting visual factors from the memes [52]. After analysing all the networks, VGG-16 is finalized for image feature extraction [53-55]. VGG-16 is a pre-trained object detection model on an image net dataset having 14 million images and nearly 1,000 objects classification categories [56]. The basis of every image processing technique is made on the actuality that the model should be trained in a way that it can understand every object existing in the visual. This fine-tuned VGG-16 model is applied to train the system for varied object categories in less time. There might be certain objects such as knives, guns, people, etc. that are responsible for social intimidation. All the visuals were resized into 224x224 pixels and then passed to the VGG-16 model which has a total of 16 layers, out of which 13 are convolutional layers and three fully connected (dense) layers. Visual feature extraction is done by VGG-16 as directed by [57]. An image is passed to convolutional layers with 3x3 filters and max-pooling layer followed by flatten layer and four fully connected layers. The first and second fully connected layers are of size 4096, third layer is of size 128, and fourth one of size 3 to have multi-way classification. The configuration of the VGG network is fixed till the second last dense layer. The weights of the second to last layer and output layer are adjusted by passing the visual via a network. The activation functions are applied at the output layer in order to generate the feature vectors of the visual. Once the visual characteristics are extracted, it is passed to the classification module. From the above steps the text and visual features are individually passed to the further steps but in some cases as shown above when both combined, indicating some features may be contradicting each other. Therefore, features obtained from text as well as from visual are concatenated to obtain the combined features set of 4,864 dimensions. This multimodal representation afterward fed as input to a classification module for calculation of overall hate prediction score and for testing the severity of input post.

3.3.3. Feature collection from Video and Audio Posts

Video posts also cause large problems in terms of both emotional and psychological means. Video posts on social media consist of two things i.e., running visual and audio. Audio is an extremely essential part of video as it expresses the emotional information independently and helps to understand facial expressions more easily. The useful voice after pre-processing is converted from speech to text using the pyaudio package of python and then handled in a similar way as (3.3.1). The dataset used in this research for video posts also consist of associated comments of users which are handled in a similar manner as section (3.3.1). Video feature collection is not that different from image feature collection. For the images, feature extractors are used to collect important features and then send them to the classifier for further classification. In videos prior to feature collection, frames are to be extracted and then similar steps are to be followed as for image classification. Video social intimidation detection is done using shot frontier detection algorithm [58]. Videos are made by physical related frame sequences and based on the complexity of the features in the video, where multiple key frames are to be extracted. To overcome the challenges of video posts initially video is broken into scenes, shots, and frames and therefore using the shot frontier detection algorithm, key frames are extracted. The extracted frames are proceeded to the feature collection module and processed in a similar manner as the visual posts, described in section (3.3.2).

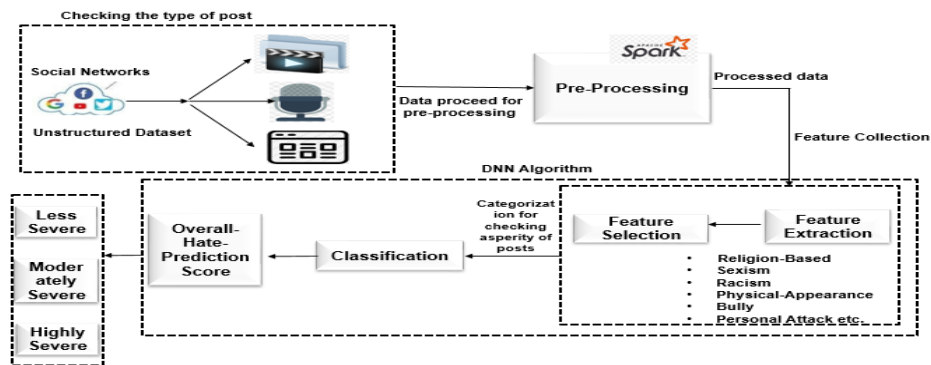


Figure 1. Architecture of proposed methodology

3.4. Feature Optimization and Classification

From the above step important features from all categories of posts are collected and passed to the feature optimizer and classification module in order to perform variable selection and to calculate hate prediction score. Based on the hate prediction score the post is classified under three severity categories, less severe, moderately severe, and highly severe. Feature vectors attained from all categories of post are afterward fed as input into Bi-directional recurrent neural networks (Bi-RNN) with ReLU activation function [59,60]. The last single neuron was added which has no activation function at the end in order to

forecast the hate prediction score. On the basis of overall hate prediction score the scale is set in order to categorize the post under three severity categories. If the score is from 0.0-0.3 it comes in the less severe category, from 0.4-0.7 moderately severe, and from 0.8-1.0 highly severe.

4. Experimental Work and Analysis

This section outlines the results of the various experimental combinations for the social intimidation detection from multimodal post categories by applying proposed methods. To evaluate the performance of the proposed system, metrics such as precision, recall, and f1-score is calculated.

4.1. Text Post as Input

Today everyone finds memes, video, and audio posts more interesting than textual posts as visuals in the posts add more wealth to the information, but it doesn't mean that these categories of posts can replace textual posts. So, to make the machine understand the difference between every word, the MLMA dataset is used. Table 2 and 3 shows the performance of the proposed approach in calculating the severity of textual posts. In Table 2, the highlighted words are the important features extracted by the machine which refers to the words used for intimidation.



Table 2. Hate prediction score and severity category for textual posts as input

| Input Post | Category | Hate Prediction score | Severity Category | Performance metrics (Average) |
|---|----------------------------|-----------------------|-------------------|---|
| Post: fXX hate lanky ginger cunt joseph brady | Appearance related +Sexism | 0.85 | Highly Severe | Precision = 0.95 Recall = 0.93 F1-Score= 0.93 |
| Post: America another 8 years Obama's ideology via hillary we'd well way shithole country | Political | 0.43 | Less Severe | |

4.2. Image Based Meme as Input

Based on the above-mentioned methodology for image-based memes as input on MMHS150K, hateful memes and Instagram dataset, the system is trained and tested for multiple topics of harassment i.e., racism, sexism, religion-based, personal attack, homophobic and physical appearance. Table 3 shows the hate prediction score and severity category for some sample results for above criteria.

Table 3. Hate prediction score and severity category for image based meme as input

| Input Post | Category | Hate-Prediction Score based on Text in Meme | Hate-Prediction Score Based on Visual in Meme | Overall Hate Prediction Score | Severity Category | Performance metrics (Average) |
|---|---------------------|---|---|-------------------------------|-------------------|---|
|  | Personal Attack | 0.24 | 0.56 | 0.43 | Less-Severe | Precision = 0.94 Recall = 0.90 F1-Score= 0.91 |
|  | Physical-Appearance | 0.32 | 0.69 | 0.58 | Moderately Severe | |

4.3. Video Post as Input

Video based communication is gaining huge popularity recently in various social media platforms as it represents the actual target for investigating social intimidation. To target video posts, the system is trained and tested on a vine dataset which provides the opportunity to detect intimidation from video along with comments associated with audio which helps to get emotions behind the voice, as shown in Table 4.

Table 4. Hate prediction score and severity category for video post as input

| Input Post | Category | Overall Hate Prediction Score | Severity Category | Performance metrics (Average) |
|------------|----------|-------------------------------|-------------------|---|
| | Sexism | 0.61 | Moderately Severe | Precision = 0.94 Recall = 0.92 F1-Score= 0.92 |
| | Racism | 0.82 | Highly-Severe | |

5. Conclusion

With the growing acceptance of various social media platforms, social intimidation has become more common and begun to uplift serious consequences. Although most of the previous efforts have been based primarily on textual posts on one social media forum and for one or two topics of harassment i.e., racism and sexism. Therefore, these methods are not flexible in today's modern social media platform and thus fail to consider multimodal nature of social media. The proposed framework for this research is based on the belief that multimodal details can offer more valuable insights for social intimidation detection and can complement and extend the previous work. This paper aims to solve all the challenges left in existing work and to take one step ahead by targeting multi-modal social media posts like textual, memes, video and audio posts for multiple topics of harassment such as religion-based, racism, physical appearance, personal attack, sexism, etc. using data from three social media platforms, namely Instagram, Facebook, and Twitter. The key concern of this research is to test the severity of every incoming post on the basis of hate prediction score in order to mark high priority posts on which further actions are to be taken in future. To propose an innovative social intimidation model, a combination of several big data and deep learning technologies were used. In the future, we try to enlarge this work by scrapping user personal information, in order to take further actions on moderately and highly severe posts. Furthermore, multilingual multi-modal posts are also used to enhance its flexibility.

References

1. Van Hee, C., Jacobs, G., Emmery, C., Desmet, B., Lefever, E., Verhoeven, B., De Pauw, G., Daelemans, W., and Hoste, V. Automatic Detection of Cyberbullying in Social Media Text. *PLoS one*, vol. 13, no. 10, pp. 0203794, 2018.
2. Singh, V.K., Ghosh, S., and Jose, C. Toward Multimodal Cyberbullying Detection. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, pp. 2090-2099, 2017.
3. Van Bruwaene, D., Huang, Q., and Inkpen, D. A Multi-Platform Dataset for Detecting Cyberbullying in Social Media. *Language Resources and Evaluation*, vol. 54, no. 4, pp.851-874, 2020.
4. Mahendru, M. and Dubey, S.K. Performance Analysis of Various Classifiers for Social Intimidating Activities Detection. In *International Conference on Advances in Computing and Data Sciences*, Springer, Cham, pp. 512-527, 2021.
5. Cheng, L., Li, J., Silva, Y.N., Hall, D.L., and Liu, H. Xbully: Cyberbullying Detection within a Multi-Modal Context. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, pp. 339-347, 2019.
6. Aluru, S.S., Mathew, B., Saha, P., and Mukherjee, A. Deep Learning Models for Multilingual Hate Speech Detection. *arXiv preprint arXiv:2004.06465*, 2020.
7. Mossie, Z. and Wang, J.H. Vulnerable Community Identification using Hate Speech Detection on Social Media. *Information Processing & Management*, vol. 57, no. 3, pp. 102087, 2020.
8. Florio, K., Basile, V., Polignano, M., Basile, P., and Patti, V. Time of Your Hate: The Challenge of Time in Hate Speech Detection on Social Media. *Applied Sciences*, vol. 10, no. 12, pp. 4180, 2020.
9. Rajamanickam, S., Mishra, P., Yannakoudakis, H., and Shutova, E. Joint Modelling of Emotion and Abusive Language Detection. *arXiv preprint arXiv:2005.14028*, 2020.
10. Samghabadi, N.S., Patwa, P., Pykl, S., Mukherjee, P., Das, A., and Solorio, T. Aggression and Misogyny Detection using BERT: A Multi-Task Approach. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pp. 126-131, 2020.

11. Kumar, A. and Sachdeva, N. Multimodal Cyberbullying Detection using Capsule Network with Dynamic Routing and Deep Convolutional Neural Network. *Multimedia Systems*, pp. 1-10, 2021.
12. Adikara, P.P., Adinugroho, S., and Insani, S. Detection of Cyber Harassment (Cyberbullying) on Instagram using Naïve Bayes Classifier with Bag of Words and Lexicon based Features. In *Proceedings of the 5th International Conference on Sustainable Information Engineering and Technology*, pp. 64-68, 2020.
13. Kumari, K., Singh, J.P., Dwivedi, Y.K., and Rana, N.P. Towards Cyberbullying-Free Social Media in Smart Cities: A Unified Multi-Modal Approach. *Soft computing*, vol. 24, no. 15, pp. 11059-11070, 2020.
14. Agrawal, S. and Awekar, A. Deep Learning for Detecting Cyberbullying across Multiple Social Media Platforms. In *European conference on information retrieval*, Springer, Cham, pp. 141-153, 2018.
15. Kumari, K. and Singh, J.P. Identification of Cyberbullying on Multi-Modal Social Media Posts using Genetic Algorithm. *Transactions on Emerging Telecommunications Technologies*, vol. 32, no. 2, pp. 3907, 2021.
16. Rezvani, N., Beheshti, A., and Tabebordbar, A. Linking Textual and Contextual Features for Intelligent Cyberbullying Detection in Social Media. In *Proceedings of the 18th International Conference on Advances in Mobile Computing & Multimedia*, pp. 3-10, 2020.
17. Zinovyeva, E., Härdle, W.K., and Lessmann, S. Antisocial Online Behavior Detection using Deep Learning. *Decision Support Systems*, vol. 138, pp. 113362, 2020.
18. Mohaouchane, H., Mourhir, A., and Nikolov, N.S. Detecting Offensive Language on Arabic Social Media using Deep Learning. In *2019 sixth international conference on social networks analysis, management and security (SNAMS)*, IEEE, pp. 466-471, 2019.
19. Das, A., Wahí, J.S., and Li, S. Detecting Hate Speech in Multi-Modal Memes. *arXiv preprint arXiv:2012.14891*, 2020.
20. Madukwe, K., Gao, X., and Xue, B. In Data We Trust: A Critical Analysis of Hate Speech Detection Datasets. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pp. 150-161, 2020.
21. Ousidhoum, N., Lin, Z., Zhang, H., Song, Y., and Yeung, D.Y. Multilingual and Multi-Aspect Hate Speech Analysis. *arXiv preprint arXiv:1908.11049*, 2019.
22. Gomez, R., Gibert, J., Gomez, L., and Karatzas, D. Exploring Hate Speech Detection in Multimodal Publications. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 1470-1478, 2020.
23. Kiela, D., Firooz, H., Mohan, A., Goswami, V., Singh, A., Ringshia, P., and Testuggine, D. The Hateful Memes Challenge: Detecting Hate Speech in Multimodal Memes. *Advances in Neural Information Processing Systems*, vol. 33, pp. 2611-2624, 2020.
24. Hosseinmardi, H., Mattson, S.A., Ibn Rafiq, R., Han, R., Lv, Q., and Mishra, S. Analyzing Labeled Cyberbullying Incidents on the Instagram Social Network. In *International conference on social informatics*, Springer, Cham, pp. 49-66, 2015.
25. Hosseinmardi, H., Rafiq, R.I., Han, R., Lv, Q., and Mishra, S. Prediction of Cyberbullying Incidents in a Media-based Social Network. In *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, IEEE, pp. 186-192, 2016.
26. Rafiq, R.I., Hosseinmardi, H., Mattson, S.A., Han, R., Lv, Q., and Mishra, S. Analysis and Detection of Labeled Cyberbullying Instances in Vine, a Video-based Social Network. *Social network analysis and mining*, vol. 6, no. 1, pp. 1-16, 2016.
27. Rafiq, R.I., Hosseinmardi, H., Han, R., Lv, Q., Mishra, S., and Mattson, S.A. Careful What You Share in Six Seconds: Detecting Cyberbullying Instances in Vine. In *2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, IEEE, pp. 617-622, 2015.
28. Adnan, K. and Akbar, R. An Analytical Study of Information Extraction from Unstructured and Multidimensional Big Data. *Journal of Big Data*, vol. 6, no. 1, pp. 1-38, 2019.
29. Mayer, R. and Jacobsen, H.A. Scalable Deep Learning on Distributed Infrastructures: Challenges, Techniques, and Tools. *ACM Computing Surveys (CSUR)*, vol. 53, no. 1, pp. 1-37, 2020.
30. Sergeev, A. and Del Balso, M. Horovod: Fast and Easy Distributed Deep Learning in Tensorflow. *arXiv preprint arXiv:1802.05799*, 2018.
31. Zhang, Q., Yang, L.T., Chen, Z., and Li, P. A Survey on Deep Learning for Big Data. *Information Fusion*, vol. 42, pp. 146-157, 2018.
32. Dai, J.J., Wang, Y., Qiu, X., Ding, D., Zhang, Y., Wang, Y., Jia, X., Zhang, C.L., Wan, Y., Li, Z., and Wang, J. Bigdl: A Distributed Deep Learning Framework for Big Data. In *Proceedings of the ACM Symposium on Cloud Computing*, pp. 50-60, 2019.
33. Kumar, A., Sangwan, S.R., and Nayyar, A. Multimedia Social Big Data: Mining. In *Multimedia big data computing for IoT applications*, Springer, Singapore, pp. 289-321, 2020.
34. Paulin, H., Milton, R.S., and JanakiRaman, S. Efficient Pre Processing of Audio and Video Signal Dataset for Building an Efficient Automatic Speech Recognition System. *International Journal of Pure and Applied Mathematics*, vol. 119, no. 16, pp. 1903-1910, 2018.
35. Chaki, J. and Dey, N. A Beginner's Guide to Image Preprocessing Techniques. CRC Press, 2018.
36. Effrosynidis, D., Symeonidis, S., and Arampatzis, A. A Comparison of Pre-processing Techniques for Twitter Sentiment Analysis. In *International Conference on Theory and Practice of Digital Libraries*, Springer, Cham, pp. 394-406, 2017.
37. Zečević, P., Slater, C.T., Jurić, M., Connolly, A.J., Lončarić, S., Bellm, E.C., Golkhou, V.Z., and Suberlak, K. Axs: A Framework for Fast Astronomical Data Processing based on Apache Spark. *The Astronomical Journal*, vol. 158, no. 1, pp. 37, 2019.
38. Sontakke, M.D. and Kulkarni, M.S. Different Types of Noises in Images and Noise Removing Technique. *International Journal of Advanced Technology in Engineering and Science*, vol. 3, no. 1, pp.102-115, 2015.
39. Singh, H. Advanced Image Processing using Opencv. In *Practical Machine Learning and Image Processing*, Apress, Berkeley,

- CA, pp. 63-88, 2019.
40. Hamilton, M., Raghunathan, S., Annavajhala, A., Kirsanov, D., Leon, E., Barzilay, E., Matiach, I., Davison, J., Busch, M., Oprescu, M., and Sur, R. Flexible and Scalable Deep Learning with MMLSpark. In *International Conference on Predictive Applications and APIs*, PMLR, pp. 11-22, 2018.
 41. Ahuja, R., Chug, A., Kohli, S., Gupta, S., and Ahuja, P. The Impact of Features Extraction on the Sentiment Analysis. *Procedia Computer Science*, vol. 152, pp. 341-348, 2019.
 42. Fortuna, P. and Nunes, S. A Survey on Automatic Detection of Hate Speech in Text. *ACM Computing Surveys (CSUR)*, vol. 51, no. 4, pp. 1-30, 2018.
 43. Salminen, J., Hopf, M., Chowdhury, S.A., Jung, S.G., Almerexhi, H., and Jansen, B.J. Developing an Online Hate Classifier for Multiple Social Media Platforms. *Human-centric Computing and Information Sciences*, vol. 10, no. 1, pp. 1-34, 2020.
 44. Liang, H., Sun, X., Sun, Y., and Gao, Y. Text Feature Extraction based on Deep Learning: A Review. *EURASIP journal on wireless communications and networking*, vol. 2017, no. 1, pp. 1-12, 2017.
 45. He, C., Chen, S., Huang, S., Zhang, J., and Song, X. Using Convolutional Neural Network with BERT for Intent Determination. In *2019 International Conference on Asian Language Processing (IALP)*, IEEE, pp. 65-70, 2019.
 46. Jogin, M., Madhulika, M.S., Divya, G.D., Meghana, R.K., and Apoorva, S. Feature Extraction using Convolution Neural Networks (CNN) and Deep Learning. In *2018 3rd IEEE international conference on recent trends in electronics, information & communication technology (RTEICT)*, IEEE, pp. 2319-2323, 2018.
 47. Safaya, A., Abdullatif, M., and Yuret, D. BERT-CNN for Offensive Speech Identification in Social Media. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation", Barcelona (online)", International Committee for Computational Linguistics*.
 48. Zhang, J., Li, Y., Tian, J., and Li, T. LSTM-CNN Hybrid Model for Text Classification. In *2018 IEEE 3rd Advanced Information Technology, Electronic and Automation Control Conference (IAEAC)*, IEEE, pp. 1675-1680, 2018.
 49. Rhanoui, M., Mikram, M., Yousfi, S., and Barzali, S. A CNN-BiLSTM Model for Document-Level Sentiment Analysis. *Machine Learning and Knowledge Extraction*, vol. 1, no. 3, pp. 832-847, 2019.
 50. Mahendru, M., Dubey, S.K., and Gaur, D. Deep Convolutional Sequence Approach Towards Real-Time Intelligent Optical Scanning. *International Journal of Computer Vision and Image Processing (IJCVIP)*, vol. 11, no. 4, pp. 63-76, 2021.
 51. Vishwamitra, N., Hu, H., Luo, F., and Cheng, L. Towards Understanding and Detecting Cyberbullying in Real-World Images. In *2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA)*, 2021.
 52. Khan, A., Sohail, A., Zahoor, U., and Qureshi, A.S. A Survey of the Recent Architectures of Deep Convolutional Neural Networks. *Artificial intelligence review*, vol. 53, no. 8, pp. 5455-5516, 2020.
 53. Tammina, S. Transfer Learning using Vgg-16 with Deep Convolutional Neural Network for Classifying Images. *International Journal of Scientific and Research Publications (IJSRP)*, vol. 9, no. 10, pp. 143-150, 2019.
 54. Savoiu, A. and Wong, J. Recognizing facial expressions using deep learning. *Recognizing Facial Expressions Using Deep Learning*, 2017.
 55. Alashhab, S., Gallego, A.J., and Lozano, M.Á. Hand Gesture Detection with Convolutional Neural Networks. In *International Symposium on Distributed Computing and Artificial Intelligence*, Springer, Cham, pp. 45-52, 2018.
 56. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., and Fei-Fei, L. Imagenet: A Large-Scale Hierarchical Image Database. In *2009 IEEE conference on computer vision and pattern recognition*, Ieee, pp. 248-255, 2009.
 57. Simonyan, K. and Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv preprint arXiv:1409.1556*, 2014.
 58. Gadicha, A.B., Sarode, M.V., and Thakare, V.M. Empirical Approach Towards Video Analysis using Shot Frontier Detection and Key-Frame Mining. In *2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS)*, IEEE, pp. 1844-1848, 2018.
 59. Kumar, A. and Sachdeva, N. Multi-Input Integrative Learning using Deep Neural Networks and Transfer Learning for Cyberbullying Detection in Real-Time Code-Mix Data. *Multimedia systems*, pp. 1-15, 2020.
 60. Banga, M., Bansal, A., and Singh, A. Proposed Intelligent Software System for Early Fault Detection. *International Journal of Performability Engineering*, vol. 15, no. 10, pp. 2578, 2019.