ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 8, Issue 4, 2019

Cyberbullying Detection Using Contrastive LanguageImage Pretraining

Dr. T. Santhi Sri,

Professor, Department of Computer Science and Engineering Koneru Lakshmaiah Education Foundation Vaddeswaram, Guntur, INDIA sri santhi2003@yahoo.com

Dr. RaviSankar Malladi,

Professor, Department of Computer Science and Engineering Institute of Aeronautial Engineering (Auto) Dundigal, Hyderabad, INDIA m.ravisankar@iare.ac.in

Dr. Padmaja Grandhe,

3Professor, Department of Computer Science and Engineering

Potti sriramulu chalavadhi mallikarjunarao college of engineering and technology Vijayawada, Andhra Pradesh, India ,padmajagrandhe@gmail.com

Dr. K. S. R. Radhika

4Professor, Department of Computer Science and Engineering TKR College of Engineering and Technology Hyderabad , Teleangana, India ksrradhika@tkrcet.com

ABSTRACT

Cyberbullying has become a nudge in several online platforms like Facebook, Instagram, and Twitter. These platforms allow bullies to send their content through messages in differ- ent modalities such as images and text to attack victims. With the harmful consequences of bullying on victims, it is necessary to detect them. There exists a plethora of cyberbullying models which study bullying detection for a single modality: images (pretrained CNNs) or text (pretrained language models). However, previous works fail to explore the joint mod- eling of text and images to perform bullying detection. This work studies the effectiveness of the multi-modality encoder, CLIP, a contrastive language image pre-training model that provides the joint alignment between text and image to identify cyberbullying content better. Further, we introduce linear probing on CLIP to investigate the effectives of text, image, and text+image features in cyberbullying detection. Experiments on two standard cyberbullying datasets, Facebook hateful memes and NIT Warangal, provided the following insights. We find that the linear probing on the CLIP model shows substantial improvement in detect- ing the bullying content and outperforms the unimodality models. The supremacy of linear probing on the CLIP model indicates that joint representation of text and image from CLIP

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 8, Issue 4, 2019

helps better understand bullying content. We reported the highest F-score of 83 on cyber-bullying data and 64 on the hate memes dataset. The code of the paper is available at 1

Keywords: Cyberbullying, CLIP, Linear Probing, CNNs, Language Models.

Introduction

With the massive growth of the internet, users are attracted to social media daily. New applications are coming daily that change the way people express their opinions. People are sharing their opinions and personal information on social sites. These sites are helpful for people to send text and images to other people. On the other hand, these sites are used by peopleto express their hatred toward other people, thus resulting in online cyberbullying (Cohen-Almagor, 2018; Nikolaou, 2017; Karan and Šnajder, 2019).

Cyberbullying is a form of online harassment to the victims by sending messages and pictures on social media. Also, these messages and images are responsible for the negative consequences for the victim. With the development of online media platforms like Facebook, Twitter,Instagram, and media platforms, it is easy for bullies to spread cyberbullying content easily. The impact of cyberbullying has attracted researchers.

For cyberbullying problems, we can use supervised machine learning algorithms. The input for these algorithms is text or Image feature representations, and the output is whether that particular input is related to cyberbullying content or not. In this paper, we addressed this problem with two different available datasets. We explored different feature representations from the successful CLIP model with two standard machine learning algorithms.

The main reason for using CLIP architecture is to get the joint representations for feature representations as it will help capture more meaningful content from the data. By using these representations, our model achieved good performance. Our work is the first attempt to use CLIP architecture for cyberbullying data. Our contributions can be outlined as follows:

- We performed linear probing on text and image features obtained from the CLIP model.
- We compared our representations from CLIP with Word2Vec and GloVe trained on the two datasets.
- We develop a multi-model encoder system for detecting cyberbullying.

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 8, Issue 4, 2019

- we present a detailed analysis of results and compare the generated feature representations with two different datasets.

Related Work

Several researchers worked continuously on Natural Language Processing (NLP) tasks over the past decade. Many tasks in NLP are successful and become useful applications for hu- Yue and Cardie, 2010), emotion identification (Ekman, 1992; Plutchik, 2001), identification of harmful content like sexism (Waseem and Hovy, 2016), toxicity (Kolhatkar, Wu, Cavasso, Francis, Shukla and Taboada, 2020),and cyberbullying (Nahar, Li and Pang, 2013; Zhao and Mao, 2016).

In (Dadvar, Trieschnigg, Ordelman and Jong, 2013), the authors investigated cyberbullying detection by considering the user content. The three different features used are content based, cyberbullying, and user-based. These three features are passed as input to the SVM classifier and classified the test samples as yes(have cyberbullying content) or no(do not have cyberbullying content. Gender based information (Dadvar et al., 2013) is added to the feature set to improve the predictions of cyberbullying content. The features, along with gender based information, are passed to the SVM classifier to make the predictions more effective.

Cyberbullying has attracted a lot of research attention due to increased social media content. In (Rosa, Pereira, Ribeiro, Ferreira, Carvalho, Oliveira, Coheur, Paulino, Simão and Trancoso, 2019), the authors introduced an automatic system for detecting cyberbullying content. They used different textual features and three classifiers and reported the evaluation values. Word based representations are widely used in detecting cyberbullying content (Pericherla and Ilavarasan, 2021). Here different combinations of neural based representations are used for cyberbullying detection.

Datasets Description

This section describes the details related to two cyberbullying datasets. For this work, we used two available cyberbullying datasets. The first is the cyberbullying dataset, and the second is the hateful memes challenge set.

Cyberbullying dataset

We used the cyberbullying data from [2]. This data is collected from various social media

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 8, Issue 4, 2019

platforms like Facebook, Twitter, and Instagram. There are a total of 2100 posts. Each post contains an image and a related comment for it. Table 1 describes the detailed statistics of the dataset.

Cyberbullying Dataset	Class	# Samples
Image	Bullying (464), Non-bullying (1636)	2100
Comment	Bullying (884), Non-bullying (1216)	2100
Image and Comment	Bullying (1481), Non-bullying (619)	2100

Table 1: Statistics of cyberbullying dataset.

The Hateful Memes Challenge Set

We used the hateful memes dataset from (Kiela, Firooz, Mohan, Goswami, Singh, Ringshia and Testuggine, 2020), and it is also called challenge set data. The dataset contains an image and corresponding text, which are considered memes. The team created this dataset at Facebook AI. The dataset details are described in Table 2.

Hateful Memes Dataset	Class	# Samples
Memes	Hate Memes (3298), Non-Hate Memes (5698)	10,000

Table 2: Statistics of Facebook hateful memes dataset.

Proposed Method

This section explains the proposed architecture, the general algorithm we used to train the logistic and LightGBM classifiers, the training procedure for logistic and LightGBM classifiers, and the different methods we used to handle the class imbalance problem of the dataset.

Linear Probing on CLIP

For this paper, we investigated linear probing using CLIP-based feature representations. Fig-

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 8, Issue 4, 2019

ure 1 shows the summary of the proposed linear probing architecture. We took the advantage from CLIP architecture (Radford, Kim, Hallacy, Ramesh, Goh, Agarwal, Sastry, Askell, Mishkin, Clark et al., 2021) for feature generation. As shown in the figure, the input is text and images. We have a text encoder for text input, and for image input, we have an image encoder. We will get the textual representations from the text encoder. The image encoder achieved Image-related feature representations. We can get three different combinations of feature representations: (i) text feature representations from text encoder, (ii) image feature representations from image encoder, and (iii) combined text+image feature representations. We passed these three representations as input to our logistic and LightGBM models to make better predictions on test data.

Genral Training Algorithm

Algorithm 1 explains the general training process followed by machine learning algorithms.

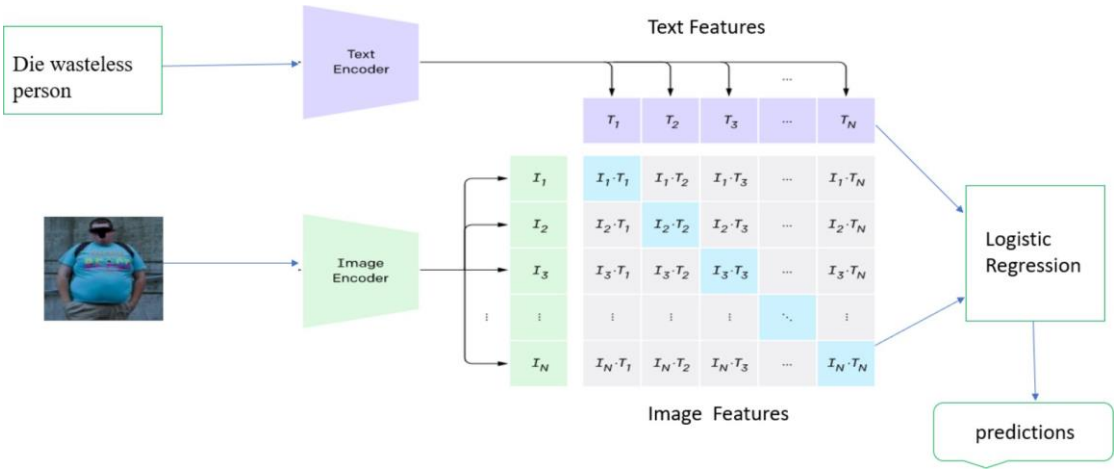

Figure 1: Linear Probing on CLIP: (a) Text encoder provides the latent representation of text, (b) Image encoder provides the latent representation of an image, (c) CLIP models provide the both text and image representations by performing the contrastive loss between text and image features from (a) and (b). Finally, we use Logistic Regression and LightGBM models inthe linear probing step to perform the cyberbullying detection.

Algorithm 1 General Training Process of Machine Learning Models

X =Word embeddings from Word2Vec, GloVe, CLIP and its variations $Y\ Pred$ =predictionsfrom the Logistic Regression classifier

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 8, Issue 4, 2019

X_Train ←training samples from the dataset

Y_Train ←corresponding labels of training data

X_Test ← test samples from the dataset

Y_Test ← corresponding labels of test data

W ← Weight vector

for every instance in X_Train **dofor** <i in *length*(X_Train)>**do**

<Update weights>

end for

<Calculate error>

end for

Repeat until error is smallreturn the Y_Pred

calculate Precision, Recall and F1-score

Training Machine Learning Algorithms

Training of Logistic Regression (LR): To perform the linear probing on the CLIP model, we use the standard classification model, Logistic Regression, to perform bullying detection. Here,the input to the model is (i) text, (ii) image, or (iii) image + text obtained from CLIP, and the target output is the class label (binary class). We set the hyper-parameters such as (i) regularization parameter: C =1.0, (ii) penalty as L2, (iii) solver as lbfgs, and (iv) maximum iterations as 100. The model predictions on test data are measured using macro-average precision, recall, andF1-score.

Training of LightGBM: Light Gradient Boosting method is one of the successful machine learning techniques in the tree-based boosting algorithms. We choose the LightGBM model as one of our training methods because of its high speed and consumption of less memory on large datasets. To train the LightGBM model, we used the exact input feature representations of the LR model. We pass the hyper parameters: max depth as 5, number of decision trees as 1000, boosting method as gradient boosting when training the model. The target class labels are used based on task data that we used in training. The model predictions on test data are measured using macroaverage precision, recall, and F1-score.

Class Imbalance Methods

To handle the class imbalance problem in the classification setting, we perform two sampling methods: (i) over-sampling (OS) and (ii) under-sampling (US).

over-sampling (OS): For solving the overs-sampling problem, we used SMOTE (Synthetic

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 8, Issue 4, 2019

Minority Oversampling Technique) (Chawla, Bowyer, Hall and Kegelmeyer, 2002) based over- sampling. This method includes the selection of random examples from the minority class. The algorithm replaces and supplements the training data with multiple copies of this instancefor the selected examples. A single instance may appear multiple times in the data and solvesthe class imbalance problem.

under-sampling (US): For solving the under-sampling problems, we use the technique present in (Yen and Lee, 2006). Under-sampling is the opposite of the over-sampling technique. This method randomly selects and removes some instances from the majority class. The number of instances in the majority class will be reduced and thus solves the class imbalance problem.

Feature Representation Methods

Feature representations are the key to understanding data. The performance of an algorithm is highly dependent on the correct selection of feature representation methods. In this section,we explain the feature representation methods used in this paper.

Word2vec

Word2Vec (Mikolov, Sutskever, Chen, Corrado and Dean, 2013) is the first neural feature representation from the text. Here, the technique is to learn high-quality word vectors from a huge text corpus. In Word2vec, there are two techniques — CBOW(Continuous bag of words) and SG (Skip-gram) model. These two techniques are shallow neural networks that map word(s) to the target variable, a word(s). Both of these techniques learn weights that act as word vec- tor representations. The word representations from Word2Vec are dense and have reduced dimensions. For our experiments, we used 300 dimension vectors for each word. From the Spacy library, we used the available pretrained Word2Vec embeddings.

Glove

GloVe (Pennington, Socher and Manning, 2014) is a well-known feature representation method that learns word vectors from the co-occurrence information. The GloVe is a count-based model that depends on the co-occurrence statistics. The main idea in the GloVe model is to utilize the statistics from the whole input data. For learning word vectors, GloVe considers global information into consideration. This model is trained on the word-word co-occurrence matrix. For our experiments, we used 300 dimension vectors for each word. From the Spacy library, we used the available pretrained GloVe embeddings.

CLIP-Text Features

CLIP (Radford et al., 2021) is a language model trained on neural network architecture. It

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 8, Issue 4, 2019

can be trained on image and text data, hence called a multi-modal language model. The main idea is to predict the relevant text based on images, thus producing effective feature representations. The CLIP model is highly effective as it learns from noisy and corrupted data. For our experiments, we used 512 dimension text features extracted from the CLIP model.

CLIP-Image Features

The CLIP architecture is more flexible and general. It produces image feature representations. We used 512 dimension image features extracted from the CLIP model for our experiments.

CLIP-Text+Image Features

The combination of text and image feature representation dimension is 1024. We used this combination of features for our experiments.

Experimental Setup & Training

To measure the performance of our datasets on cyberbullying detection, we perform the following experiments: (i) First, we create the baseline results by using pretrained word embedding methods such as Word2Vec-Te and GloVe-Te. (ii) Second, we create models on CLIP text features, CLIP image features, and CLIP-Text+Image features to compare the performance with the baseline models for the cyberbullying detection task. For our training process, we use two different machine learning classification models. One is simple Logistic Regression

	Logistic Regression									**LightGBM**								
Sampling Method⟶	NS			OS			US			NS			OS			US		
Feature set↓	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Word2Vec	75	78	76	76	78	77	75	78	76	81	80	79	81	78	79	78	80	79
GloVe	71	73	72	72	74	73	72	74	73	78	76	77	79	77	77	74	77	75
CLIP-TextFeatures	**81**	**80**	**81**	**82**	**82**	**82**	**78**	**79**	**78**	**83**	**82**	**82**	83	**81**	**83**	**80**	**80**	**80**
CLIP-ImageFeatures	74	71	72	65	68	66	67	73	67	80	70	74	75	66	68	69	76	71
CLIP-Text+ImageFeatures	71	72	72	73	73	73	72	75	72	77	70	72	**84**	75	78	73	77	74

NS = No-Sampling, OS = Over-Sampling, US = Under-Sampling P = Precision, R = Recall, F1 = F1-score

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 8, Issue 4, 2019

Table 3: Linear probing results for cyber bullying dataset: Different feature sets classification comparison for Logistic Regression, and LightGBM classifier using different sampling methodswith No/Over/Under-Sampling.

	Logistic Regression									**LightGBM**								
Sampling Method→	NS			OS			US			NS			OS			US		
Feature set↓	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Word2Vec	55	52	46	57	57	57	56	56	56	58	53	46	58	54	48	56	55	54
GloVe	54	52	46	55	55	55	56	55	55	59	54	46	57	54	48	56	55	54
CLIP-TextFeatures	58	56	52	57	57	56	56	56	56	57	53	46	57	54	49	55	54	53
CLIP-ImageFeatures	58	56	54	**60**	**60**	**60**	59	59	59	60	58	55	**64**	**61**	**59**	61	61	61
CLIP-Text+ImageFeatures	**61**	**59**	**57**	59	59	59	**61**	**61**	**61**	**64**	**59**	**55**	62	59	56	**64**	**64**	**64**

NS = No-Sampling, OS = Over-Sampling, US = Under-SamplingP = Precision, R = Recall, F1 = F1-score

Table 4: Linear probing results for hateful memes dataset: Different feature sets classification comparison for Logistic Regression, and LightGBM classifier using different sampling methodswith No/Over/Under-Sampling.

(LR) (Yu, Huang and Lin, 2011), and the other is a popular, recent successful tree-based technique called Light Gradient Boosting method (LightGBM) (Ke, Meng, Finley, Wang, Chen, Ma,Ye and Liu, 2017).

Evaluation Metrics

we use standard macro-average precision, recall, and F1-score as the evaluation metrics. Precision value tells us how accurate our model performance is. Precision is the ratio of actual positive samples by total predictive samples by the classifier. Recall value tells us how many are actual positives out of the classifier predicted positives. The recall is the ratio of classifier predicted positives by total actual positives. F1-score is the balance between precision and recall. We considered macro-average precision, recall, and F1-score for our experiments as

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 8, Issue 4, 2019

they give equal weight to each class to evaluate algorithm performance.

AUC-ROC Curves

To evaluate the model performance, we plot AUC-ROC curves. These are called Area Under Curve Receiver Operator Characteristic. These curves show the performance of a classifier for several thresholds. For this work, Figures 2, 3, and 4 report the ROC plots for hateful memes dataset. The X-axis of the ROC curve represents the false positive rate, and the Y-axis represents the true positive rate. Figures, 2, 3, and 4 show that logistic regression is doing a better job for cyberbullying task.

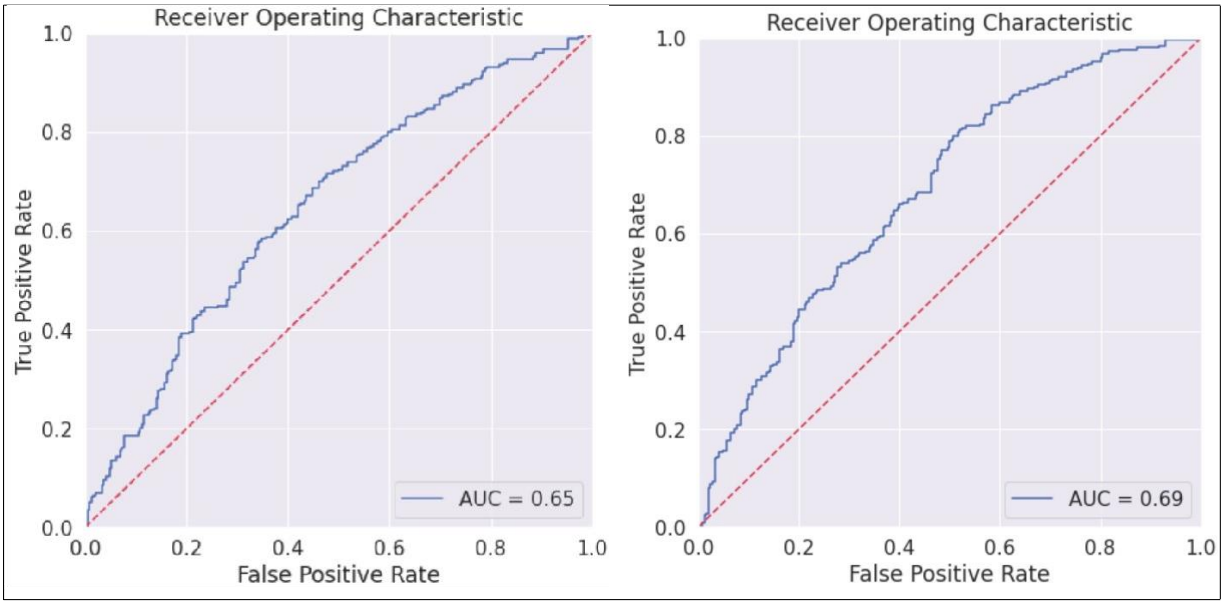

Figure 2: Left: AUC ROC curve for Logistic regression classifier with Text + Image features.Right: AUC ROC curve for LightGBM classifier with Text + Image features.

ISSN PRINT 2319 1775 Online 2320 7876

Research paper

 UGC CARE Listed (Group -I) Journal Volume 8, Issue 4, 2019

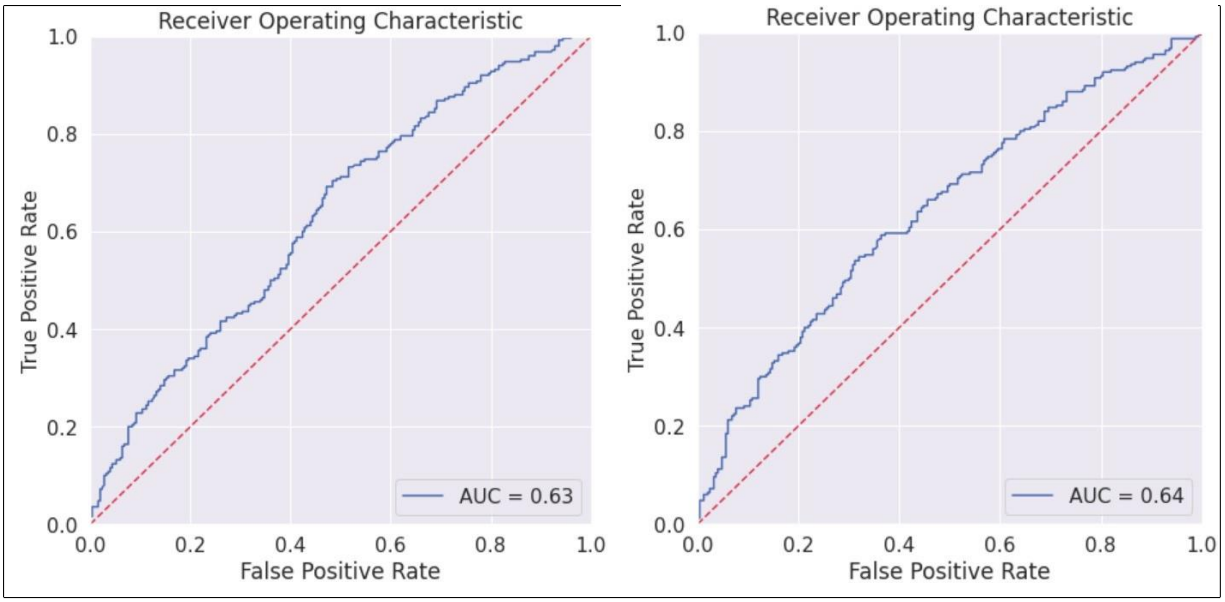

Figure 3: Left: AUC ROC curve for Logistic regression classifier with Image features. Right: AUC ROC curve for LightGBM classifier with Image features.

Conclusion

This paper studies the effectiveness of linear probing on the CLIP model for cyberbullying detection. We find that the linear probing on the CLIP model shows substantial improvement in detecting the bullying content and outperforms the unimodality models. Our experiments on two cyberbullying datasets, NIT and Facebook hateful memes, led to interesting insights and

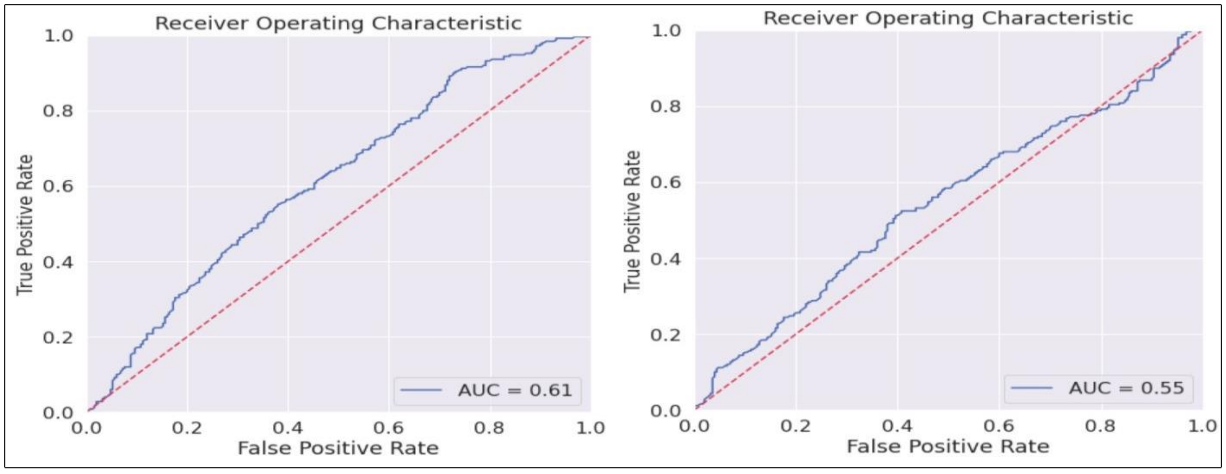

Figure 4: Left: AUC ROC curve for Logistic regression classifier with Text features. Right: AUC ROC curve for LightGBM classifier with Text features.

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 8, Issue 4, 2019

improved results. These insights indicate that the CLIP model reveals that image features have information about the text and vice versa, which shows a better understanding of cyberbullying in different context modalities. We plan to extend this work for audio data as part of future work.

References

[1] Chawla, N. V., Bowyer, K. W., Hall, L. O. and Kegelmeyer, W. P. 2002. Smote: synthetic minority over-sampling technique, *Journal of artificial intelligence research* **16**: 321–357.

[2] Cohen-Almagor, R. 2018. Social responsibility on the internet: Addressing the challenge of cyberbullying, *Aggression and violent behavior* **39**: 42–52.

[3] Dadvar, M., Trieschnigg, D., Ordelman, R. and Jong, F. d. 2013. Improving cyberbullying detec- tion with user context, *European Conference on Information Retrieval*, Springer, pp. 693–696.

[4] Ekman, P. 1992. An argument for basic emotions, *Cognition & emotion* **6**(3-4): 169–200.

[5] Karan, M. and Šnajder, J. 2019. Preemptive toxic language detection in wikipedia comments using thread-level context, *Proceedings of the Third Workshop on Abusive Language On-line*, pp. 129–134.

[6] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q. and Liu, T.-Y. 2017. Lightgbm: A highly efficient gradient boosting decision tree, *Advances in neural information processingsystems*, pp. 3146–3154.

[7] Kiela, D., Firooz, H., Mohan, A., Goswami, V., Singh, A., Ringshia, P. and Testuggine, D. 2017. The hateful memes challenge: Detecting hate speech in multimodal memes, *Advances inNeural Information Processing Systems* **33**: 2611–2624.

[8] Kolhatkar, V., Wu, H., Cavasso, L., Francis, E., Shukla, K. and Taboada, M. 2019. The sfu opin- ion and comments corpus: A corpus for the analysis of online news comments, *Corpus Pragmatics* **4**(2): 155–190.

[9] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S. and Dean, J. 2013. Distributed represen- tations of words and phrases and their compositionality, *Advances in neural information processing systems*, pp. 3111–3119.

[10] Nahar, V., Li, X. and Pang, C. 2013. An effective approach for cyberbullying

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 8, Issue 4, 2019

detection, *Com- munications in information science and management engineering* **3**(5): 238.

[11] Nikolaou, D. 2017. Does cyberbullying impact youth suicidal behaviors?, *Journal of health economics* **56**: 30–46.