International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 1
DOI: https://doi.org/10.17762/ijritcc.v11i1.5981
***Article Received**: 10 November 2022 **Revised**: 14 December 2022 **Accepted**: 26 December 2022*

CROD: Context Aware Role based Offensive Detection using NLP/ DL Approaches

T Purnima[1], Ch Koteswara Rao[2]
[1]School of Computer Science and Engineering, VIT-AP University,
Amaravathi 522237, Andhra Pradesh, INDIA.
purnima.21phd7093@vitap.ac.in
[2]Assistant Professor Sr. Grade, School of Computer Science and Engineering,
VIT-AP University, Amaravathi 522237, Andhra Pradesh , INDIA
koteswararao.ch@vitap.ac.in

Abstract—With the increased use of social media many people misuse online platforms by uploading offensive content and sharing the same with vast audience. Here comes controlling of such offensive contents. In this work we concentrate on the issue of finding offensive text in social media. Existing offensive text detection systems treat weak pejoratives like 'idiot' and extremely indecent pejoratives like 'f***' as same as offensive irrespective of formal and informal contexts . In fact the weakly pejoratives in informal discussions among friends are casual and common which are not offensive but the same can be offensive when expressed in formal discussions. Crucial challenges to accomplish the task of role based offensive detection in text are i) considering the roles while classifying the text as offensive or not i) creating a contextual datasets including both formal and informal roles. To tackle the above mentioned challenges we develop deep neural network based model known as context aware role based offensive detection(CROD). We examine CROD on the manually created dataset that is collected from social networking sites. Results show that CROD gives better performance with RoBERTa with an accuracy of 94% while considering the context and role in data specifics.

Keywords-context , role , machine learning , deep learning , BERT , RoBERTa.

I. INTRODUCTION

Social media influences the people across all ages and paves the way for sharing and acquiring information globally. According to Global social media Statistics given by a Strategy Consultancy Kepios around 4.65 billion social media users exist in April 2022. On an average WhatsApp users send 42 million messages, Twitter has 511,200 tweets and Facebook users share 150,000 messages every minute. This shows massive generation of textual data in social media which is nothing but the information that social media users share publicly including metadata like location of user, language spoken, biographical data etc. On one side, with this information Social media became the leading channel for marketing & advertising where marketers looking for customer insights may increase sales or politicians may conduct political campaigns to win votes. On the other side some malevolent users misemploy social media by posting offensive content to torment others unethically The approach of using digital communication tools to abuse or bully someone typically by sending messages of an intimidating or threatening nature is Cyberbullying. Cyberbullying may exist in different forms along with Offensive language and hate speech, which are prevalent during textual communication happening online.

A. Defining Cyberbullying, Hatespeech , Offensive

• Cyberbullying is sending vicious, abusive or threatening messages using digital technologies through social media to impersonate someone.

• Hate speech can be any talk that attacks or diminishes and arouse violence against people, based on certain characteristics such as physical appearance, religion, origin, sexual orientation, gender identity etc.

• Offensive describes rude or hurtful behavior including swear words or blunt insults. We can find offensive language in text messages, social media comments, message forums, and even in online games. Offensive sentences always contain pejoratives, profanities, or obscenities. Exposure to offensive content can cause anxiety, depression, and other stress-related disorders in humans. Hence we focus on Offensive text detection to identify potentially harmful messages in social media efficiently.

Conversational Vs non-conversational text - Existing offensive detection systems dealt mostly with Non-conversational text . For example, - Non-conversational text:

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 1
DOI: https://doi.org/10.17762/ijritcc.v11i1.5981
***Article Received**: 10 November 2022 **Revised**: 14 December 2022 **Accepted**: 26 December 2022*

Existing offensive detection systems dealt with Nonconversational text . Figure 1 shows an example of Non-conversational text. Fig.1(a) and fig.1(b) are sample offensive texts that contains no information of involving personality. fig.1(a) is offensive in formal conversations and not offensive in informal conversation. fig.1(b) is offensive in both formal and informal conversations.

TABLE I: various forms of online abuse

Category	Examples
Offensive	abusive posts, profanity
Hate speech	religion, racism
cyberbullying	threatening, harassment

.

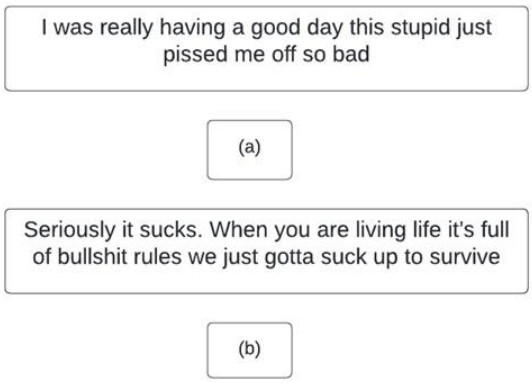

Fig. 1: Examples of non-conversational text

• conversational text:

Identification of offensive content while distinguishing the formal and informal contexts will improve the integrity of offensive detection and reasonable. As a matter of fact, conversations are based on role. Indeed we are more formal and mostly utter decent words in formal context & We often use some weak pejoratives in casual discussions with friends which are not offensive in informal context but offensive in formal context. Our work is motivated by considering the conversational context and the prevalent use of weak and strong pejoratives while detecting offensive text. For example, Fig. 2 demonstrates the instances of offensive congruence in social media. Fig. 2(a) depicts the formal - offensive form in which the weak pejorative ‘stupid‘ is offensive in formal context which is actually not offensive in informal conversation among friends. Fig. 2(b) shows informal - offensive analogy where the strong pejorative ‘f***‘ is offensive in both formal & informal contexts. Fig. 2(c) tells the formal - Not-offensive analogy of a fair discussion between two which is not offensive in both contexts. Fig. 2(d) leverages the informal Not-offensive analogy which is a casual discussion between two friends which is not offensive in informal context and is offensive when it is uttered between two formal people. In Fig. 2(a), the word ‘stupid‘ is not offensive while uttered with friends ,but the same is offensive when uttered with formal people. In Fig. 2(b), the words ‘sucks‘ and ‘bullshit‘ in formal context is offensive , but the same is not offensive in informal context. Non-conversational systems are not able to differentiate formal and informal context while detecting offensive content in text.Considering all the above mentioned cases, we develop context-aware Offensive Detection (CROD) to detect offensive text in social media.

B. Novelty of CROD

Detailed comparison highlighting the novelty of CROD. Table II describes the features of CROD and various methodologies proposed.

TABLE II: Feature interpretation of CROD

features	CROD	[1]	[2]	[3]
Role based	yes	no	no	no
Conversation based	yes	no	no	no

Razavi, Amir H., et al.[1] used linear support vector machines to detect cyberbullying-related English and Dutch posts . four roles are identified in cyberbullying interactions that are illustrated in the dataset inclusive of bully,victim and two classes of bystanders. If the conversation has harmful expressions of cyber bullying, then the candidate role gets identified. The model gives fi scores of 64% for english and 61% for dutch.

Bretschneider, Uwe .,et al.[2] focused on detecting offensive statements towards refugees and foreigners. To accomplish this,Two human experts annotated three datasets indicating offensive expressions, the severity of offense and the target collected from famous social site Facebook in which one dataset is for training and the remaining two are for testing . This pattern-based sequence model approach gives substantial precision of 75.26%.

Wiedemann, Gregor, et al.[3] proposed Automatic classification of offensive language on German Twitter data performing the tasks of binary classification of tweet that has offensive text and multi-class classification of the same tweet into either ‘insult‘, ‘profanity‘, ‘abuse‘, or other. They used sequentially combined BiLSTM-CNN which leads an accuracy of 77.5%.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 1
DOI: https://doi.org/10.17762/ijritcc.v11i1.5981
***Article Received**: 10 November 2022 **Revised**: 14 December 2022 **Accepted**: 26 December 2022*

Fig. 2: Samples of offensive text in social networking sites. (a), (b), (c) ,(d) are examples of formal Offensive , informal Offensive , formal Not-Offensive and informal Not-Offensive

Prior work does not attempt to discern between formal and informal contexts while classifying a conversation as offensive or not. On examining many conversations from both online and offline it is observed that , situated on involving roles we can categorize the same context as Offensive in formal conversation and Not-offensive in informal conversation and vice versa. CROD is a context-aware approach on account of role while identifying offensive text inn contextual dataset.

There have been a number of research studies on text-based sentiment algorithms to identify abusive language that have been that have been published over the past few years. One of them is [10] that uses sentiment analysis to spot the bullying of tweets and utilizes Latent Dirichlet Allocation (LDA) model to find relevant topics within these texts. In recent times distributed word representations, often known as embeddings of words, are being considered for similar reasons [13]. Deep learning methods are currently being utilized for text classification and sentiment analysis with the paragraph2vec method.

This paper present an approach to develop an algorithm for machine learning that can recognize the characteristics of harmful language. We focus on detecting offensive and hateful text on the Twitter datasets. With the help of publicly available Twitter datasets, we develop our classifier models using n-gram as well as terms frequency-inverse document frequencies (TFIDF) for features. We then analyze it for metrics scores. We conduct a comparative analysis of the results that are obtained with classifier models. Our results indicate that the model are superior to all others and that the models that we propose have better characteristics in dealing with data with adversity.

II. PRIOR RESEARCH

Social networking sites and other online websites are trying to erode offensive text which is a major concern to maintain healthy online environment. Number of solutions were proposed to erode offensive behavior. Some of the works are presented in this section.

[4] Focuses on automatic detection of cyberbullying in English and Dutch social media text posted by bullies. Researchers explore the feasibility of automatic recognition of cyberbullying making use of linear support vector machines. Analysis unveils that n-grams of both word & character types and sentiment lexicons are useful features for cyberbullying detection. The model gives 64% F1 score for english and 61% F1 score for Dutch.

Prior work [5] classifies text as Hate speech, offensive and Neither. Authors used pre-annotated Twitter dataset and other hatespeech dataset from Crowdflower to create new dataset to overcome class imbalance problem. LSTM and BI-LSTM classifiers were developed using single LSTM layer,stacked LSTM layers, Bi-LSTM layer , stacked Bi-LSTM layers. LSTM model with GLOVe embeddings gives an accuracy of 86%.

In This paper [6] Marcos Zampieri.et, proposed OLID dataset which was hierarchically annotated into different levels to differentiate whether the text is offensive or not, type of offense, and the target of offensive content in the text. The classification is efficient with CNN having 80% of F1 macro.

Authors [7] concentrate on building a classifier to differentiate toxic & non-toxic comments and develop a multi-headed model to detect various types of toxicity comprises insults, threats, obscenity, identity hate and toxic. As a supervised learning algorithm they used Logistic Regression for classification. Wikipedia‘s comments are used to create dataset by Jigsaw for toxic and non-toxic data.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 1
DOI: https://doi.org/10.17762/ijritcc.v11i1.5981
***Article Received**: 10 November 2022 **Revised**: 14 December 2022 **Accepted**: 26 December 2022*

TABLE III: Summary of related work

References	Model used	Classifier	Dataset	Type of abuse detected	Limitations
4	Machine learning & Deep Learning	SVM, BiLSTM, CNN	OLID dataset	Offensive	Not considered the role of personage while classifying the text as offensive & it is not conversational based.
5	Machine learning	Logistic Regression	Wikipedia comments	Toxic, obscene, Severe-Toxic,threat , Identity-hate, Insults	&Not used the context while classifying the text as toxic, obscene, threat and insult .
6	Machine learning	SVM	Twitter data	hatespeech	ignores the conversational context while detecting hatespeech
7	Deep learning	RNN	Twitter data	Racism , Sexism	Ignores the concept of role and lacks context while classifying the twitter data which is not conversational.
8	Deep Learning	Bi-LSTM, BERT, DistilBERT	Twitter data	Hatespeech, Offensive	Didn‘t evolve the concept of role and conversational context.

Prior work [8] differentiates racism and sexism messages on twitter dataset. the tendency towards offensive type and word frequency vectors from the text are used for classification of hatred content. With the assumption of obfuscating offensive terms in social media, word frequency vectorization representation is used to represent offensive terms with short dialects. The incorporation of user‘s behaviour into the classification gives F-score of 0.9295.

III. METHODOLOGY

A Detailed description of CROD and classifiers are described here.

A. Problem Formulation

In the present section, we define the problem of offensive text detection technically. Offensive.The text is considered offensive if the conversation conveys any vulgar , rude, offensive, obscene or prejudicial information against someone or something(e.g., race, gender, religion). If not it is treated as non-offensive . The goal of CROD is to evolve the concept of role involved in the conversational context , given a group of conversations T = {c1, c2, ..., cn} in social media , in which each conversation is composed of two dialogues ti = {s1, s2}, roles are R={Formal, Informal} , tasks are A = {OFF, NOT } and the classes are C1 ={Formal-Off ,Formal-NotOff} , C2={ Informaloff ,informal-NotOff}. Our task is to formulate a classification problem, where the conversational text is assigned with class labels from C and D. ie., given an unlabeled text tk, classification problem will assign one of the classes {a, b, c,d } such that a, b and c, d belongs to C1 and C2 respectively.

B. Proposed system framework of CROD

In this part, we demonstrate the CROD system to inscribe the problem of offensive detection. Fig. 3 describes the overview of the CROD which encompass various units.

1) data Preprocessing: We preprocess the data by removing the punctuation marks and stop words which do not encompass any helpful information for the clasification of text. Symbols like ,# and $ were eliminated from the text. Subsequently the the text was exposed to tokenization .

Lemmatization groups the various forms of words together to be analyzed as a single thing. Lemmatization gives context to the words.

2) Feature engineering: Our proposed model uses three feature engineering techniques namely WORD2VEC , FASTTEXT and BERT.

• Word2Vec is to learn word embeddings using neural network. Word2Vec can recognize word’s context once the word’s semantic and structural similarities are identified with its related words. Word2Vec generates a feature vector in the text corpus for each identical word.

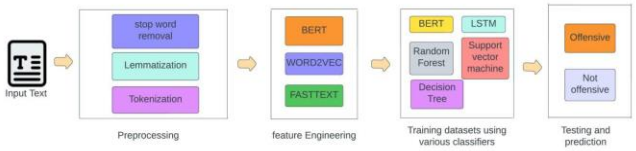

Fig. 3: The process of offensive text detection

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 1
DOI: https://doi.org/10.17762/ijritcc.v11i1.5981
***Article Received**: 10 November 2022 **Revised**: 14 December 2022 **Accepted**: 26 December 2022*

• FastText is another method to generate word embeddings which splits words into number of n-grams to feed the Neural Network. The performance of FastText is much better when compared with Word2Vec because of its capability in representing rare words accordingly.

• Bidirectional Encoder Representations from Transformers(BERT) Tokenizer tokenize sentences or segments into wordpieces to give input to the BERT model. As BERT needs the size of the input text to be fixed, the text which exceed the size are tackled by a bert Trimmer which trims to a preset size. We can collectively combine the trimmed segments to generate a combined tensor. [CLS] is the beggining token and [SEP] is the ending token that were used by the BERT and a RaggedTensor implies to correlate the tokens in the combined Tensor. The tokenization sample was shown in fig.4.

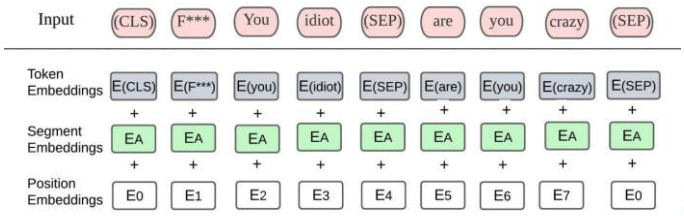

Fig. 4: BERT input format

3) training a classifier: We used both machine learning and deep learning classifiers for optimal classification performance. We evaluate different ML models namely Support Vector Machine (SVM),Logistic Regression (LR), Random Forest (RF), Decision Trees and also used deep learning algorithms namely Long Short Term Memory Networks (LSTMs) and BERT.

a) Logistic Regression : In Logistic Regression, the input values are connected linearly using weights to determine the output Training data is used to predict the weights applying the maximum-likelihood estimation learning algorithm which minimizes the error while predicting the probabilities. Logistic regression calculates probabilities between 0 and 1 where offensive text was labeled as 1 and non-offensive text as 0 in both formal & informal contexts, and determined the coefficients of the logistic function using the Tf-idf vectors.

b) SVM : SVM is one of the supervised algorithms used for various classification problems. SVM fits the data, and returns a best fitting hyper-plane that divides the data points into two or more classes.Given an input corpus of texts (x), the SVM determines the correlation between the input and output pairs to classify the output classes. The target function takes the decision of target class to be classified. The text is represented as an input vector and the identified class is 1 if the text is Offensive and 0 if Not-Offensive. The objective of the model is to decide the relative absolute mapping of input and output pairs which gives minimum error.

c) Decision Tree : Decision tree is a tree like structure in which each internal node represents a condition with parameter name, association of features represent branches which leads to leaf nodes nothing but class labels. The route from the top node which is root to leaf node represent rules for classification. DT can be constructed by parting the data set rooted on different conditions. They can solve complex problems easily. ID3 algorithm is the basic one used to build decision trees which uses greedy approach while selecting the best attribute. Best attribute is the one which has the maximum information gain.

d) Random Forest : It used to categorize the large amounts of data and contains numerous decision trees on subsets of the dataset given and works by combining many decision trees in training data. It takes average predictions for better performance. It makes its predictions on various decision trees which are combined and apllied training on the datasets. It takes identical size of training sets known as bootstraps. Once the tree is constructed, bootstraps, which are not in original dataset is used as test set. RF can handle non-linearly correlated data and noise . RF has an inherent feature selection preceding to the classification step, to reduce variable space.

e) Long short-term memory (LSTM) : LSTM is a type of recurrent neural network(RNN) and also has learning capability to handle long-term dependencies in text. LSTM retains the information of previous text sequences and performs the feature extraction. The LSTM model has a memory cell known as ‘cell state‘ that preserve its state progressively with time. LSTM has the capability to select the information to remember and the information to forgot. We applied BERT embeddings for the deep neural network based on LSTM. Figure 5 depicts the LSTM model build for the detection of Offensive text. The model have one input layer, a BERT embedding layer, LSTM layer , dense layer, and one output layer.The layers of the LSTM model are:

• Input Layer: The input layer of the network is the sentences of word sequences in the given document . The document has many sentences and each sentence consists of several words . All the words are represented as fixed-size vectors from the word embeddings which are pretrained.

• Embedding Layer: Embedding layer maps the various word indices to various embedding vectors. The resultant embedding vector is a dense vector .

• LSTM Layer:The input to this layer comes from embedding layer and generates the output of all hidden states as the output of the LSTM network.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 1
DOI: https://doi.org/10.17762/ijritcc.v11i1.5981
***Article Received**: 10 November 2022 **Revised**: 14 December 2022 **Accepted**: 26 December 2022*

• Dense Layer: It gets the input from the previous layer i.e LSTM layer. This layer helps in defining the relationship among the values of the data on which the model works.

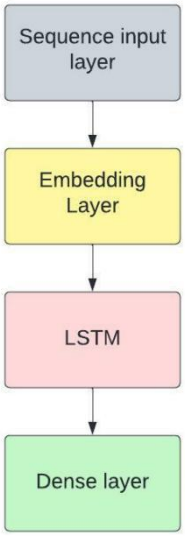

Fig. 5: LSTM neural network layered architecture

f) Bidirectional Encoder Representations from Transformers(BERT) : BERT is a deep learning framework that can be applicable to natural language processing (NLP) tasks. By using Transformer and attention mechanism, BERT learns contextual relations among words of sentences in a text. By looking at all the neighbouring words, the Transformer provides the BERT to understand the context of the word which makes the model better. BERT makes use of masked language model(MLM). MLM utilizes left context and right context and operates based on the principle of masking the random words from the given text and it finds the word with the help of its context. BERT has various variants and some of them are BERT-Large, BERT-Base,RoBERTa,ALBERT and DistilBERT. On these BERT-Base and BERT-Large are varied in their sizes, computation power and processing time. BERT-large consumes more time for processing but capable of processing large datasets. BERT architecture was depicted in figure 6 which has embedding layer as iput layer, fully connected layer and classification layer.

• Input Layer: The input layer has the sentences of various word sequences in the given text.

• Embedding Layer: The Embedding layer has block of vectors. Each vector is associated to one of the tokens in the word indices vocabulary.

• Fullyconnected Layer: It gets the input from the embedding layer and generates the output of all hidden states as the output.

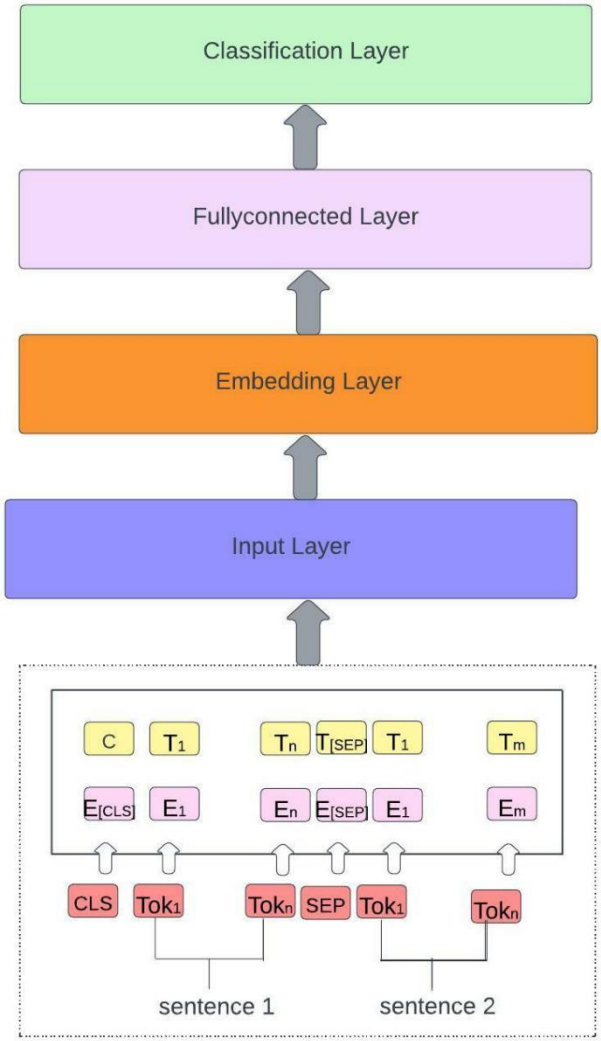

Fig. 6: BERT Architecture

• Classification Layer: It gets the input from the previous layer i.e fullyconnected layer. This layer defines the relationship among the data values and represent the classification labels.

Fine-Tuning of BERT-Large architecture have large feedforward-networks with 1024 hidden and 16 attention heads. BERTLarge has 8 attention heads & 512 hidden units. It has 340M parameters that can handle up to 512 tokens in the input text sequence. Individual layers were used in the BERT model to represent various features during the classification of offensive text. Input representation is generated for each token with all the word embeddings accompanying to the token of corresponding segment and the position of the token. The final state of the first token is considered as the text format To reduce overfitting different learning rates are used during fine-tuning . It is observed From the experiments, during BERT fine-tuning, learning rate lr, 2.5e-5 works well and shows better performance.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 1
DOI: https://doi.org/10.17762/ijritcc.v11i1.5981
***Article Received**: 10 November 2022 **Revised**: 14 December 2022 **Accepted**: 26 December 2022*

g) RoBERTa: RoBERTa stands for Robustly Optimized BERT is a pretrained model and a variant of BERT which is developed to enhance the training process. RoBERTa was developed to process larger data of long sequences and large minibatches. RoBERTa uses Dynamic Masking, Next Sentence Prediction (NSP) and Large Mini-Batches for Robust training. The original BERT- base is trained with just 256 sequences of batch size for 1 million training steps but RoBERTa is trained with batch size of 2k sequences in 125k steps. With the dynamic masking technique the input sequences are multiplied by RoBERTa to increase the number of sequences and 15% of the total sequences are randomly masked. This technique allows the model to read various distinct masking patterns in the same sequence which leads to reduction in number of instances for training that inherently enhances training procedure. RoBERTa's architecture is same as BERT's architecture, but RoBERTa uses a byte-level BERT Pretrained Embeddings.

IV. DATA

In the present part , we discussed the datasets which were collected from social media sites. We notice that social media platforms like Twitter and youtube contain offensive text. Collecting Set of offensive conversational contexts on social network platforms was a difficult task and we notice that Existing offensive textual datasets are not conversative and not discern formal & informal contexts because formal conversations are more polite in which never include any slur and informal conversations are casual which may include some weak pejoratives. Hence, we created our the datasets for the assessment of the CROD model. we collect four categories of data in our dataset ,formal - offensive , formal - Not-Offensive,Informal - Offensive and Informal - Not-Offensive for our study. For each conversation S1 and S2 comments are collected. We note that all the conversations are diversified and rarely repeated.we manually annotate the label for each context and we checked and find that the dataset we prepared is balanced

V. EXPERIMENTS & EVALUATION

The purpose of the use of TFIDF is to lessen the impact of less informative tokens which occur frequently in the corpus of data. Tests are performed with numbers ranging between one and three words. The formula used to calculate the TFIDF for the term t in document d

$$tf\ idf(d, t) = tf(t) * idf(d, t)$$

In addition, both the (L1 as well as L2) (Euclidean) regularization for TFIDF is taken into consideration when performing tests. Normalization of L1 can be defined by:

Vnorm = v/ v1+v2+…..vn

where n in the total number of documents. Similarly, L2 normalization is defined as:

$$Vnorm = v/\ Sqrt(v^2_1+v^2_2 + v^2_3 \ldots + v^2_n)$$

First, for testing the effectiveness of the generation of representations of code for vulnerability detection, our research performs analyses with ELMo against Word2vec, FastText, GloVe and BERT models, which are developed of NLP fields, but ignoring the techniques, like Code2Vec [16 as well as code semantic representation generator [17], that are designed to generate representations of code. Our next research will help bridge this gap by incorporating most recent codes that use semantic representation generation to determine if these techniques will produce more efficient representations of code with more semantic information that is preserved and ultimately leading to better security detection capabilities.

Effective code embedding strategies with more-explicit models are needed to better capture the intricate and pliable patterns. The method is based on the Bi-LSTM structure that was found to be less effective than many model languages that were trained in terms of conceptual understanding and contextual learning.

In the series of experiments we performed, we use 70%, 20%, and 10% of the dataset for the training, testing and validation respectively. LR, SVM, RF, DT are taken as Base classifiers . LSTM and BERT pretrained models are used with FAST TEXT and WORD2VEC & BERT embeddings. BERT-large (uncased), RoBERTa-large models were tested. We assess the work of the CROD framework on real-world dataset, Offensive Langiage Identification (OLID). We examine the work of CROD and baseline to detect the offensive text on social network sites. In addition we study CRODs efficacy in offensive text detection that ´ includes role as formal or informal. The evaluation result present that the CROD attains significant performance while detecting offensive text. We perform fine-tuning the BERT model with supervised training to improve efficiency .For the classification of offensive text we add classification layer to the core model. Fig.6 depictss the BERT model for offensive text classification. BERT has different configurations, BERT-Base is the most basic model with 12 encoder layers. We used the BERT-Large model which has 24 transformer layers, 16 self-attention heads, 1024 hidden layers and RoBERTa with an additional number of layers. The BERT-Large While finetuning, the model parameters are LEARNING RATE = 2e-5, and BATCH SIZE = 16. The CROD framework is trained for 30 epochs which gives the parameter values recommended by the literature for sequence classification tasks.

VI. EFFECTIVENESS OF CORD

The experimental analysis of CROD in discerning offensive and non-offensive texts, the metrics we have used for the

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 1
DOI: https://doi.org/10.17762/ijritcc.v11i1.5981
***Article Received**: 10 November 2022 **Revised**: 14 December 2022 **Accepted**: 26 December 2022*

classification are Accuracy, F1 Score. The experiment results are shown in table 4 and table 5 which gives the accuracy, F1 score results on OLID dataset and on CROD dataset for the Offensive text classification task using LR, SVM, RF and DT classifiers with word2vec , fastText and BERT embeddings. The results demonstrate that FastText and BERT embeddings give good results with Decision Tree classifier. DNN-based classifier LSTM perform better than Decision Tree. Finally BERT fine-tuning gives the best performance. We notice that the CROD performs well when compared with baseline classifiers in all evaluation metrics on both the datasets. We present various classification results for the OLID dataset. The proposed Deep learning based RoBERT yields a maximum accuracy of 94% with BERT embeddings. This experiment demonstrates the significance of finding the interrelations among various word embeddings combinations that are more achievable in BERT when compared with other models.

Fig 7 : Word cloud visualization of our predefined insulting words, i.e., insulting seeds.

TABLE IV : Accuracy on OLID Dataset

Accuracy Scores on OLID Dataset		
Classification Model	Accuracy	F1 Score
Word2Vec + LR	72	57
Word2Vec + SVM	77	62
Word2Vec + RF	77	59
Word2Vec + DT	78	63
FastText + LR	75	61
FastText + SVM	83	68
FastText + RF	85	72
FastText + DT	84	70
BERT + LSTM	85	72
BERT + BERT(Large)	90	76
BERT + RoBERTa	91	78

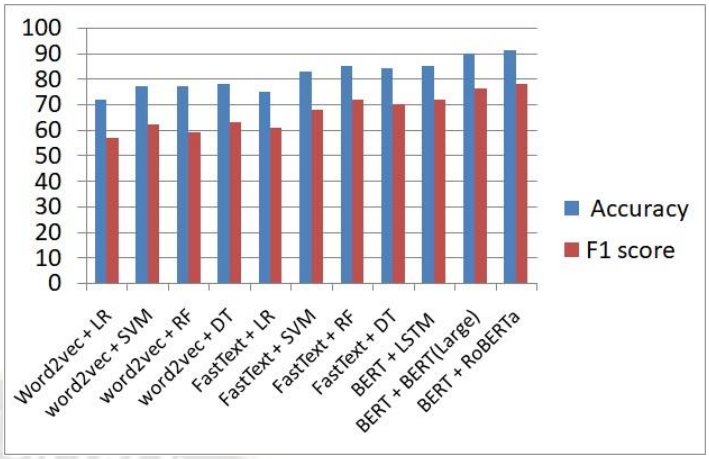

Fig 8 : Accuracy of OLID Dataset

TABLE IV : Accuracy on OLID Dataset

Accuracy Scores on OLID Dataset		
Classification Model	Accuracy	F1 Score
Word2Vec + LR	73	60
Word2Vec + SVM	78	67
Word2Vec + RF	75	64
Word2Vec + DT	79	69
FastText + LR	78	68
FastText + SVM	82	72
FastText + RF	83	73
FastText + DT	82	71
BERT + LSTM	89	76
BERT + BERT(Large)	91	77
BERT + RoBERTa	94	81

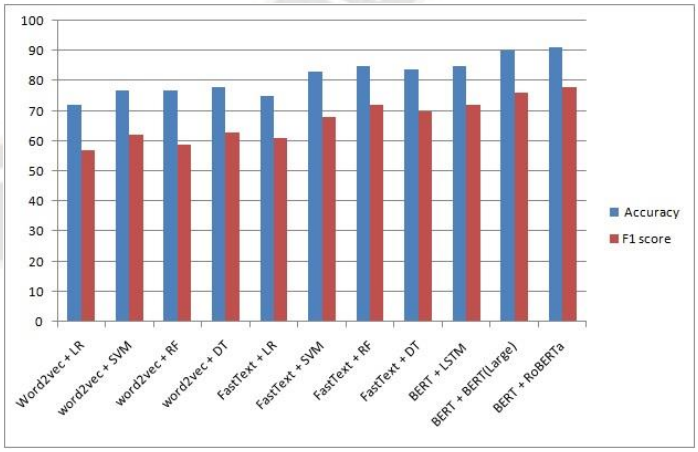

Fig 9 : Accuracy of CORD Dataset

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 1
DOI: https://doi.org/10.17762/ijritcc.v11i1.5981
***Article Received**: 10 November 2022 **Revised**: 14 December 2022 **Accepted**: 26 December 2022*

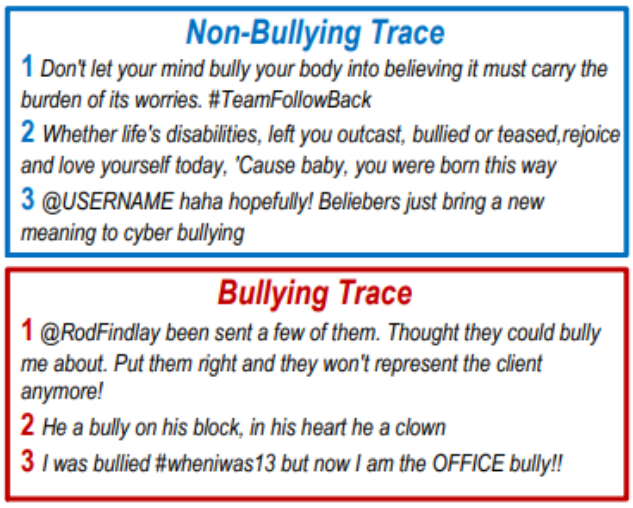

Fig 10 : Offensive Text from the dataset - Some examples from Datasets. Three of them are non-bullying traces. And the other three are bullying traces

VII. Conclusion

A. Selecting a Template (Heading 2)

In this study, we proposed CROD, a deep neural network based scheme to deal with offensive text classification on social networking sites. The novelty of this model is classifying offensive text based on the concept of role and creating a dataset which includes formal and informal roles. We evaluated the existing OLID dataset and the CROD dataset which we collected from social media.

REFRENCES

[1]. ”Razavi, Amir H., et al. ”Offensive language detection using multi-level classification.” Canadian Conference on Artificial Intelligence. Springer, Berlin, Heidelberg, 2010. ”

[2]. ”Bretschneider, Uwe, and Ralf Peters. ”Detecting offensive statements towards foreigners in social media.” Proceedings of the 50th Hawaii International Conference on System Sciences. 2017.”

[3]. ”Kocon, Jan, et al. ”Offensive, aggressive, and hate speech analysis: From data-centric to human-centered approach.” Information Processing & Management 58.5 (2021): 102643.”Yadav, Shashank H., and Pratik M. Manwatkar.

[4]. ”An approach for offensive text detection and prevention in Social Networks.” 2015 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS). IEEE, 2015.”

[5]. ”Bisht, Akanksha, et al. ”Detection of hate speech and offensive language in twitter data using lstm model.” Recent trends in image and signal processing in computer vision. Springer, Singapore, 2020. 243-264.”

[6]. ”Zampieri, Marcos, et al. ”Predicting the type and target of offensive posts in social media.” arXiv preprint arXiv:1902.09666 (2019).”

[7]. ”Ozoh, P. A., M. O. Olayiwola, and A. A. Adigun. ”Identification and classification of toxic comments on social media using machine learning techniques.” International Journal of Research and Innovation in Applied Science (IJRIAS) (2019).”

[8]. ”Pitsilis, Georgios K., Heri Ramampiaro, and Helge Langseth. ”Detecting offensive language in tweets using deep learning.” arXiv preprint arXiv:1801.04433 (2018).”

[9]. ”Van Hee, Cynthia, et al. ”Automatic detection of cyberbullying in social media text.” PloS one 13.10 (2018): e0203794.”

[10]. ”Pitsilis, Georgios K., Heri Ramampiaro, and Helge Langseth. ”Effective hate-speech detection in Twitter data using recurrent neural networks.” Applied Intelligence 48.12 (2018): 4730-4742. ”

[11]. ” Wiedemann, Gregor, et al. ”Transfer learning from lda to bilstm-cnn for offensive language detection in twitter.” arXiv preprint arXiv:1811.02906 (2018).”

[12]. ”Kocon, Jan, et al. ”Offensive, aggressive, and hate speech analysis: From data-centric to human-centered approach.” Information Processing & Management 58.5 (2021): 102643.”

[13]. ”d’Sa, Ashwin Geet, Irina Illina, and Dominique Fohr. ”Classification of Hate Speech Using Deep Neural Networks.” Revue d’Information Scientifique & Technique 25.01 (2020).”

[14]. ”Shang, Lanyu, et al. ”Aomd: An analogy-aware approach to offensive meme detection on social media.” Information Processing & Management 58.5 (2021): 102664.”

[15]. ”Bisht, Akanksha, et al. ”Detection of hate speech and offensive language in twitter data using lstm model.” Recent trends in image and signal processing in computer vision. Springer, Singapore, 2020. 243-264.”

[16]. ”Tesfaye, Surafel Getachew, and Kula Kakeba. ”Automated Amharic Hate Speech Posts and Comments Detection Model Using Recurrent Neural Network.” (2020).”

[17]. ”El-Alami, Fatima-zahra, Said Ouatik El Alaoui, and Noureddine En Nahnahi. ”A multilingual offensive language detection method based on transfer learning from transformer fine-tuning model.” Journal of King Saud University-Computer and Information Sciences 34.8 (2022): 6048-6056.”

[18]. ”Sajid, Tauqeer, et al. ”Roman urdu multi-class offensive text detection using hybrid features and svm.” 2020 IEEE 23rd International Multitopic Conference (INMIC). IEEE, 2020.”

[19]. ”Bestgen, Yves. ”A simple language-agnostic yet strong baseline system for hate speech and offensive content identification.” Forum for Information Retrieval Evaluation (Working Notes)(FIRE), CEUR-WS. org. 2021.”

[20]. ”Ameur, Mohamed Seghir Hadj, and Hassina Aliane. ”AraCOVID19-MFH: Arabic COVID-19 Multi-label Fake News & Hate Speech Detection Dataset.” Procedia Computer Science 189 (2021): 232-241.”

[21]. ”Raj, Mitushi, et al. ”An application to detect cyberbullying using machine learning and deep learning techniques.” SN computer science 3.5 (2022): 1-13.”

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 1
DOI: https://doi.org/10.17762/ijritcc.v11i1.5981
***Article Received**: 10 November 2022 **Revised**: 14 December 2022 **Accepted**: 26 December 2022*

[22]. ”Akram, Muhammad Hammad, and Khurram Shahzad. ”Violent Views Detection in Urdu Tweets.” 2021 15th International Conference on Open Source Systems and Technologies (ICOSST). IEEE, 2021.”

[23]. ”Rana, Toqir A., et al. ”An Unsupervised Approach for Sentiment Analysis on Social Media Short Text Classification in Roman Urdu.” Transactions on Asian and Low-Resource Language Information Processing 21.2 (2021): 1-16.”

[24]. ”Rizwan, Hammad, Muhammad Haroon Shakeel, and Asim Karim. ”Hate-speech and offensive language detection in roman Urdu.” Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP). 2020.”

[25]. ”Sai, Siva, and Yashvardhan Sharma. ”Towards offensive language identification for Dravidian languages.” Proceedings of the first workshop on speech and language technologies for Dravidian languages. 2021.”

[26]. ”Vasantharajan, Charangan, and Uthayasanker Thayasivam. ”Towards offensive language identification for tamil code-mixed youtube comments and posts.” SN Computer Science 3.1 (2022): 1-13.”

[27]. ”Wiedemann, Gregor, Seid Muhie Yimam, and Chris Biemann. ”UHH-LT at SemEval-2020 task 12: Fine-tuning of pre-trained transformer networks for offensive language detection.” arXiv preprint arXiv:2004.11493 (2020).”

[28]. ”Mossie, Zewdie, and Jenq-Haur Wang. ”Vulnerable community identification using hate speech detection on social media.” Information Processing & Management 57.3 (2020): 102087.”

[29]. ”Liu, Ping, Wen Li, and Liang Zou. ”NULI at SemEval-2019 Task 6: Transfer Learning for Offensive Language Detection using Bidirectional Transformers.” SemEval@ NAACL-HLT. 2019.”

[30]. ”Sigurbergsson, Gudbjartur Ingi, and Leon Derczynski. ”Offensive language and hate speech detection for Danish.” arXiv preprint arXiv:1908.04531 (2019).”

[31]. ”Desrul, Dhamir Raniah Kiasati, and Ade Romadhony. ”Abusive language detection on Indonesian online news comments.” 2019 International Seminar on Research of Information Technology and Intelligent Systems (ISRITI). IEEE, 2019.”

[32]. ”De Souza, Gabriel AraAºjo, and M ˜ A¡rjory Da Costa-Abreu. ”Automatic offensive language detection from Twitter data using machine learning and ˜ feature selection of metadata.” 2020 International Joint Conference on Neural Networks (IJCNN). IEEE, 2020.”

[33]. ”Pradhan, Rahul, et al. ”A review on offensive language detection.” Advances in Data and Information Sciences (2020): 433-439.”

[34]. ” Febriana, Trisna, and Arif Budiarto. ”Twitter dataset for hate speech and cyberbullying detection in Indonesian language.” 2019 International Conference on Information Management and Technology (ICIMTech). Vol. 1. IEEE, 2019.”

[35]. ” Chetty, Naganna, and Sreejith Alathur. ”Hate speech review in the context of online social networks.” Aggression and violent behavior 40 (2018): 108-118.”

[36]. ”Al-Hassan, Areej, and Hmood Al-Dossari. ”Detection of hate speech in social networks: a survey on multilingual corpus.” 6th International Conference on Computer Science and Information Technology. Vol. 10. 2019.”

[37]. ”Zhao, Yingjia, and Xin Tao. ”ZYJ123@ DravidianLangTech-EACL2021: Offensive language identification based on XLM-RoBERTa with DPCNN.” Proceedings of the first workshop on speech and language technologies for dravidian languages. 2021.”

[38]. ”Kogilavani, S. V., et al. ”Characterization and mechanical properties of offensive language taxonomy and detection techniques.” Materials Today: Proceedings (2021).”

[39]. ”De Souza, Gabriel AraAºjo, and M ˜ A¡rjory Da Costa-Abreu. ”Automatic offensive language detection from Twitter data using machine learning and ˜ feature selection of metadata.” 2020 International Joint Conference on Neural Networks (IJCNN). IEEE, 2020.”

[40]. ”Garca-Diaz, J. A., Salud Maria Jimenez-Zafra, and Rafael Valencia-Garcia. ”Umuteam at meoffendes 2021: Ensemble learning for offensive language identification using linguistic features, fine-grained negation and transformers.” Proceedings of the Iberian Languages Evaluation Forum (Iber-LEF 2021), CEUR Workshop Proceedings. CEUR-WS. org. 2021.”

[41]. ”Saitov, Kamil, and Leon Derczynski. ”Abusive Language Recognition in Russian.” Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing. 2021.”

[42]. ”Wu, Liang, and Huan Liu. ”Tracing fake-news footprints: Characterizing social media messages by how they propagate.” Proceedings of the eleventh ACM international conference on Web Search and Data Mining. 2018.”

[43]. ”Mridha, Muhammad F., et al. ”L-Boost: Identifying Offensive Texts From Social Media Post in Bengali.” Ieee Access 9 (2021): 164681-164699.”

[44]. ”Qasim, Rukhma, et al. ”A fine-tuned BERT-based transfer learning approach for text classification.” Journal of healthcare engineering 2022 (2022).”

[45]. ”Sherstinsky, Alex. ”Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network.” Physica D: Nonlinear Phenomena 404 (2020): 132306.”

[46]. ”Pennington, Jeffrey, Richard Socher, and Christopher D. Manning. ”Glove: Global vectors for word representation.” Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). 2014.”

[47]. ”Sadiq, Saima, et al. ”Aggression detection through deep neural model on twitter.” Future Generation Computer Systems 114 (2021): 120-129.”

[48]. ”Castorena, Carlos M., et al. ”Deep neural network for gender-based violence detection on Twitter messages.” Mathematics 9.8 (2021): 807.”

[49]. ”Chen, Junyi, Shankai Yan, and Ka-Chun Wong. ”Verbal aggression detection on Twitter comments: convolutional

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 1
DOI: https://doi.org/10.17762/ijritcc.v11i1.5981
***Article Received**: 10 November 2022 **Revised**: 14 December 2022 **Accepted**: 26 December 2022*

neural network for short-text sentiment analysis." Neural Computing and Applications 32.15 (2020): 10809-10818."

[50]. "Elouali, Aya, Zakaria Elberrichi, and Nadia Elouali. "Hate Speech Detection on Multilingual Twitter Using Convolutional Neural Networks." Rev. d'Intelligence Artif. 34.1 (2020): 81-88."

[51]. "Basile, Valerio, et al. "Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter." Proceedings of the 13th international workshop on semantic evaluation. 2019."

[52]. "Duwairi, Rehab, Amena Hayajneh, and Muhannad Quwaider. "A deep learning framework for automatic detection of hate speech embedded in Arabic tweets." Arabian Journal for Science and Engineering 46.4 (2021): 4001-4014."

[53]. "Petrolito, Ruggero, and Felice DellOrletta. "Word embeddings in sentiment analysis." Turin, Italy (2018).

[54]. "de Pelle, Rogers Prates, and Viviane P. Moreira. "Offensive comments in the brazilian web: a dataset and baseline results." Anais do VI Brazilian Workshop on Social Network Analysis and Mining. SBC, 2017.. 2007, pp. 57-64, doi:10.1109/SCIS.2007.357670.