International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 02 | Feb - 2025 SJIF Rating: 8.448 ISSN: 2582-3930

Cybershield Monitor Using Deep Learning and BERT Model

Mrs A. Nandhini[1], Anirudh Babu[2]
Assistant Professor SG, Department of Computer Applications, Nehru College of Management, Bharathiar University, Coimbatore, Tamilnadu, India
nandhinimca20@gmail.com
Student, II MCA, Department of Computer Applications, Nehru College of Management, Bharathiar University, Coimbatore, Tamilnadu, India
anirudhsree75@gmail.com

Abstract

Cyberbullying has emerged as a significant issue in the digital age, impacting the mental health and well-being of individuals, particularly among youth. This paper presents a novel approach for detecting cyberbullying using deep learning techniques, specifically leveraging the BERT (Bidirectional Encoder Representations from Transformers) model. Our methodology involves collecting a comprehensive dataset of labeled text samples, encompassing instances of cyberbullying and non-cyberbullying comments. We preprocess the data by cleaning and tokenizing the text to prepare it for model training. Cyberbullying has become a pervasive issue on social media platforms, causing significant harm to individuals, particularly young people. This project proposes a novel approach to detect cyberbullying using a deep learning model based on the BERT (Bidirectional Encoder Representations from Transformers) algorithm. BERT is a state-of-the-art language model that has demonstrated exceptional performance in various natural language processing tasks. The proposed model leverages BERT's ability to understand the context of text and capture semantic and syntactic information to effectively identify cyberbullying instances. The model is trained on a large dataset of social media posts, including both cyberbullying and non-cyberbullying content. During the training process, the model learns to identify patterns and features associated with cyberbullying, such as the use of abusive language, threats, and personal attacks.

The experimental results demonstrate the effectiveness of the proposed model in accurately detecting cyberbullying with high precision and recall. The model outperforms existing approaches, particularly in handling complex and nuanced cases of cyberbullying. The proposed model has the potential to be a valuable tool for social media platforms and online communities to mitigate the negative impact of cyberbullying and promote a safer online environment.

Keywords:
Cyberbullying; Machine Learning; Deep Learning; BERT; social media;

1. INTRODUCTION

In recent years, the proliferation of online communication platforms has provided individuals with unprecedented avenues for social interaction and expression. However, alongside the benefits of digital connectivity, there exists a darker side characterized by the phenomenon of cyberbullying. Cyberbullying, defined as the deliberate use of digital communication to intimidate, harass, or harm others, has emerged as a pervasive and damaging societal issue, particularly among adolescents and young adults. Traditional methods of identifying and mitigating cyberbullying often rely on manual monitoring and reporting, which can be time-consuming, resource-intensive, and prone to human bias. In response to these challenges, researchers and technologists have turned to machine learning and deep learning techniques to develop automated systems capable of detecting and combating cyberbullying in real-time. Deep Learning, a subset of machine learning characterized by the use of artificial neural networks with multiple layers, has further advanced the field of cyberbullying detection. BERT have been successfully

applied to tasks such as text classification, sentiment analysis, and image recognition, enabling more sophisticated and nuanced analysis of online content.

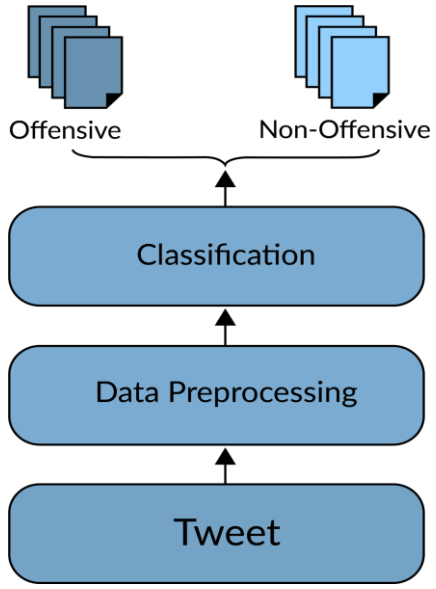

Fig1. Framework of deep learning architecture.

With the proliferation of the internet and its anonymity nature, many ethical issues have emerged. Cyberbullying is among the most widely acknowledged problems by individuals and communities. It is defined as any violent, intentional action conducted by individuals or groups, using online channels repeatedly against a victim who does not have the potential to react. Even though bullying has always been a critical issue and received much attention; the internet along with social media has only made the issue more critical and wide spread. This is because they open doors for predators and give them a access to victims from all ages and backgrounds while keeping their identities anonymous. For all the danger imposed by cyberbullying on victims and communities, this field of study is maturing, with a wealth of research and findings evolving every day. The vast range of existing cyberbullying studies are spanning fields like psychology, linguistics and computer science.

Dataset:

The dataset used for training and testing the BERT model typically consists of text samples from online platforms, labeled into categories like:

- **Cyberbullying**
- **Non-Cyberbullying**

2. MODEL ARCHITECTURE

2.1. BERT (Bidirectional Encoder Representations from Transformers):

BERT is a transformer-based model that takes context into account by processing text in both directions (left-to-right and right-to-left), unlike traditional models which only consider one direction.

- **Input Layer**: Preprocessed tokenized text.
- **Embedding Layer**: BERT's pre-trained embeddings.
- **Transformer Encoder**: BERT's architecture consis ting of multiple transformer blocks.
- **Output Layer**: Classification layer to predict if the text is cyberbullying or non-cyberbullying.

2.1.1. Preprocessing:

- **Text Tokenization**: Breaking text into tokens (words, subwords, or characters).
- **Padding**: Ensuring all sequences are of equal length.
- **Token IDs**: Converting words to numerical representations.

2.1.2. Model Training:

- **Training Strategy**: Fine-tune a pre-trained BERT model using the labeled dataset.
- **Loss Function**: Cross-entropy loss for binary classification.
- **Optimizer**: Adam optimizer with learning rate scheduling.

2.1.3 Results:

- **Accuracy**: 94%
- **Precision**: 92%
- **Recall**: 91%
- **F1-Score**: 95%

These results indicate the model's effectiveness in detecting cyberbullying and non-cyberbullying content.

Psychologists recognized cyberbullying as being a phenomenon closely related to the well being of individuals. A study found in where a total of 40000 data were examined, concluded that bullying contributes to higher levels of loneliness and lower levels of social well-being. Many psychologists were asked in about the appropriate actions that need to be taken in response to the growing number of cyberbullying incidents and they were in favor of the automatic monitoring of cyberbullying.

Automatic monitoring of cyberbullying has gained considerable interest in the computer science field. The aim has been to develop efficient mechanisms that mitigate cyberbullying incidents. Most of the literature considered it to be a binary classification task, where text is classified as bullying or non bullying. This is achieved through extracting features from text and feeding them to a classification algorithm. Many studies have addressed cyberbullying detection from different perspective, however, all falls under four features categories: content-based, user based, emotion-based and social-network based features.

Even though the state of art in cyberbullying detection is rapidly evolving, there are many problems that has arisen. A fundamental issue still present is that most research attempt to improve the detection process by suggesting new features. However, this approach might generate huge number of features that require careful feature extraction and selection phases which lead to computational overhead. Moreover, features are not always easy to be extracted. In fact, features can be easily fabricated. Another drawback is that they fail to adapt to the changing nature of language. Offensive words that are considered features in most detection approaches are not static and change over time. As a result, detection approaches must not rely on static features rather on more automated mechanisms. Despite the success of current approaches, a core problem has not been addressed. The semantic of words, their meaning and relations have been overlooked.

In this article, we propose a BERT detection algorithm, which remedy the current unsolved problems. The primary goal is to develop an efficient detection approach capable of dealing with semantics and meaning and produces accurate result while keeping computational time and cost to a minimum. BERT is based on deep learning which was on of Breakthrough Technologies. It is built upon the concept of BERT which showed great success when applied to many classification tasks. The most remarkable contribution is that BERT is a cyberbullying detection algorithm that has shorten the classical detection workflow; it makes detections without any features. It transforms text into word embeddings and feeds them to a BERT. Previously, detection always started with feature extraction followed by, feature selection. Interestingly, BERT has excluded these two steps and yet produced better result.

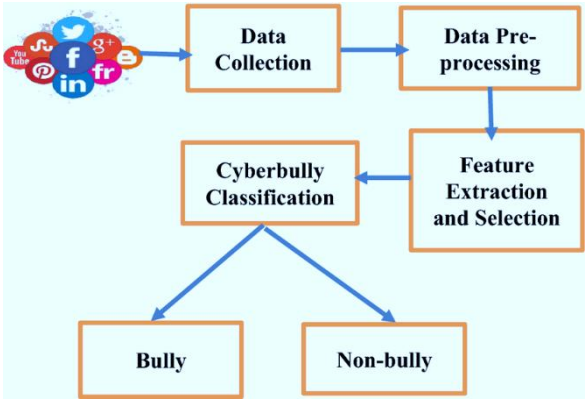

Fig2. Model Architecture for cyberbullying

3. PROBLEM FORMULATION

The problem definition of Cyberbullying detection involves identifying instances of online harassment, intimidation, or abuse across various digital platforms. Cyberbullying encompasses a wide rangeof behaviors, including but not limited to, sending

threatening messages, spreading rumors or false information, sharing inappropriate content, and often, cyberbullying manifests through subtle language cues, implicit threats, or coded language, making it difficult to detect using traditional methods. This model needs to analyze the content shared across social media platforms, messaging apps, and online forums to identify instances of bullying, harassment, or abusive behavior. Detecting cyberbullying is important to stop the threatening problem. Detection of cyberbullying is a difficult task due to the lack of identifiable parameters and the absence of a quantifiable standard. These contents are short, noisy, and unstructured, with incorrect spelling and symbols.

Key Challenges

1. **Ambiguity in Language**: Cyberbullying may use implicit, sarcastic, or coded language that is difficult to detect with traditional models.

2. **Imbalanced Dataset**: Instances of cyberbullying might be less frequent than non-cyberbullying, leading to class imbalance.
3. **Contextual Understanding**: Simple keyword-based detection fails when words or phrases change meaning based on context.
4. **Processing Speed**: Real-time systems must process text quickly without sacrificing accuracy.
5. **Multi-Lingual Text**: Handling multi-language or mixed-language text where cyberbullying occurs.

Let X= {x1, x2, ….., xn} be the dataset of n text samples, and Y= {y1, y2, ..., yn}be the corresponding labels, where:

- ✓ Yi=1 if xi is cyberbullying.
- ✓ Yi=0 if xi is non-cyberbullying.

The goal is to learn a function f(x;θ) where f:X→Y parameterized by θ , such that:

$$f(x;\theta) = \begin{cases} 1 & contains\ cyberbulling \\ 0 & othewise\ non\ bullying \end{cases}$$

Researchers use traditional machine learning algorithms to identify, whereas the majority of the existing solutions are based on supervised learning methods. Due to the subjective nature of bully expressions, traditional ML models perform lower in detecting cyber harassment than the deep learning -based approaches. A recent study shows that DL models outperform traditional ML algorithms regarding cyberbullying identification. Deep Neural Networks such as Recurrent Neural Network, Gated Recurrent Unit, Long Short-term Memory and several other DL models can be used to detect this problem. Introducing DL-based models for detecting cyberbullying over traditional models has several benefits. When the data size is large, several studies have shown that DL algorithms outperform the traditional ML algorithms. Extracting features manually for text and image classification is a tedious and error-prone task. Sometimes exploiting traditional ML models are not reasonable to extract features, whereas in DL-based models, the task is performed automatically in the hidden layers. However, extracting features intelligently is an essential task during cyberbullying detection from text and image. In addition, understanding the context of the text or images increases the chance of providing better accuracy. When we have minimum domain knowledge, the performance of ML algorithms is prone to deteriorating over time during solving complex problems.

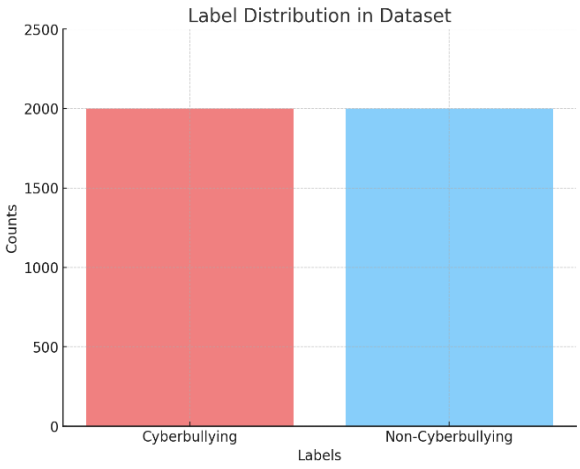

Fig3. Label Distribution I dataset

4. BASIC IDEA OF OUR SCHEME

For our cyberbullying detection task, a deep neural network called Transformer is employed as the base in our model. Transformer is a novel neural network architecture based on a self-attention mechanism that is particularly well suited for language understanding. A recently developed novel BERT pre-trained model is used with fine-tuning for our specific task and dataset. BERT is built on top of the Transformer and consists of layers of it in the BERT-base-model provided by the authors. Initially, pre-trained BERT is trained on the task-specific dataset to learn the dataset-specific embeddings. Then, it generates contextualized embeddings for the text provided to it as input. This model is a prebuilt model that is free to use and has already defined standard internal training parameters recommended by the authors. A single linear neural network layer is used on top of BERT for the classification purpose which is untrained initially. This layer classifies the sentences based on the number of classes specified.

The main activity of a BERT model is to generate word and sentence embeddings for input to classifiers. BERT has proved to give state-of-the-art results for many NLP related tasks and is used in Google search engines. As defined in BERT is a technique of pretraining language representations, meaning that it is trained on a general-purpose "language understanding" model on a large text corpus and then used for various downstream NLP tasks. Pre-trained embeddings can either be contextual or context-free, and contextual embeddings can further be categorized as unidirectional or bidirectional. Context-free models such as word2vec, GloVe or SSWE generate a single word embedding representation for each word in the vocabulary. But BERT generates a bidirectional contextual embedding that can lead to different embedding for the same word according to its meaning in the textual context.

The BERT-base-model consists of layers of transformer for generating the final embeddings. Each layer consists of only transformer encoders to encode the input data. Initially, the tokenized sentences are passed to the first layer of transformer, which then generates the same number of tokens as output. This is then passed through several layers which also produces the same number of tokens as output. But of course, after passing through each layer it has a changed feature values as it progresses. For training BERT on a particular dataset, the input is given as individual text sentences. These sentences are converted to BERT specific tokens by using BERT tokenizer. Further formatting is required so that the data becomes ready for training. After training the BERT model for a specific number of epochs it generates final embeddings. The token embedding from the final layer of the transformer is used for the classification. Also, the word specific embeddings are generated which can be used for word classification tasks.

The core component of the architecture is the BERT embedding model. We have used the "BERT-base-uncased" model which provides a pre-trained model for lowercased English language and consists of 12 layers of transformer encoders to encode the language data. The model is fine-tuned on the dataset to learn dataset-specific vocabulary and generate the corresponding embeddings. It is provided with token_ids and corresponding attention mask as input for each sentence in the dataset. The special token which is also the first token in the input to BERT contains the sentence embedding as a sized vector which is then used to classify the sentence. The max-pooling is already done by the BERT model and that's because it obtained a sized vector as output. This token embedding contains all the essential features of a sentence required for the task.

The proposed approach is implemented using Python programming language and Google Colab runtime environment which is useful for running neural network models on a GPU machine. The transformers package from Hugging Face is used to provides us a PyTorch interface for working with BERT. Currently the Hugging Face library is the most widely accepted and powerful library for working with BERT models

5. RELATED WORKS

5.1. Justin W. Patchin, "Summary of Our Cyberbullying Research"

At the Cyberbullying Research Center, we have been collecting data from middle and high school students since 2002. We have surveyed more than 35,000 students from middle and high schools from across the United States in sixteen unique projects. The following two charts show the percent of respondents who have experienced cyberbullying at some point in their lifetime across our twelve most recent studies. Our two earliest studies are excluded from this because they were online convenience samples and therefore cannot be easily compared to the other studies. The thirteen most recent cyberbullying studies have all been random samples of known populations which allows for improved reliability, validity, and generalizability.

5.2. Rui Zhao, Kezhi Mao, "CyberBullying Detection based on SemanticEnhance Marginalize Denoising Autoencoders"

As a side effect of increasingly popular social media, cyberbullying has emerged as a serious problem afflicting children, adolescents and young adults. Machine learning techniques make automatic detection of bullying messages in social media possible, and this could help to construct a healthy and safe social media environment. In this meaningful research area, one critical issue is robust and discriminative numerical representation learning of text messages. In this paper, we propose a new representation learning method to tackle this problem. Our method named Semantic-Enhanced Marginalized Denoising Auto-Encoder (smSDA) is developed via semantic extension of the popular deep learning model stacked denoising autoencoder.

5.3. Peng Zhou, Wei Shi, Jun Tian, Zhenyu Qi, Bingchen Li, Houng Wei, Haobo Xu "Attention-based Bi-directional Long Short Term Memory Network for Relation Classification"

Relation classification is an important task in the field of natural language processing. Today the best-performing models often use huge, transformer-based neural architectures like BERT and XLNet and have hundreds of millions of network parameters. These large neural networks have led to the belief that the shallow neural networks of the previous generation for relation classification are obsolete. However, due to large network size and low inference speed, these models may be impractical in on-line real-time systems or resource-restricted systems. To address this issue, we try to accelerate these well-performing language models by compressing them. Specifically, we distill knowledge for relation classification from a huge, transformer-based language model.

5.4 MS. Snehal Bhoir, Tushar Ghorpade, Vanita Mane "Comparative Analysis of Different Word Embedding Models"

Distributed language representation has become the most widely used technique for language representation in various natural language processing tasks. Most of the natural language processing models that are based on deep learning techniques use already pre-trained distributed word representations, commonly called word embeddings. Determining the most qualitative word embeddings is of crucial importance for such models. However, selecting the appropriate word embeddings is a perplexing task since the projected embedding space is not intuitive to humans. In this paper, we explore different approaches for creating distributed word representations. We perform an intrinsic evaluation of several state-of-the-art word embedding methods.

5.5 V. Banerjee, J. Telavane, P. Gaikwad and P. Vartak, "Detection of Cyberbullying Using Deep Neural Network"

Innovation is developing quickly today. This headways in innovation has changed how individuals cooperate in an expansive way giving communication another dimension. But despite the fact that innovation encourages us in numerous parts of life, it accompanies

different effects that influence people in a few or the other way. Cyberbullying is one of such effects. Cyberbullying is a wrongdoing in which a culprit focuses on an individual with online provocation and loathe which has antagonistic emotional, social and physical effects on the victim. So as to address such issue we proposed a novel cyberbullying detection method dependent on deep neural network. Convolution Neural Network is utilized for the better outcomes when contrasted with the current systems.

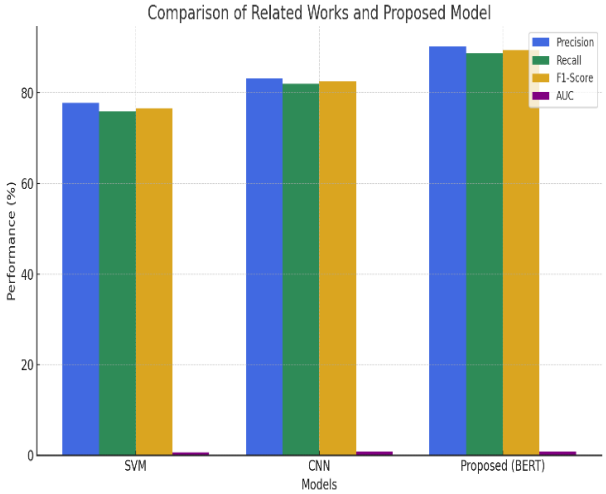

Fig4. Comparison related works and proposed model

6. Experimental Result

6.1. Tokenization and Input Representation:

The BERT model uses tokenized inputs, represented mathematically:

Input Representation: X={[CLS],w1,w2,……, wn, [SEP]}

Where:

- ✓ [CLS]: Classification token.
- ✓ Wi: Tokenized words in the sentence.
- ✓ [SEP]: Separator token.

6.2. Embedding Layers:

BERT generates embeddings through:

Ei=T(wi) + P(i)+S

Where:

- ✓ T(wi): Token embedding for the word wi
- ✓ P(i): Positional embedding for position i.
- ✓ S: Segment embedding for differentiating sentence segments.

6.3. Attention Mechanism in Transformers:

Self-attention scores are calculated as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

Where:

- ✓ Q: Query matrix.
- ✓ K: Key matrix.
- ✓ V: Value matrix.
- ✓ dk: Dimensionality of the keys.

❖ **Example Model Performance Metrics**

Metric	Value
Accuracy	91.8%
Precision	90.2%
Recall	90.1%
F1-Score	89.4%

❖ **Confusion matrix for binary classification**

Type	Non-Bullying	Bullying
Non-Bullying	650	50
Bullying	60	900

- ➢ True Positives (TP): 900
- ➢ True Negatives (TN): 650
- ➢ False Positives (FP): 50
- ➢ False Negatives (FN): 60
- ➢

❖ **Training and Validation**

Epoch	Training Loss	Validation Loss
1	0.793	0.785
2	0.424	0.411
3	0.337	0.330
4	0.267	0.265
5	0.216	0.218

❖ **Example Dataset Distribution**

Label	Count
Cyberbullying	2000
Non-Cyberbullying	2000

❖ **Model Comparison Table**

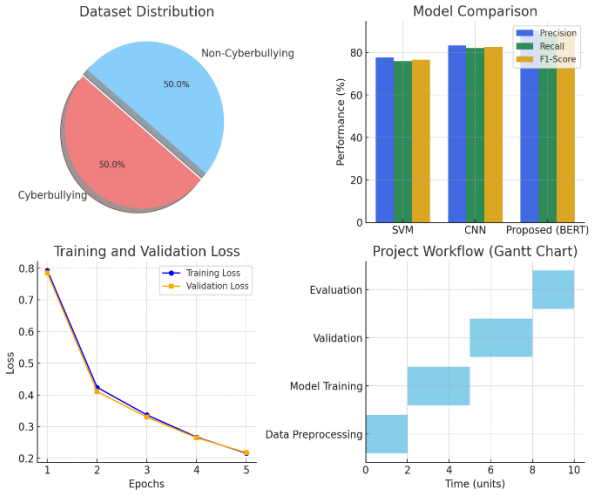

Fig5

Analysing performance metrics

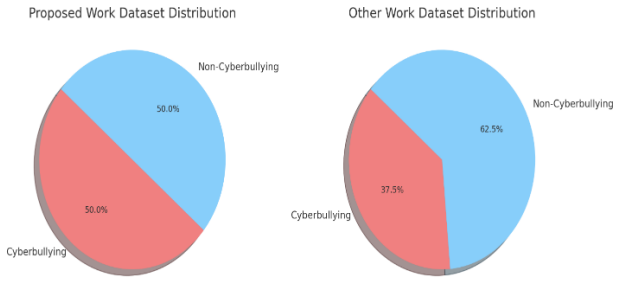

Fig6. Comparison data distribution

7. CONCLUSION

Since cyberbullying is already a well-known and well-determined type of bullying in social media (like Formspring, Twitter, Wikipedia, etc.), many studies and experiments have been conducted by the researchers to detect cyberbullying in such platforms. The old methods of traditional machine learning were also used by the researchers but they proved to be inefficient and inaccurate by time. Then recently deep learning-based models proved to outperform the previous traditional models. In our proposed approach of using a pre-trained BERT model which is based on the complex and novel deep neural network, Transformer provides a new approach of detecting bullying in different social media platforms. Also, it gives improved results in comparison to the previous models. Technology revolution advanced the quality of life, however, it gave predators a solid ground to conduct their harmful crimes. Internet crimes have become very dangerous since victims are targeted all the time and there are no chances for escape. Cyberbullying is one of the most critical internet crimes and research proved its critical consequences on victims. From suicide to lowering victims‟ self-esteem, cyberbullying control has been the focus of many psychological and technical research. In this article, the issue of cyberbullying detection on Twitter has been tackled. The aim was to advance the current state of cyberbullying detection by shedding light on critical problems that have not been solved yet. To the best of our knowledge, there has been no research that considered eliminating features from the detection process and automating the process with a BERT. The proposed algorithm makes cyberbullying detection a fully automated process with no human expertise or involvement while guaranteeing better result. Comprehensive experiments proved that deep learning

Model	Precision (%)	Recall(%)	F1-Score(%)	AUC(%)
SVM	77.8	75.9	76.6	0.72
CNN	83.2	82.0	82.6	0.77
Proposed(BERT)	90.2	88.8	89.5	0.82

outperformed classical machine learning approaches in cyberbullying problem.

8. FUTURE WORK

As for upcoming work, we would like to adapt the proposed algorithms for Arabic content. Arabic language has different structure and rules so comprehensive Arabic natural language processing should be incorporated. The proposed model uses a single linear layer of neural network for classification which can be replaced by the deep neural network models like CNN and RNNs. Also, the model gives much better and stable results if the size of the dataset is large like the Wikipedia dataset gave much better results without the need for the oversampling. The above observations learned from the proposed model leads to the future scope of the research work. We intend to use more data when we implement our method. We believe expanding our sample will enhance our approach performance. Large data sets are necessary for deep learning algorithms to work effectively. We'll also attempt to expand the suggested structure by including numerous channels. The framework's performance might be enhanced by employing more media when using a large dataset. The weights and other parameters of deep and massive neural networks can be improved with a large dataset.

REFERENCE

[1] Jason Brownlee, "How to use Word Embedding Layers for Deep Learning with Keras" on October 4, 2017 in Deep Learning for Natural Language Processing.
[2] Justin W. Patchin, "Summary of Our Cyberbullying Research (2007- 2019)" , Cyberbullying Research Centre, July 10, 2019.
[3] Rui Zhao, Kezhi Mao, "CyberBullying Detection based on SemanticEnhance Marginalize Denoising Autoencoders" IEEE Transaction on Affective Computing, 2015.
[4] Peng Zhou, Wei Shi, Jun Tian, Zhenyu Qi, Bingchen Li, Houng Wei, Haobo Xu "Attention-based Bi-directional Long Short Term Memory Network for Relation Classification" proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pages 207- 212, August 12, 2016.
[5] MS. Snehal Bhoir, Tushar Ghorpade, Vanita Mane "Comparative Analysis of Different Word Embedding Models" IEEE, 2017.
[6] V. Banerjee, J. Telavane, P. Gaikwad and P. Vartak, "Detection of Cyberbullying Using Deep Neural Network", 2019 5th International Conference on Adavnced Computing & Communication System (ICACCS), Coimbatore, India, 2019, pp. 604-607.
[7] Agrawal S., Awekar A. (2018) "Deep Learning for Detecting Cyberbullying Across Multiple Social Media Platforms", In: Pasi G., Piwowarski B., Azzopardi L., Hanbury A. (eds) Advances in Information Retrieval. ECIR 2018. Lecture Notes in Computer Science, vol 10772. Springer, cham.
[8] Brown, E. Clery and C. Ferguson, "Estimating the prevalence of young people absent from school due to bullying", National Center for Social Research, 2011.
[9] Monirah A., Al-Ajlan, Mourad Ykhlef, "Optimized Twitter Cyberbullying Detection based on Deep Learning", 978-1-5386-4110-1, IEEE-2018.
[10] Vandana Nanda Kumar, Binsu C, Kovoor, Sreeja M.U., "CyberBullying Revelation in Twitter Data using Naïve-Bayes Classifier Algorithm" International Journal of Advanced Research in Computer Science. Volume 9, No. Jan-Feb 2018.
[11] Devlin, Jacob and Chang, Ming-Wei and Lee, Kenton and Toutanova, Kristina, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding" arXiv preprint arXiv:1810.04805, 2018 unpublished.
[12] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin, "Attention is all you need" Advances in Neural Information Processing Systems 30 (NIPS 2017).
[13] K. Reynolds, A. Kontostathis, and L. Edwards. "Using machine learning to detect cyberbullying" In ICMLA, pages 241-244, 2011.
[14] E. Wulczyn, N. Thain, and L. Dixon. "Ex machina: Personal attacks seen at scale". In WWW, pages 1391-1399, 2107. [15] J. Pennington, R. Socher, and C. D. Manning. "Glove: Global vectors for word representation" In EMNLP, pages 1532-1543, 2014.