Comparison of State – of – Art Deep Learning Algorithms for Detecting Cyberbullying in Twitter

Janice Marian Jockim [1], Meghana K [1] and Karthika S [1]

[1] *Department of Information Technology, SSN College of Engineering, Chennai, Tamil Nadu, India*

Abstract

As the prominence of networking through social media is intensifying, the people's interest is also tuned towards extensive usage of social media for showcasing their current activities. An important kind of online data threat is cyberbullying which is described as the intentional and repeated use of technology by a person or group of individuals to upset or harm a person's or community's social-psychological attitude. Generally, cyberbullying and prejudice based on gender, ethnicity, physical and mental disabilities and religion are frequently linked from the text, image, audio and videos disclosed by the user. Cyberbullying might lead to many negative consequences like high risks of losing self-confidence, depression, disclosure of sensitive private information leading to self-harm and suicide. These impacts necessitate the need for analyzing the harmful bullying and discriminative social media posts and support the users by protecting them from regrets and depression. This research work is a comparison of the state-of-art models that can be used for identifying the cyberbullying content posted on a social media platform and classifies the severity of the content. The user generated content (UGC) is highly varied from Twitter. Two Machine Learning models namely SVM (Accuracy – 84%) and Naïve Bayes (Accuracy – 83%) were tested. The accuracies were found to be low due to the instability and the complexity of the model, and the highly dynamic nature of variables. Hence, Deep Learning models were tested as they use Natural Language Processing and Predictive Modeling which gives high accuracies. Six Deep Learning models namely BiLSTM + Fasttext (Accuracy – 84.83%), BiLSTM + GloveTwitter (Accuracy – 85.83%), BERTBase (Accuracy – 89%), RoBERTa (Accuracy – 89.14%), DistilBERT (Accuracy – 87.09%), BERTweet (Accuracy – 93%) were tested and BERTweet was found to have the highest accuracy since BERTweet model has been trained with data specific to Twitter. This research work identified that the Universal Sentence Encoder Model was the most efficient with the final accuracy of 96.08%.

Keywords

Cyberbullying, Twitter, Machine Learning, Deep Learning, BERT, BiLSTM, Universal Sentence Encoder

1. Introduction

As the prominence of networking through social media is intensifying, the people's interest is also tuned towards extensive usage of social media for showcasing their current activities [7]. Online socialising involves sharing links, exchanging information, images and videos through mobile and internet platforms [9]. Though sharing information in social media gives pleasure, the reach of the message and the consequences has unlimited bounds or could not be limited as the user believes, leading to harmful repercussions. One such consequence is cyberbullying. The government has taken steps to ensure that cyberbullying is not tolerated by enacting the Anti-Cyberbullying Act. According to Section

ACM-2022: Algorithms Computing and Mathematics Conference, August 29 – 30, 2022, Chennai, India.
EMAIL: janicemarianjockim18040@it.ssn.edu.in (Janice Marian Jockim)
ORCID: 0000-0001-9832-1686 (Janice Marian Jockim)

67 of the legislation, publishing or sending inappropriate videos in electronic form is punishable by up to five years in jail and a fine of ten lakh rupees.

Deep neural network-based (DNN) models have recently been used to detect cyberbullying. DNN models have been used by certain authors to detect cyberbullying, and their models have been broadened to include numerous social media sites. Their models beat typical ML models based on their reported results, and more crucially, the authors have indicated that they used transfer learning, which suggests that their generated models for detecting cyberbullying can be extended and used on other datasets. The authors of this research work suggest that modern contextual language models like BERT and USE can be used to more effectively detect cyberbullying.

The motivation of this work is the fact that cyberbullying is often linked with discrimination based on gender, race, faith, sexual orientation, physical and mental disabilities from the text, image, audio and videos disclosed by the user. Cyberbullying might lead to many negative consequences like high risks of losing self-confidence, depression, disclosure of sensitive private information leading to self-harm and suicide. These impacts necessitate the need for analyzing the harmful bullying and discriminative social media posts and support the users by protecting them from regrets and depression.

The problem identified is that an important kind of online data threat is cyberbullying which is described as the intentional and repeated use of technology by a person or group of individuals to upset or harm a person's or community's social-psychological attitude. Cyber Bullying is one of the biggest issues in today's world. There is no age where cyberbullying is accepted, nor does it stop. Cyberbullying is now much harder to control as it is done through social networking sites. The problem faced is to develop a technical method that can aid in the identification of cyberbullying. Modern algorithms for recognising and reporting incidents of bullying on social media platforms will be examined and compared by the authors.

The objective is to identify the cyberbullying content posted on a social media platform, identify the tweet nature – Offensive or Not Offensive and to compare state-of-art algorithms and to prove Universal Sentence Encoder is the best performing algorithm.

2. Existing Works

In the following section, the authors introduce various papers that have proposed enhanced methodologies for detection of cyberbullying.

In the study [1], Twitter, Wikipedia talk pages (a collaborative knowledge repository), and Formspring (a Q&A forum) are all used. (A microblogging service). These datasets are all available to the public and each one has been carefully labelled. This work effectively replicated the reference literature for detecting cyberbullying incidents on social media sites using DNN-based models. The majority of the source codes and papers were logically arranged and simple to find. The work was expanded by using a new social media dataset—YouTube—to examine the models' transferability and adaptability to the new dataset as well as to compare the performance of the DNN models to the conventional ML models that had previously been used in studies on the YouTube dataset for cyberbullying detection.

In the work of the authors of paper [2], based on some features, a supervised machine learning method for categorising the severity of cyberbullying via Twitter was developed. The study used PMI-semantic orientation, embedding, sentiment, and lexicon features. To apply the retrieved features, the techniques Naive Bayes, KNN, Decision Tree, Random Forest, and Support Vector Machine were utilised. In terms of Kappa, classifier accuracy, and f-measure metrics, experiments employing the provided framework in a multi-class scenario as well as in a binary setting show promise. These results imply that the proposed framework is an effective way to identify cyberbullying and the severity of the problem in online social networks. The results of several machine learning approaches were then

compared to the suggested and baseline attributes. The comparison's outcomes demonstrate that the proposed features are important in detecting cyberbullying.

In the paper [3], three popular deep learning algorithms—CNN, LSTM, and BiLSTM—were compared in the comparison analysis. The results of the proposed study show that, despite a significant difference in training time, BiLSTM outperforms other models in terms of accuracy (0.9745). In addition to having slightly lower test accuracy, the BiLSTM model is 65 times slower than the 1D-CNN model. It is evident that 1D-CNN can be applied in circumstances when computational resources are constrained. The 1D-CNN and RNN models are effective at identifying tokenized words.

On two real-world cyberbullying datasets, the paper [4] presents a special neural network framework with parameter optimization and a comparative analysis of eleven classification techniques using algorithms, including four typical machine learning techniques and seven shallow neural networks. The main outcomes of this study are that bidirectional neural networks and attention models generate high classification results. The best classifier among the established machine learning classifiers was found to be Logistic Regression. Term FrequencyInverse Document Frequency (TFIDF) routinely achieves good accuracy using standard machine learning approaches. The performance of Global Vectors (GloVe) is improved by neural network models. The two most effective neural networks were BiGRU and BiLSTM.

Extensive tests were conducted in this study [5] utilising three real-world datasets: Wikipedia, Twitter, and Formspring (12k posts each) (100k posts). The tests offer some insightful information on how to identify cyberbullying. This is the first study to extensively analyse the detection of cyberbullying across several SMPs using deep learning-based models and transfer learning.

The goal of this paper [6] is to address some of the difficulties brought up in order to improve the problem of identifying cyberbullying. It is suggested to 1) use contemporary contextual language models, such as BERT, to detect cyberbullying, and 2) create better representations of datasets linked to cyberbullying using slang-based word embeddings. The results show that BERT outperforms cutting-edge deep learning models and cyberbullying detection techniques. The results show that deep learning models initialised with slang-based word embeddings outperform deep learning models initialised with conventional word embeddings.

The contributions of the paper [7] are a dual facet: first, it is empirically shown that character n-gram features can improve the effectiveness of the state-of-the-art RNN techniques for abusive language identification.

The goal of this paper [8], building a character-level model that learns to anticipate embeddings for unknown words solves the problem of intentionally noisy input. On three datasets from two different domains, namely Twitter and Wikipedia talk pages, the combination of this model with character-enhanced RNN techniques advances the state of the art in abuse identification.

In the work by the authors of paper [9], eight datasets covering a range of behaviours that meet the general definition of cyberbullying were selected. Numerous of these datasets include labels that designate particular types of behaviours that are either absent from or have distinct definitions in other datasets. Each dataset was used to train deep neural network systems, which were then used to test how well they might be applied to different domains. Finally, different approaches to creating ensemble models by mixing classifiers were researched.

In the paper [10], it is suggested to use supervised machine learning to identify and stop cyberbullying. Several classifiers are used to teach and identify bullying behaviours. The recommended method outperforms SVM on the cyberbullying dataset, with an accuracy of 92.8 percent compared to 90.3 percent for SVM. NN outperforms other classifiers that have performed comparable work on the same dataset.

The goal of the study in [11] by maximising the resource-intensive training of ML models for Natural Language Processing, is to decrease the number of necessary experiment iterations. It relies on earlier work with FD to enhance the effectiveness of classifier training while also assessing the effectiveness of a number of linguistically-based feature pre-processing techniques for dialogue categorization, particularly for the identification of cyberbullying.

Table 1: Summary of algorithms applied for cyberbullying detection

S. NO	AUTHORS	DATASET	METHODOLOGY
1	Maral Dadvar et al. [1]	Formspring, Wikipedia talk pages and Twitter	Deep Learning, Neural Networks, Social Networks, Transfer Learning
2	Bandeh Ali Talpur et al. [2]	The leading annotated corpus for study on harassment	Embedding, Sentiment, and Lexicon features along with PMI-semantic orientation, Naïve Bayes, KNN, Decision Tree, Random Forest, and Support Vector Machine algorithms.
3	Sourodip Ghosh et al.[3]	The dataset, which consists of 159, 571 text instances from 6 different target classes, was contributed by the research team at Jigsaw LLC.	CNN, LSTM, BiLSTM
4	Chahat Raj et al. [4]	Attack data from Wikipedia and web toxicity data from Wikipedia	Machine learning, Neural Networks, Deep Learning, Natural Language Processing
5	Sweta Agrawal et al. [5]	Twitter (16k posts), Wikipedia (100k posts), and Formspring (12k posts)	Deep Learning
6	Fatma Elsafoury et al. [6]	Reddit, Wikipedia Talk Pages, FormSpring, Ask.FM, MySpace, YouTube, Vine, Twitter, Instagram, and Yahoo News	Deep Learning
7	Pushkar Mishra et al. [7]	Twitter and Wikipedia talk page	RNN
8	Marc-Andr Larochelle et al. [8]	Eight datasets were chosen, and they were gathered from platforms using various message formats, such as short-form tweets, question-and-answer sets, and forum discussions	Natural Language Processing, Deep Learning, Cross-Domain Generalization
9	John Hani et al. [9]	The authors Kelly Reynolds et al. collected and categorised a Kaggle dataset on cyberbullying	Machine Learning, Neural Network
10	Juuso Eronen et al. [10]	Reynolds et al(2011).'s Kaggle Formspring Dataset for Cyberbullying Detection	Feature Density, Machine Learning, Deep Neural Networks

3. Proposed System Design

The system's overall design is addressed in this section. It is introduced and briefly discussed how the workflow works. Input Data, Data Preprocessing, Feature Extraction, Model Training, Optimization, and Model Evaluation phases make up the workflow.

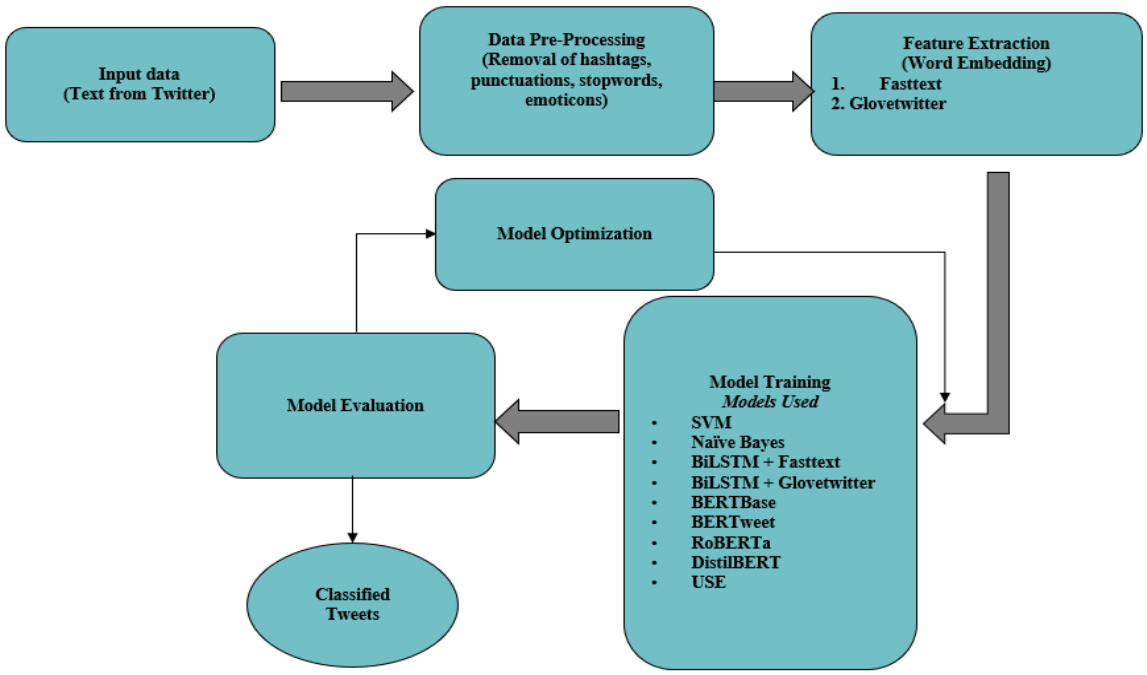

Figure 1: Proposed System Design for Cyberbullying Detection Framework

3.1. Input Data

The dataset study is presented in Table 2. Two datasets from Mendeley and four datasets from Kaggle were studied. Two datasets had social media content from all the platforms while four datasets had social media content specific to Twitter.

Table 2: Analysis of Kaggle and Mendeley Cyberbullying Datasets

S. No	Name and Source	Size	Platform	No. of Features
1	Toxicity – Mendeley	1 lakh 50k content	All platforms	3
2	Twitter parsed - Mendeley	16k tweets	Twitter	2
3	Toxic Comment Classification - Kaggle	1 lakh 60 k tweets	Twitter	4
4	Cyberbullying tweets - Kaggle	49 k tweets	Twitter	5
5	Malignant train - Kaggle	1 lakh 60 k content	All platforms	6
6	Classified tweets - Kaggle	20 k tweets	Twitter	4

The final dataset chosen is Cyberbullying tweets from Kaggle. This dataset was balanced since it had unequal number of tweets for each class. The characteristics after balancing are presented in the Table below.

Table 3: Features of the Dataset Before and After Balancing

Features	Before Balancing	After Balancing
Non Cyberbully	7,945	32,095
Cyberbully	39,747	39,747
Age	7992	7992
Gender	7973	7973
Religion	7998	7998
Ethnicity	7961	7961
Other	7823	7823
Total	47,692	71,842

3.2. Data Pre – Processing

The dataset required for creating a model for cyberbullying detection is difficult to obtain due to ethical and privacy concerns. Thus, for this research work, a dataset of cyberbullying tweets has been collected through a data collection platform, Kaggle. The dataset obtained is combed through in search for outliers and is modified into a manner easily understood by a machine learning or deep learning model. Label encoding and standardization of specific columns have been done to ensure so. To train several models and evaluate their effectiveness on the test set, this altered dataset is divided into a train set and a test set. The steps for data pre – processing is shown in the figure below.

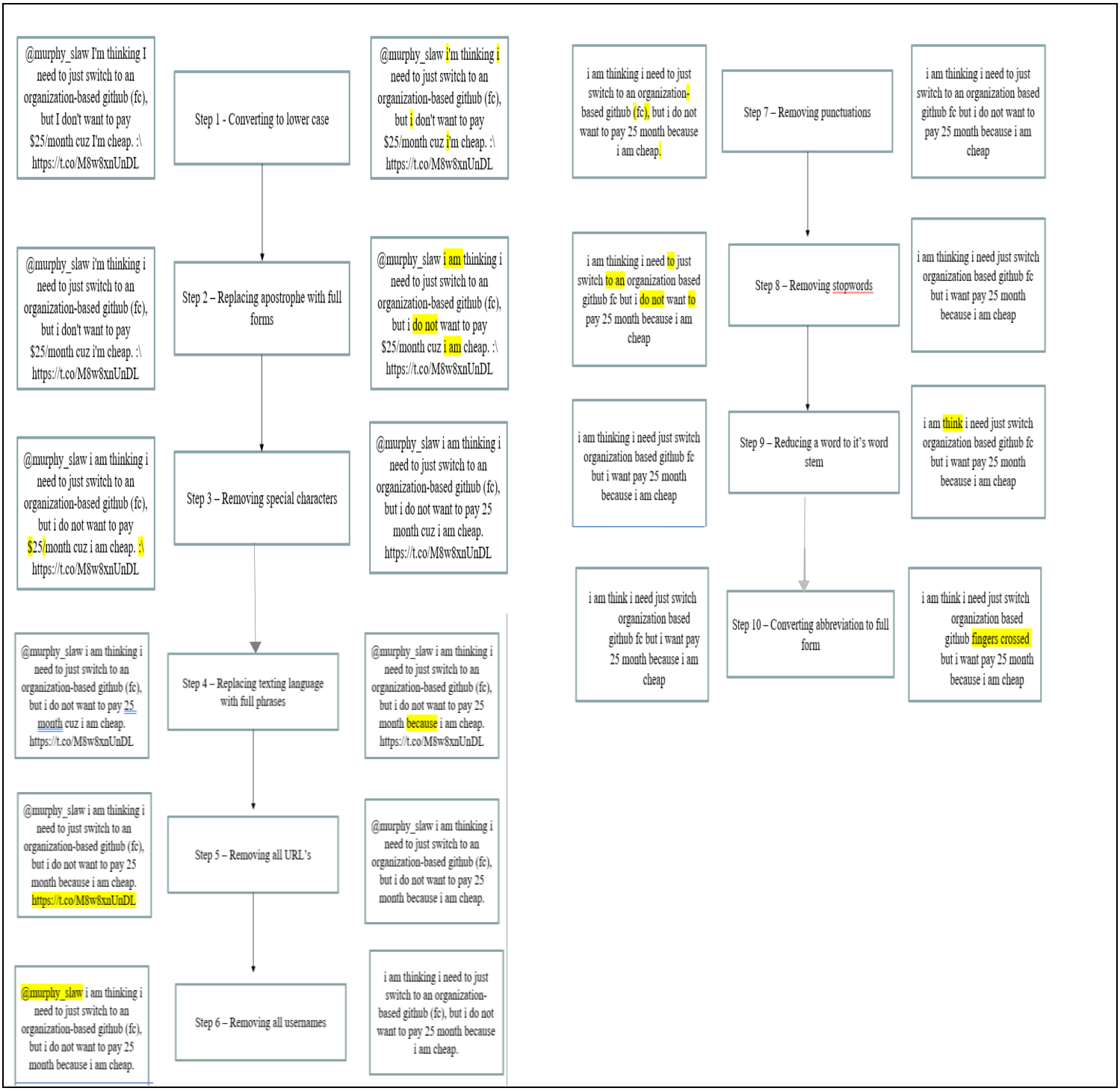

Figure 2: Pre – Processing of a Sample Tweet

3.3. Feature Extraction

Words with similar meanings are stored in a word embedding, which is a representation of text that has been learned. This method of expressing words and documents may be one of the major breakthroughs of deep learning on challenging natural language processing challenges. Word embeddings are a class of methods in which individual words are represented as real-valued vectors in a specified vector space. Each word is given its own vector, and the vector values are learned in a manner like a neural network. A real-valued vector with tens or hundreds of dimensions encodes each word. On the other hand, sparse word representations demand thousands or millions of dimensions. In

this research work, two types of Word Embedding Algorithms have been used: i. Fasttext – A word embedding algorithm which is pre-trained with data majorly from news articles [15] [16]. ii. GloveTwitter – A word embedding technique which is pre-trained with data majorly from twitter [17] [18].

3.4. Models

For this research work, two Machine Learning models (SVM, Naïve Bayes), six Deep Learning models (BiLSTM + Fasttext, BiLSTM + GloveTwitter, BERTBase, RoBERTa, DistilBERT and BERTweet) and one Transfer Learning Model (Universal Sentence Encoder) have been used.

3.4.1. Machine Learning Models

The main idea of SVM is to find separators that can best identify different classes in a search space [11]. SVM is a supervised learning technique for pattern recognition that can categorise both linear and non-linear data [12]. Support vectors are data points that use crucial training tuples to distinguish one or more hyperplanes. SVM has traditionally been used for binary classification.

One of the most effective and efficient inductive learning algorithms is naive Bayes, which has been used as a classifier in a number of social media studies [13]. It is frequently employed in document categorization projects and has the ability to categorise any type of data, including text, network attributes, phrases, and others. The most basic Nave Bayes classifier was used in this study to classify textual traits and word embeddings [14]. The probability using Naïve Bayes theorem was calculated as shown in equation 1.

$$P(A|B) = P(B|A) * P(A) / P(B) \quad (1)$$

Where
P(A) = the probability that event A occurs,
P(B) = the probability that event B occurs,
P(B|A) = the probability that event B occurs, given A has already occurred,
P(A|B) = the probability that event A occurs, given B has already occurred

3.4.2. Deep Learning Models

The BiLSTM is a more advanced version of the LSTM. BiLSTM will receive input in two different ways, which is the main difference between the two [19]. The first is from the start to the end of a series, while the second is from the finish to the start. The model may now save data from the past as well as the future. When dealing with text data, this extra feature improves the LSTM's performance. Many long text sequences provide essential information at the end, and BiLSTM are the ideal solution in these cases [20].

By pre-training against a sizable amount of unlabeled textual input, BERTBase is a bi-directional transformer for learning a language representation that can be tailored for particular machine learning tasks [21]. On a variety of challenging tasks, the bidirectional transformer, special pre-training tasks like Masked Language Model and Next Structure Prediction, as well as a significant amount of data and Google's compute power, BERTBase beat the state-of-the-art in NLP [22].

RoBERTa, a retrained version of BERT, was introduced on Facebook with improved training methodology, 1,000 times more data, and 1,000 times more processing power [25]. Larger batch sizes were also found to increase the effectiveness of the training method. In RoBERTa, dynamic masking, which changes the masked token throughout training epochs, substitutes the Next Sentence Prediction (NSP) task from BERT's pre-training [34]. 160 GB of text, including 16 GB each from the Books

Corpus and the English Wikipedia—both of which were used in BERT—are used by RoBERTa for pre-training [26]. Other data sets included the Web text corpus (38 GB), Common Crawl Stories, and CommonCrawl News dataset (63 million items, 76 GB) (31 GB). Combining this with the utilisation of 1024 V100 Tesla GPUs for a day led to pre-training.

DistilBERT is a distilled (roughly) version of BERT that uses only half the parameters (paper) while maintaining 97 percent of the performance [35]. Using a process called distillation, which swaps out the massive neural network for a smaller one, DistilBERT approximates Google's BERT [27]. Only half of the layers from Google's BERT are retained, and token-type embeddings and poolers are absent. A smaller network can be used to estimate the entire output distributions of a larger neural network once it has been trained [28]. This resembles posterior approximation in certain respects. It has also been used here. Kulback Leiber divergence is a crucial optimization function in Bayesian statistics for posterior approximation.

A combination of two corpora to build the training data for BERTweet was used [23]. The first corpus contains tweets from January 2012 to August 2019. The second corpus contains COVID19-related tweets from January 2020 to March 2020. The pre-trained language model developed as a result of this research study comprises knowledge from both the pre-covid19 world and the covid19 pandemic world [24]. This opens doors to a wide range of applications that use this technology. For example, if this model performs as well as it is stated, it will be easier to distinguish COVID19-related tweets from generic tweets. The BERTweet model is built on the same architecture as BERT-Base.

The Universal Sentence Encoder converts text into high-dimensional vectors for use in natural language applications such as text classification, semantic similarity, clustering, and others [29]. On Tensorflow-hub, the pre-trained Universal Sentence Encoder is publicly accessible [30]. It is available in two varieties: one that uses a Deep Averaging Network (DAN) and the other that uses a Transformer encoder [33]. Regarding accuracy and the use of computer resources, there is a trade-off between the two. Creating an encoder that converts each sentence into a 512-dimensional embedding is the ultimate objective [31]. The sentence embedding used by the authors of this study is updated based on mistakes it makes and is used for a range of applications. The information that is pertinent will be the only thing that is captured because the same embedding must perform many generic functions [32].

4. Results And Discussion

In this section, the results obtained from the models used, and their respective inferences observed are discussed.

4.1. Machine Learning Algorithms

The two classes in the SVM model are 0 for non-cyberbullying and 1 for cyberbullying (Cyberbullying). The term "precision" describes how accurate and precise your model is in terms of how many of the expected positives really materialise as positive. 51 and 91 percent of 0s and 1s in SVM, respectively, are positive. By classifying it as Positive, this model's recall determines how many Actual Positives it captures (True Positive). As a consequence, the SVM model correctly detects 55% of 0s and 89% of 1s, respectively. The F1 Score is necessary to achieve a balance between recall and precision. So, for the SVM model, the balance between Precision and Recall for 0 and 1 is 53% and 90%, respectively. The model's total accuracy is 84%.

The two classifications in the Naive Bayes Model are 0 (non-cyberbullying) and 1. (Cyberbullying). The term "precision" describes how accurate and precise your model is in terms of how many of the expected positives really materialise as positive. In Nave Bayes, 50% and 91% of 0s and 1s, respectively, are actually positive. By classifying it as Positive, this model's recall determines how many Actual Positives it captures (True Positive). The Nave Bayes model consequently correctly captures 56 and 89 percent of 0s and 1s, respectively. The F1 Score is necessary to achieve a balance between recall

and precision. For the Nave Bayes model, the balance between Precision and Recall for 0 and 1 is 53% and 90%, respectively. The model's total accuracy is 83%.

4.2. Deep Learning Algorithms

The parameters and the accuracies of the deep learning models used are given in Table 4.

Table 4: Comparison of DL algorithms used in cyberbullying detection

Model	Epochs	Train-Test Split	Batch Size	Accuracy
Bilstm+Fasttext (a)	40	0.8-0.2	256	84.83%
Bilstm+Fasttext (b)	30	0.8-0.2	256	78.59%
Bilstm+Fasttext (c)	40	0.7-0.3	256	57.71%
Bilstm+Fasttext (d)	40	0.8-0.2	128	78.43%
Bilstm+Glovetwitter (e)	40	0.8-0.2	256	85.35%
Bertbase	4	0.9-0.1	32	89%
Roberta	1	0.8-0.2	8	89.14%
Distilbert	2	0.8-0.2	32	87.09%
Bertweet	2	0.8-0.2	80	93%
USE (f)	4	0.8-0.2	128	96.08%

The graphs plotted for the BiLSTM and USE models are shown below. They depict the relationship between epoch and accuracy.

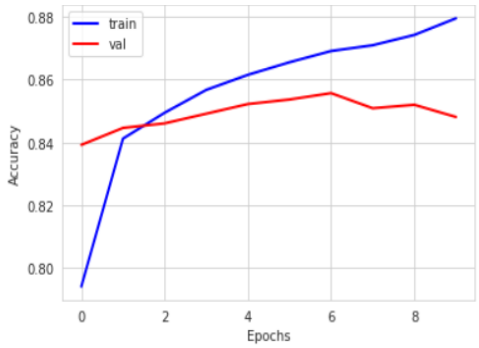

Figure 3: BiLSTM+Fasttext (a)

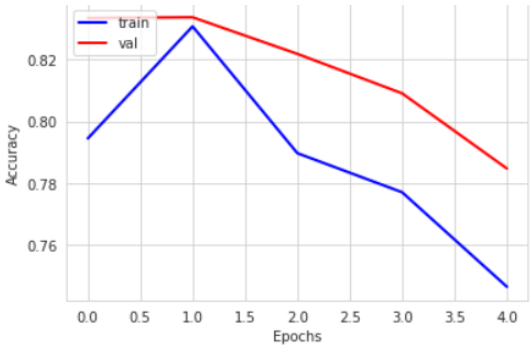

Figure 4: BiLSTM+Fasttext (b)

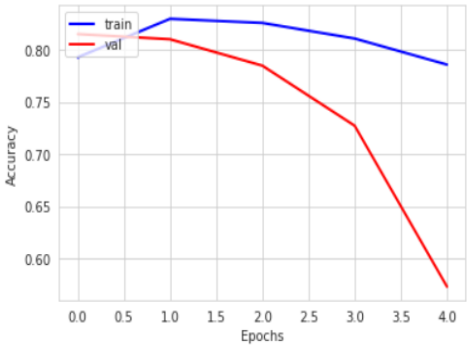

Figure 5: BiLSTM+Fasttext (c)

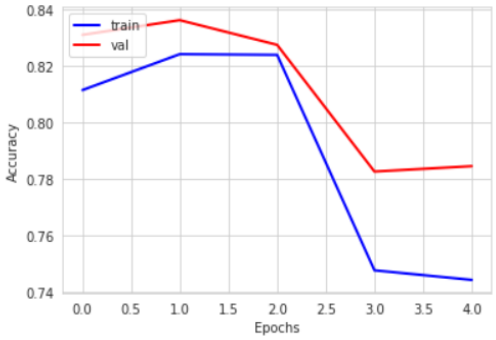

Figure 6: BiLSTM+Fasttext (d)

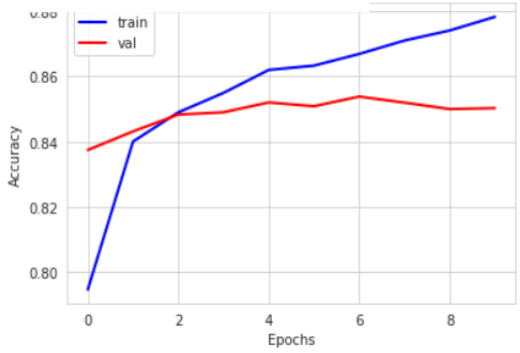

Figure 7: BiLSTM+GloveTwitter (e)

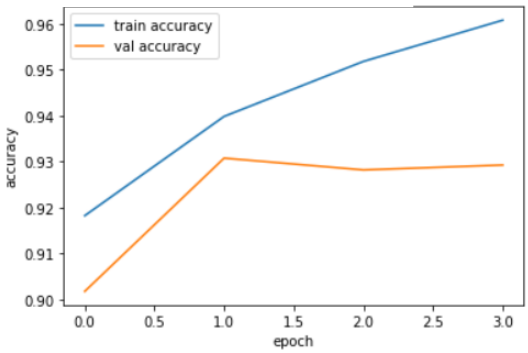

Figure 8: USE (f)

For the BiLSTM + Fasttext models (a, b, c, and d), it can be deduced that the curve shifts from underfitting to optimum as additional weights are altered in the neural network. Since the BiLSTM is a slow learning model, in some circumstances lowering the number of epochs lowers the model's capacity to learn. By using the memory content and tokens from the input data, BiLSTM is a potent method for simulating the sequential dependencies between words and phrases in both forward and backward directions, and it thus exhibits great accuracy.

According to the graph, the BiLSTM + GloveTwitter (e) model's accuracy rises with the number of epochs because, as the epoch count rises, the neural network's weights are modified more frequently and the curve moves from underfitting to optimum. GloveTwitter has a high level of accuracy because it was developed using Twitter-specific data.

The Universal Sentence Encoder (f) model's accuracy is inferred from the graph to grow with the number of epochs because, as the epoch count rises, the weights in the neural network are altered more frequently and the curve shifts from underfitting to optimum. The Universal sentence encoder was pretrained exclusively for sentence embedding, making it a better choice right out of the box for text similarity tasks.

5. Conclusion And Future Scope

The rapid rise in cyberbullying instances as a result of extensive social media use is a major source of concern. The need of the hour is to combat these harmful consequences using Machine Learning and Deep Learning models. The usage of Transfer Learning Models like the Universal Sentence Encoder Model can also help identify cyberbullying tweets. This research work examines the performance of individual ML models like Support Vector Machine and Nave Bayes, DL models like BiLSTM, BERTBase, RoBERTa, DistilBERT, BERTweet, and Transfer Learning Models like USE on a

cyberbullying tweet dataset of roughly 1 lakh records. The Universal Sentence Encoder with learning from the highest achieving BERT model and BiLSTM model offered the highest prediction efficiency of 98%.

As a result, this research provides a significant contribution to the prevention of cyberbullying on Twitter. This research could be improved by using the model as a tool for detecting cyberbullying across all social media platforms, not only Twitter, and scaling it up by assigning more resources for processing and testing. Finally, it can be stated that even a tiny step toward improved cyberbullying detection can help to prevent a slew of harmful consequences for today's youth.

6. References

[1] Dadvar, Maral, and Kai Eckert. "Cyberbullying detection in social networks using deep learning based models; a reproducibility study." arXiv preprint arXiv:1812.08046 (2018).

[2] Talpur, Bandeh Ali, and Declan O'Sullivan. "Cyberbullying severity detection: A machine learning approach." PloS one 15, no. 10 (2020): e0240924.

[3] Ghosh, Sourodip, Aunkit Chaki, and Ankit Kudeshia. "Cyberbully Detection Using 1D-CNN and LSTM." In Proceedings of International Conference on Communication, Circuits, and Systems, pp. 295-301. Springer, Singapore, 2021.

[4] Raj, Chahat, Ayush Agarwal, Gnana Bharathy, Bhuva Narayan, and Mukesh Prasad. "Cyberbullying Detection: Hybrid Models Based on Machine Learning and Natural Language Processing Techniques." Electronics 10, no. 22 (2021): 2810.

[5] Agrawal, Sweta, and Amit Awekar. "Deep learning for detecting cyberbullying across multiple social media platforms." In European conference on information retrieval, pp. 141-153. Springer, Cham, 2018.

[6] Elsafoury, Fatma, Stamos Katsigiannis, Zeeshan Pervez, and Naeem Ramzan. "When the timeline meets the pipeline: A survey on automated cyberbullying detection." IEEE Access 9 (2021): 103541-103563.

[7] Mishra, Pushkar, Helen Yannakoudakis, and Ekaterina Shutova. "Tackling online abuse: A survey of automated abuse detection methods." arXiv preprint arXiv:1908.06024 (2019).

[8] Richard, Khoury, and Larochelle Marc-André. "Generalisation of cyberbullying detection." arXiv preprint arXiv:2009.01046 (2020).

[9] Hani, John, Mohamed Nashaat, Mostafa Ahmed, Zeyad Emad, Eslam Amer, and Ammar Mohammed. "Social media cyberbullying detection using machine learning." Int. J. Adv. Comput. Sci. Appl 10, no. 5 (2019): 703-707.

[10] Eronen, Juuso, Michal Ptaszynski, Fumito Masui, Aleksander Smywiński-Pohl, Gniewosz Leliwa, and Michal Wroczynski. "Improving classifier training efficiency for automatic cyberbullying detection with Feature Density." Information Processing & Management 58, no. 5 (2021): 102616.

[11] Cherkassky, Vladimir, and Yunqian Ma. "Practical selection of SVM parameters and noise estimation for SVM regression." Neural networks 17, no. 1 (2004): 113-126.

[12] Jakkula, Vikramaditya. "Tutorial on support vector machine (svm)." School of EECS, Washington State University 37, no. 2.5 (2006): 3.

[13] Rish, Irina. "An empirical study of the naive Bayes classifier." In IJCAI 2001 workshop on empirical methods in artificial intelligence, vol. 3, no. 22, pp. 41-46. 2001.

[14] Metsis, Vangelis, Ion Androutsopoulos, and Georgios Paliouras. "Spam filtering with naive bayes-which naive bayes?." In CEAS, vol. 17, pp. 28-69. 2006.

[15] Santos, Igor, Nadia Nedjah, and Luiza de Macedo Mourelle. "Sentiment analysis using convolutional neural network with fastText embeddings." In 2017 IEEE Latin American conference on computational intelligence (LA-CCI), pp. 1-5. IEEE, 2017.

[16] d'Sa, Ashwin Geet, Irina Illina, and Dominique Fohr. "Bert and fasttext embeddings for automatic detection of toxic speech." In 2020 International Multi-Conference on:"Organization of Knowledge and Advanced Technologies"(OCTA), pp. 1-5. IEEE, 2020.

[17] Hayashi, Toshitaka, and Hamido Fujita. "Sentence-level sentiment analysis using feature vectors from word embeddings." In New Trends in Intelligent Software Methodologies, Tools and Techniques, pp. 749-758. IOS Press, 2018.
[18] Mao, Junhua, Jiajing Xu, Kevin Jing, and Alan L. Yuille. "Training and evaluating multimodal word embeddings with large-scale web annotated images." Advances in neural information processing systems 29 (2016).
[19] Xu, Guixian, Yueting Meng, Xiaoyu Qiu, Ziheng Yu, and Xu Wu. "Sentiment analysis of comment texts based on BiLSTM." Ieee Access 7 (2019): 51522-51532.
[20] Chen, Tao, Ruifeng Xu, Yulan He, and Xuan Wang. "Improving sentiment analysis via sentence type classification using BiLSTM-CRF and CNN." Expert Systems with Applications 72 (2017): 221-230.
[21] Khadhraoui, Mayara, Hatem Bellaaj, Mehdi Ben Ammar, Habib Hamam, and Mohamed Jmaiel. "Survey of BERT-Base Models for Scientific Text Classification: COVID-19 Case Study." Applied Sciences 12, no. 6 (2022): 2891.
[22] Geetha, M. P., and D. Karthika Renuka. "Improving the performance of aspect based sentiment analysis using fine-tuned Bert Base Uncased model." International Journal of Intelligent Networks 2 (2021): 64-69.
[23] Nguyen, Dat Quoc, Thanh Vu, and Anh Tuan Nguyen. "BERTweet: A pre-trained language model for English Tweets." arXiv preprint arXiv:2005.10200 (2020).
[24] Macrì, Tommaso, Freya Murphy, Yunfan Zou, and Yves Zumbach. "Classifying Tweet Sentiment Using the Hidden State and Attention Matrix of a Fine-tuned BERTweet Model." arXiv preprint arXiv:2109.14692 (2021).
[25] Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. "Roberta: A robustly optimized bert pretraining approach." arXiv preprint arXiv:1907.11692 (2019).
[26] Murarka, Ankit, Balaji Radhakrishnan, and Sushma Ravichandran. "Detection and Classification of mental illnesses on social media using RoBERTa." arXiv preprint arXiv:2011.11226 (2020).
[27] Büyüköz, Berfu, Ali Hürriyetoğlu, and Arzucan Özgür. "Analyzing ELMo and DistilBERT on socio-political news classification." In Proceedings of the Workshop on Automated Extraction of Socio-political Events from News 2020, pp. 9-18. 2020.
[28] Dogra, Varun, Aman Singh, Sahil Verma, N. Z. Jhanjhi, and M. N. Talib. "Analyzing DistilBERT for Sentiment Classification of Banking Financial News." In Intelligent Computing and Innovation on Data Science, pp. 501-510. Springer, Singapore, 2021.
[29] Mohammad, AL-Smadi, Mahmoud M. Hammad, A. Sa'ad, AL-Tawalbeh Saja, and Erik Cambria. "Gated Recurrent Unit with Multilingual Universal Sentence Encoder for Arabic Aspect-Based Sentiment Analysis." Knowledge-Based Systems (2021): 107540.
[30] Fu, Qunchao, Cong Wang, and Xu Han. "A CNN-LSTM network with attention approach for learning universal sentence representation in embedded system." Microprocessors and Microsystems 74 (2020): 103051.
[31] Cer, Daniel, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant et al. "Universal sentence encoder." arXiv preprint arXiv:1803.11175 (2018).
[32] Cer, Daniel, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant et al. "Universal sentence encoder for English." In Proceedings of the 2018 conference on empirical methods in natural language processing: system demonstrations, pp. 169-174. 2018.
[33] Majumder, Soumayan Bandhu, and Dipankar Das. "Detecting Fake News Spreaders on Twitter Using Universal Sentence Encoder." In CLEF (Working Notes). 2020.
[34] Kayastha, Tanay, Pranjal Gupta, and Pushpak Bhattacharyya. "BERT based Adverse Drug Effect Tweet Classification." In Proceedings of the Sixth Social Media Mining for Health (# SMM4H) Workshop and Shared Task, pp. 88-90. 2021.
[35] Guo, Yuting, Xiangjue Dong, Mohammed Ali Al-Garadi, Abeed Sarker, Cecile Paris, and Diego Mollá Aliod. "Benchmarking of transformer-based pre-trained models on social media text classification datasets." In Proceedings of the The 18th Annual Workshop of the Australasian Language Technology Association, pp. 86-91. 2020.