Comparative Analysis of Deep Learning Techniques to detect Online Public Shaming

Mehdi Surani [1*], *Ramchandra Mangrulkar*[2]

[1]D.J.Sanghvi College Of Engineering, Mumbai, imsurani15@gmail.com
[2]Department of Computer Engineering, D.J.Sanghvi College Of Engineering, Mumbai, ramchandra.mangrulkar@djsce.ac.in

Abstract. Over the past years the exponential growth of social media usage has given the power to every individual to share their opinions freely. This has led to numerous threats allowing users to exploit their freedom of speech, thus spreading hateful comments, using abusive language, carrying out personal attacks, and sometimes even to the extent of cyberbullying. However, determining abusive content is not a difficult task and many social media platforms have solutions available already but at the same time, many are searching for more efficient ways and solutions to overcome this issue. Traditional models explore machine learning models to identify negative content posted on social media. Shaming categories are explored, and content is put in place according to the label. Such categorization is easy to detect as the contextual language used is direct. However, the use of irony to mock or convey contempt is also a part of public shaming and must be considered while categorizing the shaming labels. In this research paper, various shaming types, namely toxic, severe toxic, obscene, threat, insult, identity hate, and sarcasm are predicted using deep learning approaches like CNN and LSTM. These models have been studied along with traditional models to determine which model gives the most accurate results.

1 INTRODUCTION

The act of instant emotional expression through social media is becoming a common phenomenon in today's world. Platforms like Facebook, Twitter, Instagram, etc. allow one to freely communicate and express their thoughts and opinions. This can simply go from informing people about some accidental incident over the highway to divert traffic to commenting ill about some famous politician's speech. Social media can therefore range from being informative to offensive.

Maintaining decency and avoiding informal content is the current area of concern for all social media platforms. Such unhealthy acts have indirect effects on an individual's health. Lifelong traumatic effects like anxiety, depression, mental illness, and other psychological effects are seen to some extent. This causes users to isolate themself from such social platforms and can inculcate suicidal thoughts. Therefore, there is a need for establishing classification techniques and blocking mechanisms.

Many social platforms are working on this current issue where a user can be reported/ restricted or blocked if it is observed by many users around them. Such acts of shaming individuals on social media platforms need to be controlled sufficiently.

There is a need for automation in this area, which can help companies classify the different shaming categories and take actions accordingly. Many machine learning techniques have been deployed in the past years to work on this topic. However, this paper would like to implement deep learning techniques to find better results for the detection of Online Public Shaming. This research paper includes sarcasm as a shaming category where an ironic or satirical remark tempered by humour is mainly used to say the opposite of what's true to make someone look or feel foolish. Such comments are difficult to detect. Therefore, it is necessary to include this as a part of other shaming categories, namely, toxic, severe toxic, threat, obscene, insult, and identity hate.

The rest of the paper is organized as follows. The literature review is present in section 2. The Methodology is given in section 3. Experimentation and Results are present in section 4. Detailed discussion is given in section 5. Concluding remarks and future scope is given in section 6.

2 LITERATURE REVIEW

Most prior work in the area of online public shaming prediction has been spread across several overlapping fields. This can cause some confusion as different works may tackle specific aspects of shaming language; some may define the term differently or apply it to specific online domains only (Twitter, Online forums, Facebook, YouTube, etc.) to further complicate the comparison between approaches, nearly all previous work uses different evaluation sets. One of the contributions of this paper is to provide a public dataset to better move the field ahead. A public dataset that includes not only the existing shaming categories but also aspects where some users would post sarcastic comments in the same voice as the people that were producing toxic or identity hate content. Although online public shaming has been carried over a long time, their detection using Machine Learning has taken off only in recent years; in-depth research has been carried out on their detection in recent years. Hence, the authors have conducted an extensive survey to study the current findings and identify the gaps.

In "Online Public Shaming on Twitter: Detection, Analysis, and Mitigation" Rajesh Basak et al. Created a

* Corresponding author: imsurani15@gmail.com

web application that is based on the classification and categorization of shaming tweets. Here classification is performed using Support Vector Machine (SVM) and desired results are achieved. This [1] paper has the potential solution for countering the menace of online public shaming on Twitter by categorizing shaming comments. They have performed detection on a small dataset. Currently, the data available over the internet is in millions and hence we need to tackle large datasets. SVM, however, doesn't perform well when a huge dataset is taken, there is a need for a between model to work with a large amount of data.

In "A deeper look into sarcastic tweets using deep convolutional neural networks" S. Poria et al. For sarcasm detection, the system proposed by Cambria et al. [2] uses deep convolution neural networks. They have a deep understanding and study of emotions as well as sentiments for the detection of sarcasm. This is done by using predefined models for the extraction of features. Both sentiment and emotion feature for sarcasm detection and discussed the use of predefined models for feature extraction. This paper [2] identifies sarcastic text using the CNN model. But the performance of the model on large corpus and other domain-dependent corpora has not yet been explored. Other shaming categories have not been incorporated for sentiment classification.

In "An Effective Approach for Detection of Sarcasm in Tweets" Sreelakshmi K et al. proposed a system for sarcasm detection using Support Vector Machine (SVM) and Decision Trees for modelling the proposed system. Various features of the text like lexical, pragmatic, context incongruity, topic, and sentiment were taken into consideration. Sreelakshmi K et al. [3] have worked with SVM Models which again limits the usage of large datasets. Large corpora must be explored along with decision trees in order to find better results.

In "Classification of Abusive Comments in Social Media using Deep Learning", Mukul Anand et al [4] propose deep learning methods like CNN (Convolution Neural Network) and LSTM (Long Short Term) categorizing various shaming types like toxic, severe toxic, obscene, threat, insult, and identity hate. Online abuse is only identified by the user's report there is a need for automation in the detection of abusive comments in social networks which are carried out using deep learning techniques in this research work. Deep learning techniques were used for detecting online public shaming however there is a need to explore more shaming types and automation

In "Abusive Language Detection in Online User Content" Chikashi Nobata et al. [5] have proposed a system to identify hate speech on online user comments from two domains that outperform a state-of-the-art deep learning approach. They achieved this using standard NLP features along with different syntactic and embedding features. Naïve NLP techniques are implemented; in [5] there is a need for enhanced and new machine learning models to detect precise and accurate shaming categories.

In "Tackling Toxic Online Communication with Recurrent Capsule Networks" Soham et al. [6] have implemented Recurrent Neural Network and Capsule network as its backbone and captures contextual information for toxic online communication. Traditional methods have been overcome using neural networks. There is a need for multi-label toxicity categorization for different types of shaming.

In "Imbalanced Toxic Comments Classification using Data Augmentation and Deep Learning" Mai Ibrahim et al. [7] present various data augmentation techniques to overcome the data imbalance issue in the Wikipedia dataset. An ensemble of three deep learning models namely CNN, LSTM, and GRU (Gated Recurrent Units) were used to detect the type of toxicity present in the toxic comments dataset. Using multiple deep learning models can help us in understanding which technique will prove the best for online shame detecting. Findings from this can be of good contribution to this research work.

In "Audio and Video Toxic Comments Detection and Classification" Sangita et al. [8] model aim to apply the text-based Convolution Neural Network (CNN) with word embedding, using fastText word embedding technique improving the detection of different types of toxicity to improve the social media experience. Deep learning techniques were used for detecting online public shaming however there is a need to explore more shaming types and automation along with other models like RNN, LSTM, to detect which model gives better results.

In "Online Public Shaming Approach using Deep Learning Techniques" Mehdi et al. [9] have performed data visualization using text-based Convolution Neural Network (CNN) as the proposed model. They aim to categorize the dataset based on the different multi-labels, top words, confusion matrix, ROC curves for each label type, and classification report for the same. A deep understanding of the dataset is carried out in [8] using Convolution Neural Network.

In "A deep understanding of the dataset is carried out in [10] using Convolution Neural Network." Manav Kohli et al. used a wide range of Recurrent Neural Network models for the task of classifying abusive comments by users of Wikipedia. Findings include non-neutral baseline models based on TFIDF sentences. Gated Recurrent Unit (GRU) and LSTM results were compared which showed similar outcomes.

3 METHODOLOGY

This paper aims at detecting online public shaming using deep learning algorithms. Deep Learning is a subset of Machine Learning where data preprocessing and feature selection can be minimized without impacting the performance of algorithms. Deep learning models can extract relevant features on their own. The research is implemented in the following way. Machine or deep learning algorithms cannot process plain text as inputs. The system does not understand the human language. It becomes difficult or rather impossible for the algorithms to process such raw input data. To solve this issue there are Word embedding techniques available. Word embedding techniques provide a solution for this issue

by transforming the plain text into a vector format which can be used by the machine learning model. Such transformations can also be used for finding semantic relationships between the associated words. The proposed model is using GloVE (global vectors for word representation) as a word Embedding Technique. The GloVE works by aggregating the global word-word co-occurrence matrix from a corpus. GloVE stresses that the frequency of co-occurrences is vital information and should not be "wasted" as additional training examples. Acquiring less training time GloVE outperforms Bag-of-Verbs and TFIDF. The position and semantic aspect of the word is also looked into.

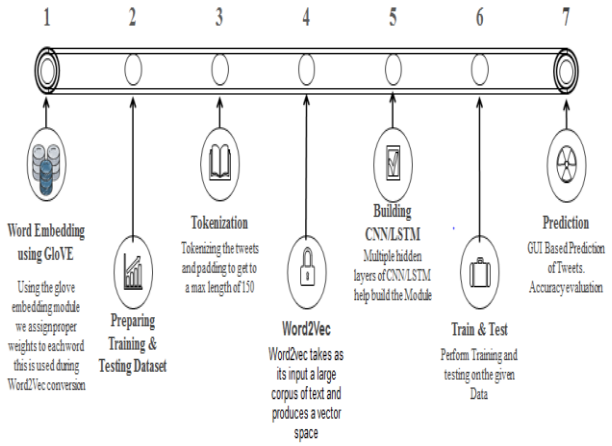

Fig. 1. Overview of Methodology

Word embedding is done using the GloVe (Global Vectors for Word Representation) embedding module. Here weights are assigned to each word this will then be used during word2Vec conversion. Training and testing data are separated and then prepared for tokenization. Tokenizing is the process of taking the tweets and padding them accordingly to maintain uniformity in the data set created. As it is known that the input set is very large and therefore there is a need to create vectors from the large text corpus hence forming a vector space. Once the vector space is formed, connecting the model is created using the following 2 algorithms: CNN and LSTM. Training and testing are performed on the above algorithms and accuracy evaluation along with a GUI Based prediction of Tweets is created. Further detailed understanding of the Deep Learning modules with their hidden layers is shown below

3.1. CNN (Convolution Neural Network)

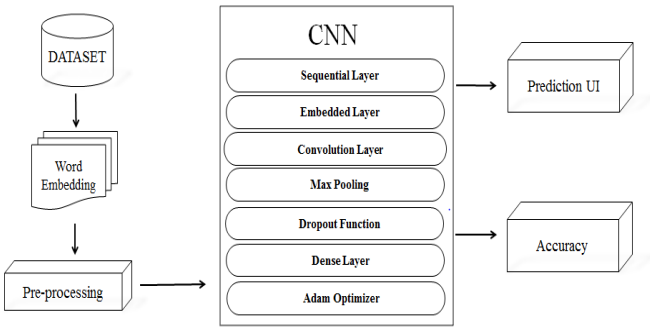

Fig.2. Convolution Neural Network and its Hidden Layers

3.2 LSTM (Long Short Term Memory)

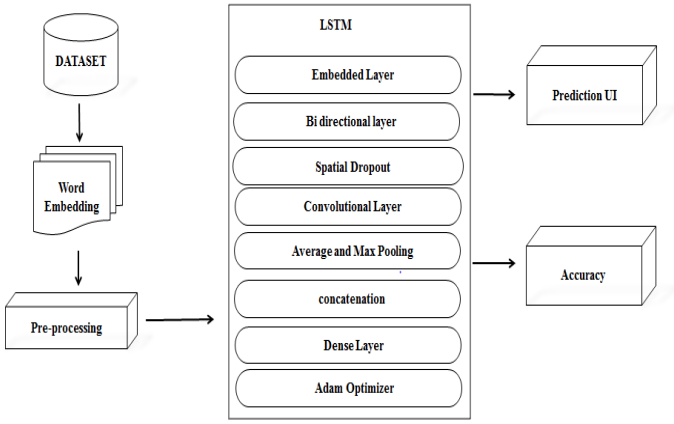

Fig.3. Long Short Term Memory (LSTM) and its Hidden Layers

With the recent advancements that are happening in data science, it is found that for nearly all those sequence prediction problems, LSTMs are observed as the most effective solution. LSTMs have a foothold over conventional feed-forward neural networks compared to RNN in some ways. This is often due to their property of selectively remembering patterns for long period (memory). LSTM's are nothing but improved RNNs as the data flows through a mechanism referred to as cell states and in this way, LSTMs can selectively remember or forget things. The data at a specific cell state has three different dependencies. These dependencies are often generalized as the previous cell state (i.e. the data that was present within the memory after the previous time step) The previous hidden state (i.e. this is often an equivalent as the output of the previous cell) The input at the present step (i.e. the new information that's being fed in at that moment). Every information from the previous state can be carried forward to the next state and whether that information is relevant to be kept or forget during training can be adjusted using the gates. For this very purpose, a Bi-Directional Layer is added, connecting two hidden layers generating a single similar output layer. Through this form of deep learning, a model which has information from the past and the future simultaneously can be achieved giving a more precise outcome.

4 DETAILED DISCUSSIONS

Online public shaming which targets to humiliate people over the internet publically is a form of individual harassment which is legally not acceptable. The serious consequences include ruining reputations, careers, and psychological damages. Social media being a vast platform makes it difficult to track each nuisance created by the audience. Naive Machine learning methods have helped achieve limiting these behaviors by blocking, restricting accounts or comments. The research work presented by Basak et al [1] and Shreelaxmi et al. [3] detected online shaming using the SVM model restricting the experiment to a small dataset. The research and implementation presented in this paper use advance deep learning techniques like CNN and LSTM which help explore large datasets, provide better results and accuracy, combine multiple datasets and predict multi-class labels which serve the objectives of this research. The Kaggle data set is combined along with the sarcasm dataset which helps us to expand the dataset and therefore achieve our first objective of using multiple datasets. This data set is then applied to the CNN and LSTM models to predict multiple shaming labels using a Graphical User Interface. This is trained and tested for more than 162070 entries. The model is trained in such a way that any new entry or tweet added to the user interface will predict the shaming categories hence, achieving the next objective of the research work. The shaming categories include toxic, severe toxic, threat, obscene, insult, identity hate, and sarcasm. Future implementations aim at identifying the severity of the shaming category and blocking/restricting the comment using deep learning models. Another evaluation matrix includes the Accuracy calculation. Logistic Regression used in [8] is a machine learning technique that is compared to the deep learning model of CNN and LSTM and a comparative analysis is carried out. Section 5 will discuss the results and experiment details. The main motive of the research aims at implementing the latest technologies available to achieve better results and find out how they outperform traditional methods of predicting online shaming.

5 EXPERIMENT AND RESULTS

The research paper shows the implementation of CNN and LSTM based approaches for detecting online public shaming. The experiment was carried out to achieve better accuracy results and predict the shaming labels. After the model is trained, Graphical User Interface was created which takes in tweets or comments as input and predicts the multi-label shaming categories. The resultant graph shows the sentiments namely toxic, severe toxic, obscene, threat, insult, identity hate, and sarcasm. This is carried out for CNN and LSTM models separately showing results in Tables 2 and 3.

The GUI-based model has an input screen where the tweet is added and on submission, it returns a chart that represents the multi-label shaming categories each representing the intensity of that sentiment. Let us understand this with an example. The tweet "*Trump Administration send 300 million NOTHING to Puerto Rico victims "*is a pure example of sarcasm along with mere toxicity against Donald Trump. The model implemented in this paper predicts the severity of the shaming type. A multi-label shaming category graph is created.

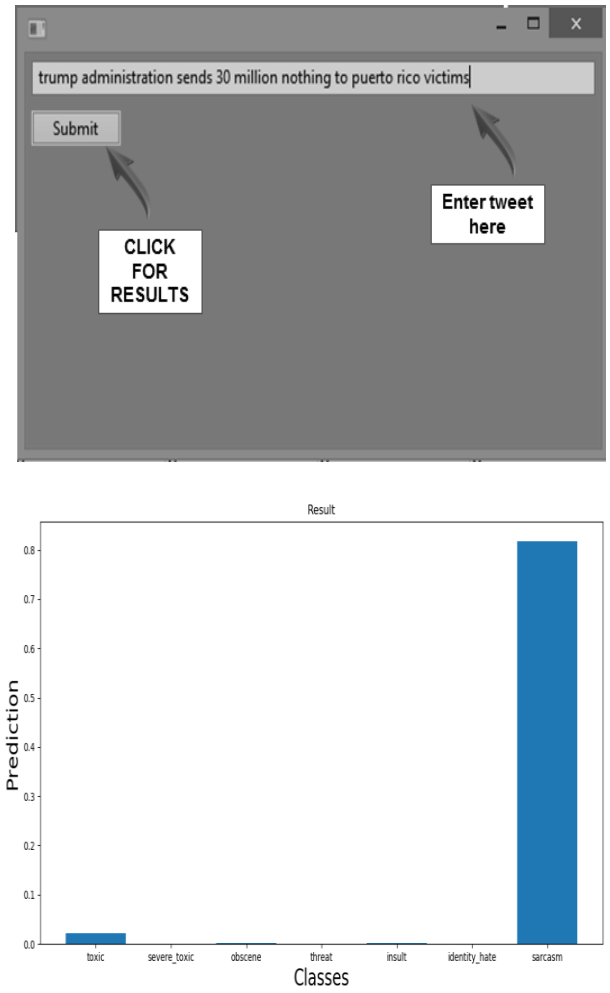

Fig.4.GUI Based shaming category prediction using LSTM

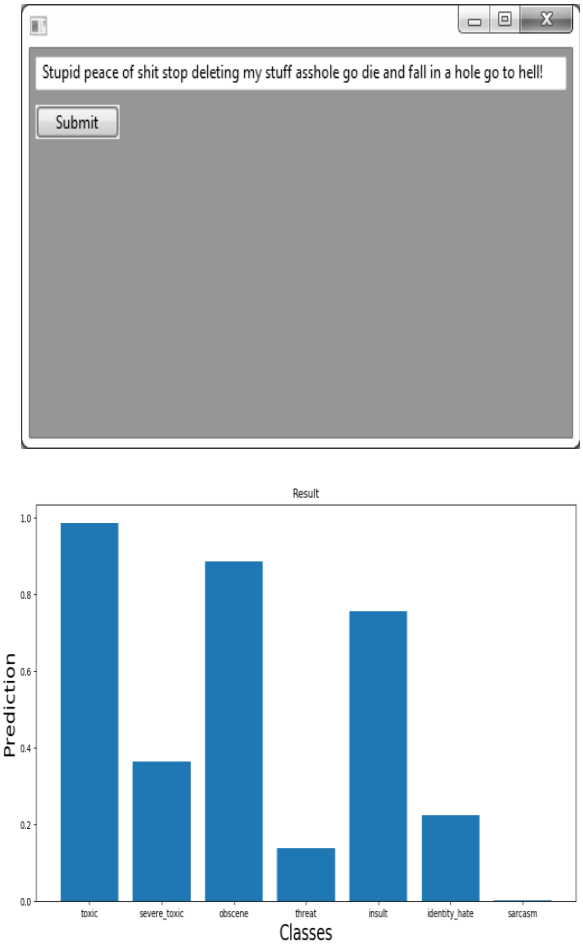

Fig.5.GUI Based shaming category prediction using CNN.

Sr.No.	Tweets from Dataset
1	Trump Administration send 300 million NOTHING to Puerto Rico victims
2	Are you threatening me for disputing neutrality? I know in your country it's quite common to bully your way through a discussion and push outcomes you want. But this is not Russia.
3	There must be some chemical imbalance in your brains. Are you ok? The gibberish you guys are writing concerning makes me feel sorry for you. Don't ever go to law school. You would do the legal profession badly
4	The government is so good, it ensures the citizens happily die of corona!
5	You are gay or antisemmitian?
6	F*CK YOUR FILTHY M*TH*R IN THE A*S, DRY!
7	I was merely apologising for my interference ;)

Table 1. Tweets selected to test CNN model.

Sr. No.	**Toxic**	**Severe toxic**	**Obscene**	**Threat**	**Insult**	**Identity Hate**	**Sarcasm**
1	0.021	0	0	0	0.009	0	0.819
2	0.305	0.001	0.001	0	0.002	0.0007	0
3	0.504	0	0.064	0.006	0.182	0.028	0.004
4	0.112	0	0.008	0.003	0.013	0.003	0.009
5	0.794	0.038	0.394	0.043	0.450	0.133	0.009
6	0.990	0.453	0.966	0.079	0.804	0.187	0
7	0.002	0	0	0	0	0	0.144

Table 2. Shaming type detection using CNN model.

Sr.No.	Tweets from Dataset
1	The JEWZ are Market Manipulators Hate them freaking Jews !! To bad Hitler did not nish them o !!!
2	300 missiles are cool! Love to see um launched into TelAviv! Kill all the gays there!
3	mom wants to know if you'll be free if she visits 14 months from now
4	Stupid peace of shit stop deleting my stuff asshole go die and fall in a hole go to hell!
5	You are gay or antisemitismstops?
6	F*CK YOUR FILTHY M*TH*R IN THE A*S, DRY!
7	Stupid peace of shit stop deleting my stuff asshole go die and fall in a hole go to hell!

Table 3. Tweets selected to test LSTM model.

Sr. No.	**Toxic**	**Severe toxic**	**Obscene**	**Threat**	**Insult**	**Identity Hate**	**Sarcasm**
1	0.767	0.004	0.769	0.004	0.2	0.694	0.01
2	0.642	0	0.018	0.056	0.053	0.56	0.005
3	0	0	0	0	0	0	0.951
4	0.986	0.361	0.885	0.137	0.756	0.2218	0
5	0.900	0.391	0.098	0.050	0.645	0.818	0
6	0.991	0.505	0.940	0.006	0.864	0.051	0
7	0.997	0.258	0.920	0.212	0.932	0.0344	0.006

Table 4. Shaming type detection using LSTM model.

Here the algorithm is applied on the Kaggle dataset which is 67MB. Combining this dataset with that of sarcasm we get a total of 7 shaming categories. This approach gives an aggregate accuracy of 98% in predicting the shaming type for the CNN and LSTM model. The authors aim to detect online shaming carried over Twitter and at the same time, increase the overall accuracy of classification through their approach. Figure 6 shows the accuracy achieved using CNN Model. Here an accuracy of 98.42% is achieved whereas figure 7 shows accuracy of 98.57% achieved by LSTM Model. A collective resultant of Machine Learning model and Deep Learning model is shown in figure 8. It clearly shows that CNN and LSTM outperform Logistic Regression and provide better results.

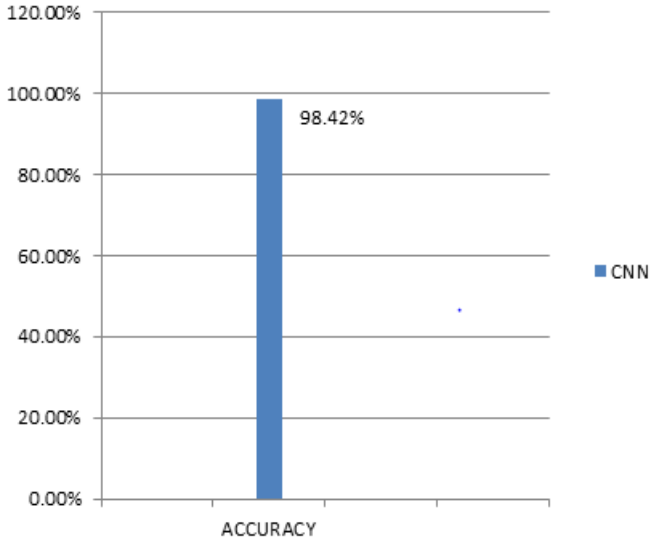

Fig6. Accuracy prediction using CNN.

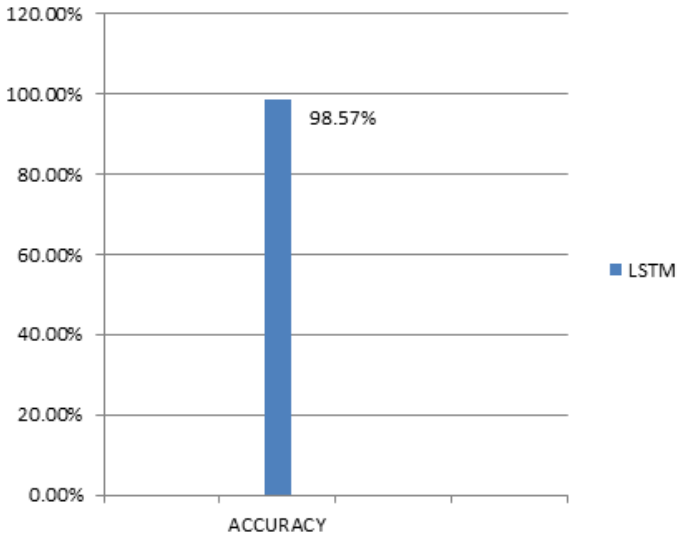

Fig.7. Accuracy prediction using LSTM.

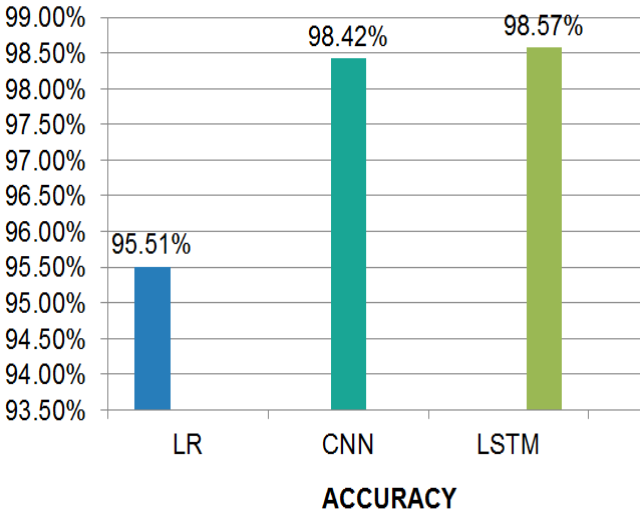

Fig.8. Accuracy comparison LR, CNN, LSTM.

The above is a comparative analysis of the 3 Models namely Logistic Regression [8], Convolution Neural Network, and Long Short Term Memory. The findings clearly show that Deep Learning models perform better than other machine learning models with respect to accuracy estimation

6 IMPACT AND CASE STUDIES

Being a victim of online public shaming can have multiple ill effects on a person's mental, physical, and emotional health. People are taking their own lives and affecting the lives of their fellow family members and friends. The ugly side to this digital world is that it gives easy access to people to pass on their opinion without giving them a second thought. Social media has given the platform for people to troll the latest memes circulating and comments from unknown users in an enormous volume on the victim's image take no time. Let's have a look at few case studies which had immediate actions taken for their action.

1. **Case Study 1: Justine Sacco, 2013.**

In the year 2013, Justine Sacco, a Public Relations person in an American Internet Company tweeted the below: *"Going to Africa. Hope I don't get AIDS. Just kidding. I'm white!"*. With a mere 170 followers, Justine faced criticism at an enormous rate and became one of the most discussed topics over the internet. Before she even landed, she lost her job.

2. **Case Study 2: Aamir Khan, 2015.**

In 2015, India's most famous Actor Aamir Khan commented on the rising intolerance in India and spoke about his wife Kiran Rao suggesting to move out of the country. The response from the general public over

Twitter and fellow actors seems to be very grinding and the consequences faced were that he was removed as Brand Ambassador of SnapDeal.

3. Walter Palmer, Cecil the lion's killer dentist, endures the latest onslaught from social media mob, 2016.

In the endless feedback loop of social media's shaming machine, Walter Palmer suddenly found himself with nothing to shoot and no place to hide. As soon as word got out that the Minnesota dentist had shot and killed an African lion. He received a lot of threats and hate over Twitter. Social media shaming goes so viral that even other users with the name Walter Palmer received more than 15 threat full tweets.

4. Case Study 3: Melania Trump, 2016.

In 2016 a Twitter user pointed out Melania Trump, spouse of the US President for plagiarism in one of her campaign speeches. There were media coverage and tweets regarding her speech being plagiarized from Michelle Obama. There was huge criticism and negative media coverage encountered immediately.

7 CONCLUSIONS AND FUTURE WORK

In this research implementation, it is observed that the deep learning approach for online public shaming has been carried out using CNN and LSTM. A comparative analysis with [1] it is observed that the LSTM model works better in predicting online public shaming. The graphical user interface also helps predict the shaming type, especially for sarcastic tweets that are difficult to predict. It is also seen that a large amount of data that could not be solved using SVM and other machine learning modules have shown better results using Deep Learning. This research shows that Deep Learning-based models outperform machine learning-based model. In the future complex deep learning models and hybrid algorithms can be studied and implemented and various Transformers can be used for the prediction of online public shaming.

8 REFERENCES

[1]Rajesh Basak, Shamik Sural, Niloy Ganguly, and Soumya K. Ghosh, "Online Public Shaming on Twitter: Detection, Analysis, and Mitigation.", IEEE Transaction -2329-924X,2019.

[2]S. Poria, E. Cambria, D. Hazarika, and P. Vij, "A deeper look into sarcastic tweets using deep convolutional neural networks," Creative Commons Attribution 4.0,arXiv preprint DOI:1610.08815, 2016.

[3]Sreelakshmi K Rafeeque P C, "An Effective Approach for Detection of Sarcasm in Tweets", International CET Conference on Control, Communication, and Computing (IC4)DOI-978-1-5386-4966-4,2018.

[4]Mukul Anand, Dr.R.Eswari, " Classification of Abusive Comments in Social Media using Deep Learning", IEEE DOI-978-1-5386-7808-4,2019.

[5]Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, Yi Chang, "Abusive Language Detection in Online User Content", IW3C2, DOI-10.1145/2872427.2883062,2019.

[6]Soham Deshmukh, Rahul Rade, "Tackling Toxic Online Communication with Recurrent Capsule Networks", IEEE-DOI: 10.1109/INFOCOMTECH.2018.8722433 2018.

[7]Mai Ibrahim, Marwan Torki and Nagwa El-Makky, "Imbalanced Toxic Comments Classification using Data Augmentation and Deep Learning", 2018.

[8]Sangita Holkar, S. D. Sawarkar, Shubhangi Vaikole, "Audio and Video Toxic Comments Detection and Classification", International Journal of Engineering Research & Technology (IJERT), Vol. 9 Issue 12, IISSN: 2278-0181,2020.

[9]Mehdi Surani, Ramchandra Mangrulkar," Online Public Shaming Approach using Deep Learning Techniques", Journal of the University of Shanghai for Science and Technology ISSN: 1007-6735,2021.

[10]Manav Kohli, Emily Kuehler and John Palowitch. "Paying attention to toxic comments." Stanford University,2018.