Online Public Shaming Approach using Deep Learning Techniques

D.O.I - 10.51201/Jusst12675
http://doi.org/10.51201/Jusst12675

Mehdi Surani, Ramchandra Mangrulkar

Department of Computer Engineering, D.J.Sanghvi College of Engineering, Mumbai, India.

Abstract: Public shaming on social media platforms like Twitter / Instagram / Facebook etc. have recently increased from the past years. This results in affecting an individual's social, political, mental and financial life. The impact can range from mild bullying to severe depression. With the growing leniency on these social platforms, many people have started misusing the opportunity by turning to online bullying and hate speech. When something is posted online, it stays there forever and it becomes extremely hard taking something out of the digital world. Manually locating and categorizing such comments is a lengthy procedure and just cannot be relied upon. To solve this challenge, automation was performed to identify and classify the shamers. This has been done using the classic SVM model which worked on a given quantity of data. To identify the negative content being posted and discussed online, this paper further explores the deep learning system which can successfully classify these content pieces into proper labels. The text-based Convolution Neural Network (CNN) is the proposed model in this paper for this analysis.

Keywords: — Social Media Analysis, Natural Language Processing, Embedding System,Tag Cloud, CNN.

1. INTRODUCTION

In today's digital world, most of the conversations we have are through some or the other social forum. It allows one to communicate and express their thoughts and opinions freely. They are also the conversation starters for various topics ranging from informational content to simply putting your own voice out there. However, some people find it increasingly difficult to maintain decency and conduct while putting their thoughts out. This is mainly because they are facing a screen, instead of a real person, making their bad behavior much easier to navigate. Abusive content, harassment and cyber-bullying have unfortunately become a part and parcel of being a part of the digital culture. You are either subjected to it, or bear witness to it. This has significantly increased its unhealthy effects on an individual's health. Which can be mental, psychological or physical health in some cases. This can lead to harmful and life-long traumatic effects on a person. When an individual is subjected to such situations, it can traumatize them and lower their self-esteem, leading to them holding back from expressing their opinions online and in real life. They might resort to alienating themselves and stop oneself from receiving help from people who are willing to help. Many social platforms have been working on finding solutions to strain out these comments by establishing classification techniques and user blocking mechanisms.

Automation in this area can thus help companies save time and manual efforts which go in classifying and detecting comments. Unknown victims are put to shame in an enormous volume by the other users whom generally give their opinion regarding them. For example, when in 2016 a twitter user pointed out on Melania Trump spouse of the US President for plagiarism in one of her campaign speech. There was huge criticism and negative media coverage encountered immediately

Machine Learning Technique has been deployed in the past to work on this topic. However, this paper would like to explore deep learning algorithms and also introduce feature enhancement suggestion further in the report.

The rest of the paper is organized as follows. Proposed algorithm is explained in section 2. Experimental results are presented in section 3. Concluding remarks and future scope are given in section 4.

2. PROPOSED ALGORITHMS

• Support Vector Machine

For classification, six Support Vector Machine (SVM) classifiers namely abusive, comparison, passing comment, religious, joke were used on a limited amount of data. For enhanced performance large annotated datasets should be experimented with.

But as it is known SVM algorithm is not suitable for large data sets. Also SVM does not perform very well, when the data set has more noise i.e. target classes are overlapping. Hence there is a need for Deep Learning Approaches to be analyzed.

• Logistic Regression

Logistic Regression is basically a statistical model which works on binary logics. In the case of online public shaming as there are multiple labels the logic of 1 v/s rest is used for finding the various estimation values like total data present, top words, word clouds for different labels, confusion matrix, ROC curve and classification report for each shaming category.

• Convolution Neural Network

Convolution Neural Network (CNN) is a feed forward neural network that is generally used to analyse visual images by processing data with grid like topology. Here, instead of an image, the input to CNN is sentences or documents represented as a matrix. The beauty of CNN is that large amount of data can be stored in a matrix format for complex classification & can work on large data. If you don't have a good GPU they are quite slow to train (for complex tasks).

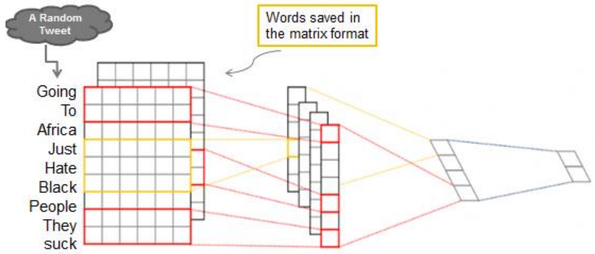

Fig.1. Example of matrix representation in CNN of textual data

Understanding the design architecture of CNN

Being a regularized version of multilayer perceptron's, word embedding is the first layer of CNN which converts them into a low dimension vector. The next layer performs Convolutions are performed over the embedded word vectors in the next layer mistreatment different filter sizes. For example sliding over 3 to 5 words at a time.

The next step followed is max-pooling. After convolution layer the pooling layers reduce the dimensions of the data, while max pooling uses the maximum value from each cluster, which in our case is the cluster of embedded words. The results of max-pooling are then kept in a feature vector which is then fed to the next layer called as dropout regularization.

By using the dropout regularization process, it will also significantly improve the training speed and make the model more practical for utilization. As is the norm, the research paper uses the Softmax activation in the last layer of the neural net owing to its useful ability of converting the output of the last layer into out neural network. The research paper can classify the results by putting the Softmax layer to use and deploy a look-up table for each word from the word embedding's. These multi-dimensional embedding's square measure are initialized at random in the start and then updated later. These word representations are square measured and the text-representation is then averaged. Objective can be achieved by feeding the embedded tokens into first the layer of convolution followed by having a maximum pooling layer from which the least relevant values can be removed using the dropout regularization function.

Convolutional neural networks avoid over fitting by using large amount of training data, which is present in plenty. However, it can change the network parameters by fine-tuning the network weights leading to the convolutional network successfully dealing with smaller training sets. Figure.2 represents the entire working model of the CNN based Public shaming Model.

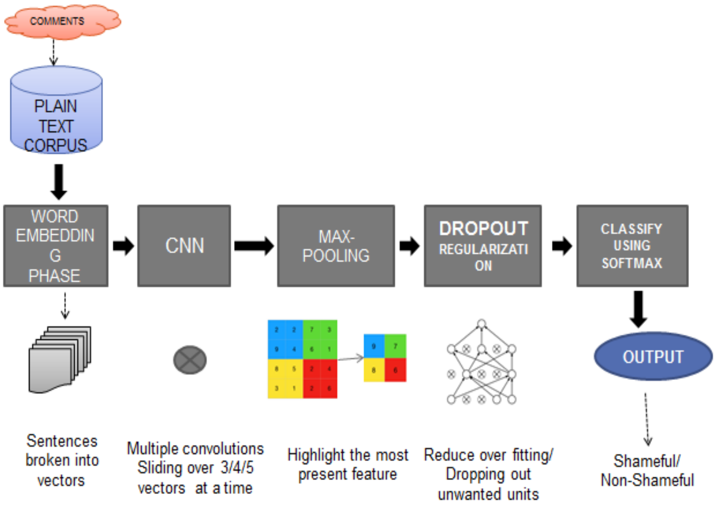
COMMENTS
PLAIN TEXT CORPUS
WORD EMBEDDIN G PHASE
CNN
MAX-POOLING
DROPOUT REGULARIZATI ON
CLASSIFY USING SOFTMAX
OUTPUT
Sentences broken into vectors
Multiple convolutions Sliding over 3/4/5 vectors at a time
Highlight the most present feature
Reduce over fitting/ Dropping out unwanted units
Shameful/ Non-Shameful

Fig.2. Design Architecture using CNN.

3. EXPERIMENT AND RESULT

The research paper shows the steps involved in implementation of Logistic Regression based approaches for data visualization of online public shaming labels.

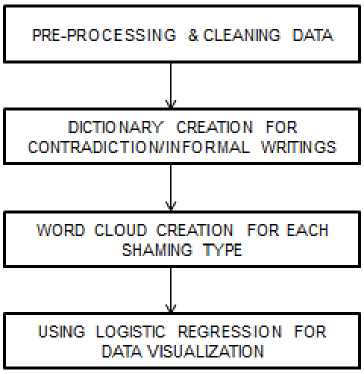
PRE-PROCESSING & CLEANING DATA
DICTIONARY CREATION FOR CONTRADICTION/INFORMAL WRITINGS
WORD CLOUD CREATION FOR EACH SHAMING TYPE
USING LOGISTIC REGRESSION FOR DATA VISUALIZATION

Fig.4.Steps involved in proposed algorithm.

Here the algorithm is applied on Kaggle dataset which is of 67MB .A total of 18 machine learning graphs from the given dataset.

The following data visualizations are created for different shaming labels.

- Total data
- Top Words
- Tag words
- ROC Curve
- Confusion Matrix

- Classification report

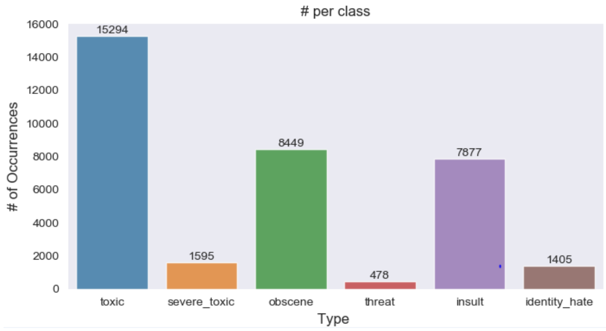

Fig.3. Figures shows the total number of labels in the data set

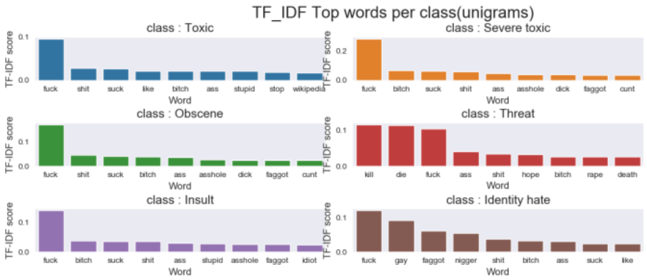

Fig.4. Figures shows the top words per class label

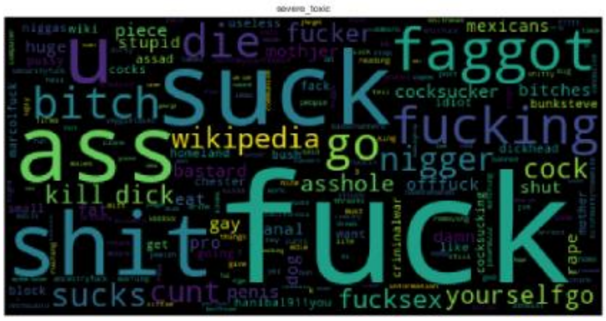

Fig.5. Figures shows the total number of severe toxic labels in the data set.

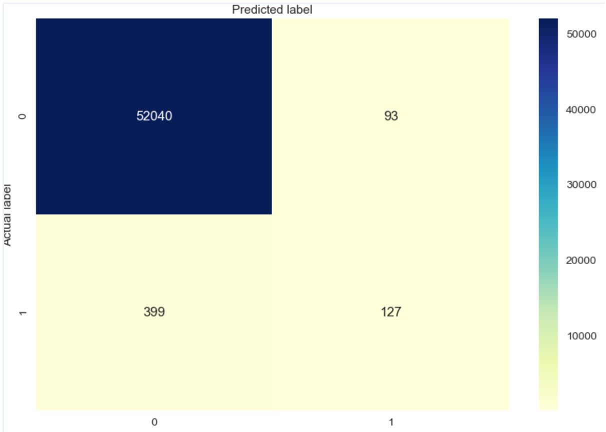

Fig.6. Figures shows the confusion matrix for severe toxic label in the data set.

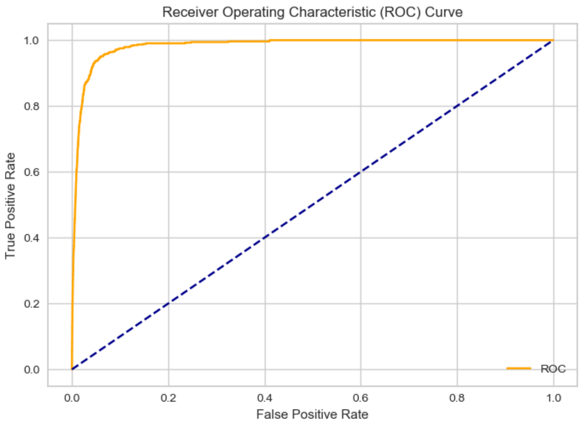

Fig.7. Figures shows the ROC curve for severe toxic label in the data set.

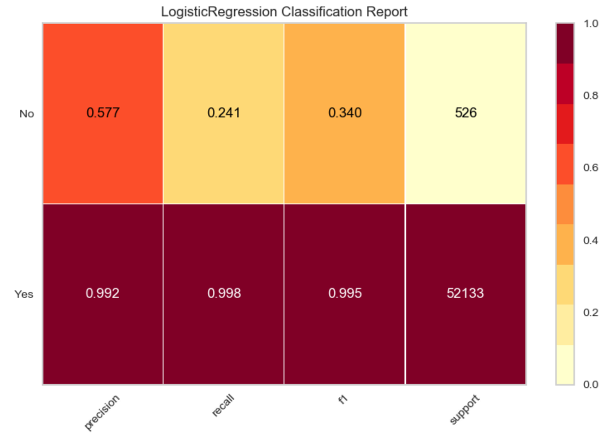

Fig.8. Figures shows the classification report for severe toxic label in the data set.

CONCLUSION

In this research work the deep learning approach for online public shaming has been discussed .Data visualization is performed on the data set. This helps us understand the data sets more precisely. It is also seen that data present on the internet is in large amount for which SVM algorithm does not give result. Hence other machine learning algorithms have been studied and also various deep learning algorithms have been understood in order to find better results.

4. FUTURE WORK

This research paper aims at having a comparative analysis of based on machine learning algorithms for prediction of online public shaming and data visualization. In the future deep learning based detection of public shaming can be considered for prediction in order to prevent any hazardous acts in the near future. A tool for detection of shaming comments can be implemented using deep learning and hybrid techniques.

5. REFERENCES

[1] Rajesh Basak, Shamik Sural , Niloy Ganguly, and Soumya K. Ghosh, : “Online Public Shaming on Twitter: Detection, Analysis, and Mitigation.”,2019.

[2] Anukarsh G Prasad, Sanjana S, Skanda M Bhat, B S Harish, “Sentiment Analysis for Sarcasm Detection on Streaming Short Text Data”, 2017.

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

[4] Mukul Anand, Dr.R.Eswari,: “ Classification of Abusive Comments in Social Media using Deep Learning”, 2019.

[5 Sreelakshmi K Rafeeque P C,:"An Effective Approach for Detection of Sarcasm in Tweets",2018.

[6] Soham Deshmukh, Rahul Rade,:"Tackling Toxic Online Communication with Recurrent Capsule Networks", 2018.