Cyberbullying Detection using Recursive Neural Network through Offline Repository

Nidhi Chandra[1], Sunil Kumar Khatri[2], Subhranil Som[3]

[1]*Amity School of Engineering and Technology, Amity University Uttar Pradesh, India*
[2,3]*Amity Institute of Information Technology, Amity University Uttar Pradesh, India*
[1]*nsrivastava5@amity.edu,* [2]*skkhatri@amity.edu,* [3]*ssom@amity.edu*

Abstract: **The objective of this paper is to predict the user behaviour based on his posts on social networking sites specifically Twitter. The data sets are captured through secured API's exposed by these social network sites and stored in data lakes or NoSQL databases. Tensorflow API's have been used to do predictive analysis of this stored data through recursive networks. This paper is to demonstrate the identification of specific text from the data which is available in various forms – structured and unstructured and coming from various online sources in real time, posted by users worldwide. The online sources referred to in this paper are social networking sites, twitter etc. where multiple users collaborate with each other and post contents. To demonstrate the approach, the Information is captured from these sites through API exposed by these vendors which is captured in NoSQL databases and then NLP is applied to break the text which is then applied against text corpus to identify analogous data. From programming perspective data strictures are used to store the data at run-time. Specific WordNet API is being leveraged for their capabilities to find synonyms.**

Keywords: **Semantic Similarity, Deep Learning, Recursive Neural Network and Tensor flow.**

I. INTRODUCTION

Internet hides user identity and in a sense, it provides anonymity to user. With the rise in social media among users, societies across the world are now closely connected, sharing their views and ideas in a form of comments, tagging and video sharing. Many of these views are direct views in a form of user opinions or indirect views in a form of pictures and videos, audios and sometimes music (religious or otherwise) where in it is difficult to understand opinion through programmatic means. These contents many of times are misused by unscrupulous elements to brainwash masses against political institutions, creating instability in the society and on many times leading to online radicalization. Such a trend is quite dangerous as it is leading to brain washing of masses without physical presence of such persons. This paper has tried to identify solution to check on such posts on various social media platforms such as twitter, Facebook etc. leveraging Information Technology capabilities.

From technology perspective it is quick to detect antisocial feeds and posts from the users and even getting the identity of the users in some cases. In other words it helps in automatic detection and processing of such information.

This paper focusses on Twitter as data source from social media perspective where in user creates their handles and publish their opinions which could be offensive and racist as well. To access data on their platforms these vendors (twitter, Facebook etc.) exposes their API's to public to be consumed in applications. These API's leverages platform independent and language agnostic REST technology such that they could be consumed from any language Implementations. The payload is in the form of JSON for which each language offers standard parsers to process.

Areas such as Natural Language Understanding (NLU), Language Translation, Image recognition, machine learning and deep semantic analysis plays an important role here. For example, many of the feeds might be in any standard Internationalization language format which needs to be translated into format understood by data dictionary. Similarly, if some image is being posted by some individual, it could be analyzed from image recognition perspective.

Internet Terrorism

Refers to provocations certain section of people, society or ethnic groups to take up arms or rebellion against the government or religious minorities leading to mass killings or devastation of society in one or the other forms. This is a tool taken up by enemy countries or terrorist groups to attack countries worldwide for their political or religious motives. Many governments have collaborated to fight this kind of terrorism. This also involves hacking major government's servers and putting their banners onto it to show their prowess. Many major IT organizations have come up with SIEM (Security Information and Event Management) Solutions. These IT tools provides real-time analysis of security alerts which are being generated by applications and network hardware and helps preventing hacking of such critical servers and might have an impact to bring down economy of a country for some hours leading to huge financial losses and hence an economic Internet terrorism.

Countering Internet Terrorism – Technical Challenges

The key challenge to counter Internet terrorism from technical perspective is Volume, Velocity and Algorithms. The volume at which data is getting generated 24x7x365 is tremendous and that too in huge speed by internet users across the world. The challenge is how to keep this data? Second challenge is how to

978-1-5386-4692-2/18/$31.00 ©2018 IEEE

classify this data and apply right set of algorithms to get appropriate results out of it as applying wrong algorithm would create wrong results and hence disastrous implications.

II. BACKGROUND

The language is a construct of multiple text where in text is composed of sequence of words from a vocabulary. The vocabulary contains multiple set of words. NLP can help extracting important concepts from the texts and assigning them to slot in a certain template. Information extraction comprises of named entity recognition wherein named entities are universal known words in dictionary.

The phonetic and phonology is another area of NLP which comprises of study of linguistic sounds and their relations to words. One can leverage morphology to identify internal structure of words and how they can be modified. This is helpful in identifying the double meaning or synonymous meaning words and sentences. This involves parsing complex words into their components. One of the key to understand sentences is to study structural relationships between words and literals used in making those sentences. NLP has special packages to help do this syntactic analysis.

Syntactic analysis is done in collaboration with semantic analysis which helps study the meaning of words and how these are combined to form the meaning of sentence.

The fundamental challenge of natural language processing is to understand the meaning of a piece of text. However, judging whether a computer program "understands" a piece of text is an ambiguous task. We thus distill the abstract task of understanding text to the concrete problem of determining whether two pieces of text have very similar or distinct meanings. This is the challenge of "semantic similarity." Although the problem of semantic similarity has a very simple statement, it has broad applications. This fact is a testament to both the importance and the difficulty of this problem. Application:

Automated Short-Answer Grading

Grading short-answer questions on tests by hand is very time-consuming and expensive. Consequently, researchers have recently been exploring methods of automated essay grading (Dikli 2006). A functional semantic similarity evaluator could automatically grade short-answer questions by evaluating the similarity of a student answer with the corresponding correct answer.

Application: Machine Translation

Although machine translation models have become very successful in recent years, it is not entirely clear how they should be evaluated. Possible evaluation metrics for machine translation include the Word Error Rate, the NIST score, the BLEU score, and the Translation Error Rate (Callison- Burch 2006, Vilar 2007). However, these schemes are ad hoc and provide only rough notions of what constitutes a "good translation." We propose that a semantic similarity system could evaluate a machine translator by measuring the similarity between the machine-produced translation and the gold standard.

Application: Image Captioning

In the past few years, automated image captioning has become a much-studied area in computer vision and machine learning (Pan 2004). Many such systems are generative; the system takes as input an image and produces a number of possible outputs. A functional semantic similarity evaluator would be useful in the training of automated image captioning systems by providing a system with a measure of goodness of the captions it produces.

Machine learning plays an important role with NLP. There are multiple methods that can be used to get the best results out of Natural Language Processing like Document Classifiers, Word Sense Disambiguation. In NLP Structured Models Tagging, Parsing, Extraction can be achieved. Unsupervised learning helps in generalization and structure induction.

Deep Learning

Deep Learning is a subfield of machine learning which depends upon layer to layer hierarchal representations, where the elements of the lower-level layer correspond to the elements of the upper level layer.

It is a method of abstractions through which learning is done on multiple levels. This kind of a process is helpful for the analyzation of images, audio and text. The Deep Learning methodology is simple an Artificial Neural Network setup alongside Multilayer Perceptron's which contain the hidden data fields.

Convolutional Neural Networks

Convolutional Neural Networks are networks which with help of multiple hidden layers provide the program an excellent ability to learn. The learning comprises of data description and visualization/classification.

The Learning process within a Convolutional Neural Network algorithm can be achieved by following a technique known as pre-learning. In this technique, hordes of data are fed into the program and the algorithm learns to realize certain aspects of images which are being input.

The reason why Convolutional Neural Network algorithms are in demand is because of their features, which are:-Structure simplicity, Limited training parameters, Adaptability

Normally, the Convolutional Neural Network algorithms contain two parts:

Extraction Layer:

The input field of each input synapse (also known as a neuron) is connected to the head of another neuron, thus making a

neurological chain. This chain connects the lower receptive layers to the upper receptive layers. This is used for extracting localized features and determining positional relationship.

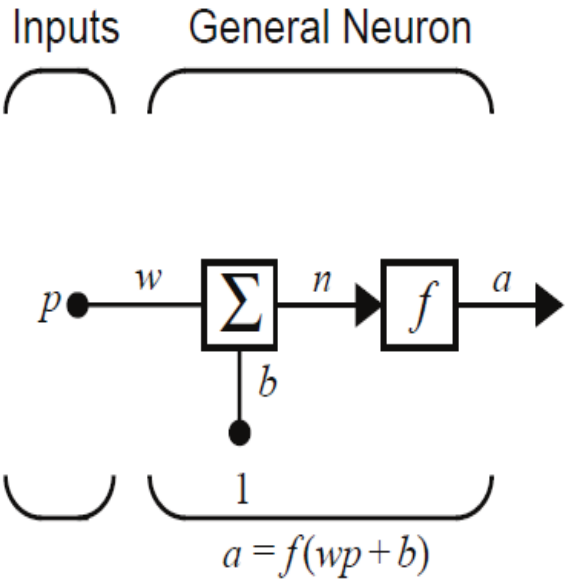
Feature Map Layer:

The Feature Map Layer is essentially a plane where each factor is given an equal defining weight. It uses the Sigmoid Function to begin processing the inner workings of the convolutional network, causing shift invariance to arise within the characteristic feature map.

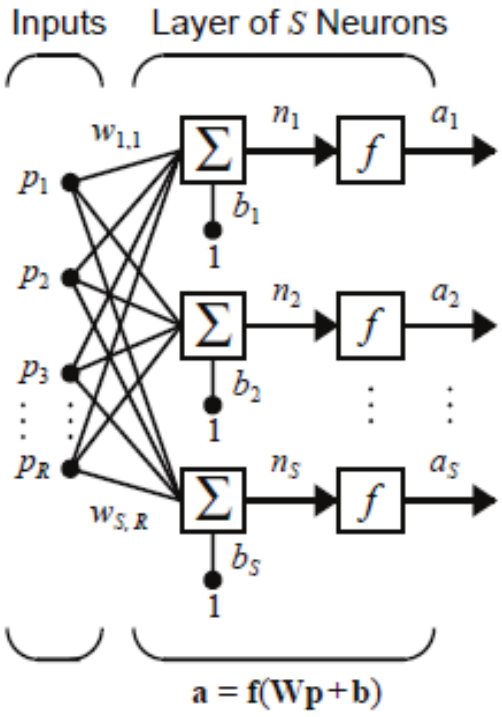
The general strategy followed by the conventional Convolutional Neural Network algorithms is that firstly, we extract simple features as the full (or high) resolution, and then transform them into lower resolution complex features.

The Deep Learning Architecture and NLP

Deep learning comprises of machine learning technologies which utilize 'deep' artificial neural networks, such as deep neural networks (DNN), convolutional neural networks (CNN), and recurrent neural networks (RNN).The current NLP systems are fragile because of their atomic symbol representation. There is a need for distributional and distributed representation as it helps enormously in NLP as they provide a powerful similarity model for words. Convolutional Neural Networks (CNNs) have proven to be adept at finding structure in raw text data. Historically, the text modeling field has been dominated by excessive amounts of preprocessing techniques to get the input text aligned and transformed into a form that modeling techniques could better handle. Neurons are the key elements of all forms of neural networks. They are represented mathematically as below

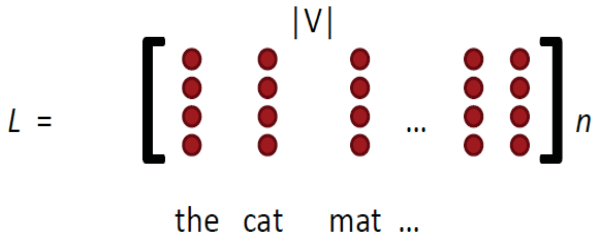


These neurons collectively help form a multi-layer neural network wherein they forms a fabric by passing information to the other connected layer with each layer responsible for specific processing.



Where each layer has its own weight matrix W, its own bias vector b, a net input vector p and an output vector a. When it comes to NLP the input comprises of set of words in a form of matrix which is referred to as word embedding matrix $\boldsymbol{L} \in \boldsymbol{R}^{n \times |V|}$ represented as below:

Word Embedding Matrix:

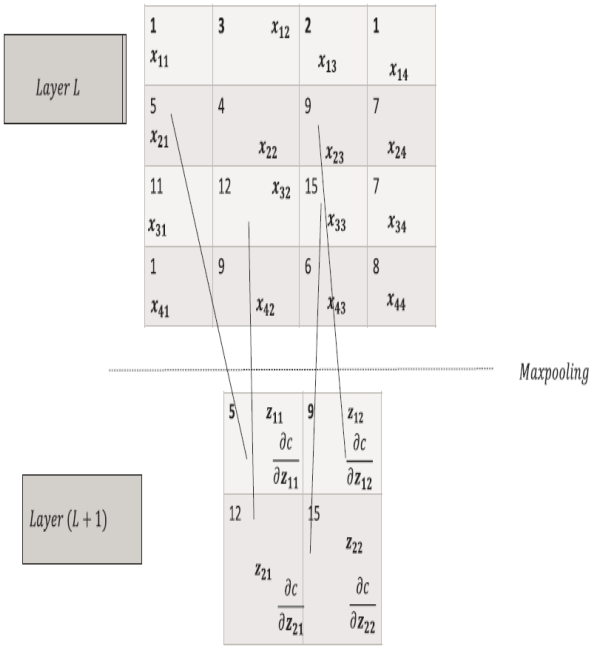


Prediction also plays an important role in NLP as it helps predict the future behavior of the person who is writing the text in different forums. To do predictions one needs to use convolution layers which leverages ordered set of items containing set of words in sentences or set of sentences inside the document.

Convolution is represented by the below equation:

$$f\left(x\left(n_1,n_2\right)\right)=\sum_{k_2=-\infty}^{+\infty}\sum_{k_1=-\infty}^{+\infty}x\left(k_1,k_2\right)f\left(\delta\left(n_1-k_1,n_2-k_2\right)\right)$$

The backpropagation in convolution network helps in refining the prediction results by applying feedbacks on the results processed by the layers. The layers are considered to be flip of gradient matrix which is cross correlation of gradient at (L+1) layers and the output at Layer L of feature map.



III. RELATED WORK

Internet Anti-Social Behaviour includes trolling, flaming, and griefing. Trolling has been defined as a person that engages in "negatively marked online behavior" (Hardaker 2010), or a user who initially pretends to be a legitimate participant but later attempts to disrupt the community (Donath 1999). Trolls have also been characterized as "creatures who take pleasure in upsetting others" (Kirman, Lineham, and Lawson 2012), and indeed, recent work

Has found that sadism is strongly associated with trolling tendencies (Buckels, Trapnell, and Paulhus 2014). Finally, Some literature instead provides taxonomy of deviant behavior (Suler and Phillips 1998).

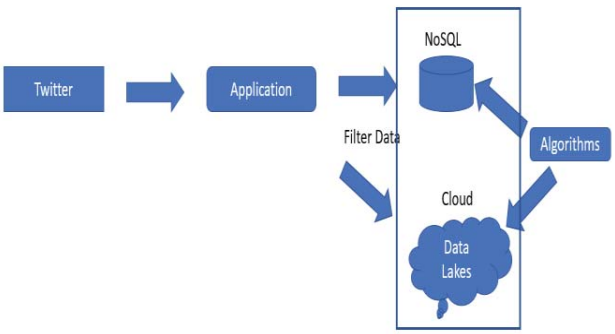
Cyberbullying, Cyber Trolling is recognized as an issue since year 2003 when Internet has evolved with social media applications as MySpace (2003), Orkut (2004), Facebook (2004), and Twitter (2005).Social Media and Cyber Forums then declared Cyberbullying, Anti-social Behaviour as a major concern area for research and prevention.

P. Badjatiya et. al (2011) identified racist and anti-social behaviour using Naïve Bayes, SVM, J48 and JRip techniques using offline tweets.[6] . Char-n-grams, word n-grams has been used as a data feature for identification. Tweeter tweets have been used as a data set. N. Djuric et. al. (2011) worked on Distributed representations of comments to classify speech as hate speech and clean speech. Word embeddings have been used to convert paragraph to vector.

[3] S. Hinduja and J. W. Patchin [2010] emphasized on terms as Bullying, Cyberbullying, and suicide and their behaviour and their similar pattern. [4]. this problem likes in classifier problem. D. Karthik et. al [2011] Modeled the detection of textual cyberbullying using TF-IDF and list of swear words as features to detect racist and inappropriate behaviour using Naive Bayes, SVM, J48 and JRip algorithm. [6] Swear words have been used as a corpus for detection. C. Nobata et. al. used Yahoo News feeds to detect Abusive and Non Abusive language in online user content.[9] word and character N-grams, Linguistic features, Syntactic and Distributional Semantics have been used as a metric for analysis.[9]. Regression model have been used for prediction .K. Reynolds et. al.[2011] Used machine learning techniques to detect cyberbullying. number of "bad" words (NUM), density of "bad" words have been used as a metric. [12] J48 and JRip simulated to detect cyberbullying. C. Van Hee, et. al.[2015] worked on Automatic detection and prevention of cyberbullying via word unigram and bigram bags-of words, character trigram bag-of-words, sentiment lexicon features(comment2vec).[15] Data have been modelled via SVM and categorized the text in Threat, Insult, Defense, Sexual Talk, Defamation, Encouragements and Swear categories.

IV. PROPOSED METHODOLOGY

To handle these technical challenges, one would need leverage cloud Infrastructure to handle volume and velocity because it provides appropriate scalability for data lakes where this data could be stored. One could use stream analytics to filter out invalid data to manage the cost of the overall solution. Various cloud vendors provides GPU driven hardware to run ML and Deep Neural Networks to handle complex interrelated data which would mostly originate from social platforms.



Data Collection Methodology

Every social network vendor exposes a secured API which can be consumed using secret tokens and secret keys provided by this vendor. This secret token and keys helps maintain encryption of data and isolation during transit. The data transfer is based on Push methodology which means that once API makes a connection real time data is pushed onto the client. This data can be collected in NoSQL database as it comes in JSON format or a data lake where it could be processed further.

V. EXPERIMENT

In the last several years, training deep learning algorithms on large corpora of text has emerged as a general, powerful approach, performing as well as hand-designed algorithms based on many years of research and tuning.

Contemporary deep learning algorithms used to determine the similarity of two general pieces of text. Although we could build a better performing system by training on a particular task we instead seek to build a system which can evaluate the similarity of any two arbitrary sentences. The reason for this is two-fold. First, the lack of a very large publicly available dataset for any one specific task makes it difficult to use deep learning methods, which require very large datasets for training. More importantly, we feel that training for a particular task would not show the extent to which deep learning methods can truly evaluate semantic similarity of two sentences gathered from an arbitrary domain.

The texts of the semantic similarity challenge are quite short, only consisting of individual sentences and sometimes of very short sentences. Thus, any models trained only on the provided text are quite limited. To get around this limitation, we represent each word w by a vector L[w]. We construct these word vectors, a weighted bilinear model which produces distributed word representations based on co-occurrence counts in a large corpus. In particular we have used 50- or 100-vectors pre-trained on Anti-Social data corpus. We use pre-trained word vectors because the relatively small size of our dataset prevents us from learning word vectors directly. Recurrent neural networks take as input a structured set of words or tokens. Although each model treats its input differently, model has been trained to report the probability that a given pair of sentences belongs to each similarity category.

Although each model treats its input differently, we train the models to report the probability that a given pair of sentences belongs to anti-social similarity category.

Recurrent Neural networks

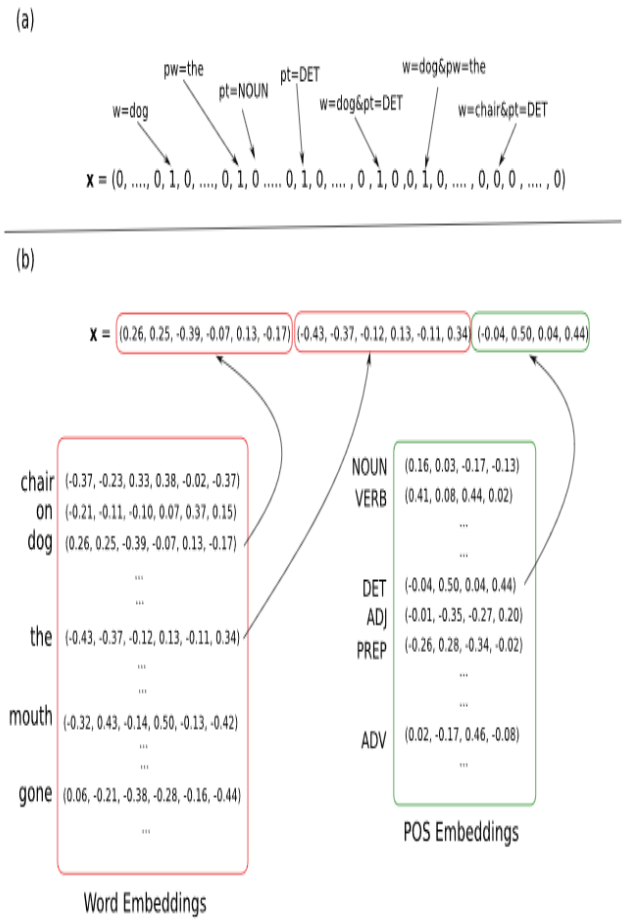
Recurrent neural networks are powerful deep learning models which take as input a sequence of Tokens, each of which is used to update a hidden state.

Recursive Neural networks

Recursive neural networks are deep learning models frequently used in applications of deep learning to natural language processing because, unlike recurrent neural networks, they can take advantage of known linguistic structure. Rather than working on a sequence of tokens, recursive neural networks take as input a (binary) tree. The leaves of this tree correspond to the words in an input sentence, and we infer the binary tree structure using a calculated parse of the sentence. For each leaf with word w, we assign the vector L[w].

Data Sets

The datasets are generated as JSON document which are tweets pushed by twitter through its API in NoSQL data base and then transformed into binary tuples also referred to as **one hot representation** which is further represented as binary tree. The handles used to capture data are #paki and #chinki.



Logistic Regression models are used to achieve high accuracies by applying twitter data sets to its variables. The macro is defined to calculate the score by averaging Precision and Recall.

Model

We leveraged feed forward neural network for model as a function NN(x) which takes as input a din (subscript)

dimensional vector x and produces a dout(subscript) dimensional output vector. We use this function as classifier and assign the input vector x a degree of membership through multiple dout(subscript) classes. As a model that is applied to neural nets we leveraged CBOW (continues bag of words) as mentioned below:

$$WCBOW(f_1, ..., f_k) = \frac{1}{\sum_{i=1}^{k} a_i} \sum_{i=1}^{k} a_i v(f_i)$$

And the embedding layer is evaluated as below

$$\begin{aligned} \mathbf{x} = c(f_1, f_2, f_3) &= [v(f_1); v(f_2); v(f_3)] \\ NN_{MLP1}(\mathbf{x}) &= NN_{MLP1}(c(f_1, f_2, f_3)) \\ &= NN_{MLP1}([v(f_1); v(f_2); v(f_3)]) \\ &= (g([v(f_1); v(f_2); v(f_3)]\mathbf{W^1} + \mathbf{b^1}))\mathbf{W^2} + \mathbf{b^2} \end{aligned}$$

Baseline performance for models can be achieved by training them by leveraging three capabilities:

a) Bag of word vector

b) GloVe vector

c) Jaccard similarity

One of the main components of RNN approach is the use of embedding which represents the feature as a vector in a low dimensional space. These vectors come through approaches such as - Random Initialization, Supervised Task-specific Pre-training, and Unsupervised Pre-training. Given a word w and its context c, different algorithms formulate different auxiliary tasks.

In all cases, each word is represented as a d-dimensional vector which is initialized to a random value. Training the model to perform the auxiliary tasks well will result in good word embeddings for relating the words to the contexts, which in turn will result in the embedding vectors for similar words to be similar to each other.

VI. RESULTS

The problem statement lies in the category of cyber bullying and anti-social classification. This experiment presents effective mechanisms for classifying cyberbullying and cyber trolling as antisocial behavior. Tensorflow contribution library estimator function is used to generate the prediction from data set. We prepared the training and test set and converts them into NumPy array to keep them in Pandas Data Frame format. For Example written natural text “Smash you wanna try” comes under the threat category. Another example “Commit suicide, you call urself a Stand up Comedian” falls under the category of Curse word where Average classifier precision is (0.67) to predict cyberbullying.

VII. FUTURE WORK

In this experiment we have explored the recurrent neural networks, one using GRU units in the hidden layers and one using ReLU units in the hidden layers. In future, we may explore Recursive neural Network for performing additional pretraining on the word vectors.

REFERENCES

[1] P. Badjatiya, S. Gupta, M. Gupta, and V. Varma, “Deep learning for hate speech detection in tweets, ” 26th International World Wide Web Conference, 2017 .pp759-760, 2017.

[2] R. Johnson and T. Zhang, ”Supervised and semi-supervised text categorization using lstm for region embeddings, ” ICML'16 Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48 pp. 526-534, 2016.

[3] N. Djuric, J. Zhou, R. Morris, M. Grbovic, V. adosavljevic, and N. Bhamidipati, ”Hate speech detection with comment embeddings, ” 26th International World Wide Web Conference, 2015 pp 29, 30, 2015.

[4] Y. Tang and A.-R. Mohamed, “Multiresolution deep belief networks, ” in IEEE Transactions on Neural Networks and Learning Syatems, VOL. 25, NO. 12, Dec 2014.

[5] Y. Bengio, A. Courville, and P. Vincent, “Representation learning: A review and new perspectives, ” IEEE Transaction on Pattern Analytics and Machine Intelligence, vol. 35, no. 8, pp. 1798–1828, Aug. 2013.

[6] A. Mohamed, G. E. Dahl, and G. Hinton, “Acoustic modeling using deep belief networks, ” IEEE Transaction on Audio, Speech, Language Processing, vol. 20, no. 1, pp. 14–22, Jan. 2012.

[7] N. Srivastava and R. Salakhutdinov, “Multimodal learning with deep Boltzmann machines, ” in Proceedings of Advance Neural Infrastructure Process and System, vol.1 2012, pp. 2231.

[8] A. Krizhevsky, I. Sutskever, and G. Hinton, “Imagenet classification with deep convolutional neural networks, ” in Proceedings of Advance Neural Infrastructure Process and System, vol.1 2012, pp.4.

[9] M. Baccouche, F. Mamalet, C. Wolf, C. Garcia, and A. Baskurt, “Spatiotemporal convolutional sparse auto-encoder for sequence classification, ”in British Machine Vision Conference, 2012. pp 63, 72

[10] D. Ciresan, U. Meier, and J. Schmidhuber, “Multi-column deep neural networks for image classification, ” in Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Feb. 2012, pp. 3642–3649.

[11] D. C. Cire¸san, U. Meier, J. Masci, L. M. Gambardella, and J. Schmidhuber, “Flexible, high performance convolutional neural networks for image classification, ” in Proceeding of International Joint Conference of Artificial Intelligence, 2011, pp. 1237–1242.

[12] D. Karthik, R. Roi, and L. Henry, ”Modeling the detection of textual cyberbullying”. In Workshop on the Social Mobile Web, ICWSM, 2011.

[13] Yunos, Z. and Hafidz Suid, “ Protection of critical national information infrastructure against cyber terrorism: Development of strategy and policy framework, ” In International Conference on Intelligence and Security Informatics (ISI), 2010 IEEE. pp169.

[14] S. Hinduja, J. W. Patchin, " Bullying, cyberbullying, and suicide" Archives of suicide research, Vol.14 (3):pp206-221, 2010.

[15] A. Graves and J. Schmidhuber, "Framewise phoneme classification with bidirectional lstm and other neural network architectures, " IEEE Transaction on Neural Network., vol. 18, nos. 5–6, pp. 602–610, 2005.

[16] Ajay Vikram Singh, Vandana Juyal, Ravi Saggar, "Trust based Intelligent Routing Algorithm for Delay Tolerant Network using Artificial Neural Network", Wireless Networks (WINE), Springer Publication, US, Volume-22, Issue-135 pp 1-10, DOI 10.1007/s11276-015-1166-y Print ISSN 1022-0038, Online ISSN 1572-8196.

[17] Nidhi Chandra, Sunil Kumar Khatri, **Subhranil Som**, (2017) "Anti-Social Comment Classification based on kNN Algorithm", 6th International Conference on Reliability, Infocom Technologies and Optimization (ICRITO) (Trends and Future Directions), IEEE Conference, indexed with SCOPUS Sep. 20-22, 2017, Amity University Uttar Pradesh, Noida, India.

[18] Shivani Chowdhary, **Subhranil Som**, Vipul Tuli, Sunil Kumar Khatri (2017), "Security Solutions for Physical Layer of IoT", International Conference on Infocom Technologies and. Unmanned Systems (Trends and Future Directions) (ICTUS'2017), December 18-20, 2017, IEEE Conference, indexed with SCOPUS**,** Amity University Dubai Campus, Dubai International Academic City, Dubai.

[19] Fatma Al Shuhaimi; Manju Jose; Ajay Vikram Singh, "Software defined network as solution to overcome security challenges in IoT", 2016 5th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO) at AUUP, NOIDA, India, September 07-09, Year: 2016 Pages: 491 - 496, DOI: 10.1109/ICRITO.2016.7785005.