Journal of
Data Acquisition and Processing

ISSN: 1004-9037
https://sjcjycl.cn/
DOI: 10.5281/zenodo.777768

# AUTO ENCODER ARCHITECTURE TO DETECT AND ENCRYPT THE UNWANTED COMMENT EXPLOITING IN SOCIAL MEDIA NETWORK

**Dr. P. Kavipriya**
Associate Professor, Department of Computer Science KPR College of Arts Science and Research, Coimbatore, India
kavipriya.p@kprcas.ac.in / kavipriya.rajen@gmail.com

**Dr . D. Maheshwari**
Associate Professor, Department of Computer Science with Data Analytics, KPR College of Arts Science and Research, Coimbatore, India
maheshwari.d@kprcas.ac.in / maheshgkrish@gmail.com

## Abstract

Online Social media is exploiting with large number of abusive comment on proliferation of the user-generated contents. Social media platform has been empowered user to create, share, exchange content with each for socializing but it leads to various challenges to network user on propagation of abusive content from cyber bullies, and haters to their post as it spread to various audience. However many conventional approaches using machine learning model has been proposed to handle these challenges but it provides wrong misinterpretation. In order to mitigate those challenges, a new deep learning architecture entitled as Autoencoder is designed to detect and encrypt the unwanted content in social media. In this architecture, dataset is collected from twitter. It contains comments for various event and news and it is transformed as CSV file. Transformed file is preprocessed using data normalization, stop word removal and stemming process for removing number and emojis. Preprocessed data is projected to feature extraction model, which process the data on computation of the feature from part of speech tagging. Computed feature vector is employed to deep learning model along BERT based sentiment analysis architecture to identify the polarity of the term. Positive polarity of the vector is classified as normal comment and negative polarity of the feature vector is classified as unwanted comment using deep learning model. Classified content is encodedusing encoderin deep learning model represent the abusive content in the encoded form to other audience. Experimental analysis of the proposed approach is compared with conventional approaches to evaluate the performance. Further performance of the approaches is computed using performance metric such as precision, recall and f measure. Proposed model produces 98% efficiency is identifying the abusive comment compared state of art approaches.

**Keywords: Unwanted Review Classification, Autoencoder, Deep learning, BERT, Principle Component Analysis, Sentiment analysis**

## 1. Introduction

Online Social media is exploiting with large number of unwanted review on proliferation of the user-generated contents. Social media platform has been empowered user to create, share and exchange content with each for socializing but it leads to various challenges to network user on propagation of abusive content from cyber bullies and haters to their post as it spread to various audience. However many conventional approaches have been proposed to handle these challenges but it provides wrong misinterpretation.

In order to mitigate those challenges, a new deep learning architecture entitled as Autoencoder is designed to detect and encrypt the unwanted reviews content in social media. In this architecture, dataset is collected from twitter. It contains comments for various product reviews and it is transformed as CSV file. Transformed file is preprocessed using data normalization, stop word removal and stemming process for removing number and emojis. Preprocessed data is projected to Principle Component Analysis, which process the data in matrix form to generate the eigen vector with eigen value.

Computed feature vector is employed to deep learning model along sentiment analysis BERT architecture to identify the polarity of the term. Positive polarity of the vector is classified as normal comment and negative polarity of the feature vector is classified as unwanted review using autoencoder classifier. Classified content is encrypted using encoder technique to represent the unwanted review in the encrypted form to other audience.

Rest of the paper is organized as follows, section 2 represents the review of literatures and section 3 represents the design of the proposed Autoencoder model to twitter dataset on incorporation of the sentiment analysis named as BERT in the model. Section 4 evaluates the performance of the proposed model against the conventional model on accuracy measure in the mentioned experimental setup.

## 2. Related work

In this section, various review of literature using machine learning model related to unwanted reviewclassification and filtering is analyzed and detailed on the functional specification is as follows

### 2.1. Unwanted Message Filtering

In this literature, unwanted message in social media is filtered using machine learning algorithms. it uses decision tree algorithm such as naive bayes to detect the unwanted message from the content posted in the user walls of twitter and facebook. Particular model process the dataset with preprocessing , feature extraction and feature classification step. It produce accuracy of 84% in classifying the unwanted message from normal message in the user walls

## 3. Proposed model

In this part, detailed specification of the proposed Autoencoder architecture along sentiment analysis BERT architecture to identify the polarity of the extracted word and basic data processing step of the twitter dataset is presented as follows

### 3.1. Data Transformation

Initially dataset extracted from the twitter is converted into CSV file format, as it is high reliable data format for matrix operation of feature extraction and feature classification. In addition, it is supported format to sentiment analysis.

### 3.2. Data Preprocessing

Transformed dataset is preprocessed with normalization, stop word removal, emojis removal and stemming process for processing the data for classification and encryption on detection of abusive terms in the processed data.

- Stop Word Removal : It is to remove the stop words like and , is , was , were , are etc as it is considered non actionable words

- Stemming process : It is to remove "ing" and "ed" terms in the comment as it is also considered as non actionable words
- Emoji removal: it is to remove emoji in the comment as it is considered non actionable information for data classification
- Number Removal: It is to remove numbers in the comments

### 3.3. Tokenization

It is to split the words as tokens. The classifier easily processes token. Token considered in words. Tokenization also removesthe punctuation, whitespaces, and delimiters from the comments as it is considered as meaningless.

### 3.4. Feature Extraction

Feature extraction is carried out principle component analysis. It is processed using data in matrix form. Correlation matrix compute the frequent occurring words in the particular review sentences and covariance matrix compute the non frequent occurring words in the entire review content is collected as feature vector.

### 3.5. Sentiment analysis -BERT model

Sentiment is calculated to the feature vector. In this BERT model, feature vector is used to analyze the sentiment. It provide the polarity to the features as positive and negative on contextual analysis of the terms. Model contains sentiment base and modifier base along rule base to compute the polarity to the word. It is high utilized NLP model.

### 3.6. Classification -Autoencoder

Autoencoder is a classifier employed to classify the polarity of the word. It uses the fully connected layer to process the polarity of the features to classify the comment. The feature vector with positive polarity is considered as normal comment and feature vector with negative polarity is considered as unwanted comment.  Figure 1 represents the architecture of the proposed model.

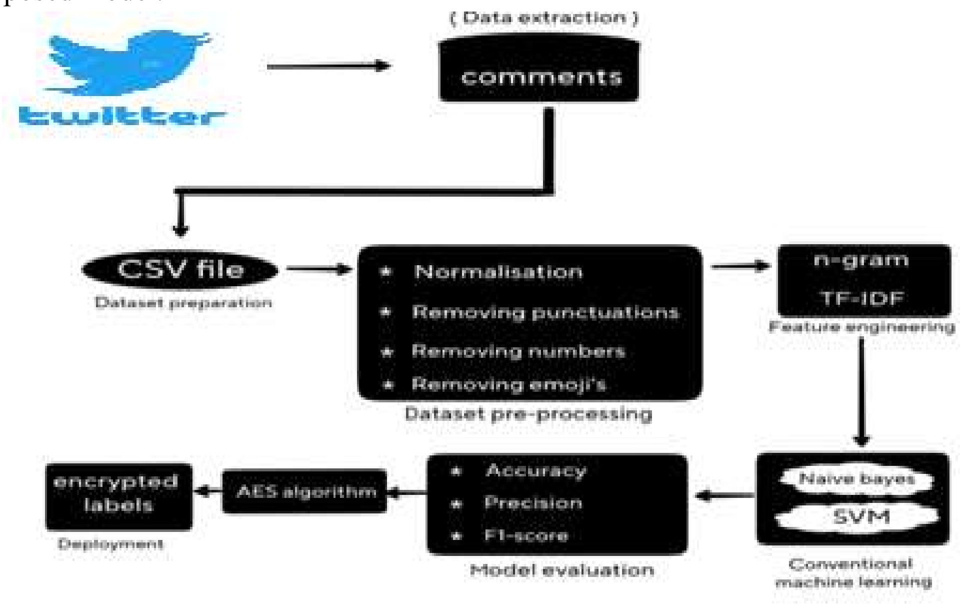


**Figure 1: Architecture of the Proposed model**

### 3.7: Encoding

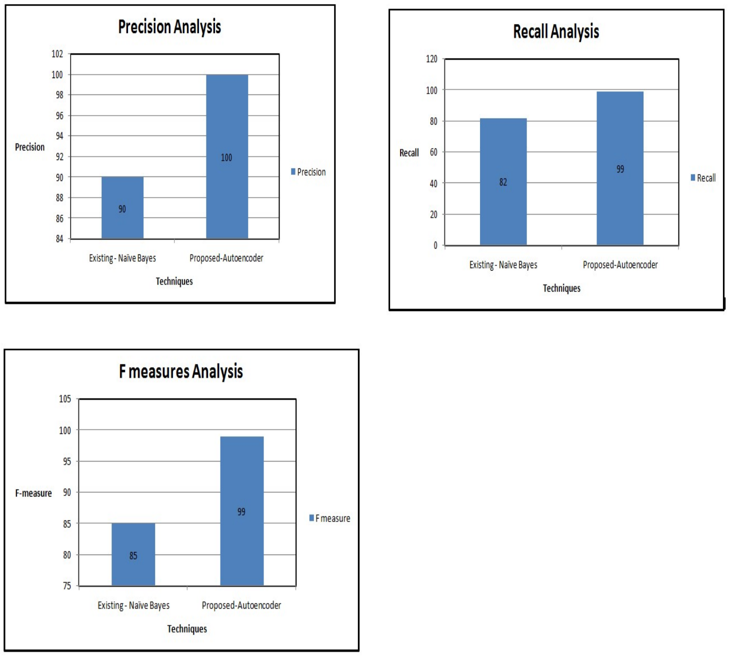
On detection of the unwanted word by auto-encoder classifier, encoding layer encrypts the unwanted words. It transforms the text to encrypted form and display it in the user wall. It displays encrypted form to audience of the user whereas user will be displayed with same content.

**Algorithm 1: Unwanted Review Detection and encoding**

```
Input: User Review
Output: Unwanted Review
Process:
Transformed Data = CSV( data extracted )
Preprocess()
$S_w$ = Stop Word Removal (Transformed data)
  $S_t$ = Stemming ($S_w$)
$R_t$= Remove (Emoji on $S_t$)
    T = Tokenize($R_t$)
     Compute feature vector_PCA
Correlation Matrix = frequency of words in the review sentence.
Covariance Matrix = Non frequency of words in the review sentences
   Feature Vector FV= Eigen Vector ( Covariance and Correlation Matrix)
BERT ()
Sentiment polarity P = Sentiment base ( FV)
Classify Polarity
       Class C = Autoencoder (p)
          Class = { Normal , Unwanted Review }
if (Comment = Unwanted or Negative)
   Encrypt the Review
        Cipher text = Encoding ( Unwanted Review)
Display EncodedReview.
Else
 Display Normal Review
```

## 4. **Experimental results**

Experimental of the proposed approach is carried out on the twitterdataset[10] and performance of the current approach is evaluated using the various performance measures such as precision, recall and Fmeasure to evaluate the efficiency and accuracy against the conventional approaches.



(a) (b) (c)

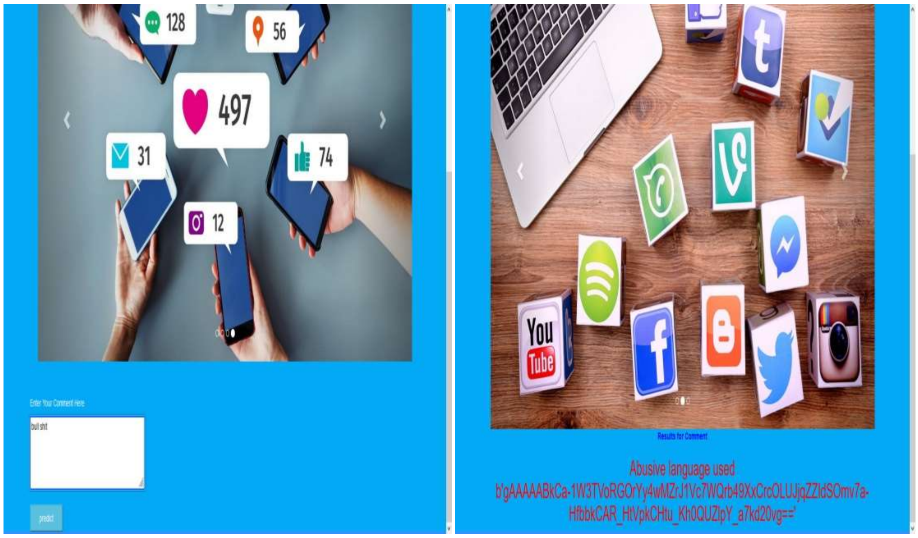
**Figure 2: Performance evaluations of the Abusive Detection in comment**

Proposed approach generates high Performance as it incorporates sentiment analysis on the classifier with deep learning model technique. Unwanted review detection accuracy is computed and it is depicted in the Figure 2 and Table 2.

**Table 2: Performance Evaluation**

| **Technique** | **Precision** | **Recall** | **Fmeasure** |
|---|---|---|---|
| Naive Bayes –Existing | 90 | 82 | 85 |
| Autoencoder –Proposed | 100 | 99 | 99 |

As the proposed approach possessed higher accuracy when compared with the existing approach, we opted for the proposed approach to assign labels for the review as unwanted and normal. On detection of unwanted review, it is encoded and it is displayed on the user wall. Figure 3 represent output of the model.



Enter Your Comment Here

bull shit

Results for Comment

Abusive language used

```
b'gAAAAABkCa-1W3TVoRGOrYy4wMZrJ1Vc7WQrb49XxCrcOLUJjqZZIdSOmv7a-HfbbkCAR_HtVpkCHtu_Kh0QUZIpY_a7kd20vg=='
```

(a) (b)

**Figure 3: Output of the model (a) Review Input (b) comment encoded form**

## Conclusion

We designed and implemented a deep learning model asAutoencoder to detect the unwantedreview in the twitter content of user. Particular model incorporates the BERT model for sentiment analysis of content as itreduce the complexity of the data processing and increase the detection accuracy of the comment with classification of the abusive content and further it encrypt the abusive comment and display on the user wall of youtube. On experimental analysis, it is considered to more accurate compared to the conventional approaches. It is required to incorporate the constraints to reduce the over fitting issues is considered as future direction of the particular work.

## References

[1] E. F. Cardoso, R. M. Silva, and T. A. Almeida, ``Towards automatic filtering of fake reviews," Neurocomputing, vol. 309, pp. 106116, Oct. 2018.

[2] L. Da Xu,W. He, and S. Li, ``Internet of Things in industries: A survey," IEEE Trans. Ind. Informat., vol. 10, no. 4, pp. 22332243, Nov. 2014.

[3] Kumar, E. Boopathi, and M. Sundaresan. "Edge detection using trapezoidal membership function based on fuzzy'smamdani inference system." 2014 International Conference on Computing for Sustainable Global Development (INDIACom). IEEE, 2014.

[4] N. Jindal and B. Liu, ``Opinion spam and analysis," in Proc. Int. Conf. Web Search Web Data Mining (WSDM), 2008, pp. 219230.

[5] Yookesh, T. L., et al. "Efficiency of iterative filtering method for solving Volterra fuzzy integral equations with a delay and material investigation." Materials today: Proceedings 47 (2021): 6101-6104.

[6] L. Li,W. Ren, B. Qin, and T. Liu, ``Learning document representation for deceptive opinion spam detection," in Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data. Nanjing, China: Springer, 2015, pp. 393404.

[7] H. Aghakhani, A. Machiry, S. Nilizadeh, C. Kruegel, and G. Vigna, ``Detecting deceptive reviews using generative adversarial networks," in Proc. IEEE Secur. Privacy Workshops (SPW), May 2018, pp. 8995.

[8] Kumar, E. Boopathi, and V. Thiagarasu. "Color channel extraction in RGB images for segmentation." 2017 2nd International Conference on Communication and Electronics Systems (ICCES). IEEE, 2017.

[9] Reddy, C. S., Yookesh, T. L., & Kumar, E. B. (2022). A Study On Convergence Analysis OfRunge-Kutta Fehlberg Method To Solve Fuzzy Delay Differential Equations. Journal of Algebraic Statistics, 13(2), 2832-2838.

[10] R. Y. K. Lau, S. Y. Liao, R. C.-W. Kwok, K. Xu, Y. Xia, and Y. Li, ``Text mining and probabilistic language modeling for online review spam detection," ACM Trans. Manage. Inf. Syst., vol. 2, no. 4, pp. 130, Dec. 2011

[11] Y. Ren and Y. Zhang, ``Deceptive opinion spam detection using neural network," in Proc. 26th Int. Conf. Comput. Linguistics: Tech. Papers (COLING), 2016, pp. 140150.

[12] Kumar, E. Boopathi, and M. Sundaresan. "Fuzzy inference system based edge detection using fuzzy membership functions." *International Journal of Computer Applications* 112.4 (2015).

[13] A. Mukherjee, V. Venkataraman, B. Liu, and N. Glance, “Fake review detection: Classification and analysis of real and pseudo reviews," Univ. Illinois Chicago, Chicago, IL, USA, Tech. Rep. UIC-CS-03-2013, 2013.

[14] R. Yafeng, J. Donghong, Z. Hongbin, and Y. Lan, ``Deceptive reviews detection based on positive and unlabeled learning," J. Comput. Res. Develop., vol. 52, no. 3, p. 639, 2015.

[15]A. Heydari, M. Tavakoli, and N. Salim, ``Detection of fake opinions using time series," Expert Syst. Appl., vol. 58, pp. 8392, Oct. 2016.