Annals of R.S.C.B., ISSN: 1583-6258, Vol. 25, Issue 5, 2021, Pages. 4578 - 4586

Received 25 April 2021; Accepted 08 May 2021.

A Live Suspicious Comments Detection using TF-IDF and Logistic Regression

Dr.S.Gnanavel[1], N.Duraimurugan[2], M.Jaeyalakshmi[3], M.Rohith[4], B.Rohith[5], S.Sabarish[6]

[1, 2, 3]Assistant Professor (SG), [4, 5, 6]UG Student

[1, 2, 3,4,5,6]Department of SomputerScience and Engineering, Rajalakshmi Engineering College, Chennai

[1]gnanavel.s@rajalakshmi.edu.in

ABSTRACT

The internet has changed the lives of everyone since its arrival. It gave birth to social networking platforms and online forums where people can share their thoughts and it also contains vast amount of information and it is become an effective and convenient communication tool for people to convey their thoughts. Although internet paves the way for many benefits but it also has its own drawbacks. One of the drawbacks is the threat of abuse and harassment for sharing our thoughts. Many platforms fail to moderate user comments and toxic behaviors which restrict people from expressing themselves. In this concern, we specialize in creating a machine learning model using logistic regression algorithm to implement the model in a live chat to detect abusive and harmful comments in real time. This research detects different types of toxicity levels and classifies them into toxic, threat, severe toxic, obscene, insults and identity-based hatred and to list out the names of the abusers and toxic users.

Keywords

Machine Learning, Logistic Regression, Toxic, Threat, Severe Toxic, Obscene, Insults and Identity-Based Hatred, Convolutional Neural Networks

1. Introduction

The Internet is a huge network that connects computers all over the globe. Social networking is a crucial part of the cyberspace.Internet and social networking have been growing vast in the 21st century. Internet has become a very successful and suitable communication platform for users to share their view.Expressing opinions and thoughts are the major benefits of social platforms. Reaching vast audience in a short period of time with almost no cost can only be achieved through social platforms. Building brands and reaching specific target audience has been made easier through social media and networking sites. social networking platform is the huge source of content, covering wide variety of content ranging from informative to entertainment.

Easy access to the Internet, social media, chat forum, other online platforms are producing a huge volume of text contents in recent years. A number of these contents aren't genuine, authentic and even suspicious [3]. People may have difference in opinions which leads to debates among groups in social media platforms. Sometimes people tend to have non-healthy debates which may lead to the usage of offensive language and toxic comments which results in user getting abused and harassed [1]. A toxic comment is an ill-mannered, discourteous comment which may force some of the users to leave the discussion or the conversation [4]. This refrains users from sharing their honest thoughts and opinions. In result of those verbal abuses communities are forced to restrict and limit user comments or even they are forced to turn off the comments section completely. This leads to lower user interaction which causes users to lose interest in these platforms and this affects advertisement and subscription revenue of those social media platforms. Social media platforms like twitter, Facebook, and Instagram has monitoring tools for healthy communication but these models are prone to error and poorly efficient. Our objective is to make people express their thoughts and opinions without hesitation and to keep

Annals of R.S.C.B., ISSN: 1583-6258, Vol. 25, Issue 5, 2021, Pages. 4578 - 4586

Received 25 April 2021; Accepted 08 May 2021.

the users away from toxicity and negativity. In our work, we developed a system to effectively detect and classify toxic comments into six types. That is listed as threat, toxic, obscene, severe toxic, identity-based hatred and insults. Based on the toxicity level we list out the toxic users and send to the administrator.

2. Related Works

Navoneel Chakrabarty proposed a model that uses algorithms and methods such as removal of punctuation, lemmatization, removal of stop words, tf-idf and word count vectorizer to create a Support Vector Machine Model with Linear Kernel. He used this model for finding toxic comments and its types that are toxic, threat, severe toxic, obscene, insults and identity-based hatred. His aim is to have a fair online discussion and share ideas on social media [1].

Omar Sharif et al created a model "Detecting Suspicious Texts Using Machine Learning Techniques". The researchers created a system to detect suspicious comments or texts in Bengali. According to the researchers, they were the ones who set up the system to investigate suspicious text detection in Bengali language. They developed a dataset containing 7000 text documents in Bengali and used multiple algorithms likeLogistic regression, Decision tree, stochastic Random forest, gradient descent, and Multinomial Naïve Bayes. Results obtained from these algorithms were compared with each other, existing systems and human experts (baseline). Finally, researchers used those above methods to distinguish a text whether it is suspicious or non-suspicious [3].

Megha K.B et al presented a theory "Monitoring of Suspicious and Fraudulent Activities on Online Forums" that uses sentimental analysis which includes four phases such as tokenization (splitting the text corpus into separate elements), stop word removal, stemming & lemmatization (converting a text into its base form) and naive bayes classification algorithm which are used to categorize user text into negative, positive and neutral. Text will be accepted or ignored based on the result [5].

Julian Risch et alproposed "Toxic Comment Detection in Online Discussions" focusing on comments of online news platforms. To classify the toxic comments neural network architecture has been implemented. According to the researchers three layers of Neural network architecture has been used namely Long Short-Term Memory (LSTM), Gated Recurrent Unit (GRU) to classify toxic comments. They also implemented Naive Bayes algorithm for sentimental analysis. Using these algorithms, they classified the text comments into Profanity, Insults, Threats, Hate Speech, Otherwise Toxic. This paper also includes discussion about fine-grained classifications [4].

Bavane A.B et al presented the paper "Monitoring Suspicious Discussion on Online Forum by Data Mining" in which they use text mining algorithms such as Stop word selection, Affix Stripping, Suffix Stripping and Matching Algorithms to extract suspicious texts from the chat and replace the extracted text with asterisk. By using this method, they maintain a healthy chat [6].

Spiros V. Georgakopoulos et al suggested a system to classify toxic comments in this paper "Convolutional Neural Networks for Toxic Comment Classification". Kaggle's Wikipedia dataset has been used in Convolutional Neural Networks (CNN). Convolutional Neural

Annals of R.S.C.B., ISSN: 1583-6258, Vol. 25, Issue 5, 2021, Pages. 4578 - 4586

Received 25 April 2021; Accepted 08 May 2021.

Networks are multistage trainable neural networks architectures developed for categorization tasks. CNN includes Embedding Layer, Convolutional Layers, Fully-Connected Layer and Pooling Layers to classify toxic comments. CNN can perform better in classifying toxic comment than well-established methodologies [8].

A research paper published by Tanya Srivastava, et al "Monitoring of Suspicious Discussions on Online Forums Using Data Mining" suggests various algorithms and techniques that can be used for finding suspicious comments and behaviors in online forums. This paper gives importance to datamining techniques and sentimental algorithms which classifies text into six categories such as hacking, sexuality, religious, piracy, gambling, fraud. This is implemented using python's Natural Language Toolkit (NLTK) library [2].

Harshala Patil, Kalpana Khandelwal, Trupti Kini presented a system "Monitoring Online Forum for Suspicious Discussions with Text Summarization" by using Statistical Corpus-Based Approach to monitor suspicious movement and to monitor illegal activities in online forums. Stop word algorithm, lemmatization, sentimental analysis is implemented in the system. Finally, text summarization is used to get brief summary of the paragraph [7].

Dana Warmsley et al proposed a model "Automated Hate Speech Detection and the Problem of Offensive Language" which distinguishes hate speech and offensive language in twitter. Researchers used the dataset available in hatebase.org. Using twitter API, they extracted around 85 million tweets and 25k were taken as a sample randomly to classify them into hate speech, offensive language or neither. After comparing several algorithms, they decided to use logistic regression for final model. The final sample is trained using the complete dataset, and it is used to predict the label (hate speech, offensive language or neither) for each tweet [9].

Priyanka proposed a system "monitoring malicious discussions on online forums using data mining" to predict market volatility based on people's opinions in stock forums. This system uses data analysis technique and algorithms such as brute force algorithm, stop word selection, affix stemmers, stemming algorithm, emotional algorithm and matching algorithm. Using those above-mentioned techniques and algorithms the author built an early warning system that is used to recognize risk in market volatility [10][11].

The internet influencing people activities for a long time. The current system is to analyze textual comment from social media and categorize them as spam or not, usingsentimentalanalysis, stop word selection, stemmer algorithm, and lemmatization. The commonly used words ("the", "is"," has") are removed using stop word selection [12]. The suffixes of the words are removed and transformed into its base form using stemmer algorithm.Different forms of the same word are grouped together using lemmatization. Though the current systems work fine, there is still room for improvement. Existing system does not detect toxic users in real time. The existing system does not detect various types of threats and obscene language.

Human beings have innovated the ability of computersystems with their divergent overall fields, reducing size interms of time and their increasing swiftness. ArtificialIntelligence pursues building the computers or machines asbrainy as human beings [13]. The system uses RNN and CNNnetwork for action recognition either in the form of images orsignals. Combining CNN and RNN will enhance the ability torecognize different actions at varied time span [14]

Annals of R.S.C.B., ISSN: 1583-6258, Vol. 25, Issue 5, 2021, Pages. 4578 - 4586

Received 25 April 2021; Accepted 08 May 2021.

3. Working Model

The purpose of this work is to create a system based on machine learning that can detect toxic user comments in real time. It consists of four sub sections such as 3.1 The Dataset,3.2 Block Diagram, 3.3 Preprocessing of text and 3.4 Model Training.

3.1 The Dataset

The Wikipedia Dataset produced by Jigsaw is now publicly available on the Kaggle is used in this research work. The Dataset consists of 159571 comments and its respective multiple binomial labels such as threat, toxic, obscene, severe toxic, identity-based hatred and insults. Table 1 shows the sample instance of the dataset.

Id	Comment Text	Toxic	Severe Toxic	Obscene	Threat	Insult	Identity Hate
ffe897e7f7182c90	Hai.!!!	0	0	0	0	0	0
ffe8b9316245be30	Welcome home	0	0	0	0	0	0
ffe987279560d7ff	Well played	0	0	0	0	0	0
ffee36eab5c267c9	Good morning	0	0	0	0	0	0
ffee36eab5c267c9	Idiot	1	0	0	0	0	0
fff125370e4aaaf3	I will kill you	1	0	0	1	0	0

Table 1:ExampleDataset

3.2 Block diagram

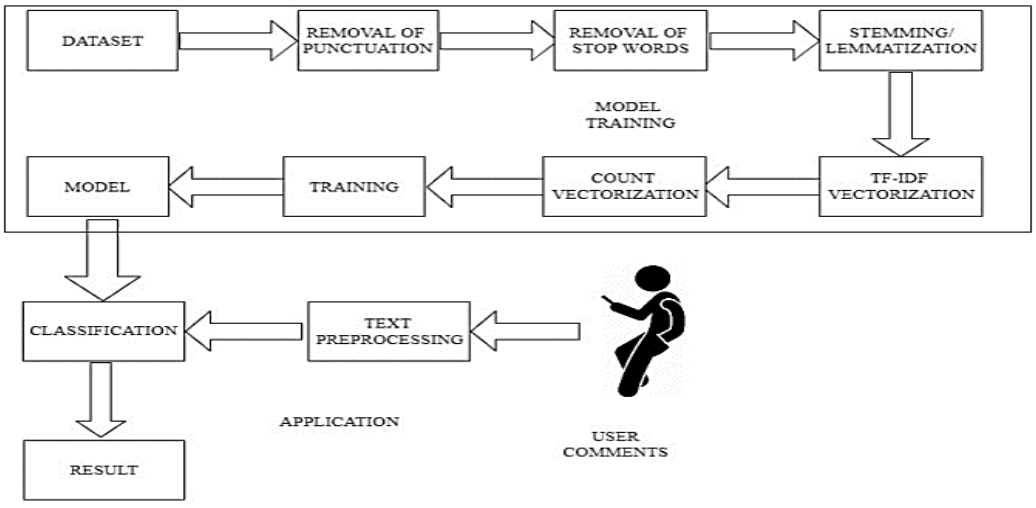

Figure 1: Block Diagram of live suspicious comments detection

Annals of R.S.C.B., ISSN: 1583-6258, Vol. 25, Issue 5, 2021, Pages. 4578 - 4586

Received 25 April 2021; Accepted 08 May 2021.

3.3 Preprocessing of text

The text preprocessing contains three stages.

- punctuation Removal – Punctuation will be removed for each text comment.
- Removal of stop words – Regularly using common words like articles and prepositions are removed using stop word selection.
- Stemming & lemmatization – Reducing words down to their base form by removing suffixes like “ed”, “ing”, “ly” is known as stemming. Different forms of the same word are grouped together using lemmatization.

3.4 Training the model

Vectorization allows computers to evaluate the meaning of words by mapping similar word meanings to similar vector spaces.

3.4.1 TF-IDF

- The expansion of TF-IDF isTerm Frequency and Inverse Document Frequency.
- TF-IDF vectorization commonly used for assigning unique vector number for each word in the dataset.
- By converting a word into a vector, machine learning process can be made faster.
- The term frequency is ratio of total number of times a word occurring in a comment divided by total number of words in the comment

$$tf(t,d) = \text{count of t in d} / \text{number of words in d} \quad \text{------(1)}$$

t-word, d-comment

- Inverse Document Frequency formula is

$$Idf = \log(N/n) \quad \text{------(2)}$$

N-Total number of comments,
N-The number of comments from which the word appeared

$$Tf(t,d)\text{-}Idf = \text{count of t in d} / \text{number of words in d} * \log(N/n) \quad \text{------(3)}$$

3.4.2 Count Vectorization

- Count Vectorizer is used to change a group of text or comments in a document to token counts. Prior to generating vector representation, pre-processing (refer 3.3) of text or comment is also achieved.
- TF-IDF assigns unique value to every word whereas Count Vectorizer counts occurrences of each word
- Table 2 shows how count vectorizer works for the given sentence below,“The lecture is at noon, please come to the lecture on time.”

Table 2: Example of count vectorizer

the	lectur	is	at	noon	pleas	come	to	on	time
2	2	1	1	1	1	1	1	1	1

Annals of R.S.C.B., ISSN: 1583-6258, Vol. 25, Issue 5, 2021, Pages. 4578 - 4586

Received 25 April 2021; Accepted 08 May 2021.

3.4.3 Model training with Logistic Regression algorithm

Logistic Regression (LR) is the appropriate regression analysis method to use when the dependent variable is binary (0 or 1).
To express values between 0 and 1, we use the sigmoid equation number 4

$$p(X) = \frac{exp(\beta_0+\beta_1 X)}{1+exp(\beta_0+\beta_1 X)} \qquad \text{------ (4)}$$

S shaped curve will be formed when we use the above equation (0 and 1)

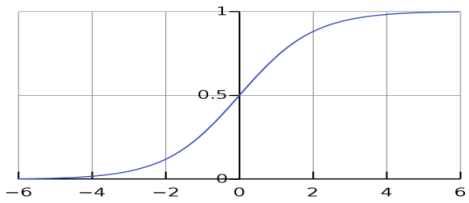

Figure 2 LRcurves

The above equation can be rewritten after some changes

$$\frac{p(X)}{1-p(X)} = exp(\beta_0 + \beta_1 X) \qquad \text{------ (5)}$$

Finally, the logit equation is formed by taking log on both sides,

$$log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X \qquad \text{------ (6)}$$

```
Accuracy of logistic regression classifier on toxic set: 0.96206
CV score for class toxic is 0.9701987371646276
Accuracy of logistic regression classifier on severe_toxic set: 0.99103
CV score for class severe_toxic is 0.985834345366384
Accuracy of logistic regression classifier on obscene set: 0.98003
CV score for class obscene is 0.9858708195981674
Accuracy of logistic regression classifier on threat set: 0.99729
CV score for class threat is 0.9823498652166861
Accuracy of logistic regression classifier on insult set: 0.97317
CV score for class insult is 0.9768841022318627
Accuracy of logistic regression classifier on identity_hate set: 0.99241
CV score for class identity_hate is 0.9750104309542947
```

Figure 3: Accuracy of each label

Finally, we train our models by providing each comment into our TF-IDF vectorizer, which turns that comment into a vector. Then we pass those TF-IDF vectors into our model. For clarity, we train six separate models independently, for each and every label (threat, toxic, obscene, severe toxic, identity based hatred and insults).

We implemented the model in real time to detect toxic comments in social media platform like YouTube. The application works by fetching all the live chat comments from users and then processes the text and classifies the text to list out the names of abusers based on toxicity level.

Annals of R.S.C.B., ISSN: 1583-6258, Vol. 25, Issue 5, 2021, Pages. 4578 - 4586

Received 25 April 2021; Accepted 08 May 2021.

4. Result and Illustration

Roc curve for the model accuracy is shown below,

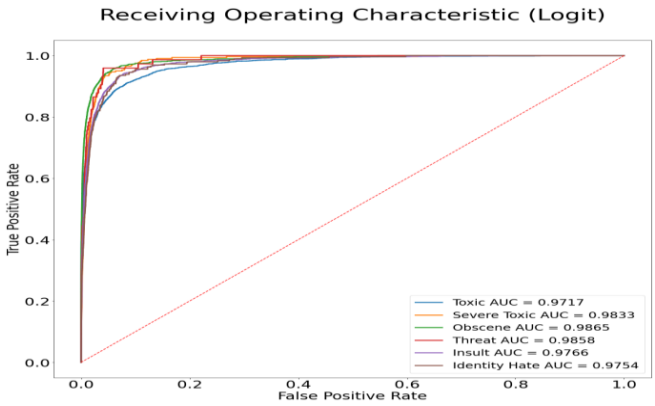

Figure 4: Receiving Operating Characteristic Curve

This work classifies the user comments into six types such as threat, toxic, obscene, severe toxic, identity-based hatred and insults. The output and bar chart representation of the output is shown below,

```
toxic 0.9415
severe_toxic 0.0856
obscene 0.1597
threat 0.9586
insult 0.1322
identity_hate 0.0372
```

Figure 5: Output of the comment "I will kill you"

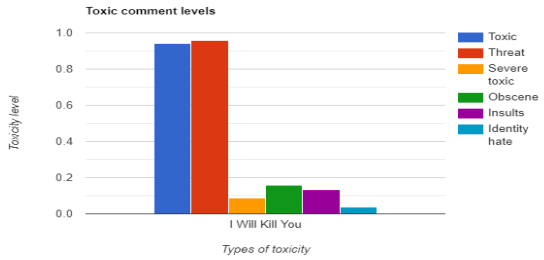

Figure 6: Bar chart representation of a comment 1

```
toxic 0.8974
severe_toxic 0.0054
obscene 0.0617
threat 0.0021
insult 0.3958
identity_hate 0.007
```

Figure 7: Output of the comment "You are an idiotic person"

Annals of R.S.C.B., ISSN: 1583-6258, Vol. 25, Issue 5, 2021, Pages. 4578 - 4586

Received 25 April 2021; Accepted 08 May 2021.

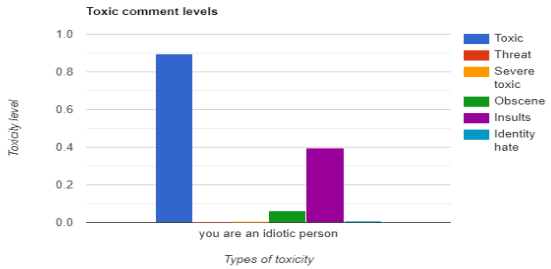

Figure 8: Bar chart representation of a comment 2

5. Conclusion

In this research work, we given a detailed explanation for real-time toxic comments detection on social media sites such as YouTube. Real-world applications such as semi-automatic comments moderation can take advantage of this works. We also focus on promoting fair online speech and sharing ideas on social media. Here we used logistic regression algorithm, pre-processing of text, TF-IDF vector for detection of toxic comments. The motivation of this research is to help people express their thoughts and ideas without hesitation and to keep users away from toxicity and negativity.

References

[1]Navoneel Chakrabarty, 2019, "A Machine Learning Approach to Comment Toxicity Classification", Advances in Intelligent Systems and Computing ,Computational Intelligence in Pattern Recognition, Volume. 999, Page number: 183-193

[2] Tanya Srivastava, R.Mangalagowri, Shailesh S.Dudala, "Monitoring of suspicious discussions on online forums using data mining", International Journal of Pure and Applied Mathematics, Volume 118 No. 22, Page number: 257-262

[3] Omar Sharif, Mohammed Moshiul Hoque, A.S. M. Kayes, Raza Nowrozy,2020, "Detecting Suspicious Texts Using Machine Learning Techniques",Applied Sciences, received: 5 August 2020; Accepted: 12 September 2020; Published: 18 September 2020, Volume. 10 No. 18, Page number: 6527, 2020

[4] Julian Risch, Ralf Krestel,2020, "Toxic Comment Detection in Online Discussions, Deep Learning-Based Approaches for Sentiment Analysis", Part of the Algorithms for Intelligent Systems book series (AIS), Page number. 85-109

[5] Megha K B, Nishan B, Nithin Thomas,2019, "Monitoring of Suspicious and Fraudulent Activities on Online Forums", International Journal of Engineering Research & Technology (IJERT), Volume. 7, Issue. 08, RTESIT - 2019 Conference Proceedings.

Annals of R.S.C.B., ISSN: 1583-6258, Vol. 25, Issue 5, 2021, Pages. 4578 - 4586

Received 25 April 2021; Accepted 08 May 2021.

[6]BavaneA.B.,Ambilwade Priyanka V., BachhavMourvika D., Dafal Sumit N., Fulari Priyanka Y,2017,"Monitoring suspicious discussions on online forum by data mining", International Journal of Advanced Engineering & Science Research (IJAES) Volume.5, Issue. 01

[7] Harshala Patil, Kalpana Khandelwal, Trupti Kini, 2019, "Monitoring Online Forum for Suspicious Discussions with Text Summarization", IJSRD - International Journal for Scientific Research & Development| Volume. 7, Issue. 02

[8] Spiros V. Georgakopoulos, Sotiris K. Tasoulis, Aristidis G. Vrahatis, Vassilis P. Plagianakos, 2018, "Convolutional Neural Networks for Toxic Comment Classification",Proceedings of the 10th Hellenic Conference on Artificial Intelligence July 2018 Article No. 35 Page number. 1–6

[9] Thomas Davidson, Dana Warmsley, Michael Macy, Ingmar Weber, 2017, "Automated Hate Speech Detection and the Problem of Offensive Language", Eleventh International AAAI Conference on Web and Social Media, Volume. 11 No. 1

[10] Priyanka B.Hulde,Prof. Priyanka Dhudhe ,2018, "Monitoring malicious discussions on online forums using data mining", International Conference on Emanations in Modern Engineering Science & Management (ICEMESM-2018)

[11] S.Gnanavel et al "HD video transmission on UWB networks using H.265 encoder and ANFIS rate controller" ,Cluster Computing the Journal of Networks, Software Tools and ApplicationsMarch 2018, Volume 21, Issue 1, pp 251–263.

[12] S.Gnanavel et al "Wireless video transmission over UWB channel using fuzzy based rate control technique",Journal of Theoretical and Applied Information Technology 28th February 2014. Volume. 60 No.3 page No: 491 to 503

[13] N.Duraimurugan et al, "Learnart: Drawing environment using convolutional neural networks" International Journal of Recent Technology and Engineering, 2019, 8(2 Special Issue 3), pp. 770–772

[14] J.Rajalakshmi, N.Duraimurugan, S.P.Chokkalingam, "Action recognition for controlling electronic appliances" International Journal of Engineering and Advanced Technology, 2019, 8(3 Special Issue), pp. 565–567