International Journal of Scientific Research in Computer Science, Engineering and Information Technology
ISSN : 2456-3307 (www.ijsrcseit.com)
doi : https://doi.org/10.32628/CSEIT2173123

Toxic Word Analyzer

Dhairya Timbadia
Computer Engineering, Rajiv Gandhi Institute of Technology, Mumbai, Maharashtra, India

Article Info
Volume 7, Issue 3
Page Number: 578-581

Publication Issue :
May-June-2021

Article History
Accepted : 10 June 2021
Published : 15 June 2021

ABSTRACT

In this generation social media has been a huge part of our lives and there is no need to say that the current generation spend a huge amount of time on their social media accounts. Apart from there being a good social media influencer there are a lot of people who spread hatred among these influencers as well as among each other. I have tried to make a speedometer which would be able to tell the toxicity of the words that are basically used in the input sentence or paragraph. The main processing that would be done on the sentence or the paragraph would be removing punctuation marks, tokenization on the words, 'Stop' word removal, bigram creation, matching tokens with predefined dictionary, generating toxicity percent using scaling.

Keywords : Toxicity, Tokenization, Speedometer

I. INTRODUCTION

Social Media have made communication easier and more convenient than ever. This era has seen all kinds of personalities, introverts, extroverts etc. but social media is a platform where any of them are not discriminated.

Social Media open up freely and communicate, put forth their thoughts and opinions on various topics by means of either sharing comments, images, videos, podcasts, etc. With such a diverse pool of opportunities available, issues like cyber bullying, abusive content, harassment, threats arise, which creates a negative environment and disturbs the audience consuming the content. No platform would want to seize the freedom of speech of their active users, but every problem has a solution and we can always use preventive measures to deal with such situations. When the Conversation AI team first built toxicity models, they found that the models incorrectly learned to associate the names of frequently attacked identities with toxicity.

Models predicted a high likelihood of toxicity for comments containing those identities (eg:- "gay"), even when those comments were not actually toxic (such as "I am a gay woman"). This happens because training data was pulled from available sources where unfortunately, certain identities are overwhelmingly referred to in offensive ways. In the comment classification approach, the goal is to classify the comments or sentences based on their toxicity levels into various categories.

By categorizing these comments, the action team can take appropriate actions to curb the occurrence and growth of negative influences created with such

activities on social platforms. The aim of this multi-label classification model is to make expressing thoughts more effective and positive on social media. To reduce the time and effort of these social platforms, the companies can use our system.

II. PROPOSED SYSTEM

A. Analysis and Framework

We have used four python libraries which are nltk for tokenization, pickle for storing bag of words, os and sys for handling system path, directions and arguments and regular expression for data pre-processing and a file words.txt file which consists of words in English Language, Hindi Language and also regular expressions in Natural Language Processing along with its search method. The libraries are used to implement a toxic word checker. It produces the output as original text and checks the toxicity of the word. Whereas, the words.txt file which consists of words in English Language and Hindi Language are used as to implement a checker. The output is the word along with whether the sentence contains toxic words or not.

III. PROBLEM STATEMENT AND OBJECTIVES

To implement a toxicity speedometer and specifier system that will check the and will notify about the words that the user enters in the text field in English, Hindi Language.

A. Objectives

1) Creating a system that can help to identify the informal language consisting of words and expressions that are not considered appropriate for formal occasions using word embedding.

2) Providing Parental control for the underage users.

Help to take strict action against users violating the law.

3) Eradicating social media bulling.

IV. LITERATURE REVIEW

A. Survey existing systems

Mukul Anand, Dr. R. Eswari, Classification of Abusive Comments in Social Media using Deep Learning

In this paper, Kaggle's toxic comment dataset is used to train deep learning model and classifying the comments in following categories: toxic, severe toxic, obscene, threat, insult, and identity hate. The dataset is trained with various deep learning techniques and analyze which deep learning model is better in the comment classification. The deep learning techniques such as long short term memory cell (LSTM) with and without word GloVe embeddings, a Convolution neural network (CNN) with or without GloVe are used, and GloVe pretrained model is used for classification.

Yoon Kim, Convolutional Neural Networks for Sentence Classification.

In this report on a series of experiments with convolutional neural networks (CNN) trained on top of pre-trained word vectors for sentence-level classification tasks. It shows that a simple CNN with little hyperparameter tuning and static vectors achieves excellent results on multiple benchmarks. Learning task-specific vectors through fine-tuning offers further gains in performance. It additionally proposes a simple modification to the architecture to allow for the use of both task-specific and static vectors. The CNN models discussed herein improve upon the state of the art on 4 out of 7 tasks, which include sentiment analysis and question classification.

Spiros V. Georgakopoulos, Sotiris K. Tasoulis, Aristidis G. Vrahatis, Vassilis P. Plagianakos, Convolutional Neural Networks for Toxic Comment Classification.

In this work, CNN approach is employed to discover toxic comments in a large pool of documents provided by a current Kaggle's competition regarding Wikipedia's talk page edits. To justify this decision CNNs is compared against the traditional bag-of-words approach for text analysis combined with a selection of algorithms proven to be very effective in text classification. The reported results provide enough evidence that CNN enhance toxic comment classification reinforcing research interest towards this direction.

V. DESIGN DETAILS

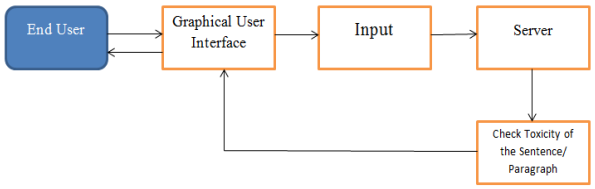

Fig 1. Design

There are basically five blocks namely end user, graphical user interface (GUI), input which is the data that we basically get from the front end of the system, the server and finally the checking of the toxicity of the phrases or the paragraph takes place.

The system works in a fashion where the end user sends data to the GUI from where the data is sent to the server where the actual checking of the toxicity takes place and the output of which is then send back to the GUI where the user can actually view the final results.

VI. RESULTS AND DISCUSSION

Input:

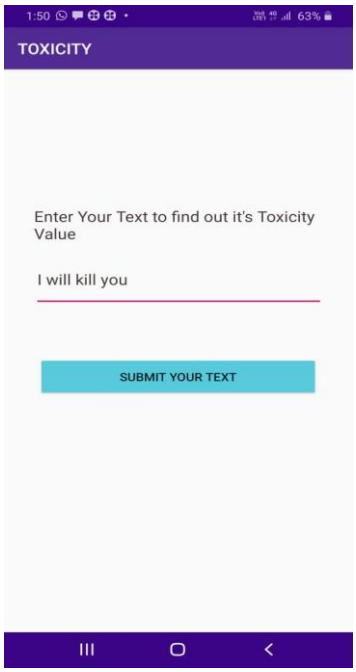

Fig 2 :-Input for medium toxicity

Output

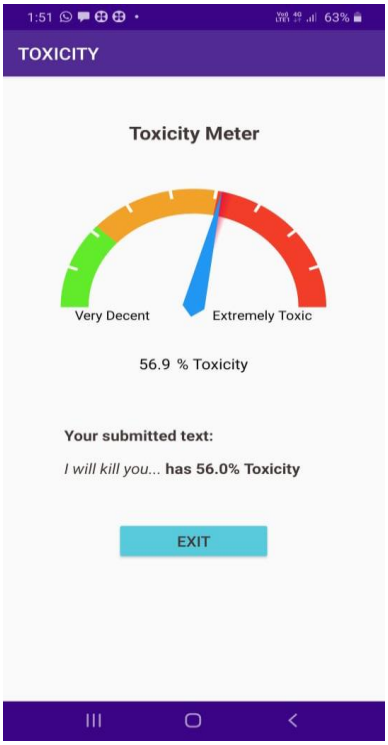

Fig 3 :- Output for medium toxicity

VII. CONCLUSION

The proposed system was developed taking in mind the benefits of the social platform users and organization. In this work we presented a framework for identifying the toxicity of the comments written on social conclaves and we intent on reducing the unintended bias while classification of toxic text. With respect to many attributes and identities. Furthermore, we focus on comments shared on applications like online forums, social media platforms. This comment contains words and expressions which ranges from positive impact to negative and toxic attacks. Our project focuses on to categorize such toxic sentences and identify such practices online so as to reduce its impact and monitor such acts.

VIII. REFERENCES

[1]. Thedora Chu, Max Wang, Kylie Jue. “Comment Abuse Classification with Deep Leraning.” Stanford University.

[2]. Karthik Diankar, Roi Riechart, Henry Lieberman. “Modeling the Detection of Textual Cyberbullying.” Massachusetts Institute of Technology, Cambridge MA 02139 USA.

[3]. Xin Wang, Yuanchao Li, Chengjie Sun, Baoxum Wang and Xialong Wang. “Polarities of Tweets by Composing Word Embeddings with Long Short Term Memory.” 7th International Joint Conference of Natural Language Processing. July-2005.

[4]. S. V. Georgakopoulus, A. G. Vrahatis, S. K. Tasoulis, V. P. Plagianakos. “Convolutional Neural Networks for Toxic Comment Classification.” arXiv:1802.099574v1 cs.CL], 27 Feb 2018.

[5]. Kevin Khieu, Neha Narwal. “Detecting and Classifying Toxic Comments.” Stanford University- CSS224N.

[6]. C. Nobata, J.Tetreault,A. Thomas, Y. Mehdad and Y.Chang. “Abusive language detection in online user content.”

Cite this article as :

Dhairya Timbadia, "Toxic Word Analyzer", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 7 Issue 3, pp. 578-581, May-June 2021. Available at doi : https://doi.org/10.32628/CSEIT2173123
Journal URL : https://ijsrcseit.com/CSEIT2173123