USER CONTENT – ANALYSIS AND CLASSIFICATION

[1]Dr. Renuka Deshpande, Associate Professor, [2]Aman Yadav, [3]Jagdish Chaudhary, [4]Amit Vhatkar
[1]Assistant Professor, [2]Student, [3]Student, [4]Student,
[1]Department of Computer Engineering,
[1]Shivajirao S. Jondhale College of Engineering, Dombivali, India.

Abstract: Internet negativity has always been a hot topic. Moderators of online discussion forums often struggle with controlling extremist comments on their platforms.

We also assess the class imbalance issues associated with the dataset by employing sampling techniques and loss. Models we applied yield high overall accuracy with relatively low cost.

In order to reduce the negative impact of toxic comment in day-to-day life we have attempted to design a toxic comment detector.

Index Terms **-** Toxic Comment Detection, Language, Intelligent System, Analysis, Content, Social Media.

I. INTRODUCTION

Content analysis is a research method used to identify patterns in recorded communication. To conduct content analysis, you systematically collect data from a set of texts, which can be written, oral, or visual. People have been seeking help from various tools to analyse text-based information so that they can identify toxic expressions from a sea of information, both efficiently, and more importantly, accurately**.**

As we all know in this developing world of social media there, are various impact on human life, of which some are positive and some are negative.

Negative comments can cause various issues such as depression, anxiety, panic attack etc. In order to reduce such things, we are trying to develop software which can detect abusive words, depressing and negative comments using computer as our asset

II. LITERATURE REVIEW

2.1 UCLA

UCLA Electronic Theses and Dissertations

Application of Recurrent Neural Networks in Toxic Comment Classification (Li, Sycuan) 2018 Internet negativity has always been a hot topic. The anonymity and the sense of distance of people's internet presence have encouraged people to express themselves freely. This freedom can sometimes lead to extreme outtakes on others people or the particular topics.

Extreme negativities has sometimes stopped people from expressing themselves or made them give up looking for different opinions online Issues like this happen almost all the time, across all platforms of discussion, and the modulators of these platforms have limited capabilities dealing with it. Needless to say the time, energy and effort these modulators have to put into controlling this negativity on their platform, people have been seeking help from various tools to analyze text-based information so that they can identify toxic expressions from a sea of information, efficiently, and more importantly, accurately, In this thesis, we will be applying word embedding techniques and recurrent neural network to perform text classification on a multi-label text dataset to identify different forms of internet toxicity. Natural language processing with deep Neural Networks is one of the most influential tools that enable researchers to extract, analyze, and classify essential features from text-based information.

2.2 Random Forest Algorithm Report

Random forests are built by combining the predictions of several trees, each of which is trained in isolation.

There are three main choices to be made when constructing a random tree. These are the method for splitting the leaf's, The type of predictor to use in each leaf, and the method for injecting randomness into the trees. Specifying a method for splitting leaf's requires selecting the shapes of candidate splits as well as a method for evaluating the quality of each candidate. In order to split a leaf, a collection of candidate splits are generated and a criterion is evaluated to choose between them. The most common for predictors in each leaf is to use the average response over the training points which fall in that leaf. Different leaf predictors for regression and other tasks, but these generalizations are beyond the scope of this paper. We consider only simple averaging predictors here.

2.3 Toxic Comment Detector Report

To identify and classify toxic online commentary, the modern tools of data science transform raw text into key features from which either learning algorithms can make predictions for monitoring offensive conversations. We systematically evaluate 62 classifiers representing 19 major algorithmic families against features extracted from the Jigsaw dataset of Wikipedia comments. We compare the classifiers based on statistically significant differences in accuracy and relative execution time. Among these classifiers for identifying toxic comments, tree-based algorithms provide the most transparently explainable rules and rank order the predictive contribution of each feature. Among 28 features of syntax, sentiment, emotion and outlier word dictionaries, a simple bad word list proves most predictive of offensive commentary.

Comment	Toxic Rating	
We suck at dealing with abuse and trolls on the platform and we've sucked at it for years	0.77	Toxic
We suck at dealing with abuse and trolls on the platform but we'll get better at it.	0.74	Toxic
We don't suck at dealing with abuse and trolls on the platform and we've never sucked at it	0.64	Unsure
We're not good at dealing with abuse and trolls on the platform and we've sucked at it for years	0.61	Unsure
We're not good at dealing with abuse and trolls on the platform but we'll get better at it.	0.35	Unlikely
We're not good at dealing with fame and fortune on the platform but we'll get better at it.	0.06	Unlikely

FROM THE ABOVE RESEARCH PAPER, WE HAVE BEEN INSPIRED TO CATEGORIZE THE COMMENT INTO DIFFERENT PARAMETERS SUCH AS INSULT, TOXICITY, ETC.

III. PROBLEM DEFINATION

The data we will be focusing on is a public dataset provided by the Conversation AI team; a research initiate co-founded by Jigsaw and Google. Jigsaw is a technology incubator created by Google, with the primary objective to "use technology to tackle the toughest geopolitical challenges, from countering violent extremism to thwarting online censorship to mitigating the threats associated with digital attacks. Jigsaw and Google launched Perspective API in February 2017, a free tool that utilizes machine learning to identify toxic comments. To improve the performance of the Perspective API, and with the belief that "collaborative problem-solving yields the best solutions the Conversation AI team hosted a "Wikipedia Talk Page Comments annotated with toxicity reasons" Kaggle competition. We will be building our deep learning classification model and monitor its performance base on the dataset provided in this competition. Currently, the model used by Conversation API performs quite well, able to provide a relatively accurate toxic score given text comments. However, the team mentioned that their model still makes errors it is unable to classify toxicity if the model has not seen the pattern before, and it may miss-classify texts that share similar patterns as toxic comments. Recurrent neural networks' ability to process sequences of documents and analyze contexts may prove useful in resolving the problems Conversation API currently encounters.

IV. OUR APPROACH

4.1 What Is NLP?

LP uses perceptual, behavioral and communication techniques to make it easier for people to change their thoughts and actions NLP relies on language processing but should not be confused with natural language processing, which shares the same acronym. For example, a central feature of NLP is the idea that a person is biased towards one sensory system, known as the preferred representational system or PRS. Therapists can detect this preference through language Phrases such as "I see your point may signal a visual PRS. Or "I hear your point" may signal an auditory PRS. An NLP practitioner will identify a person's PRS and base their therapeutic framework around it. The framework could involve rapport-building. Information-gathering, and goal-setting with them.

Figure: NLP Training

In this thesis, we will be applying word embedding techniques and recurrent neural network to perform text classification on a multi-label text dataset to identify different forms of internet toxicity.

4.2 Random Forest algorithm

Random forest is a supervised learning algorithm which is used for both classification as well as regression. But however, it is mainly used for classification problems. As we know that a forest is made up of trees and more trees means more robust forest. Similarly, random forest algorithm creates decision trees on data samples and then gets the prediction from each of them and finally selects the best solution by means of voting. It is an ensemble method which is better than a single decision tree because it reduces the over-fitting by averaging the result.

Working of Random Forest Algorithm

We can understand the working of Random Forest algorithm with the help of following steps-

Step 1 - First, start with the selection of random samples from a given dataset.

Step 2 - Next, this algorithm will construct a decision tree for every sample. Then it will get the prediction result from every decision tree.

Step 3 - In this step, voting will be performed for every predicted result.

Step 4 - At last, select the most voted prediction result as the final prediction result.

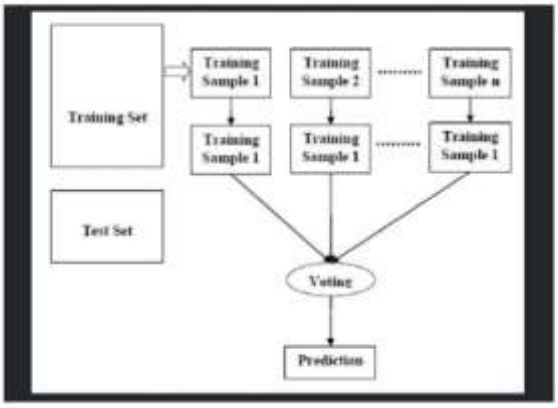


Figure: Working of Random Forest Algorithm

These labels were summarized into a total of six classes:

Toxic: A general classification of the toxicity of the comments.

Severe Toxic: Extreme toxicity.

Obscene : Indecent language.

Threat : Statements with the intention to perform hostile action.

Insult : Disrespect or verbal abuse.

Identity Hate: Sexism, racism, homophobic, etc.

4.3 Some of the important concepts used are:

Counter vectorization

The Count Vectorizer provides a simple way to both tokenize a collection of text documents and build a vocabulary of known words, but also to encode new documents using that vocabulary.

Tdif vectorization

It can be used as input to estimator. Vocabulary, Is a dictionary that converts each token (word) to feature index in the matrix, each unique token gets a feature index.

Pickles

"Pickling" is the process whereby a Python object hierarchy is converted into a byte stream, and "unpicking" is the inverse operation, whereby a byte stream is converted back into an object hierarchy.

Tokenization

In Python tokenization basically refers to splitting up a larger body of text into smaller lines, words or even creating words for a non-English language.

4.4 Basic Data Flow Diagram

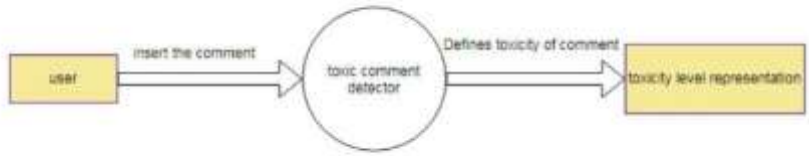


Figure: Basic Data Flow Diagram

V. REQUIREMENT ANALYSIS

Hardware Requirement

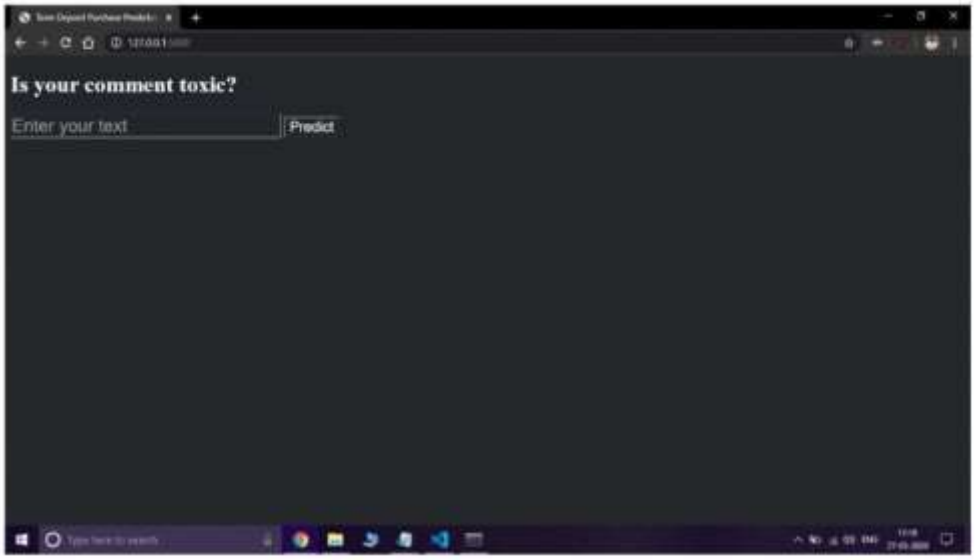
- I3 Processor
- Minimum 4 GB RAM
- Internet Connection

Software Requirement

- Windows 7 or above Operating System
- Jupyter
- Anaconda
- Flask
- Html
- Python

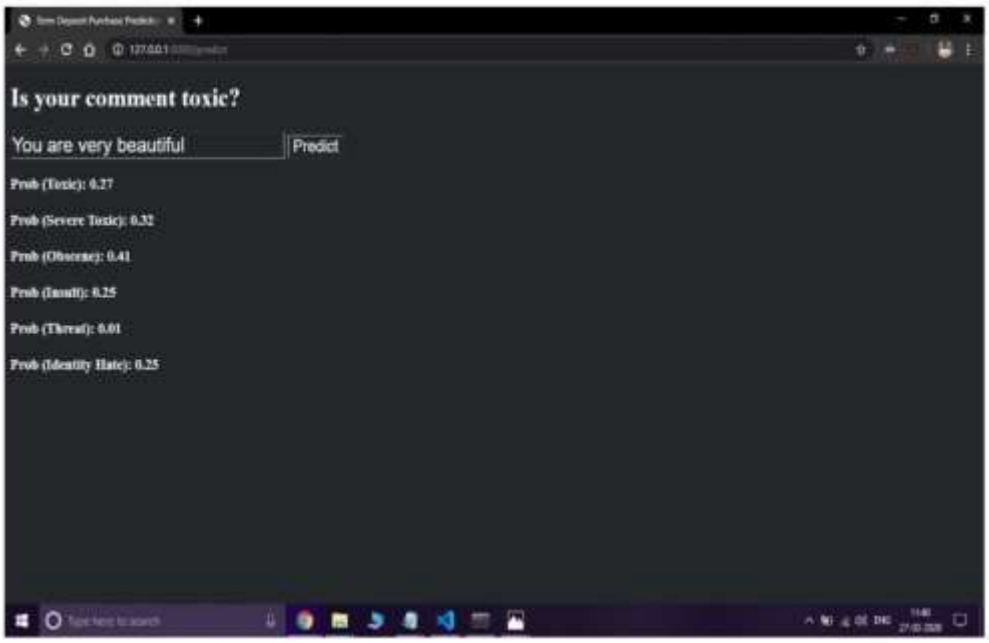
VI. RESULTS

1.



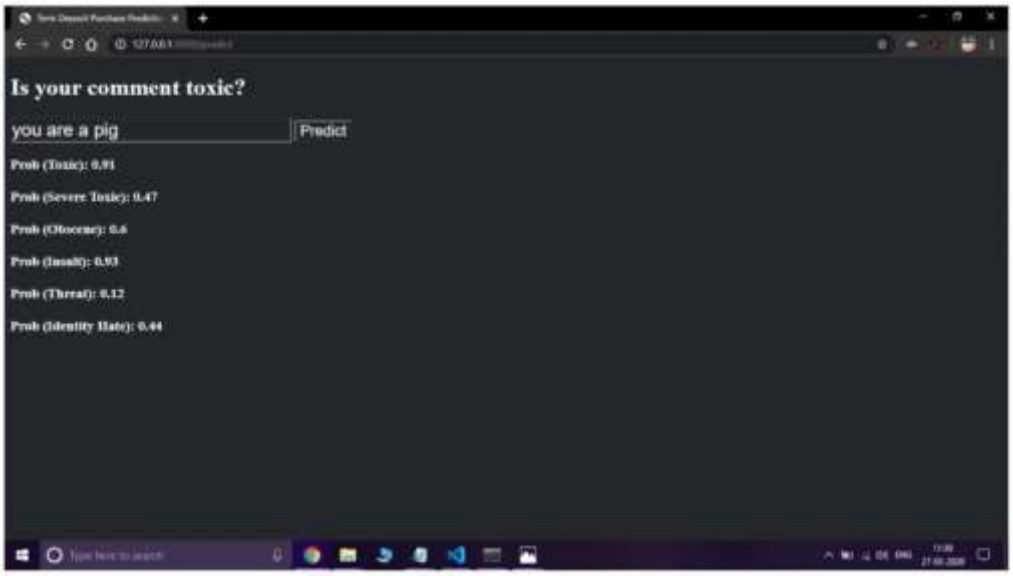
Output: This is a front end for toxic comment detector

2.



It includes all positive words in sentence. Hence all the categories in the output are minimal.

3.



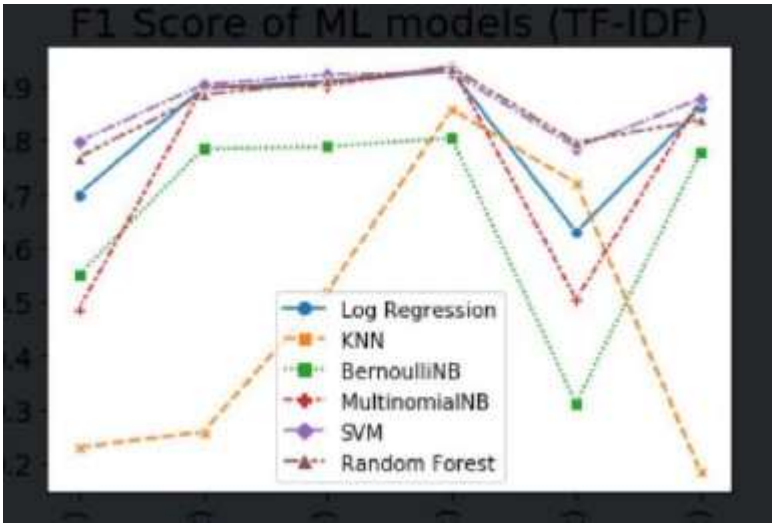
In this case word 'pig' is categorized in insult section. Hence the software has displayed the probability of insult parameter as maximum as compared to other parameters.

4.

These are some other examples of toxicity or insult.

VII. CONCLUSION

Toxic comment detector can further be used for detecting comments comment on applications which are similar to WhatsApp, Facebook, Instagram, etc. As compared to other various algorithms random forest algorithm has been proved considerably effective. It can been observed through the following graph



These observations also mean that further improvements can be made to improve our current model, which does not limit to a different overall network structure. We can also perform additional hyperparameter tuning on our model, which will most definitely prove beneficial. Nevertheless, the model's ability to process context of words proves efficient in identifying toxic texts from a large sample.

VIII. ACKNOWLEDGEMENT

We sincerely wish to thank the project guide Dr. Renuka Deshpande for her encouraging and inspiring guidance helped us to make our project a success. Our project guide makes us endure with her expert guidance, kind advice and timely motivation which helped us to determine our project.

We would like to thank our project coordinator Dr. Uttara Gogate for all the support we needed from her for our project.

We also express our deepest thanks to our HOD. Prof. P. R. Rodge whose benevolent helps us making available the computer facilities to us for our project in our laboratory and making it true success. Without his kind and keen co-operation our project would have been stifled to standstill.

Lastly, we would like to thank our college principal. Dr. J. W Bakal for providing lab facilities and permitting to go on with our project. We would also like to thank our colleagues who helped us directly or indirectly during our project.

IX. REFERENCES

[1] Prof. Ranjit Mane1, Sagar Chavan2, Trushali Birambole3, Asmita kamble4, "Fingerprint Based ATM System", International Journal for Research Trends and Innovation (**IJRTI**) , ISSN:2456-3315, Volume 4, Issue 4, 2017.

[2] SaimaRafat Bhandari1, ZarinaBegum K Mundargi2, "A Review on Securing ATM System Using Fingerprint", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (**IJSRCSEIT**), ISSN:2456-3307, Volume 3, Issue 2, 2018.

[3] Christiawan1, Bayu Aji Sahar2, Azel Fayyad Rahardian3, Elvayandri Muchtar4, "Fingershield ATM – ATM Security System Using Fingerprint Authentication". Issue 4.

[4] Vijayraj A, "A Survey on Cardless Cash Access Using Biometric ATM Security System", Scholars Journal of Engineering and Technology(SJET), ISSN 2347-9523, Issue 2.

[5] Moses Okechukwu Onyesolu, Ignatius M. Ezeani, "ATM Security Using Fingerprint Biometric Identifier : An Investigative Study", International Journal of Advanced Computer Science and Applications(IJACSA), 030412, 2012.

[6] Samayita Bhattacharya and Kalyani Mali," Fingerprint Recognition Using Minutiae Extraction Method",(International World Wide Web Conference Committee) © 2011 IW3C2, published under Creative Commons CC by 4.0 January, 2011.