International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

CLASSIFICATION OF ONLINE TOXIC COMMENTS USING MACHINE LEARNINGALGORITHMS

N.Sathvika Reddy[1], T.Prametha Reddy[2], Mrs. D Archana[3]
Department of Computer Science and Engineering, Stanley College of Engineering and Technology for Women, Telangana, India

Abstract.

Toxic comments are disrespectful, abusive, or unreasonable online comments that usually make other users leave a discussion. The danger of online bullying and harassment affects the free flow of thoughts by restricting the dissenting opinions of people. Sites struggle to promote discussions effectively, leading many communities to limit or close down user comments altogether. This paper will systematically examine the extent of online harassment and classifythe content into labels to examine the toxicity as correctly as possible. Here, we will use two machine learning algorithms and apply them to our data to solve the problem of text classification and to identify the best machine learning algorithm based on our evaluation metrics for toxic comments classification. We will aim at examining the toxicity with high accuracy to limit down its adverse effects which will be an incentive for organizations to take the necessary steps

1. Introduction

1.1 About Project

The exponential development of computer science and technology provides us with one of the greatest innovations of the "Internet" of the 21st century, where one person can communicate to another worldwide with the help of a mere smart phone and internet. In the initial days of the internet, people used to communicate with each other through Email only and it was filled with spam emails. In those days, it was a big task to classify the emails as positive or negative i.e. spam or notspam. As time flows, communication, and flow of data over the internet got changed drastically, especially after the appearance of social media sites. With the advancement of social media, it becomes highly important to classify the content into positive and negative terms, to prevent any form of harm to society and to control antisocial behavior of people. In recent times there have many instances where authorities arrest people due to their harmful and toxic social media contents. For example, one 28-year-old man was arrested in Bengal for posting an abusive comment against Mamata Banerjee on Facebook and one man from Indonesia was arrested for insulting the police of Indonesia on Facebook. Thus, there is an alarming situation and it is the need of the hour to detect such content before they got published because these negative contents are creating the internet an unsafe place and affecting people adversely. Suppose there is a comment on social media “Nonsense? Kiss off, geek. What I said is true”, it can be easily identified thatthe words like Nonsense and Kiss off are negative and thus this comment is toxic. But to mine the toxicity technically this comment needs to go through a particular procedure and then classification technique will be applied on it to verify the precision of the obtained result. Different machine learning algorithms will be used in the classification of toxic commentson the Data set of Kaggle.com. This paper includes two machine learning techniques i.e. logistic regression,SVMclassifier to solve the problem of text classification. So, we will apply all the two machine learning algorithms on the given data set and calculate and compare their accuracy, log loss and hammingloss.

1.2 Objectives of the Project

The main purpose of this classification is to systematically examine the extent of online harassment and classify the content into labels to examine the toxicity as correctly as possible.

We will aim at examining the toxicity with high accuracy to limit down its adverse effects which will be an incentive for organizations to take the necessary steps

1.3 Scope of the Project

Using Algorithms like SVM classifier and Logistic Regression we are classifying data and calculating the amount of toxicity present in comments and evaluating them through some metrices like Accuracy, Precision and hamming loss . Detecting Toxic comments has been a great challenge for the all the scholars in the field of research and development. This domain has drawn lot of interests not just because of the spread of hate but also people refraining people from participating in online forums which diversely affects for all the creators/content-providers to provide a relief to engage in a healthy public interaction which can be accessed by public without any hesitation.

2. Literature Survey

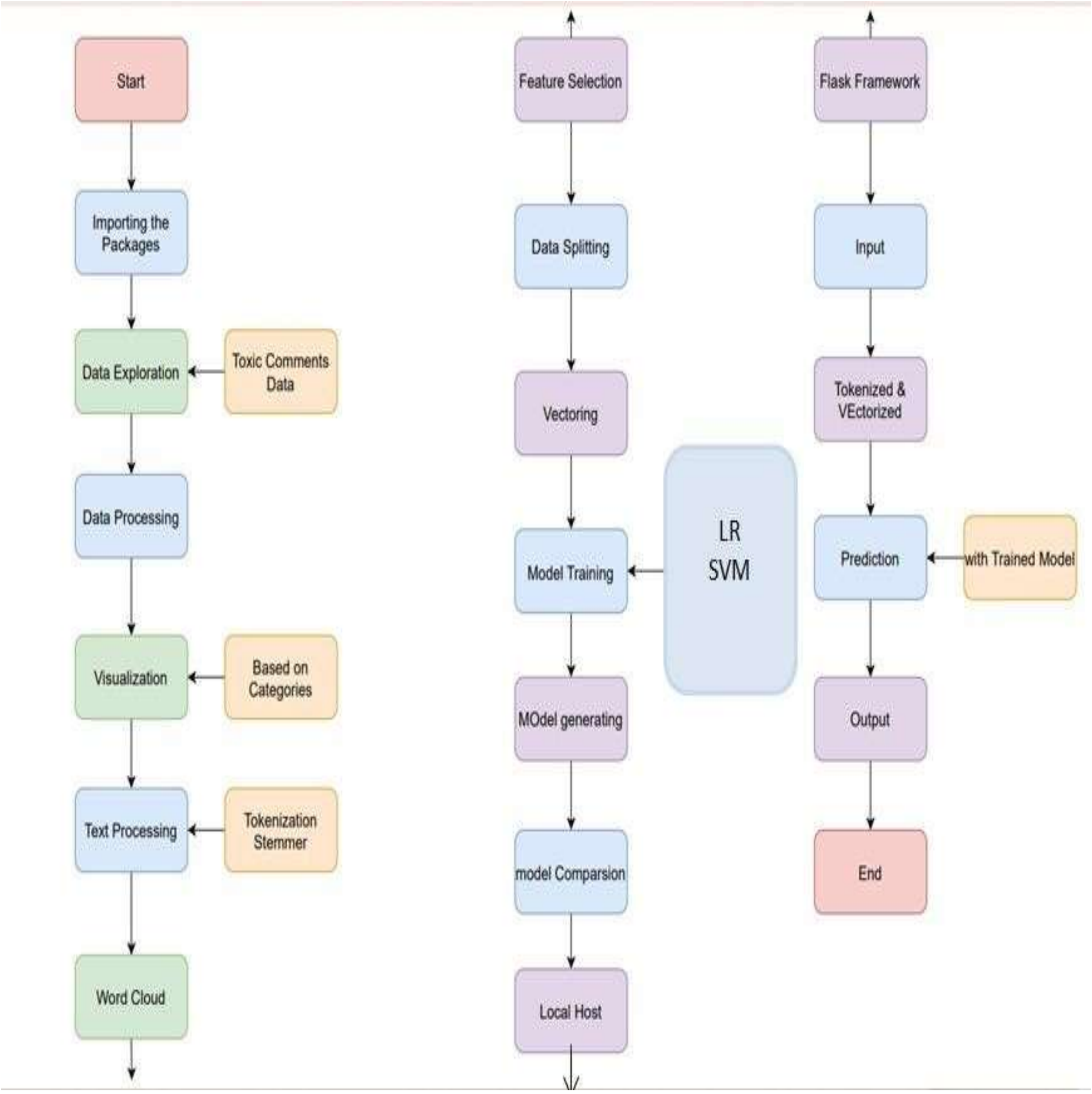
2.1 Existing System

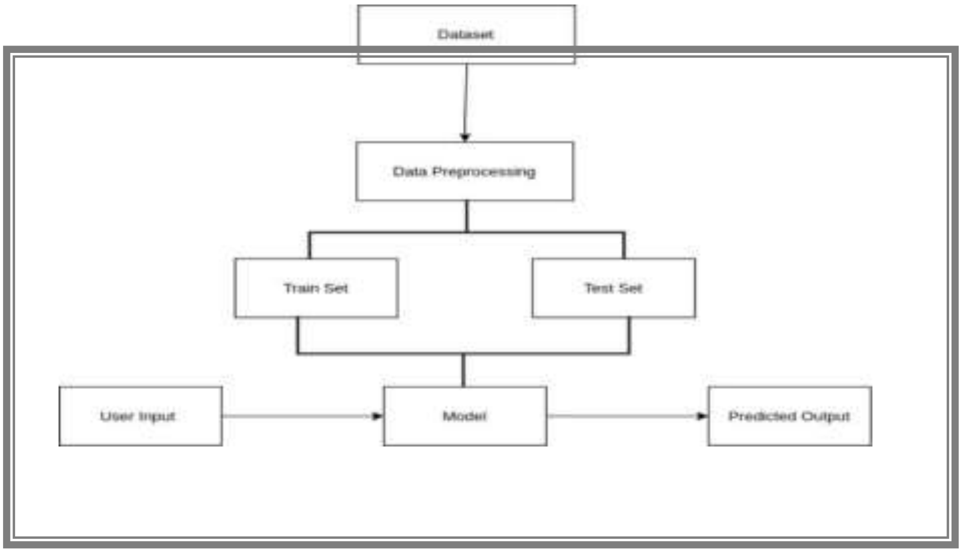
There were efforts in the past to increase the online safety by site moderation through crowd-sourcing schemes and comment denouncing, in most cases these techniques fail to detect the toxicity.

2.2Proposed System

We have to classify the data into six categories i.e. threat, insult, toxic, severe toxic, obscene, or identity hate and we can put one data value into zero, one or more than one category. Before the start of any processing on our data, our first task will be to identify whether our classification is multiclass or multi-label in nature. In multi-label classification, one data value can belong to more than one category, a given sketch of a garden may contain a tree, monument, walking path, or a combination of these and thus sketch can belong to zero, one or more than one categories. While in multi-class classification, one data value can belong to only one category, a given car can belong to Honda, Hyundai, Tata Motors , or none of the above companies and thus belongs to either1 category or of none of them.

3. Proposed Architecture





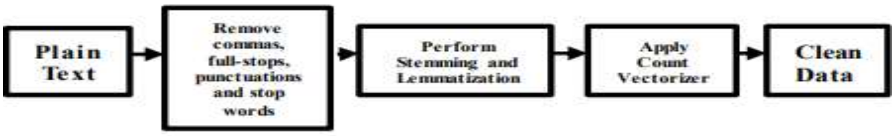

Plain Text
Remove commas, full-stops, punctuations and stop words
Perform Stemming and Lemmatization
Apply Count Vectorizer
Clean Data

Fig 1: Pre-processing steps for data cleaning.

Fig.1. Proposed Architecture

4. Implementation

4.1 Algorithm

SVM

Support-vector learning models machines (SVMs, with associated also support-vector learning algorithms networks) that are supervised analyse data for classification and regression_analysis.

SVM training algorithm builds a model that assigns new examples to one category or the other, making it a non-probabilistic binary linear classifier (although methods such as Platt scaling exist to use SVM in a probabilistic classification setting). An SVM maps training examples to points in space so as to maximise the width of the gap between the two categories.

Step-1: Load the important libraries

Step-2: Import dataset

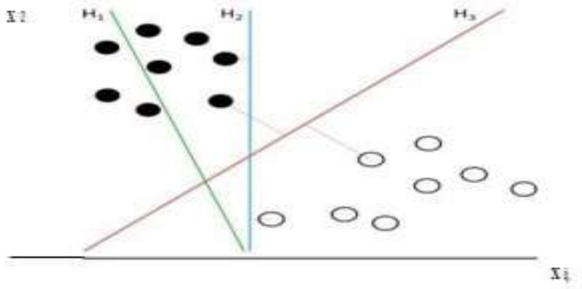
Step-3: Divide the dataset into train and test

Step-4: Initializing the SVM classifier model

Step-5: Fitting the SVM classifier model

Step-6: Coming up with predictions

Step-7: Evaluating model’s performance



LOGISTIC REGRESSION

Logistic regression is a statistical model that m its basic form uses a logistic function to model a binary dependent variable, although many more complex extensions exist. In regression analysis, logistic regression (or logit regression) is estimating the parameters of a logistic model (a form of binary regression).

Step-1: Data Pre-processing step

Step-2: Fitting Logistic Regression to the Training set

Step-3: Predicting the test result

Step-4: Test accuracy of the result(Creation of Confusion matrix)

Step-5: Visualizing the test set result

4.2 Code Implementation

Anaconda Navigator Anaconda Navigator is a desktop graphical user interface (GUI) included in Anaconda® distribution that allows you to launch applications and easily manage conda packages, environments, and channels without using command-line commands. Navigator can search for packages on Anaconda.org or in a local Anaconda Repository. It is available for Windows, macOS, and Linux. In order to run, many scientific packages depend on specific versions of other packages. Data scientists often use multiple versions of many packages and use multiple environments to separate these different versions. The command-line program conda is both a package manager and an environment manager. This helps data scientists ensure that each version of each package has all the dependencies it requires and works correctly. Navigator is an easy, point-and-click way to work with packages and environments without needing to type conda commands in a terminal window. You can use it to find the packages you want, install them in an environment, run the packages, and update them – all inside Navigator.

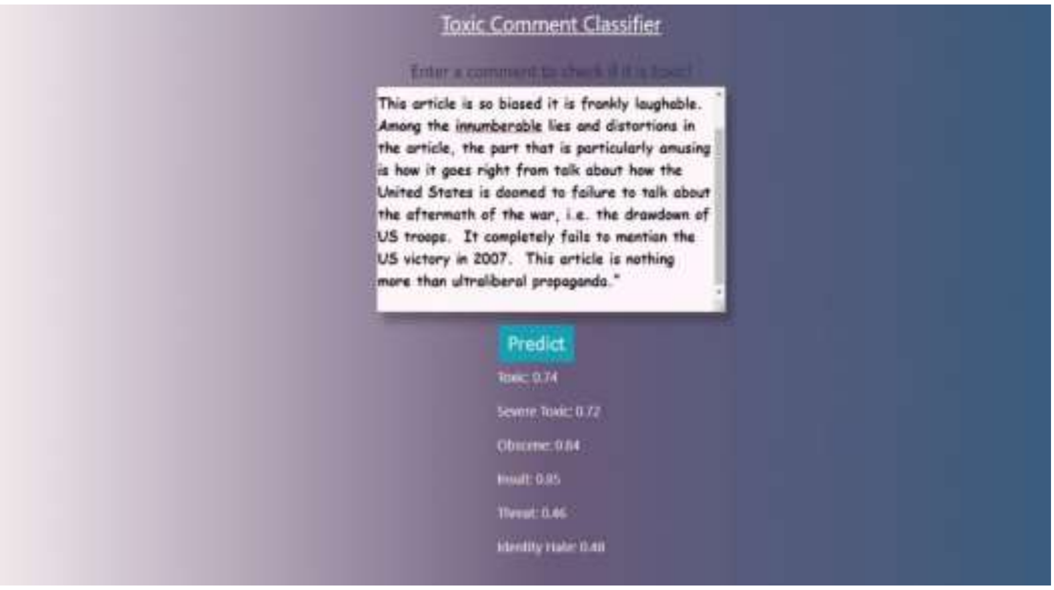
Anaconda Prompt Anaconda command prompt is just like command prompt, but it makes sure that you are able to use anaconda and conda commands from the prompt, without having to change directories or your path. When you start Anaconda command prompt, you'll notice that it adds/("prepends") a bunch of locations to your PATH.

Python 3.7. Python is broadly utilized universally and is a high-level programming language. It was primarily introduced for prominence on code, and its language structure enables software engineers to express ideas in fewer lines of code. Python is a programming language that gives you a chance to work rapidly and coordinate frameworks more effectively.

Jupyter Notebook Jupyter Notebook is an open-source**,** web-based interactive environment**,** which allows you to create and share documents that contain live code, mathematical equations, graphics, maps, plots, visualizations, and narrative text**.** It integrates with many programming languages like Python, PHP, R, C#, etc.

1. All in one place: As you know, Jupyter Notebook is an open-source web-based interactive environment that combines code, text, images, videos, mathematical equations, plots, maps, graphical user interface and widgets to a single document.
2. Easy to convert: Jupyter Notebook allows users to convert the notebooks into other formats such as HTML and PDF. It also uses online tools and nbviewer which allows you to render a publicly available notebook in the browser directly.
3. Easy to share: Jupyter Notebooks are saved in the structured text files (JSON format), which makes them easily shareable.
4. Language independent: Jupyter Notebook is platform-independent because it is represented as JSON (JavaScript Object Notation) format, which is a language-independent, text-based file format. Another reason is that the notebook can be processed by any programing language, and can be converted to any file formats such as Markdown, HTML, PDF, and others.
5. Interactive code: Jupyter notebook uses ipywidgets packages, which provide many common user interfaces for exploring code and data interactivity.

6. Result



Toxic Comment Classifier

Enter a comment to check if it is toxic!

This article is so biased it is frankly laughable. Among the innumberable lies and distortions in the article, the part that is particularly amusing is how it goes right from talk about how the United States is doomed to failure to talk about the aftermath of the war, i.e. the drawdown of US troops. It completely fails to mention the US victory in 2007. This article is nothing more than ultraliberal propaganda."

Toxic: 0.74

Severe Toxic: 0.72

Obscene: 0.84

Insult: 0.85

Threat: 0.46

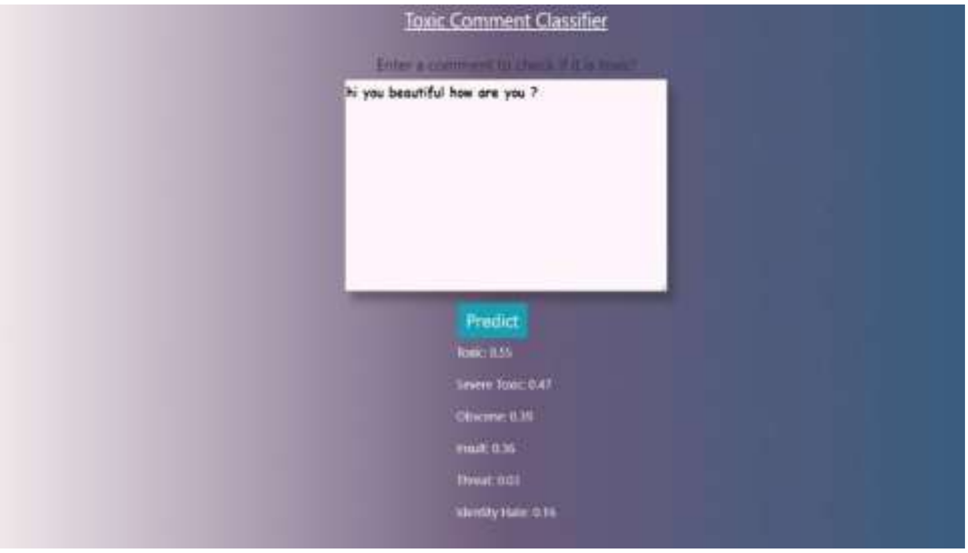
Identity Hate: 0.48

Figure 1 Output screen 1

International Journal for Innovative Engineering and Management Research

A Peer Revieved Open Access International Journal

www.ijiemr.org



Toxic Comment Classifier

Enter a comment to check if it is toxic!

hi you beautiful how are you ?

Toxic: 0.55

Severe Toxic: 0.47

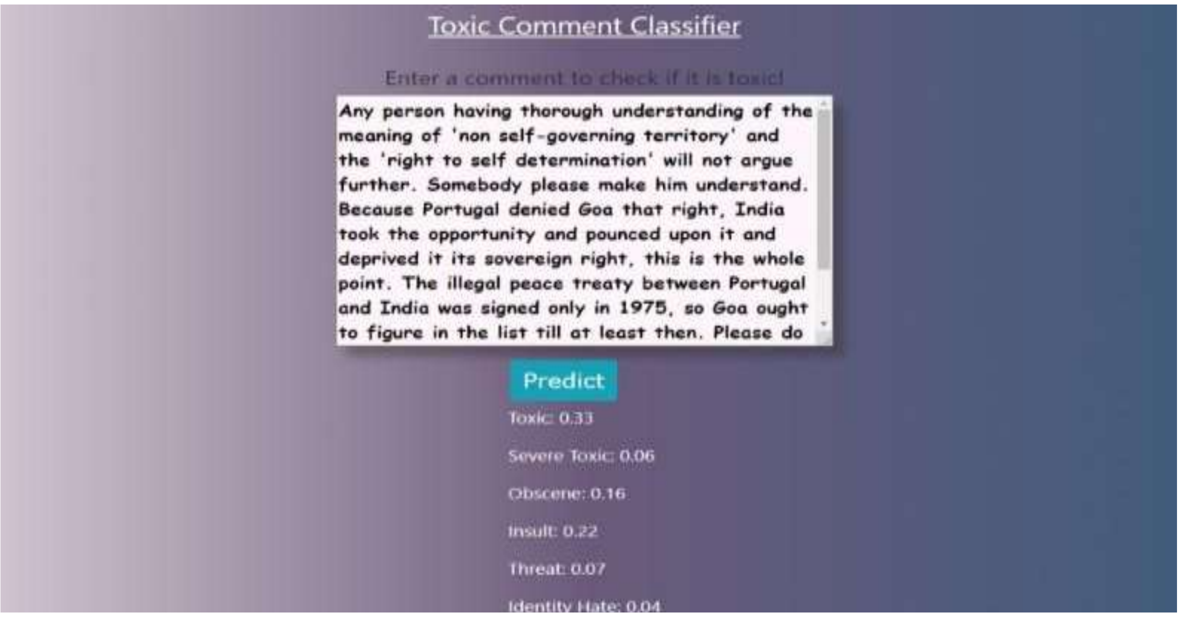
Obscene: 0.39

Insult: 0.36

Threat: 0.03

Identity Hate: 0.16

Figure 2 Output Screen 2



Toxic Comment Classifier

Enter a comment to check if it is toxic!

Any person having thorough understanding of the meaning of 'non self-governing territory' and the 'right to self determination' will not argue further. Somebody please make him understand. Because Portugal denied Goa that right, India took the opportunity and pounced upon it and deprived it its sovereign right, this is the whole point. The illegal peace treaty between Portugal and India was signed only in 1975, so Goa ought to figure in the list till at least then. Please do

Toxic: 0.33

Severe Toxic: 0.06

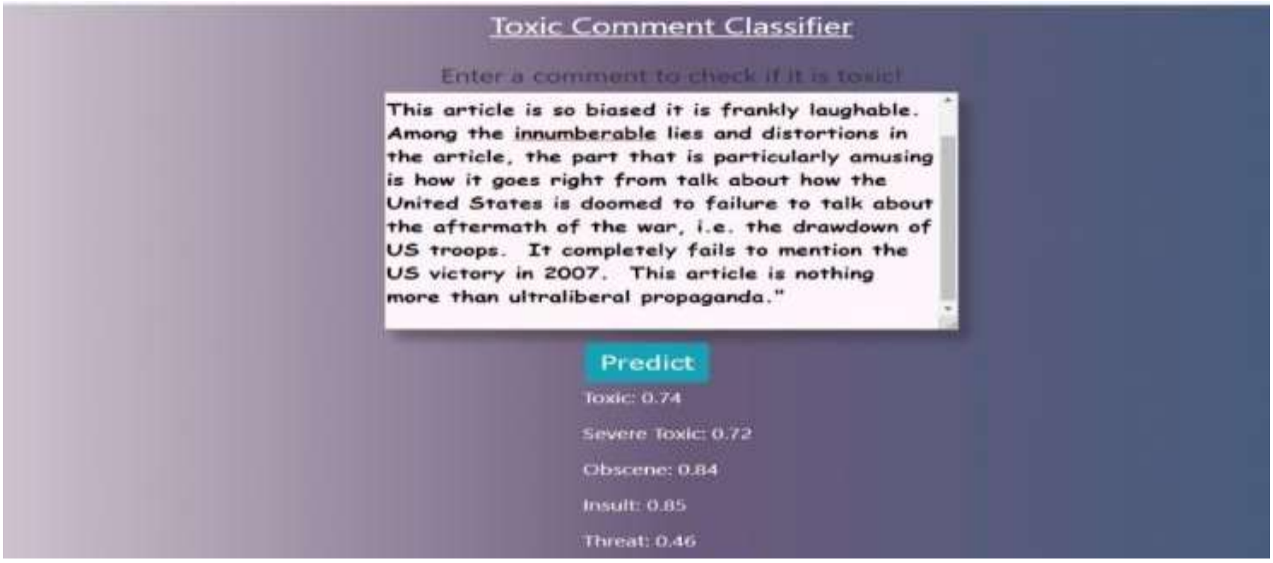
Obscene: 0.16

Insult: 0.22

Threat: 0.07

Identity Hate: 0.04

Figure 3 Output Screen 3



Toxic Comment Classifier

Enter a comment to check if it is toxic!

This article is so biased it is frankly laughable. Among the innumberable lies and distortions in the article, the part that is particularly amusing is how it goes right from talk about how the United States is doomed to failure to talk about the aftermath of the war, i.e. the drawdown of US troops. It completely fails to mention the US victory in 2007. This article is nothing more than ultraliberal propaganda."

Toxic: 0.74

Severe Toxic: 0.72

Obscene: 0.84

Insult: 0.85

Threat: 0.46

Figure 4 output screen 4

8.Conclusion

We have discussed two Machine learning techniques i.e. logistic regression, and SVM classifier, and compared their hamming loss, accuracy, and log loss in this paper. Now after proper analysis, we can say that in terms of hamming loss, logistic regression performs best because in that case, our hamming loss is least, while in terms of accuracy, logistic regression performs best because accuracy is best in that model in comparison to other ones and terms of log loss, random forest works best due to least possible log loss in that model. So, our final model selection will bebased on the combination of hamming loss and accuracy.

Since we got the maximum accuracy i.e. 89.46 % and least possible hamming loss i.e. 2.43 % in case of the logistic regression model. We will select the logistic regression model as our final machine learning technique since it works best for our data.

9. Future Scope

In further research, other machine learning models can be used to calculate accuracy, hamming loss, and log loss for better results. We can also explore some deep learning algorithms such as LSTM (long short-term memory recurrent neural network), multi-layer perceptron, and GRU. So, we can explore many other techniques which will help us to improve the obtained results

10. References

1. [1]A. Olariu, “Efficient Online Summarization of Microblogging Streams,” in Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, Volume 2: Short Papers, 2014, pp. 236–240.
2. [2]Y. Qu, C. Huang, P. Zhang, and J. Zhang, “Micro blogging After a Major Disaster in China: ACase Study of the 2010 Yushu Earthquake,” in Proceedings of the ACM 2011 Conference on Computer Supported Cooperative Work. ACM, 2011, pp. 25–34.

3. [3]P. Gamallo and M. Garcia, “Citius: A Naive-Bayes Strategy for Sentiment Analysis on English Tweets,” in Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), 2014, pp. 171–175.
4. [4]S. Kim, H. Kim, and Y. Namkoong, “Ordinal Classification of Imbalanced Data with Application in Emergency and Disaster Information Services,” IEEE Intelligent Systems, vol. 31,no. 5, pp. 50–56, 2016.
5. [5]B. E. Parilla-Ferrer, P. Fernandez, and J. Ballena, “Automatic Classification of Disaster-Related Tweets,” in Proceedings of International Conference on Innovative Engineering Technologies (ICIET), vol. 62, 2014.
6. [6]K. Stowe, M. J. Paul, M. Palmer, L. Palen, and K. Anderson, “Identifying and Categorizing Disaster-Related Tweets,” in Proceedings of The Fourth International Workshop on Natural Language Processing for Social Media, 2016, pp. 1–6.
7. [7]K. Rudra, S. Ghosh, N. Ganguly, P. Goyal, and S. Ghosh, “Extracting Situational Information from Microblogs During Disaster Events: A Classification-Summarization Approach,” in Proceedings of the 24th ACM International on Conference on Information and Knowledge Management. ACM, 2015, pp. 583–592.
8. [8]Z. Cao, W. Li, S. Li, and F. Wei, “Improving Multi-Document Summarization via Text Classification,” in Thirty-First AAAI Conference on Artificial Intelligence, 2017.
9. [9]K. Rudra, N. Ganguly, P. Goyal, and S. Ghosh, “Extracting and Summarizing Situational Information from the Twitter Social Media During Disasters,” ACM Transactions on the Web (TWEB), vol. 12, no. 3, p. 17, 2018.
10. [10]C. Kedzie, K. McKeown, and F. Diaz, “Summarizing Disasters Over Time,” in Proceedings of Bloomberg Workshop on Social Good (with SIGKDD), 2014.