Multi Label Toxic Comment Classification

(Using Machine Learning Algorithms)

Gudibanda Karthik
20BDS023/*DSAI*

S V Amarnadh Reddy
20BDS055/*DSAI*

S B Mohammad Karimullah
20BDS048/*DSAI*

V Karuna Prakash
20BDS062/*DSAI*

Abstract— **The threat of bullying and abuse on the internet obstructs the free exchange of ideas by limiting people's opposing viewpoints. Most of the Websites fail to successfully facilitate healthy conversations, leading them to either restrict or disable user comments entirely. This paper would explore the scope of online abuse and categorize them into different labels to assess the toxicity as accurately as possible using machine learning algorithms.**

Keywords: Accuracy, Multilabel Classification, Learning Algorithms, Toxic Comments

I. Introduction

The worst of the internet comes in the form of online comments and it is getting worse each year. People often comment on a news story and it seems that, no matter what the story is, someone will find a way to connect it to politics, a personal attack or a conspiracy theory — and they will have no problem posting it.The importance of maintaining civility on a web forum can't be immoderate. Cyber bullying is outlined as "willful and recurrent damage inflicted through the medium of textual matter." It involves causing degrading, threatening, and/or sexually express messages and pictures to targets via websites, blogs, instant electronic communication, chat rooms, e-mail, cell phones, websites and private on-line profiles. As a result, police investigation and deleting toxic communication from public forums could be a very important duty that's not possible for human moderators to try . The goal of the project is to create a multi-label classifier using different types of machine learning models and deep learning models. By which we can detect the severity of toxicity of comments. And finding the best models for the job mentioned above.

II. Methodology

A. About the Dataset

The dataset has been taken from a kaggle competition that was bought by the collaboration of Jigsaw and Google. This data has Wikipedia's comments and the data collected has been labeled by human raters for the toxic behavior. The toxicity types are labeled as toxic, severe_toxic, obscene, threat, insult and identity hate.

B. Exploratory Data Analysis

Exploratory Data Analysis is used to gain a better understanding of the given data and to analyze their key characteristics by using data visualization techniques.

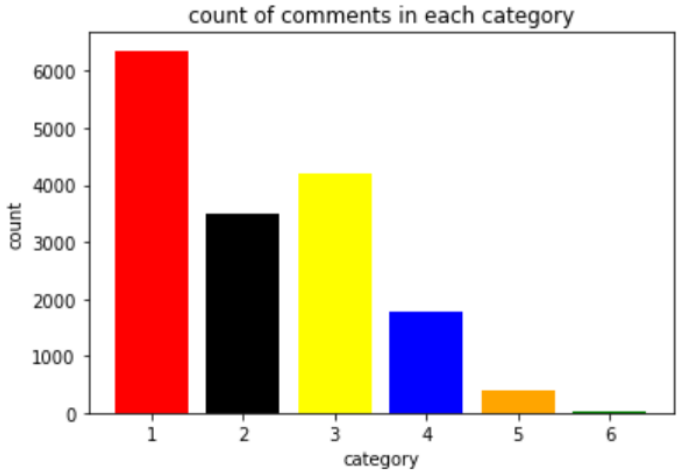


Figure 1

Plot (Figure 1) shows the number of comments having multiple labels.

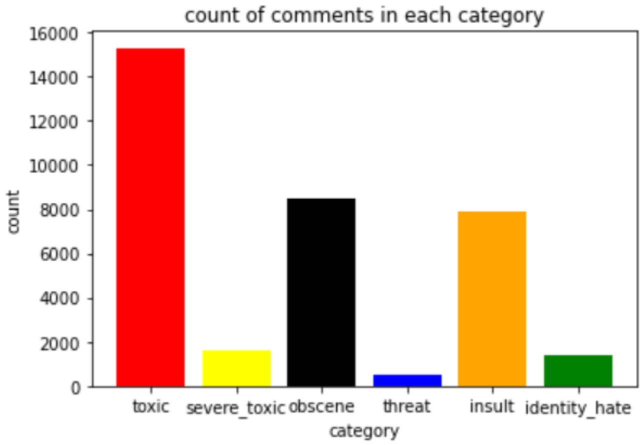


Figure 2

Plot (Figure 2) shows the number of comments that fall under each label. It can be observed that the bulk of the comments fall into the toxic category, and the threat category has the least number of comments.

C. Data Pre-Processing

The dataset consists of different comment statements, this data is cleansed and preprocessed by the means of Natural Language Processing(NLP). The steps are as follows : Lower casing the statement, removing punctuations, removing numbers from the text, tokenization, removing stop words, lemmatization, combining the sentence and vectorizing the statement using TF-IDF.

D. Applying Multi Label Classification Techniques

Since the majority of the conventional machine learning algorithms are designed for classification problems with single labels. Hence, we'll use techniques to break the multi-label problem into several single-label problems, allowing us to use the existing conventional machine learning algorithms.

1. Binary Relevance Method: The interdependence of labels is not taken into account in this process. Each label is solved separately, like a single-label classification problem.

2. Classifier Chain Method: We train the first classifier on the given data in this method, followed by each subsequent classifier being trained on the previous classifier and the input space, and so on. Hence, this approach considers the interdependence of labels and input data. Some classifiers may show dependence, such as toxic and severely toxic.

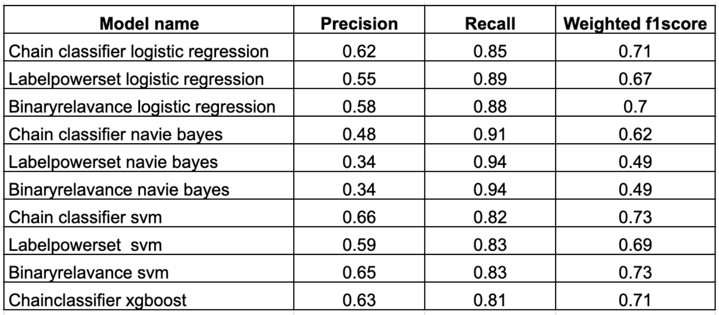
3. Label Power Set Method: This approach takes all possible label combinations into account. As a result, any specific combination can be used as a label,
breaking our multi-label problem into a multi-class classification problem.

E. Machine Learning Methodologies

With each of the above three multi label methods, we used four machine learning models for optimal results.

1. Logistic Regression
2. Multinomial Naive Bayes
3. Linear SVM
4. XGBoost

Thereafter the above methodologies, a new model is built based upon a Deep Learning Model using Artificial Neural Networks.



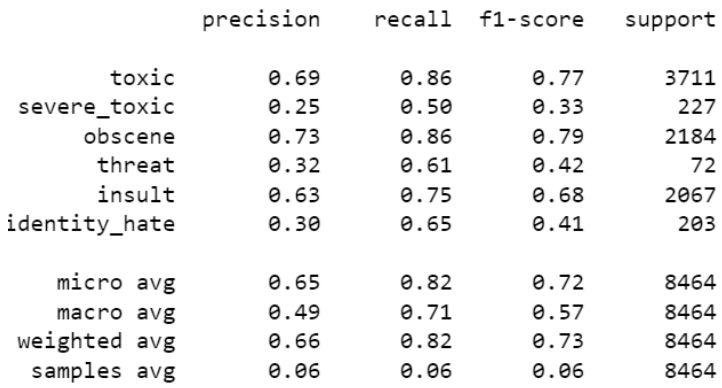
Model name	Precision	Recall	Weighted f1score
Chain classifier logistic regression	0.62	0.85	0.71
Labelpowerset logistic regression	0.55	0.89	0.67
Binaryrelavance logistic regression	0.58	0.88	0.7
Chain classifier navie bayes	0.48	0.91	0.62
Labelpowerset navie bayes	0.34	0.94	0.49
Binaryrelavance navie bayes	0.34	0.94	0.49
Chain classifier svm	0.66	0.82	0.73
Labelpowerset svm	0.59	0.83	0.69
Binaryrelavance svm	0.65	0.83	0.73
Chainclassifier xgboost	0.63	0.81	0.71

Table 1

Given above(Table 1) are the precision, recall and weighted F1 scores for the corresponding machine learning models.

III. Results And Analysis

We used three methods, i.eBinary Relevance, Classifier chain, and Label power set for each of the machine learning algorithms and the deep learning algorithm for finding the toxic severity . Out of all the algorithms the optimal models are chain classifiers(Table 2) using Linear Support Vector Machines(SVM) and Binary Relevance using SVMs(Table 3). [Weighted F1 score is used as the metric to compare the models, as we have unbalanced data]

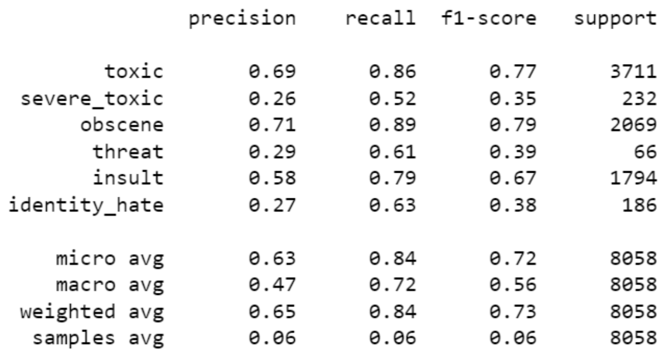


```
               precision    recall  f1-score   support

        toxic       0.69      0.86      0.77      3711
 severe_toxic       0.25      0.50      0.33       227
      obscene       0.73      0.86      0.79      2184
       threat       0.32      0.61      0.42        72
       insult       0.63      0.75      0.68      2067
identity_hate       0.30      0.65      0.41       203

    micro avg       0.65      0.82      0.72      8464
    macro avg       0.49      0.71      0.57      8464
 weighted avg       0.66      0.82      0.73      8464
  samples avg       0.06      0.06      0.06      8464
```

Table 2



```
               precision    recall  f1-score   support

        toxic       0.69      0.86      0.77      3711
 severe_toxic       0.26      0.52      0.35       232
      obscene       0.71      0.89      0.79      2069
       threat       0.29      0.61      0.39        66
       insult       0.58      0.79      0.67      1794
identity_hate       0.27      0.63      0.38       186

    micro avg       0.63      0.84      0.72      8058
    macro avg       0.47      0.72      0.56      8058
 weighted avg       0.65      0.84      0.73      8058
  samples avg       0.06      0.06      0.06      8058
```

Table 3

Acknowledgment

This project requires the collaboration and direction of many people. We consequently believe that thanking everyone who supported us in this endeavor is our top priority. We would like to express our gratitude to Dr. Animesh Chaturvedi, our professor, for motivating us and providing us with the necessary tools while we worked on this project. We appreciate the support and advice from our friends, the teaching and non-teaching staff, who helped us improve our AI Project.

References

[1] Yin, Dawei, Xue, Zhenzhen, Hong, Liangjie, Davison, Brian, Edwards, April, Edwards, Lynne. (2009), “Detection of harassment on Web 2.0”

[2] Razavi, A.H., Inkpen, D., Uritsky, S., and Matwin, S. (2010), “Offensive Language Detection Using Multi-level Classification”. Canadian Conference on AI.

[3] Maxime Rivet and Mael Tran, “Toxic comments classification”, Stanford University journal Year [2016].

[4] Spiros V. Georgakopoulos et al. “Convolutional Neural Networks for Toxic Comment Classification”, Cornell University arXiv:1802.09957 Year 2018.

[5] https://www.kaggle.com/competitions/jigsaw-toxic-comment-classification-challenge/overview

[6] https://denisechendd.github.io/Keras-Multi-Label-Text-Classification-on-Toxic-Comment-Dataset/

[7] https://medium.com/@nupurbaghel/toxic-comment-classification-f6e075c3487a

[8] https://towardsdatascience.com/journey-to-the-center-of-multi-label-classification-384c40229bff