Identification and Classification of Toxic Comment Using Machine Learning Methods

P.Vidyullatha[1], Satya Narayan Padhy[1] ,Javvaji Geetha Priya[2] , Kakarlapudi Srija[3] ,Sri Satyanjani Koppisetti[4]
[1]Associate Professor, Dept. of Computer Science and Engineering,
Koneru Lakshmaiah Education Foundation, A.P.,, India
[1,2,3,4]IVth year B.Tech Student, Dept. of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, A.P., India

[1]Latha22pellakuri@gmail.com., [1]snpadhi12@gmail.com , [2]geethaprasannaj@gmail.com , [3]srija23.kakarlapudi@gmail.com, [4]anjanikoppisetti28@gmail.com , [5]

Article History: Received: 10 January 2021; Revised: 12 February 2021; Accepted: 27 March 2021; Published online: 20 April 2021

Abstract

The increase in penetration of usage of internet services has increased exponentially in the past 4 months due to the ongoing pandemic, this has empowered an enormous number of dynamic new and old clients utilizing the web for different administrationsranging from academic, entertainment, industrial, monitoring and the emergence of a new trend in the corporate-life i.e work-from-home. Due to this sudden emergence of the crowd using the web, there has been an ascent in the number of mischievous persons too. Now it is the primary task of every online platform provider to keep the conversations constructive and inclusive. The best example can be referred to, can be twitter, a web-based media stage where people share their views. This platform has already drawn a lot of flak because of the spread of hate speech, insults, threat, defamatory acts which becomes a challenge for many such online providers in regulating them. Thus, there is active research being conducted in the field of Toxic comment classification. Here wecollatenon-identicalmachinelearning and other trivial techniques on the dataset and propose a model that outflanks all others and compares them one-on-one. We have undertaken the Kaggle dataset for the above reason which has been broadly used and one of the prime resources for scholars working in deducing the challenge of toxic comment classification. The results would help up to create an online interface where we would be able to identify the toxicity level in the given phrase or sentence and classify them into their order of toxicity.

Keywords: Binary relevance,Classification, Defamatory, Multinomial naive bayes,Support vector machine,Toxic comment.

I. Introduction

Due the penetration of the internet in all domains of life which has led to increase of people's participation actively and give remarks as an issue of communicating their concern/feedback/opinion in various online forums. Although most of the times these comments are helpful for the creator toextemporize the substance that is being provided to people, but sometimes these may be abusive and create hatred-feeling among the people. Thus as these are openly available to the public which is being viewed from various sections of the society, people in different age groups, different communities and different socio-economic background, it becomes the prime responsibility of the content-creator (the host) to filter out these comments in order to stop the spread of negativity or hatred within people.

Lately there has been many cases in which the growing menace of hate and negativity has been witnessed in the online platforms especially social media as such, many governments around the world has seen the rise of cases related to cyber bullying that has led to spread of hatred and violence.

Since the democratizationof substance creation following the dispatch of web-based media stages, every single one of us has become content makers making and distributing our own substance, which thus has made a framework where the nature of distributed substance cannot, at this point be controlled. The effect of the most recent twenty years' innovation unrest is presently affecting organizations, political frameworks, family lives, society, and individuals [1].

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.

There have been sure turns of developments in this area which includes couple of models served through API. But the models still make errors and still fail to provide an accurate solution to the problem. In this paper we have widely discussed a set of models which is utilized for text classification. These models/methods have been widely used in various fields such as economics, medical and environmental studies. We have approached a three-tier

approach in this paper. At first, we have evaluated each of the algorithm's efficiency (by tuning and tweaking with different set of parameters in pre-processing to get appealing results). Secondly, we have compared them one-on one with respect to their contrasting features. And then, we have categorized them into sequential manner based on their outcomes to emphasize their comparative predicted values.

II. Related Work#

In the wake of exploring the diverse writing indicated that there have beenseveral studies on the early papers. For example,In 2018 Revati Sharma and Meet Kumar Patel performed classifying toxic comments using Neural networks like CNN and RNN, Using the word embedding techniques and also performing an head on comparison with the primary level neural network algorithms, results with intricate Convolutional Neural Networks(CNN) and Recurrent Neural Network(RNN).

Long-Short Term Memory(LSTM) results, the obtained analysis show that the LSTM performin a way that is better than the CNNs in terms of both the precision and time execution given the same number of epoch and hence are preferable to use rather than CNN with word-level embedding's [4].In 2018 if we see Mujahed A. Saif he performed logistic regression and RNN, LSTM among these 2 LSTM layers and 4 conv layers, has got a score of 0.9645 shows best accuracy [7].

All these papers classify the toxic comments using Neural Network techniques by considering all this work. In this paper we carried out toxic comment classification using Binary relevance method with Multinomial Naïve Bayes and Support vector classifier. In this research paper we tried to classify the toxic, obscene, insult, severe-toxic, identity-hate and threat comments using BR Methods.

III. Proposed system

This section focuses on different algorithms and the various stagesthat are involved for the proposed Toxic comment classification system such as 'logistic regression', 'BR Method with Multinomial Naive Bayes classifiers' and 'BR Method with SVM classifier' which helps us in classifying the comments and results in a decisive outcome.

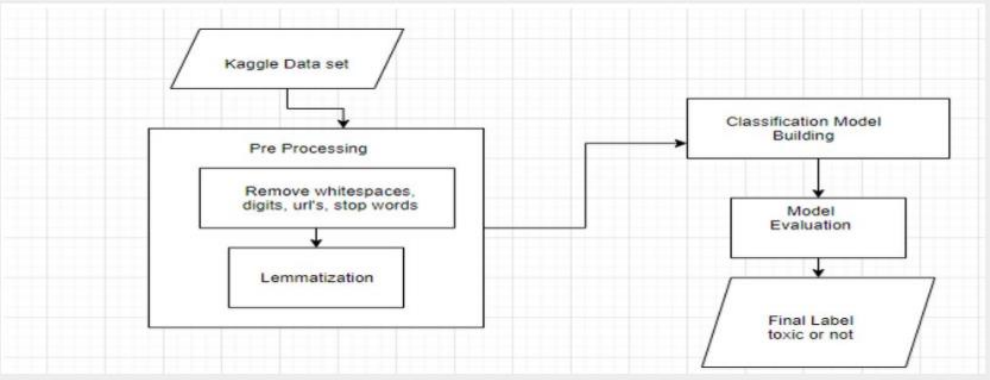

Figure 1: Proposed system.

As the aim was to find out whether the data belongs to either zero, one or more than one from the six in the list, the initial agenda before working on the problem was to audibly differentiate between multi-label and multi-class classification. In multi-class classification, we undertake one basic assumption that our data can belong to only one label out of all that are available to us. Let us say, for a given picture of a vegetable may be apotato, cabbage, or onion only and not a combination of the above. Whereas, in multi-label classification, data can belong concurrently to more than one label. For example, in this project a comment may be belonging to more than one classification concurrently, like it may be toxic, hateful, obscene and abusive at the same time it might concomitantly belong to non-toxic category and thus does not have an affinity to any of the six labels which are used for classification.

At that point we dealt with the number of comments belonging to various categories (which can be perceived from a decisive visualization). Toxic comments were highest in number, followed by obscene, insult, severe-toxic, identity-hate and threat in decreasing order.

The length of the remarks are pretty big so we performed a couple of visualizations in order to make the data more understandable. Initially, we found the count of different toxicity of comments in each of the bins.

	toxic	severe_toxic	obscene	threat	insult	identity_hate
0	0	0	0	0	0	0
1	0	0	0	0	0	0
2	0	0	0	0	0	0
3	0	0	0	0	0	0
4	0	0	0	0	0	0
5	0	0	0	0	0	0
6	1	1	1	0	1	0
7	0	0	0	0	0	0
8	0	0	0	0	0	0
9	0	0	0	0	0	0

Figure 2: Count of different toxicity of comments in each of the bins.

This analysisgivesin depthinsights about the distribution of the data in the database. The succeedingstep was to perform pre-processingof the data, as the volume of the data was good as per our requirements and end goals for this project, we have discussed further on this about the techniques and procedures that we have adopted in order to curate the data and use it further in the project.

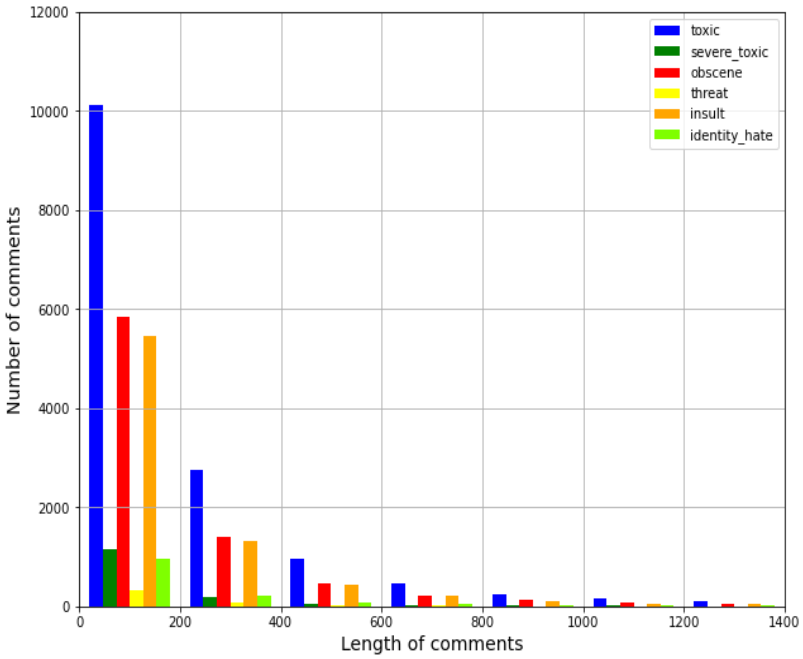

Figure3: Length of comments

Preprocessing

In preprocessing we started off with removing of punctuation and special character from the comments. We at that point recognize that we had to also remove the stop words which are useless as they do not add any value to the dataset i.e. those comments are meaningless. Further we performed stemming and lemmatizing for the words. Lastly, we applied count vectorizer and afterward split the information into preparing and testing for the further use.

We have undergone various ways of visualizing the data to get meaningful outputs, which would help us in knowing the dataset and our final goals effectively. We tried to segregate the analysis of the dataset into various categories such as based on the word lengths, the toxic words utilized and the degree of toxicity present in them. All this visualization helped us in gaining a wholistic picture which led us to finally sort down to two algorithms which was making our end goal in classifying the toxic comments efficiently.

Algorithm

Traditional algorithms are unable to handle a set of multi-label instances because such algorithms were designed to predict a single label. So scikit-multilearn library was used for implementing different methods. Each method requires a base classifier which is created for each of the label and combined in a peculiar way. Classifiers which

were used in this include Multinomial Naive Bayes, SVC. To handle multi-labels we are using binary relevance method.

In Binary Relevance Method all the labels in the dataset are partitioned into single labels and each single label are performed as single label classification problem. And the given figure 4 and figure 5 describes it.

X	Y_1	Y_2	Y_3	Y_4
$x^{(1)}$	0	1	1	0
$x^{(2)}$	1	0	0	0
$x^{(3)}$	0	1	0	0
$x^{(4)}$	1	0	0	1
$x^{(5)}$	0	0	0	1

Figure 4

X	Y_1
$x^{(1)}$	0
$x^{(2)}$	1
$x^{(3)}$	0
$x^{(4)}$	1
$x^{(5)}$	0

X	Y_2
$x^{(1)}$	1
$x^{(2)}$	0
$x^{(3)}$	1
$x^{(4)}$	0
$x^{(5)}$	0

X	Y_3
$x^{(1)}$	1
$x^{(2)}$	0
$x^{(3)}$	0
$x^{(4)}$	0
$x^{(5)}$	0

X	Y_4
$x^{(1)}$	0
$x^{(2)}$	0
$x^{(3)}$	0
$x^{(4)}$	1
$x^{(5)}$	1

Figure 5

Furthermore, the information is smidgen slanted i.e.extremely less level of thecomments are toxic so accuracy metric gives invalid results. So the best metrics for this algorithm to find the performance are Hamming loss and Log loss.

IV. Result and Discussion

The outcomes for the algorithms were as follows, if weif we compare both hamming losses, we could come to the conclusion that Naïve Bayes has a hamming loss of 3.6 and an accuracy of 87.6 whereas the hamming loss for SVM is 4.36 and the accuracy is 88.16.

This gives us a brief insight to understand the optimal algorithm that can be utilized for ordering toxic comments.

```
Hamming_loss : 3.6503650365036506
Accuracy : 87.5937593759376
Log_loss : 1.9625004170858478
```

Figure 6: BR Method with Multinomial Naive Bayes classifiers

```
Hamming_loss : 4.367936793679368
Accuracy : 88.16381638163816
Log loss : 0.4695775266638782
```

Figure 7: BR Method with SVM classifier (from scikit-multilearn)

Conclusion

Thus, from the results we can conclude that taking hamming loss as a measure of identifying the optimal algorithm to classify toxic comments we can say that Binary Relevance method with Multinomial Naive Bayes is an efficient algorithm that serves our purpose and has a hamming loss of 3.6 as compared to the hamming loss of SVM with a score of 4.36.

References:

[1] NayanBanik, Md. Hasan Hafizur Rahman," Toxicity Detection on Bengali Social Media Comments using Supervised Models", (ICIET) 23-24 December, 2019

[2] Salvatore Carta, Andrea Corriga, Riccardo Mulas, Diego ReforgiatoRecupero and Roberto Saia," A Supervised Multi-class Multi-label Word Embeddings Approach for Toxic Comment Classification",2019

[3] Hind Almerekhi, Haewoon Kwak, Bernard J. Jansen, Joni Salminen," Detecting Toxicity Triggers in Online Discussions", HT '19, September 17–20, 2019, Hof, Germany, pg no: 291 – 292.

[4] Revati Sharma , Meetkumar Patel, "Toxic Comment Classification Using Neural Networks and Machine Learning", Vol. 5, Issue 9, September 2018, DOI 10.17148/IARJSET.2018.597,pg no:47- 52

[5]Mai Ibrahim, Marwan Torki and Nagwa El-Makky. (2018), "Imbalanced Toxic Comments Classification using Data Augmentation and Deep Learning ",2018

[6] Spiros V. Georgakopoulos, Sotiris K. Tasoulis, Aristidis G. Vrahatis, Vassilis P. Plagianakos," Convolutional Neural Networks for Toxic Comment Classification" ,arXiv:1802.09957v1 [cs.CL] 27 Feb 2018.

[7] Mujahed A. Saif, Alexander N. Medvedev, Maxim A. Medvedev, TodorkaAtanasova, "Classification of Online Toxic Comments Using the Logistic Regression and Neural Networks Models",2018

[8] Fahim Mohammad, "Is preprocessing of text really worth your time for toxic comment classification", Int'l Conf. Artificial Intelligence | ICAI'18,2018, pg no: 447-480.

[9] Pooja Parekh, Hetal Patel," Toxic Comment Tools: A Case Study", Volume 8, No. 5, May-June 2017, pg no: 964 – 967

[10] NavoneelChakrabarty ," A Machine Learning Approach to Comment Toxicity Classification", 2016

[11]Pallam Ravi, Hari Narayana Batta, Greeshma S, Shaik Yaseen," Toxic Comment Classification",Volume: 3, Issue: 4, 2019