Automated Classification of Online Toxic Comments with Machine Learning Techniques

ALANKARAM SHOBITHA LAKSHMI1, PONTHAGANI KALYANI2

#1Assistant Professor, Department of CSE, PBR Visvodaya Institute of Technology and Science, Kavali

#2Assistant Professor, Department of CSE, PBR Visvodaya Institute of Technology and Science, Kavali

ABSTRACT

Toxic Comments are disrespectful, abusive or unreasonable online comments that usually make other users leave a discussion. The danger of online bulling and harassment affects the free flow of thoughts by restricting the dissenting the opinions of people. Now a days we use internet to exchange information, based on the information people leave their opinions through comments. It is need to detect and restrict the antisocial behavior over the online discussion forums. In this we use Machine Learning algorithms to classify online comments. This paper will systematically examine the extent of online harassment and classify the content into labels to examine the toxicity as correctly as possible.

1. INTRODUCTION

Internet is the greatest innovation of 21st century. By using the internet one can communicate with others using smart phones and computers. In earlier days of internet one can used to send emails, we don't know whether it is positive or negative. Example – whether the mail is spam or not. As time changes, the using of internet is increased and its origins are expanded. And one of it is finding the positive and negative in the data. In emails, negative mails are identified and it is sent to the spam. Some methodologies are used to identify whether it is spam or not.

Now a days, many apps are used for information, data is transferred through the internet. Based on the information the users leave their opinions. These opinions are referred as comments. The comments may be positive or negative. By seeing comments people may react either in positive or negative way.By using Machine Learning algorithms we classify the data into toxic and non-toxic comments and find the percentage of toxicity used in the comments.

As for real life examples, an abusive comment is posted in facebook on Mamatha Benarjee. The person is arrested for commenting in negative way in Bengal. Another example is takes place in Indonesia for posting a toxic comment against police and got arrested.

Thus,it 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.

The Toxic comment like “Nonsense! You are a narcissist.” Here nonsense and narcissist are the toxic words. By using machine learning algorithms on the given dataset, they are classified and verify the precision of the obtained result.

2.RELATED WORK

In daily life, a large amount of data is released in many sites through internet. This data can affect the human life due to the toxicity in it. The toxic comments are restricting people to express themselves. So, it’s time to find and restrict the toxic data in social media. Text classification is used to classify the data of the real-world into binary form for proper processing through the computer. Text classification can be easily applied on given dataset and set of labels by applying the data on a function, that will assign a value to each data value of dataset.

In this context, Wulczyn et al. research introduced a technique that incorporates crowdsourcing and machine learning to evaluate on-scale personal attacks. Recently, a project called Perspective was introduced by Google and Jigsaw, to detect the online toxicity, threats and offensive content with the help of machine learning algorithms. In the approach used by Y.Chen et al. , introduced a combination of a parser and lexical feature to detect the toxic language in YouTube comments to protect adolescents. In the approach used by Sulke et al., online comments are classified with the help of machine learning algorithms. So, lots of work is already done to detect and classify online toxic comments. In our research paper, we will learn from the already published work and use machine learning algorithms to detect and classify online toxic comments with better accuracy.

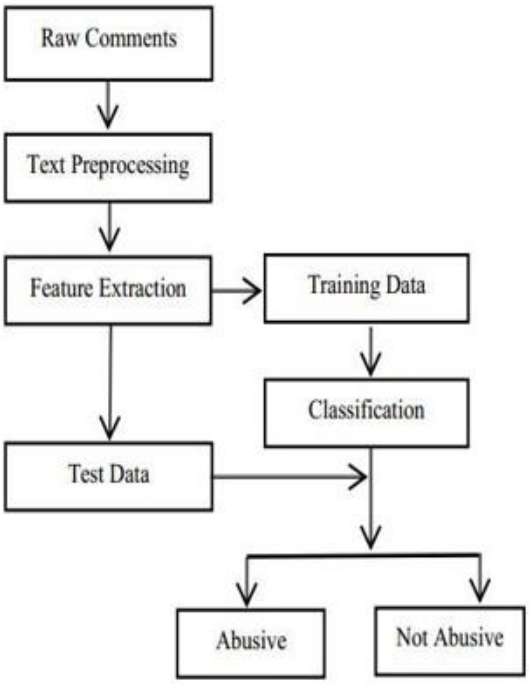

Fig : Flow chart for detecting toxic comments

3.PROPOSED SYSTEM

Classify the data into six categories threat, insult, toxic, severe toxic, obscene, and identity-hate and we can put one data value into zero, one or more than one category. First classify whether our data is multi-class or multi-label in nature.

In Multi-label classification, one data value can belong to more than one category. Example, garden – in garden it consists of trees, walking path, etc..

In Multi-label classification, one data value can belong to only one category. Example, car company – it consists of different brands like Hyundai, Suzuki, Honda, etc..

In existing system, we used six machine learning algorithms they are Support Vector Machine(SVM), K Nearest Neighbor(KNN), Naive Bayes, Decision Tree, Random Forest, Logistic Regression. Among these six Random Forest gives the best accuracy.

In proposed system, we can predict the type of toxic comment it belongs in the categories given. By comparing all the algorithms Logistic Regression gives the better accuracy.

4. RESULTS

```
l.append(data[0][4])
l.append(data[0][5])
```

```
Enter the text: In group therapy someone said that they're proud, they never got depressed, no matter what life threw at the
m. Because they're such a positive and strong person. With a person with depression sitting right next to them.
```

```
values = ['toxic','severe_toxic','obscene','threat','insult','identity_hate']
keys = l
```

```
print("Original key list is : " + str(keys))
print("Original value list is : " + str(values))
res = dict(zip(keys, values))
print("Resultant dictionary is : " + str(res))
```

```
Original key list is : [0.09048163326410724, 0.007373605916174548, 0.04846639275468747, 0.001800266476131348, 0.044466560203
530126, 0.006828479826805164]
Original value list is : ['toxic', 'severe_toxic', 'obscene', 'threat', 'insult', 'identity_hate']
Resultant dictionary is : {0.09048163326410724: 'toxic', 0.007373605916174548: 'severe_toxic', 0.04846639275468747: 'obscen
e', 0.001800266476131348: 'threat', 0.044466560203530126: 'insult', 0.006828479826805164: 'identity_hate'}
```

5.CONCLUSION

By using evaluation metrics on the dataset for Logistic Regression algorithm, we get the more accuracy when compared to existing system. i.e., 89.46% .

REFERENCES

1. H. M. Saleem, K. P. Dillon, S. Benesch and D. Ruths, "A Web of Hate: Tackling Hateful Speech in Online Social Spaces", 2017, [online] Available: http://arxiv.org/abs/1709.10159.

2. M. Duggan, "Online harassment 2017", *Pew Res.*, pp. 1-85, 2017.

3. M. A. Walker, P. Anand, J. E. F. Tree, R. Abbott and J. King, "A corpus for research on deliberation and debate", *Proc. 8th Int . Conf. Lang. Resour. Eval. Lr. 2012*, pp. 812-817, 2012.

4. J. Cheng, C. Danescu-Niculescu-Mizil and J. Leskovec, "Antisocial behavior in online discussion communities", *Proc. 9th Int. Conf. Web Soc. Media ICWSM 2015*, pp. 61-70, 2015.

5. B. Mathew et al., "Thou shalt not hate: Countering online hate speech", *Proc. 13th Int. Conf. Web Soc. Media ICWSM 2019*, no. August, pp. 369-380, 2019.

6. C. Nobata, J. Tetreault, A. Thomas, Y. Mehdad and Y. Chang, "Abusive language detection in online user content", *25th Int. World Wide Web Conf. WWW 2016*, pp. 145-153, 2016.

7. E. K. Ikonomakis, S. Kotsiantis and V. Tampakas, "Text Classification Using Machine Learning Techniques", August 2005.

8. M. R. Murty, J. V. Murthy and P. Reddy, "Text Document Classification basedon Least Square Support Vector Machines with Singular Value Decomposition", *Int. J. Comput. Appl*, vol. 27, no. 7, pp. 21-26, 2011.

9. E. Wulczyn, N. Thain and L. Dixon, "Ex machina: Personal attacks seen at scale", *26th Int. World Wide Web Conf. WWW 2017*, pp. 1391-1399, 2017.

10. H. Hosseini, S. Kannan, B. Zhang and R. Poovendran, "Deceiving Google's Perspective API Built for Detecting Toxic Comments", 2017, [online] Available: http://arxiv.org/abs/1702.08138.

11. Y. Kim, "Convolutional neural networks for sentence classification", *EMNLP 2014 - 2014 Conf. Empir. Methods Nat. Lang. Process. Proc. Conf.*, pp. 1746-1751, 2014.

12. R. Johnson and T. Zhang, "Effective use of word order for text categorization with convolutional neural networks", *NAACL HLT 2015 - 2015 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. Proc. Conf.*, no. 2011, pp. 103-112, 2015.

13. Y. Chen and S. Zhu, "Detecting Offensive Language in Social Media to Protect Adolescents", [online] Available: http://www.cse.psu.edu/~sxz16/papers/SocialCom2012.pdf.

14. A. L. Sulke and A. S. Varude, "Classification of Online Pernicious Comments using Machine Learning", October 2019.

15. N. Chakrabarty, "A Machine Learning Approach to Comment Toxicity Classification", *Adv. Intell. Syst. Comput*, vol. 999, pp. 183-193, 2020.

AUTHOR PROFILES