International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 06 Issue: 06 | June - 2022 Impact Factor: 7.185 ISSN: 2582-3930

TOXIC COMMENTS CLASSIFICATION BASED ON DEEP LEARNING

[1]AKHILESH K J,[2]Ms ASHWINI C

[1]Student, Department of Master Applications,University B.D.T College of Engineering, Davangere,Karnataka, India

[2]Assistant Professor, Department of Master Applications, University B.D.T College of Engineering, Davangere, Karnataka, India

---

Abstract - **Nowadays so many people using social media for express the feelings like uplode the photos and video on social media and this photos and videos give some comments this comments like good or toxic, Social networks sometimes become a place for threats, insults and other components of cyber bullying. A huge number of people are involved in online social networks. Hence, the protection of network users from anti-social behavior is an important activity,but some people give harassment,**

The harassment physically or direct is controlled by police and other forces but online harassment should be controlled by some models that restrict the user not to post a comment by identifying the comment toxicity level.

One of the major tasks of such activity is automated detecting the toxic comments. Toxic comments are textual comments with threats, obscene, racism etc.

To prevent this we come up with a solution, in that various techniques are used for human-free detecting the toxic comments. Bag of words statics and bag of symbols statics are the typical source of information for the toxic comments detection. Usually, the following statistics-based features are used: length of the comment, number of tokens with non-alphabet symbols, number of abusive, arguments. Aggressive, and threatening words in the comment, etc. A neural network model is used to classify the comments. In this paper, Kaggle's toxic comment dataset is used to train deep learning model and classifying the comments in following categories: toxic, severe toxic, obscene, threat, insult, and identity hate

1. INTRODUCTION

in 21st century technology is so developed and the people use the many social media like whatsapp,Instagram, facebook,twitter,etc. social networks some times become a place for threats and other components of cyberbullying, in this social media uplode so many things and some person gives bad comments this comments will hurt the innocent peoples so avoid bad comments for social media to this purpose to develop toxic comments classification project in this to find and avoid so many bad comments**,** The advances in IT technologies and generalizing virtualization all over the world has led to an unprecedented participation in social media; and there is no doubt that social media is one of the biggest hallmarks of the 21st century. According to de Bruijn, Muhonen [1] ; social media has been growing exponentially since 2004. Meanwhile, social media is a place to express individual opinions and share thoughts in line with a constructive contribution to develop a safe place for everybody practicing their rights accordingly [2]. Based on the report by Birkland [3], 'Twitter users generate 500 million tweets per day, and in 2019 they had a 14% year-over-year growth of daily usage'. However, behind the shield of computers as virtual walls, some individuals also think they can abuse and harass other people's opinions and characters. Accordingly, a jargon word has been coined recently to address such behaviors as "cyberbullying".

Online social networking sites provide a platform for people to anonymously share and express their opinions [1]. Sometimes, such opinions can be harassive, abusive, or trollicious to others and cause some individuals to stop sharing, getting depressed, or even have suicidal thoughts [2]. Therefore, an automatic system needs to be developed to avoid, remove, or flag such unhealthy contents from online platforms [3,4]. The development of such a toxicity identification system, however, is a very challenging task for online platform providers. Natural language processing (NLP) helps to identify toxicity in texts, which are expressed as posts or comments. These comments are naturally associated with multiple toxic labels such as toxic, severe toxic, obscene, threat, insult, and identity hate

2. Literature survey

Related work has been mainly into methods to detect a toxic comment, challenges faced in the process and proposals for the inclusion of new and innovative architectures for detection. One such proposed architecture for solving NLP tasks, in general, is BERT proposed by Devlin et al. (2019) in their paper. The paper primarily focused on all layers of the BERT architecture, pre-trained procedures, fine-tuning of the model and analyzing the performance of the model based on standard parameters and benchmark scores including GLUE score, MultiNLI accuracy and F1 score. The paper explains how an attention layer can greatly help in solving various challenges in the field of NLP. Yang et al. (2020) in their paper proposes an incremental iteration for the BERT model that is a generalized autoregressive pre-training method that tries to enable learning bidirectional contexts by maximizing the expected likelihood over all possible permutations of the factorization order.

It also succeeds in overcoming the limitations of BERT with its proposed autoregressive formulation. The paper also claims better scores on 20 diverse tasks over the traditional

pretrained transformer architecture. The paper proposed by Joshi et al. (2019) proposes a pre-training approach called RoBERTa that was specifically developed to overcome the shortcomings of BERT. The primary differences in the pre-training method were that the model was trained over more data, for more epochs and with a bigger batch size, and the next sentence prediction objective was removed from the process T. Gaber[1]. (2015) suggested a plant recommender method that uses 2D visual photographs of plants. This program used the methodology of attribute fusion and the process of multilabel classification. The experimental findings revealed that the function fusion method's accuracy was much higher than other individual applications. The tests showed their robustness in providing accurate recommendations .

3. Methodology

A. Type of Classification

In this paper, we will classify the given dataset (comments written by a user in an online forum) provided by Kaggle in six labels, i.e., toxic, obscene, identity hate, severe toxic, threat, or insult. The next step is to determine if the given data (comment) belongs to one or more than one or none of the mentioned six labels. For example, the given comment can be toxic and insulting, hence falling into more than one label, but the comment can also be non-toxic and not fall into any of the six labels.

B. Exploratory Data

Analysis Exploratory data analysis is a crucial step in the data analysis process. The main aim of EDA is to gain a better understanding of the given data and to analyze their key characteristics. This is achieved by using data visualization techniques**.**

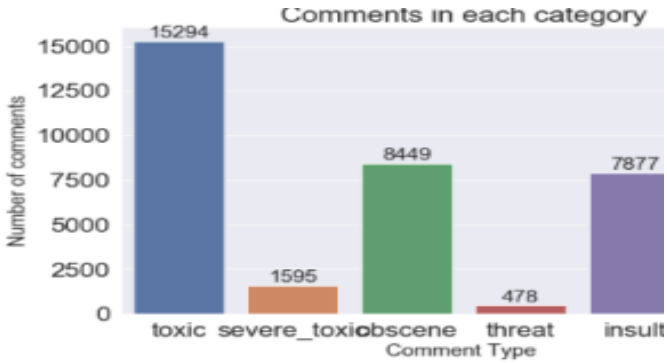

Figure 3: flow diagram

Plot 1 depicts 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

Data pre-processing is a technique used to transform the raw data into an understandable and readable format to make it suitable for building and training Machine Learning models. For our dataset, this can be achieved in 2 stages: (1) Data Cleaning (Removal of unnecessary elements from our text); (2) Feature Engineering (extracting features from data and transforming them into formats that are suitable for Machine Learning algorithms).

4. Data set

We use the Wikipedia talk pages dataset published by Google Jigsaw on Kaggle [40]. This dataset includes 223; 549 instances with six labels, namely, toxic, obscene, severe toxic, insult, threat, and identity hate. These labels define an instance as toxicity or non-toxicity. In particular, it is one of the largest datasets with class imbalance. Moreover, 201; 081 instances were assigned with a 'clear' category matching none of the above six labels. 'Threat' is the least category in the dataset as shown in table

Table:. The proposed MCBiGRU model

Label distribution for Wikipedia talk pages dataset.

Code	Label	Occurrences	%
0	Toxic	21,384	8.53
1	Severe toxic	1962	0.78
2	Obscene	12,140	4.84
3	Threat	689	0.27
4	Insult	11,304	4.51
5	Identity hate	2117	0.84
-	Clean	201,081	80.23

The CNN is widely used in the applications of image classification, image and video recognition, recommender systems, and NLP [31,32,7,22]. The CNN is passed over an input sequence with many filters in a fixed-length vector to produce new feature maps at different positions [26]. Specifically, we use a 1D-Convolution layer with many filters (W) and five different kernel sizes (h) separately for multichannel environments. Let Wi 2 Rhd be the filter for channel i in dimension d. Let Vi 2 RNd be the word embeddings for channel i with the maximum input sequence length N. Then, features mk are generated as in (1).

$$m_k = \left(\sum_{i=1}^{c} V^i[k : k+h-1] \otimes W^i + b \right)$$

$$C = [m_1, m_2, m_3, \ldots, m_k]$$

Comment	Toxic
Stupid law. Why does Quebec always have to cause a fuss over everything? Just leave people alone, for heaven's sake.	1
sorry people, i'm just bored!}}	0
fuck you white trash!!!	1
White racism leaks out through the seams of an establishment race bigot.	1
Who are you to determine fact from fiction? The general public considers the NAACP and ADL to be civil rights organizations. David Duke is regarded as a racist by most and even the US Government has investigated his actions as a domestic terrorist along with federal tax evasion.	0

Table 1: A random sample from the dataset

5. Results and Discussion

After applying all the 6 machine learning techniques over the cleaned data set of Kaggle, we will get the required result of each machine learning technique in the form of Hammingloss, Accuracy, and Log-loss. As we have to select the best machine learning model, we have to properly analyze and compare these results. Hamming-loss, accuracy, and log-loss for each machine learning algorithm are presented in table 2.

Model	Hamming loss	Accuracy	Log loss
Logistic Regression	2.432451957 809565	89.466840 05201561	2.1435876 92292937
Naive Bayes	3.764629388 816645	86.592977 893368	2.3524402 426558524
Decision Tree	3.028464094 783991	86.684005 20156047	2.2582552 11101094
Random Forest	5.436353128 16067	85.652362 37537928	0.5844382 317269664
KNN Classificatio n	3.839040601 0692093	87.121803 20762896	1.6592639 816189594
SVM classifier	2.764651043 1735623	88.697077 82814157	2.2914469 953676764

Table 2: Hamming loss, accuracy and log loss for machine learning models

The result comparison of the existing deep learning model is shown in Table 2. In [7,27,4,27,20,3,15], the authors used 159; 571 toxic comments in their research works. In [18,29,30], and our proposed work 223; 549 comments have been used for experimental study. Specifically, the Aken et al. [18] has described in-depth error analysis in this large dataset with two input word embeddings such as character and n-gram word embeddings. However, we achieve better training and testing accuracy than the existing models using only n-gram word embeddings.

6. CONCLUSION

To summarize the work, four deep learning models with different architectures were implemented on the same dataset and results were compiled. The state-of-the-art transformer model out-performed all the other models by a discernible margin. BiLSTMs with CNNs was the second best and closest to the transformer model.

We presented a multichannel convolutional bidirectional gated recurrent unit to categorize multilabel toxicities in online comments. Especially, the proposed model combines CNN and BiGRU in each channel to extract local features and long-term dependencies within comments using many filters and different kernel sizes. Our results show that the proposed MCBiGRU model outperforms the existing results. In the future, we intend to apply multichannel attention mechanisms in a distributed environment for multilabel toxic detection.

Deep learning model is trained using various deep learning techniques to classify the comments in social meida netoworks in the following categories: toxic, severe toxic, obscene, threat, insult, and identity hate. Kaggle's toxic comment dataset is used for training. In conclusion, Glove & CNN performs the best and Glove & CNN & LSTM performs

the worst in terms of training and testing, loss and accuracy. LSTM and ANN are performing the same followed by CNN and GloVe & LSTM. Recursive neural networks comprise a class of architecture that can operate on structured input. They have been prevoiusly successfully applied to model compositionality in natural language processing using parse tree based structural representation. It can be contructed by stacking multiple recursive layers. The results show thet deep RNNs outperforms the associated shallow counterpart the empoly the same number of parameters. Deep RNNs can be used for Abuse classification.

FUTURE WORK

Newer and more innovative architectures can be implemented on an even bigger dataset that contains comments from a diverse range of online forums.

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 result.

We suggest a plan to improve the NLP classifiers: first by using other algorithms which such as Support Vector Clustering (SVC) and Convolutional Neural Networks (CNN); secondly, extend the classifiers to the overall goal of Kaggle competition which is multi-label classifiers. in the current study, the problem simplified into two classes but it worth to pursue a main goal which is 7 classes of comments**.**

REFERENCES

[1] Alissa de Bruijn, Vesa Muhonen, Tommaso Albinonistraat, Wan Fokkink, Peter Bloem, and Business Analytics. Detecting offensive language using transfer learning. 2019.

[2] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. Xlnet: Generalized autoregressive pretraining for language understanding. In Advances in neural information processing systems, pages 5754–5764, 2019.

[3] Thomas A Birkland. An introduction to the policy process: Theories, concepts, and models of public policy making. Routledge, 2019.

[4] Estela Saquete, David Tomas, Paloma Moreda, Patricio Martinez-Barco, and Manuel Palomar. Fighting post-truth using natural language processing: A review and open challenges. Expert Systems with Applications, 141:112943, 2020.

[5] Sameer Hinduja and Justin W Patchin. Bullying, cyberbullying, and suicide. Archives of suicide research, 14(3):206–221, 2010.

[6] 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.

[7] M. Duggan, "Online harassment 2017," Pew Res., pp. 1–85, 2017, doi: 202.419.4372.

[8] 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.

[9] 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.

[10] 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.

[11] 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, doi: 10.1145/2872427.2883062.

[12] E. K. Ikonomakis, S. Kotsiantis, and V. Tampakas, "Text Classification Using Machine Learning Techniques," no. August, 2005.

[13] M. R. Murty, J. V. . Murthy, and P. Reddy P.V.G.D, " 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, doi: 10.5120/3312-4540.

[14] 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, doi: 10.1145/3038912.3052591.

[15] 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.

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

[17] 2. Ravi, P. (2012), "Detecting Insults in Social Commentary".

[18] Guang Xiang, Bin Fan, Ling Wang, Jason Hong, and Carolyn Rose. 2012, "Detecting offensive tweets via topical feature discovery over a large scale twitter corpus". In Proceedings of the 21st ACM international conference on

Information and knowledge management (CIKM '12). Association for Computing Machinery, New York, NY, USA, 1980–1984. DOI: https://doi.org/10.1145/2396761.2398556.

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

[20] 5. Kansara, Krishna B. and N. Shekokar. "A Framework for Cyberbullying Detection in Social Network." (2015).