2022 14th International Conference on Computational Intelligence and Communication Networks (CICN) | 978-1-6654-8771-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/CICN56167.2022.10008301

Detecting Toxic Comments Using Convolutional Neural Network Approach

Varun Mishra
Dept. of Computer Science & Engineering,
Shri Krishna University, Chhatarpur,
Madhya Pradesh, INDIA
varun.mishra97@gmail.com

Monika Tripathi
Dept. of Computer Science & Engineering,
Shri Krishna University, Chhatarpur,
Madhya Pradesh, INDIA

***Abstract*— In the most significant issue now plaguing social networking platforms and online communities is toxicity identification. Therefore, it is necessary to create an automatic hazardous identification system to block and restrict individual from certain online environments. We introduce multichannel Convolutional Neural Network (CNN) approach in this paper for the detection of toxic comments in a multi-label context. With the help of pre-trained word embeddings, the suggested model produces word vectors. Also, to model input words with long-term dependency, this hybrid model extracts local characteristics using a variety of filters and kernel sizes. Then, to forecast multi-label categories, we integrate numerous channels with three layers as fully linked, normalization, and an output layer. The results of the experiments show that the suggested model performs where we are presenting the fresh modeling CNN approach to detect the toxicity of textual content present on the social media platforms and we categorized the toxicity into positive and negative impact on our society.**

Keywords- Toxicity, Deep learning, CNN, Sentiment Analysis

I. INTRODUCTION

On a global scale, Web technology is enabling individuals to engage freely without being physically there. As a result, people can express their opinions online on any topic they want others to know about. The graph of time spent online increases as technology advances minute by minute. The Communities are seeing an increase in toxic remarks, whether people know one another. Words, phrases, or statements that indicate or suggest contempt for the other person are referred to as toxic comments [1]. The toxic statements may cause someone to feel insulted or degraded. Additionally, they could make someone feel uneasy to the point where they feel uncomfortable expressing a particular viewpoint in public. However, Eight out of ten Indians have experienced some form of online harassment, the most frequent of which are abuse and insults, claims the study, which surveyed over 1000 people and attempted to gauge the country's level of online harassment [2]. Thus, using poisonous language might have serious consequences like isolating or feeling hopeless. Since 2015, when scientists expressed concern the lack of social of published studies on hazardous as numerous subsequent researchers have worked to identify offensive language and stop cyber-bullies, to online development [3]. in part because they both threaten and cause harm.

TABLE1. WORD-BASED ATTACK ON THE PROSPECTIVE API RATING SYSTEM

Comment	Toxic Rating	
We suck at handling harassment and trolls on the platform, and we have been sucked for years.	0.77	TOXIC
We are not suck at handling harassment and trolls on the site, and we never have sucked at it.	0.64	Unsure
We are not adept at handling fame and money on the platform, but we will get improve.	0.06	Unlikely

The original comment made by the CEO of Twitter was to determine whether it was likely to be toxic (in a context of machine learning), and then to gradually change the phrase structure to lessen its severity. This experimental alteration that highlights some linguistic difficulties analogous to an adversarial assault on the Perspective API as shown in TABLE 1.

II. RELATED WORK

The goal of toxic comment classification, an NLP problem related to sentiment analysis is to classify unfavorable comments into toxic and non-toxic categories that somehow express a positive or fake opinion. Abusive text classification was done using supervised machine learning. On the features of TF-IDF, SVM was applied. *Ahuja et al.* studied that SVM is used to classify poisonous comments. 2665 English comments were gathered, and the data came from Youtube.com 1451 of the comments were either neutral or good. Using tenfold cross-validations, the remaining 1214 were identified as spam or abusive comments with an accuracy of 86.95 percent [4].

Hosseini et al. studied those attacks against the perspective API developed by Google and the Jigsaw team for a toxic comment detection system. Some sample phrases are provided on the Perspective website, and these phrases are subjected to an attack. This study demonstrates how such

978-1-6654-8771-9/22/$31.00 ©2022 IEEE
DOI: 10.1109/cicn.2022.43

Authorized licensed use limited to: Penn State University. Downloaded on March 31,2023 at 15:03:46 UTC from IEEE Xplore. Restrictions apply.

toxicity scores can be lowered to those of non-toxic terms. Additionally, this viewpoint API utility is only available for English. Convolution neural network (CNN) and LSTM are the two deep learning models that researchers have recently used to improve text classification [5].

To address the issue of class imbalance*, Ibrahim et al. [6]* offered three data augmentation methods: substitution of synonyms, random mask, and unique words augmentation. Additionally, in order to identify toxicity in user generated post, they developed a deep learning ensemble modeling approach with Bi-directional LSTM, CNN and GRU.

Anand and Eswari [7] employing both pre-trained word embeddings with LSTM and CNN to classify offensive comments. According to this study, compared to word embeddings with GloVe, CNN performs better.

The effectiveness of RNN was studied by *Pavlopoulos et al. [8]* using user comments from a Greek news portal and Wikipedia comments. In this study, GRU approach perform better than the logistic regression (LR), multi-layered perceptron (MLP), and CNN models.

Mohammad [9] used classification models logit, NBSVM, FastText- Bi-LSTM, and XGBoost to show how raw comments may be transformed. The author concludes that without any change, this model produce a quite good result.

Saeed et al.10] examined the comparison of deep neural network architecture for overlapping data with the negative sentiment. Hence, with overlapping multi-label hazardous text categorization concluded that the Bi-GRU model performed the best amongst all.

Yoon and Kim [11] used CNN-BiLSTM with multichannel lexical embeddings to boost the effectiveness of sentiment categorization.

For sentiment categorization, a multichannel convolutional-long short term memory network (CNN-LSTM) was given by Zhang et al. [12]. The suggested model beats all benchmark algorithms, according to the authors.

Thus, after collecting, annotating the datasets and designing the feature engineering is the development of the classifier. This section highlights the literature includes different approaches to tackle the difficult challenge of toxicity detection or hate speech detection. The large majority of the models previously built follow the supervised learning approach: Support Vector Machines (Ahuja et al.,2021). Deep Neural Network CNN-LSTM (Hosseini et al.,2017; Anand and Eswari,2019; Mohammad,2018; Yoon and Kim,2017; Zhang et al.,2017), Stacked Ensemble (Ibrahim et al.,2018), GradientBoosted Decision Trees (Mohammad, 2018), Logistic Regression (Pavlopoulos et al,2017) and Recurrent Neural Networks (Pavlopoulos et al., 2017; Saeed et al., 2018).

The results of the classifier per model as shown in TABLE2 discussed below:

TABLE2. REPORT ON THE MOST USED MODELS IN TOXIC COMMENT DETECTION AND THEIR CORRESPONDING ACCURACY & F1-SCORE RESULTS

Year	Research	Algorithm	Results
2021	Ahuja et al.	Support Vector Machines	98.46%
2017	Hosseini et al.	Deep Neural Network CNN-LSTM	Lowering Toxicity score
2018	Ibrahim et al.	Stacked Ensemble	0.82, F1-score
2019	Anand & Eswari	Deep Neural Network CNN-LSTM	97.26%
2017	Pavlopoulos et al.	Recurrent Neural Network, Logistic Regression	98.22%
2018	Mohammad	Deep Neural Network, Gradient Boosting Decision Trees	97.75%
2018	Saeed et al.	Recurrent Neural Network	90%
2017	Yoon and Kim	Deep Neural Network CNN-BiLSTM	89.6%
2017	Zhang et al.	Deep Neural Network CNN-LSTM	64%

In conclusion, the multichannel deep learning models were employed by the researchers to the issues with single labels and several classes. To detect multiple labels of harmful categories, a multichannel convolution bi-directional gated recurrent unit is what we recommend.

III. PROPOSED WORK

In this paper, we describe multi-label category detection with the CNN model. The proposed schematic diagram in Figure 1 represents the normal form of the hybrid model.

Here, we go over the main ideas of the suggested model.

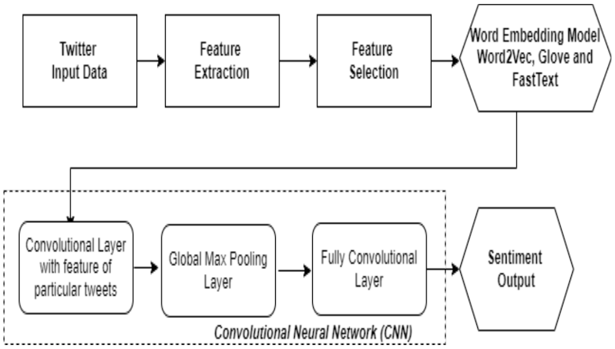
Twitter Input Data
Feature Extraction
Feature Selection
Word Embedding Model Word2Vec, Glove and FastText
Convolutional Layer with feature of particular tweets
Global Max Pooling Layer
Fully Convolutional Layer
Sentiment Output
Convolutional Neural Network (CNN)

Figure 1. Schematic diagram for Sentiment analysis with CNN approach

A. Dataset

We use the dataset from Github, a public dataset library. The dataset includes 25,000 instances with multi-label categories that count the occurrence of words in tweets and define toxicity or non-toxicity for an instance with the proposed approach.

Authorized licensed use limited to: Penn State University. Downloaded on March 31,2023 at 15:03:46 UTC from IEEE Xplore. Restrictions apply.

B. Multichannel Convolutional bi-directional gated recurrent (CNN-BiGRU) unit

A CNN-BiGRU unit represented the several versions of the same CNN model with various kernel intensities. With the help of encoding, the text can process many n-grams at once, including 1-, 2-, and 3-grams. The typical CNN model consists of various layers including word embedding, a 1-D convolutional layer, max-pooling, a dropout layer and a Bi-GRU unit [13]. This standard version has five channels defined for various n-grams. The following provides an explanation of each channel component.

1) Multichannel Word Embedding

For toxic remarks, we express each word as a number vector by removing punctuation and other special symbols. Then, with the help of pre-trained word embedding models, we extract the training data's semantic information to set up word vectors. To specifically capture the semantic meaning of words, we employ the 100-dimensional GloVe word embedding for multi-channel environment for all channels with various situations or window widths.

Multichannel word embeddings have the benefit of extracting many input features concurrently from the same training data within a model. Additionally, throughout the model training, the learnt word vectors are not updated [14].

2) Convolutional Neural Network (CNN)

The CNN is extensively taken into an account that deals with tasks like as image and video recognition, image classification, NLP and recommender systems. To create new feature maps at various places, the CNN is applied to an fixed length input vectors with numerous filters.

3) Pooling Layers

Pooling layers in CNN are a crucial component. Its major objective is to decrease the input feature maps' dimension. Particularly, each feature map receives sub-sampling from the pooling layers. The created feature map is larger than this pooled feature map. To determine the local's maximum feature value neighbourhoods for all feature map, we use the maximum pooling operations on several channels in this work.

4) Bidirectional Gated Recurrent Unit (Bi-GRU)

For the purpose of reading a document, Recurrent Neural Networks (RNN) uses a sequential word order. The RNN's long range dependencies make it challenging to train. To solve this issue, RNN variants like LSTM and GRU are introduced. LSTM model uses three gating mechanisms such as forget gate, input gate, and update gate to regulate the input sequence. As a result, the update and reset gate in GRU network regulates the input sequences [14-15].

In this study, for every channel utilizes the max-pooling layer and the Bi-GRU added to CNN. It stores in memory the text information from the present and the future. The update gate maintains the necessary memory block and the reset gate mixes the prior and new input. Additionally, the GRU operates more quickly than LSTM and updates hidden states with less processing.

C. Output Layer

The concatenation layer receives the output of each channel's convolutional bidirectional gated recurrent unit and concatenates it in a given dimension using the same input size. While the dense layer alters the dimension of vectors for updating the trainable parameters, the normalizing layer on each layer enables the model to learn more independently of this network. Then, multi-label category prediction is performed using the output layer with a sigmoid function. For each label, the proposed model uses the binary cross-entropy loss function to classify if an instance belongs to a toxic content or non-toxic.

IV. RESULTS

At Kaggle toxic dataset, we assess the proposed model in multi-label settings. The word sequences are converted to integer sequences using a tokenization approach, after that padded into a fixed-unit vector. After that, word vectors from the training data were created using the pre-trained Word2Vec, GloVe, and Fasttext word embedding methods. Additionally, for each channel, we employed various layers along with five channels as 1, 2, 3, 4, and 5 kernel sizes, respectively. These channels were combined using a concatenate layer, a fully linked and a normalization layer and with sigmoid function, output layer is presented.

This graph displays word counts for each of the three levels (0, 1, and 2) depicted in Figure 2.

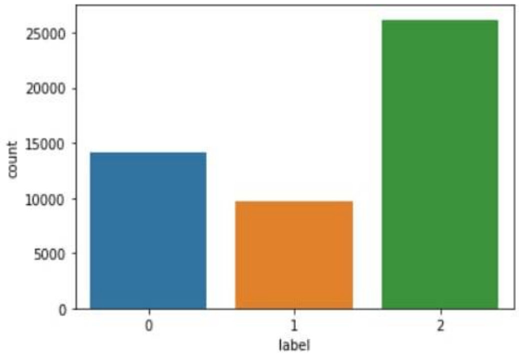

Figure 2. Number of Word Counts based on different labels.

According to word occurrence, frequency of word count is determined as shown in TABLE3.

TABLE3. FREQUENCY OF WORD WITH ITS NUMBER OF COUNTS

	Word	Count
0	Good	18556
1	Phone	17088
2	Not	12131
3	Camera	10611

Authorized licensed use limited to: Penn State University. Downloaded on March 31,2023 at 15:03:46 UTC from IEEE Xplore. Restrictions apply.

4	Batteri	8860
5	Quality	6344
6	Mobil	5186
7	Product	4827
8	Use	4463
9	Best	4406
10	Like	4391
11	One	4268
12	Samsung	3873
13	Price	3847
14	Work	3705
15	Great	3674
16	Bad	3659
17	Life	3646
18	Nice	3575
19	Also	3552

According to dataset, some examples show the results of tweets categorized as positive and negative sentiment along with the occurrence of each phrase or word as a predicted result shown in Figure 3 and Figure 4.

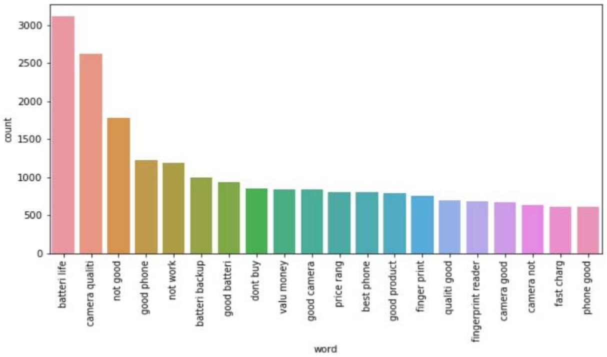

Figure 3. Word count corresponding to each mini-statement.

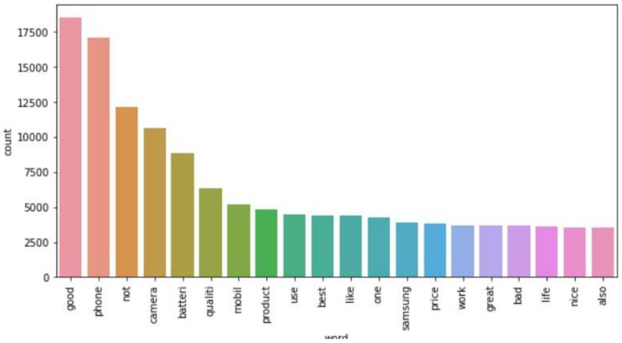

Figure 4. Word count corresponding to each word.

V. Conclusion

In this work, we presented a multichannel convolutional method to classify multi-label toxicity in online comments. To extract local characteristics and long-term relationships within the feedback using various filters and kernel intensity, the proposed model leverages CNN network in each channel. Our findings also demonstrate that the suggested model performs better than the current findings. In the future, we want to use dispersed environments with multi-channel attention techniques to detect several toxicants.

References

1. Rahul, H. Kajla., J. Hooda, G. Saini., 2020. Classification of Online Toxic Comments Using Machine Learning Algorithms. 4th International Conference on Intelligent Computing and Control Systems (ICICCS), 1119-1123. https://doi.org/10.1109/ICICCS48265.2020.9120939.
2. Singh, I., Goyal, G., Chandel, A., Alexnet Architecture Based Convolutional Neural Network for Toxic Comments Classification, Journal of King Saud University - Computer and Information Sciences (2022), doi: https://doi.org/10.1016/j.jksuci.2022.06.007
3. Djuric, N., Zhou, J., Morris, R., Grbovic, M., Radosavljevic, V., & Bhamidipati, N. (2015, May). Hate speech detection with comment embeddings. In Proceedings of the 24th international conference on world wide web (pp. 29-30). ACM.
4. Ahuja, Ravinder, Alisha Banga, and S. C. Sharma. "Detecting abusive comments using ensemble deep learning algorithms." In Malware Analysis Using Artificial Intelligence and Deep Learning, pp. 515-534. Springer, Cham, 2021.
5. Hosseini, Hossein, Sreeram Kannan, Baosen Zhang, and Radha Poovendran. 2017. Deceiving google's perspective api built for detecting toxic comments. arXiv:1702.08138
6. M. Ibrahim, M. Torki, N. El-Makky, Imbalanced toxic comments classification using data augmentation and deep learning, in: 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA), 2018, IEEE, pp. 875–878.
7. M. Anand, R. Eswari, Classification of abusive comments in social media using deep learning. in: 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC), 2019, IEEE, pp. 974–977.
8. J. Pavlopoulos, P. Malakasiotis, I. Androutsopoulos, Deep learning for user comment moderation, 2017. arXiv preprint arXiv:1705.09993.
9. F. Mohammad, Is preprocessing of text really worth your time for online comment classification?, 2018. arXiv preprint arXiv:1806.02908.
10. H.H. Saeed, K. Shahzad, F. Kamiran, Overlapping toxic sentiment classification using deep neural architectures, in: 2018 IEEE International Conference on Data Mining Workshops (ICDMW), 2018, IEEE, pp. 1361–1366.
11. J. Yoon, H. Kim, multichannel lexicon integrated CNN-BiLSTM models for sentiment analysis, in: Proceedings of the 29th Conference on Computational Linguistics and Speech Processing (ROCLING 2017), 2017, pp. 244–253.
12. H. Zhang, J. Wang, J. Zhang, X. Zhang, Ynu-hpcc at semeval 2017 task 4: using a multichannel cnn-lstm model for sentiment classification, in: Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017), 2017, pp. 796–801
13. Y. Kim, Convolutional neural networks for sentence classification, 2014. arXiv preprint arXiv:1408.5882.
14. Kumar, Ashok, S. Abirami, Tina Esther Trueman, and Erik Cambria. "Comment toxicity detection via a multichannel convolutional bidirectional gated recurrent unit." Neurocomputing 441 (2021): 272-278.
15. W. Li, Y. Xu, G. Wang, Stance detection of microblog text based on two-channel CNN-GRU fusion network, IEEE Access 7 (2019) 145944–145952.

Authorized licensed use limited to: Penn State University. Downloaded on March 31,2023 at 15:03:46 UTC from IEEE Xplore. Restrictions apply.